Improved Description of Intermolecular Interactions in Hybrid Ab Initio/Empirical Force-field Models for Crystal Structure Prediction

Benjamin Isaac Tan

A thesis submitted for the Doctor of Philosophy of Imperial College London and for the Diploma of the Imperial College

Supervisors:
Claire S. Adjiman
Constantinos C. Pantelides

The Sargent Centre for Process Systems Engineering
Department of Chemical Engineering
Imperial College London
London SW2 2AZ
United Kingdom

June, 2025

Declaration of Authorship

I, Benjamin Isaac Tan, declare that this thesis titled, 'Improved Description of Intermolecular Interactions in Hybrid Ab Initio/Empirical Force-field Models for Crystal Structure Prediction' and the work presented in it are my own. All information derived from other sources is appropriately referenced. My contributions to the curation and analysis of the CE–755 dataset from Ref [1] have been adapted in Chapter 5 of this thesis.

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International Licence (CC BY-NC-ND).

Under this licence, you may copy and redistribute the material in any medium or format on the condition that; you credit the author, do not use it for commercial purposes and do not distribute modified versions of the work.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

The prediction of polymorphism is important, since crystal physicochemical properties are crystal structure dependent. Crystal Structure Prediction (CSP) seeks to predict the putative polymorphs of a system from a minimal description of molecular connectivity. In this thesis, I study routes to improve CSP accuracy in hybrid *ab initio*/empirical force-field (HAIEFF) models, through developments in the description of intermolecular electrostatic interactions and residual intermolecular interactions.

I first seek to parametrize more-accurate empirical force-fields (FFs). To this end, I contribute to the CE-755 dataset: a chemically-diverse set of high-quality organic molecular crystal data derived using dispersion-corrected periodic Density Functional Theory. Employing the CE-755, I fit seven FFs for four electrostatic models, totaling 28 independent parametrizations. Improved energy and geometry prediction accuracy is achieved by combining new atom types with modified FF forms that mitigate the unphysical traits of the Buckingham potential. The new FFs are reasonably transferable and predictions of relative polymorph stability in the CE-755 are marginally enhanced.

The parameter estimation results indicate that describing induction is vital for HAIEFF model accuracy. As such, the Self-Consistent Electronic Response to Point-charge (SCERP) induction model was implemented into our group's CSP programs. Single-point evaluations of induction energies using SCERP and the polarizable continuum model (PCM) suggest that the former is sounder. As before, SCERP-compatible FFs are parametrized, through which an effective set of atom-pair specific Tang-Toennies damping factors are proposed.

Finally, I perform CSP studies for several hydrogen bonding molecular systems. Generally, good CSP results are obtained with the newly parametrized FFs, although further improvement from using SCERP is tenuous. Several notable deficiencies with both PCM and SCERP are uncovered, particularly as it pertains to the modeling of crystals with polar unit cells. I offer some guidance on addressing these flaws, as well as broader direction for the future development of HAIEFF models.

Publications & Conferences

Peer Reviewed CSP-Related Publications

1. Bowskill, D. H.; Tan, B. I.; Keates, A.; Sugden, I. J.; Adjiman, C. S.; Pantelides, C. C. Large-Scale Parameter Estimation for Crystal Structure Prediction. Part 1: Dataset, Methodology, and Implementation. *Journal of Chemical Theory and Computation* **2024**, *20*, 10288–10315.
2. Hunnisett, L. M. et al. The Seventh Blind Test of Crystal Structure Prediction: Structure Generation Methods. *Acta Crystallographica Section B* **2024**, *80*, 517–547.
3. Sugden, I. J.; Bowskill, D. H.; Tan, B. I.; Zhang, Y.; Adjiman, C. S.; Pantelides, C. C. Accurate and Efficient Representation of Intramolecular Energy in Ab Initio Generation of Crystal Structures. Part III: Torsional Group Partitioning. *Acta Crystallographica Section B* **2025**, *81*, 114–127.
4. Large-scale parameter estimation for Crystal Structure Prediction. Part 2: Transferable Force-Fields for Point-charge and Multipole Models (In Preparation)

CSP-Related Conferences/Workshops

1. Tan, B. I.; Bowskill, D. H.; Keates, A.; Pantelides, C. C.; Adjiman, C. S. In *33rd European Symposium on Computer Aided Process Engineering*; Kokossis, A. C., Georgiadis, M. C., Pistikopoulos, E., Eds.; Computer Aided Chemical Engineering; Elsevier, 2023; Vol. 52; pp 1153–1158.
2. Tan, B. I.; Bowskill, D. H.; Keates, A.; Pantelides, C. C.; Adjiman, C. S. Improving Transferable Force-Fields for Describing Crystal Structures Containing Hydrogen-Bonds. Poster presentation at the 33rd European Symposium on Computer Aided Process Engineering, June 2023.
3. Tan, B. I.; Bowskill, D. H.; Keates, A.; Pantelides, C. C.; Adjiman, C. S. Improving Buckingham and Mie Potentials for Crystal Structure Prediction. Poster presentation at the CECAM Crystal Structure Prediction in Materials Discovery Workshop, July 2023.
4. Tan, B. I.; Bowskill, D. H.; Keates, A.; Pantelides, C. C.; Adjiman, C. S. Development of Transferable Mie Potential Force-Fields for Crystal Structure Prediction. Poster presentation at the 22nd International Symposium on Industrial Crystallization, September 2023.

5. Tan, B. I.; Zhang, Y.; Bowskill, D. H.; Keates, A.; Pantelides, C. C.; Adjiman, C. S. Advancing the Description of Induction in Hybrid Ab Initio/Force-field Models for Crystal Structure Prediction. Oral presentation at the ChemEngDayUK24 Conference, April 2024.
6. Tan, B. I.; Zhang, Y.; Bowskill, D. H.; Keates, A.; Teng, J.; Kuminek, G.; Selbo, J.; Pantelides, C. C.; Adjiman, C. S. An Induction-Corrected Crystal Structure Prediction Study of a Simple Hydrate System. Poster presentation at the Gordon Research Conference on Crystal Engineering, June 2024.

Other Peer Reviewed Publications

1. Shah, U.; Inguva, P.; Tan, B.; Yuwono, H.; Bhute, V. J.; Campbell, J.; Macey, A.; Brechtelsbauer, C. CREATE labs — Student Centric Hybrid Teaching Laboratories. *Education for Chemical Engineers* **2021**, *37*, 22–28.
2. Tan, B. I.; Bolourinejad, P.; Eriksen, D. K.; Jackson, G.; Haslam, A. J. Application of the SAFT-γ Mie Equation of State for Reservoir Fluid Modelling in the Petroleum Industry. *Energy & Fuels* **2025**, *37*, 2435–2470.

Acknowledgements

The last four years have been a uniquely challenging, frustrating, and rewarding experience. My PhD journey would not have been the same without the support and guidance of many people. Firstly, I would like to extend my sincerest gratitude to my supervisors, Claire Adjiman and Costas Pantelides, who have provided me with invaluable guidance and enabled me to flourish throughout my time here. Likewise, the CSP team has been an unending source of inspiration and companionship during my time in the group: Yizu, Stefanos, and Leticia. A special shout-out to Isaac whose knowledge and (more importantly) affability has been a godsend since day one. I also wish to acknowledge Terrence, an absolute wizard at his craft, without whom the CSP codes might never have seen the light of day outside the group.

To the members of the MSE group, I could not have asked for a more supportive and jovial bunch of colleagues, who *constantly* try to stop me from working. These include, Griffin, Tanuj, Malak, Harry, Felipe, Gustavo, Shubhani, Riccardo, Ahmed, Thomas, Louis, Saman, Florian, Gaurav, and Lifeng. Similarly, I would like to thank Ilya and Qiaoyan, without whom the office would not be the same. I am particularly grateful to Fabian, Lilly, Lauren, and Alex who have been invaluable mentors, and even better friends, both in and outside the office. Beyond Imperial, I deeply appreciate the support and guidance provided by Adam, David, Yousif, and Rachel. While my time at Syngenta was brief, the experience was unforgettable.

To Bryan, Cason, Rico, Nina, Kuanyang, (and Jenny, Jiawei, Charmaine, Junhong, and Wailun back home) who don't have the faintest idea what my work is about, thanks for supporting me nonetheless. Last and not least, I would like to thank my parents, my brothers, and my partner Yuting (who has put up with more working weekends than is fair). Without your unwavering support, I could not have completed this journey.

Contents

1 Introduction **1**

1.1 Background and Motivation 1

1.1.1 Organic Molecular Crystals 1

1.1.2 Polymorphism in Organic Crystals 1

1.1.3 Crystal Structure Prediction 3

1.2 Thesis Scope and Outline 5

2 Literature Review **7**

2.1 Solid State Thermodynamics 7

2.2 The Hierarchical CSP Framework 8

2.3 Candidate Generation Strategies 11

2.3.1 Constraining the Search Space 11

2.3.1.1 Space Group Constraints 11

2.3.1.2 Z' Constraints 12

2.3.1.3 Molecular Flexibility Constraints 12

2.3.2 Landscape Exploration Strategies 13

2.3.2.1 Unbiased Exploration 14

2.3.2.2 Biased Exploration 14

2.4 Lattice Energy Models 15

2.4.1 Electronic Structure Methods 15

2.4.1.1 Electron Wavefunction Methods: Hartree–Fock and Post-Hartree–Fock 15

2.4.1.2 Electron Density Methods: Periodic DFT 16

2.4.1.3 Fragmentation Methods 18

2.4.1.4 Symmetry-Adapted Perturbation Theory with DFT Monomers . 19

2.4.2 Force-field Methods 19

2.4.2.1 Transferable Molecular Mechanics Force-fields 20

2.4.2.2 Tailor-made Force-fields 21

2.4.3 Hybrid Ab Initio/Empirical Force-Field Methods 22

2.4.3.1 Intramolecular Contributions 23

2.4.3.2 Electrostatic Contributions 24

2.4.3.3 Validity of the Isolated-Molecule Assumption 26

2.4.3.4 Empirical Force-Field Contribution . . . 27
2.4.3.5 HAIEFF Models in CSP . . . 28
2.4.4 SAPT(DFT)-Parametrized Intermolecular Force-Fields . . . 28
2.4.5 Machine Learning Methods . . . 29
2.5 Current State of HAIEFF Intermolecular Models . . . 30
2.5.1 Empirical Force-Field parametrization . . . 31
2.5.1.1 Reference Data Curation . . . 31
2.5.1.2 Force-field Model Definition . . . 33
2.5.1.3 Electrostatic Model Definition . . . 35
2.5.1.4 Force-field Parameter Estimation . . . 35
2.5.1.5 Force-field Deployment . . . 37
2.5.2 Induction in HAIEFF . . . 37
2.5.2.1 Implicit Models for Induction . . . 38
2.5.2.2 Explicit Models for Induction . . . 39
2.5.2.3 Previous Attempts to Incorporate Induction into HAIEFF . . . 41
2.6 Summary . . . 42

3 HAIEFF Model Formulation 44
3.1 Describing Crystal Geometries . . . 44
3.1.1 Unit Cells . . . 44
3.1.2 Atomic Coordinate Positions . . . 46
3.1.3 Body-Fixed Coordinates and Molecular Orientation . . . 46
3.1.4 Space Groups and Asymmetric Unit Cells . . . 48
3.2 HAIEFF Lattice Energy Formulation . . . 51
3.2.1 Intermolecular Interactions . . . 51
3.2.1.1 Electrostatic Interactions . . . 52
3.2.1.2 Repulsion–Dispersion Interactions . . . 59
3.2.1.3 Gradients of Isotropic Intermolecular Interactions . . . 60
3.2.1.4 Gradients of Anisotropic Intermolecular Interactions . . . 70
3.2.1.5 Truncation of Infinite-Sums . . . 72
3.2.2 Intramolecular Interactions and Flexible CSP . . . 73
3.2.2.1 Defining Molecular Flexibility . . . 74
3.2.2.2 The LAM Surrogate Model . . . 75
3.2.2.3 Overcoming LAM Local-Validity . . . 77
3.3 Summary . . . 79

4 The Imperial MSE CSP Suite 80
4.1 Supporting External Programs/Libraries . . . 80
4.1.1 NAG Libraries . . . 80
4.1.2 CCDC Suite . . . 81

4.1.3 Ab Initio Packages . . . 82
4.1.4 Packages for Electrostatic Feature Determination . . . 82
4.2 The Imperial MSE CSP Suite . . . 84
4.2.1 CrystalPredictor II . . . 84
4.2.1.1 LAM Generation . . . 84
4.2.1.2 Production Run . . . 86
4.2.1.3 Version 2.5.2.1 . . . 87
4.2.2 Crystal Structure Optimizer – Rigid Molecules (CSO–RM) . . . 89
4.2.3 Crystal Structure Optimizer – Flexible Molecules (CSO–FM) . . . 90
4.2.3.1 CSO–FM Bilevel Problem Formulation . . . 90
4.2.3.2 CSO–FM LAM Implementation . . . 91
4.2.3.3 CSO–FM Gradients with respect to Independent Dofs . . . 92
4.2.3.4 Final Polish . . . 93
4.2.4 CrystalDynamics . . . 94
4.3 Parameter Estimation for HAIEFF Force-fields . . . 94
4.3.1 Parameter Estimation Problem Formulation . . . 94
4.3.2 Derivatives for Solving the Parameter Estimation . . . 97
4.3.3 The CrystalEstimator Algorithm . . . 102
4.4 Summary . . . 104

5 Force-Field Parametrization: Enriching the Reference Data 106
5.1 Background . . . 106
5.2 Quantifying Chemical Diversity . . . 107
5.2.1 Atomic Environments . . . 107
5.2.2 Intermolecular Hydrogen Bonding . . . 108
5.3 Generating Atomic Environment Datasets . . . 109
5.3.1 DFT-D Methodology . . . 109
5.3.2 Curation of Crystal Structures . . . 110
5.4 The CE–755 Dataset . . . 112
5.4.1 Chemical Profile of the Reference Dataset . . . 113
5.4.2 Validation of the Reference Data . . . 116
5.5 Train/Test Partition . . . 121
5.6 Summary . . . 125

6 Force-Field Parametrization: Fitting Transferable Force-Fields 126
6.1 Parameter Estimation Methodology . . . 126
6.1.1 Quantum Mechanical and Electrostatic Models . . . 126
6.1.2 Force-field Formulation . . . 128
6.1.3 Programmatic Settings . . . 129
6.1.4 Parameter Bounds and Multistart Settings . . . 130

6.1.5 Parameter Estimation Termination Strategy . . . 132
6.1.6 Fitting Schemes for FF Training . . . 133
6.1.7 Assessment Metrics . . . 134
6.2 Force-field Training . . . 135
6.2.1 Parameter Estimation Residual Scores . . . 136
6.2.2 CSP Error Metrics . . . 138
6.2.2.1 Non-hydrogen Bonding (NHB) Crystal Structures . . . 138
6.2.2.2 Hydrogen Bonding (HB) Crystal Structures . . . 142
6.3 Force-field Transferability Validation . . . 145
6.4 Prediction of Relative Polymorphic Energies . . . 149
6.5 Force-field *Ad Hoc* Deployment . . . 155
6.5.1 Sensitivity to Electronic Structure Method . . . 155
6.5.2 Sensitivity to Basis Set . . . 157
6.5.3 Sensitivity to Electrostatic Method . . . 159
6.5.4 Sensitivity to PCM Settings . . . 161
6.6 Summary . . . 164

7 Induction for CSP: Implementing the SCERP Model 166
7.1 Theoretical Origin of Polarization . . . 166
7.2 The SCERP Protocol . . . 167
7.2.1 Algorithmic Implementation . . . 168
7.2.2 SCERP Process Convergence . . . 171
7.2.3 SCERP-field Convergence . . . 173
7.3 Rigid-Molecule Calculations with SCERP . . . 175
7.3.1 Electrostatic Calculations for SCERP . . . 175
7.3.1.1 Treatment of the Ewald Sum . . . 176
7.3.2 Single-point Evaluation of Induction Energy . . . 178
7.3.2.1 SCERP Process Convergence and Compute Times . . . 178
7.3.2.2 Predicted Induction Energies . . . 180
7.3.2.3 Evaluation of GDMA RADIUS . . . 183
7.3.3 CSO–RM Minimization Protocol . . . 185
7.4 Flexible-Molecule Calculations with SCERP . . . 188
7.4.1 CSO–FM Minimization Protocol . . . 188
7.4.2 Gradients for CSO–FM with SCERP . . . 189
7.4.3 Numerical Limits and Early-Stopping . . . 193
7.5 Summary . . . 195

8 Induction for CSP: Fitting SCERP-Compatible Force-fields 196
8.1 Parameter Estimation with SCERP . . . 196
8.2 Damping for Electrostatic Interactions . . . 197

8.2.1 Background . . . 197
8.2.2 Implementation of Electrostatic Damping . . . 198
8.2.3 The Damping Factor . . . 199
8.3 Force-field Parameter Estimation . . . 201
8.3.1 General Methodology . . . 201
8.3.2 Fitting with Small Training Sets . . . 203
8.3.3 Fitting with Medium Training Sets . . . 207
8.3.4 Final Parametrization . . . 212
8.3.4.1 SCERP-Embedded CSO–RM Minimization Performance . . . 212
8.3.4.2 Training Set Performance . . . 214
8.3.4.3 Test Set Performance . . . 217
8.3.4.4 Prediction of Relative Polymorphic Energies . . . 221
8.4 Summary . . . 222

9 CSP Case Studies 224
9.1 General CSP Methodology . . . 224
9.1.1 Global Search . . . 224
9.1.2 Refinement I . . . 225
9.1.3 Refinement II . . . 225
9.1.4 Structure Clustering and Matching . . . 226
9.1.5 Local Minimizations . . . 226
9.2 Rigid Monohydrate: 2,6-Diamino-4(3H)-Pyrimidinone . . . 227
9.2.1 Background and Early-Stage CSP . . . 227
9.2.2 Refinement Stage . . . 228
9.2.3 SCERP-Embedded Algorithmic Performance . . . 231
9.3 Flexible $Z' = 1$ Neutral Molecule: Urea . . . 232
9.3.1 Background and Early-Stage CSP . . . 232
9.3.2 Refinement Stage . . . 233
9.3.3 Final Periodic DFT-D Re-ranking . . . 236
9.3.4 Assessment of SCERP Gradient Approximations . . . 237
9.4 Flexible $Z' = 1$ Zwitterion: Glycine . . . 240
9.4.1 Background and Early-Stage CSP . . . 240
9.4.2 Refinement Stage . . . 241
9.4.3 Final Periodic DFT-D Re-ranking . . . 245
9.4.4 Assessment of SCERP Gradient Approximations . . . 246
9.5 Flexible $Z' = 1$ Zwitterion: Proline . . . 248
9.5.1 Background and Early-Stage CSP . . . 248
9.5.2 Refinement Stage . . . 249
9.5.3 Final Periodic DFT-D Re-ranking . . . 252
9.5.4 Assessment of SCERP Gradient Approximations . . . 253

9.6 SCERP Gradient Approximations: Discussion . . . 254
9.7 Flexible Cocrystal: Caffeine + m-Nitrobenzoic Acid . . . 256
9.7.1 Background and Early-Stage CSP . . . 256
9.7.2 Refinement Stage . . . 257
9.8 HAIEFF Model Accuracy: Discussion . . . 260
9.9 Summary . . . 262

10 Conclusions & Future Work **264**

Bibliography **268**

A Gamma Functions and their Derivatives **300**

B Trigonometric Identities used in Ewald Sums **302**

C Derivatives of (Reciprocal) Unit Cell Matrix with respect to Lattice Parameters **304**
C.1 First-Derivatives . . . 304
C.2 Double-Derivatives . . . 306

D Derivation of CP2 Flexible Gradients **311**
D.1 Constructing Molecules . . . 311
D.2 Derivatives of Atomic Cartesian Coordinates . . . 314

E CP2 Shift Algorithm **324**
E.1 Background . . . 324
E.2 The Simplest Case: Cubic Shift Algorithm . . . 324
E.3 Quadratic Solutions with Deformed (Non-Cubic) Cells . . . 332
E.4 Space group Symmetry Constraints . . . 364
E.5 Undefined Analytical Solutions . . . 373
E.6 Algorithm Implementation and Solution Selection . . . 376

F CE–755 Validation Against Experimental Sublimation Energies **382**

G Training/Test Set Partitioning of the CE–755 **386**

H Gradients Involving Tang–Toennies Damping **406**

I Parameter Bounds and Reference Potentials for Parameter Estimation **410**
I.1 Reference Potentials for CrystalEstimator . . . 410
I.2 Lower Bounds for Mie σ Parameter . . . 413

J Optimized Force-field Parameters for Point-charge and Multipole Models with/without PCM **416**

J.1 Point-charge Electrostatic Model, without PCM 416
J.2 Point-charge Electrostatic Model, with PCM 423
J.3 Multipole Electrostatic Model, without PCM 431
J.4 Multipole Electrostatic Model, with PCM 439

K Distributed Point-charges for SCERP **447**
K.1 Introduction 447
K.2 The Distributed Point-charge Model 447
K.3 DCM in SCERP 453
K.4 Conclusions 457

L Atom-pair Specific Tang–Toennies Damping Parameters **458**

M SCERP-Embedded Parameter Estimation **460**
M.1 Training Sets for Parameter Estimation 460
M.2 TP442 Parameter Estimation Performance Metrics 473
M.3 Optimized Force-field Parameters for SCERP-embedded Models 476
M.4 CE–755 Energy Accuracy without Outliers 480

N CSP Case Studies: Supplementary Information **482**
N.1 2,6-Diamino-4(3H)-Pyrimidinone Monohydrate CSP 482
N.2 Urea CSP 491
N.3 Glycine CSP 504
N.4 L-/DL-Proline CSP 514
N.5 1:1 Caffeine + *m*-Nitrobenzoic Acid CSP 523

List of Figures

2.1 The multistage, hierarchical CSP workflow. 10

3.1 Illustration of an arbitrary unit cell of a crystal in the space-fixed coordinate system. 45

4.1 Schematic of the CP2 algorithm, involving LAM Generation and the Production Run. Key stages are enumerated and detailed in the main text. 85

4.2 Schematic of the CSO–FM algorithm that employs on-the-fly LAM generation and Final Polish by CSO–RM. Key stages are enumerated and detailed in the main text. 91

4.3 Schematic of the CrystalEstimator algorithm with dynamic starting points, algorithm reinitialization, and temporary structure removal. Key stages are enumerated and detailed in the main text. 103

5.1 Tally of structures containing each pairwise atomic-interaction within the DB–445 dataset [2]. Each matrix entry corresponds to the total number of structures which contain the pair of atoms on the row and column. Cells are colored orange (< 10), yellow ($10 \leq$ and ≤ 19), pink ($20 \leq$ and ≤ 29), and blue (≥ 30) according to their structure counts. 109

5.2 (a) The proportion of crystals in the CE–755 containing each element. (b) The proportion of crystals in the CE–755 (inner ring) in the six most frequently observed space groups for organic crystals in the CSD (outer ring). (c) The distribution of molecules based on their composite number of atoms and atomic environments (as defined in Table 5.1). The opacity of the black ● reflect the frequency of compounds with a given number of atoms and atomic environments; this is also indicated by the contour lines and density curves. 114

5.3 Tally of structures containing each pairwise atomic-interaction within the CE–775 dataset. Each matrix entry corresponds to the total number of structures which contain the pair of atoms on the row and column. Cells are colored orange (< 10), yellow ($10 \leq$ and ≤ 19), pink ($20 \leq$ and ≤ 29), and blue (≥ 30) according to the structure counts. 115

5.4 $RMSD_1$ overlay of the FUZZAB molecule in its equilibrium conformation (grey) and crystalline-phase conformation (green), viewed from (a) the side of the aromatic ring's plane, (b) orthogonal to the aromatic ring's 'face', and (c) along the axis of the C–N bond. In the equilibrium conformation, the O–H is almost planar with the carbonyl oxygen while in the crystalline-phase, the O–H orientates orthogonally to the carbonyl oxygen. 120

5.5 Atomic environment matrix showing the proportion of test set structures that contain cross-interactions between atomic environments n and m, as a percentage ($p_{n,m}$) of the number of structures in the CE–755 containing atomic environments n and m (Equation 5.5.1). Colors correspond to: $p_{n,m} = 0$ or $p_{n,m} = 100$ (orange); $p_{n,m} = 50$ (pink); $p_{n,m} \in [10, 30]$ (blue); all other $p_{n,m}$ (white), and black if there are no structures in the CE–755 with that interaction. 123

5.6 Hydrogen bond length distribution for different motifs in the training set (blue), test set (orange), and entire CE–755 dataset (black). Each boxes' bounds represent the 1st and 3rd quartiles, with whiskers extending to ±1.5x the interquartile range and outliers given by ○. Medians and means are indicated by ━ and × respectively. The counts of each motif in the entire CE–755 (N_{Total}), the training set (N_{TRN}), and the test set (N_{TST}) are tabulated below. 124

6.1 The parameterization termination strategy, illustrated for the (d)BHP. Parameter estimation initializes from 512 Sobol' seeds, using $\bar{\boldsymbol{p}}^{\text{ref}} = \bar{\boldsymbol{p}}^{\text{FIT}}$. If the Minimum Sobol (MS) (red) has an unsatisfactory termination status, the parameter estimation is repeated with 128 seeds, using the preceding $\bar{\boldsymbol{p}}^{*}_{\text{MS}}$ as $\bar{\boldsymbol{p}}^{\text{ref}}$. This continues until a satisfactory termination status is achieved at the MS. Estimations using the MIP add a zeroth seed at each stage. 133

6.2 Sum of residual scores across the training set for the different fitting schemes, using (a) point-charge electrostatics, and (b) multipole electrostatics. Results derived without and with PCM are given by the blue △ and orange ▽ respectively. The gap in (a) is due to the parameter estimation failing over all attempted Sobol' points and no solution being found. 136

6.3 Histograms of $RMSD_{15}$ scores (Å) across Schemes 1–7 for (a) NHB, and (b) HB structures in the TRN set using point-charge electrostatics. The tables contain the number of failed structures, average $RMSD_{15}$, and maximum $RMSD_{15}$. Failed structures are omitted from the average. Results derived without and with PCM are given in blue and orange respectively, and are separated by '|'. Scheme 3 fails without PCM. 138

6.4 Histograms of dU_{inter} (kJ/mol) across Schemes 1–7 for (a) NHB, and (b) HB structures in the TRN set using point-charge electrostatics. The tables contain the AD, AAD, and SD statistics. Results derived without and with PCM are given in blue and orange respectively, and are separated by '|'. Scheme 3 fails without PCM. 139

6.5 Histograms of RMSD_{15} scores (Å) across Schemes 1–7 for (a) NHB, and (b) HB structures in the TRN set using multipole electrostatics. The tables contain the number of failed structures, average RMSD_{15}, and maximum RMSD_{15}. Failed structures are omitted from the average. Results derived without and with PCM are given in blue and orange respectively, and are separated by '|'. 140

6.6 Histograms of dU_{inter} (kJ/mol) across Schemes 1–7 for (a) NHB, and (b) HB structures in the TRN set using multipole electrostatics. The tables contain the AD, AAD, and SD statistics. Results derived without and with PCM are given in blue and orange respectively, and are separated by '|'. 141

6.7 Histograms of RMSD_{15} scores (Å) across Schemes 1, 5, and 7 for (a) NHB, and (b) HB structures using point-charge electrostatics. The tables contain the *change* in average and maximum RMSD_{15}. TST set results derived without and with PCM are given in blue and orange respectively, and are separated by '|'. TRN set results are overlaid as unfilled bars. 146

6.8 Histograms of RMSD_{15} scores (Å) across Schemes 1, 5, and 7 for (a) NHB, and (b) HB structures using multipole electrostatics. The tables contain the *change* in average and maximum RMSD_{15}. TST set results derived without and with PCM are given in blue and orange respectively, and are separated by '|'. TRN set results are overlaid as unfilled bars. 146

6.9 Histograms of dU_{inter} (kJ/mol) across Schemes 1, 5, and 7 for (a) NHB, and (b) HB structures using point-charge electrostatics. The tables contain the *change* in AD, AAD, and SD. TST set results derived without and with PCM are given in blue and orange respectively, and are separated by '|'. TRN set results are overlaid as unfilled bars. 147

6.10 Histograms of dU_{inter} (kJ/mol) across Schemes 1, 5, and 7 for (a) NHB, and (b) HB structures using multipole electrostatics. The tables contain the *change* in AD, AAD, and SD. TST set results derived without and with PCM are given in blue and orange respectively, and are separated by '|'. TRN set results are overlaid as unfilled bars. 147

6.11 ΔU_{poly} (kJ/mol) when using a point-charge model (a) without PCM, and (b) with PCM. Each polymorphic pair is labeled with their CSD Refcodes ordered as 'More Stable/Less Stable'. $\Delta U_{\text{poly}}^{\text{ref}}$ values are given in black, while $\Delta U_{\text{poly}}^{\text{pred}}$ obtained using the Scheme 5 and 7 models are in orange and blue respectively. Hatched bars indicate that a member of the pair is from the test set. NHB and HB systems are on the left and right of the dashed line respectively. 151

6.12 ΔU_{poly} (kJ/mol) when using a multipole model (a) without PCM, and (b) with PCM. Each polymorphic pair is labeled with their CSD Refcodes ordered as 'More Stable/Less Stable'. $\Delta U_{\text{poly}}^{\text{ref}}$ values are given in black, while $\Delta U_{\text{poly}}^{\text{pred}}$ obtained using the Scheme 5 and 7 models are in orange and blue respectively. Hatched bars indicate that a member of the pair is from the test set. NHB and HB systems are on the left and right of the dashed line respectively. 152

6.13 Parity plots of (a) RMSD_{15} scores (Å), and (b) dU_{inter} (kJ/mol) between the fitted (PBE0/6-311G(d,p), MIP, rank 4 multipoles, $\varepsilon_{\text{PCM}} = 11$) and extrapolated models, for NHB (left) and HB (right) test set structures. The extrapolated models substitute PBE0 with: HF (black ○), PBE (orange □), B3LYP (blue ◇), and MP2 (grey △) during the QM calculations. The blue ◆ in the left plots corresponds to CLMETH03, which failed minimization with B3LYP. 156

6.14 Parity plots of (a) RMSD_{15} scores (Å), and (b) dU_{inter} (kJ/mol) between the fitted (PBE0/6-311G(d,p), MIP, rank 4 multipoles, $\varepsilon_{\text{PCM}} = 11$) and extrapolated models, for NHB (left) and HB (right) test set structures. The extrapolated models substitute 6-311G(d,p) with: 3-21G (black ○), 6-31G(d,p) (orange □), and 6-311+G(d,p) (blue ◇) during the QM calculations. 158

6.15 Parity plots of (a) RMSD_{15} scores (Å), and (b) dU_{inter} (kJ/mol) between the fitted (PBE0/6-311G(d,p), MIP, rank 4 multipoles, $\varepsilon_{\text{PCM}} = 11$) and extrapolated model, for NHB (left) and HB (right) test set structures. The extrapolated model substitutes rank 4 GDMA multipoles with rank 0 G09 HLYGAt atomic-charges (black ○). 160

6.16 Parity plots of (a) RMSD_{15} scores (Å), and (b) dU_{inter} (kJ/mol) between the fitted (PBE0/6-311G(d,p), MIP, rank 4 multipoles, $\varepsilon_{\text{PCM}} = 11$) and extrapolated models, for NHB (left) and HB (right) test set structures. The extrapolated models substitute $\varepsilon_{\text{PCM}} = 11$ with: no PCM (black ○), and PCM with $\varepsilon_{\text{PCM}} = 3$ (orange □) and $\varepsilon_{\text{PCM}} = 7$ (blue ◇) during the QM calculations. When optimized without PCM, CLMETH03 cannot achieve a valid RMSD_{15} score (black ●). When $\varepsilon_{\text{PCM}} = 3$, the rigid optimization of CLMETH03 failed (orange ■). 162

7.1 Schematic of the SCERP process. Key stages are enumerated and detailed in the main text. 169

7.2 $Q_{Ii,\text{field}}^{\text{total,pc}}$ for atoms in NAPOAC01 over the SCERP process (Table 7.1), as a function of iteration number. Each line/color represents a different atom. 171

7.3 $Q_{Ii,\text{field}}^{\text{total,pc}}$ and $\Delta Q_{Ii,t}^{\text{total,pc}}$ for atoms in NAPOAC01 using decelerated SCERP, as a function of iteration number. (a) and (b): $t_{\text{slow}} = 3$; (c) and (d): $t_{\text{slow}} = 6$. Each line/color represents a different atom. Test parameters are given in Table 7.1. The dashed line in (b) and (d) shows $\varepsilon_{\text{scrp}} = 1 \times 10^{-3}$ a.u. $\Delta Q_{Ii,t}^{\text{total,pc}}$ between 'Static' and 'Iteration 0' are omitted. 172

7.4 $\Delta Q_{Ii,t}^{\text{total,pc}}$ for atoms in NAPOAC01 using decelerated SCERP, as a function of iteration number when (a) $t_{\text{slow}} = 3$, and (b) $t_{\text{slow}} = 6$. Each line/color represents a different atom. Test parameters are given in Table 7.1, but $\varepsilon_{\text{scrp}} = 1 \times 10^{-4}$ a.u. and $t_{\text{max}} = 20$. The dashed line shows $\varepsilon_{\text{scrp}} = 1 \times 10^{-4}$ a.u. $\Delta Q_{Ii,t}^{\text{total,pc}}$ between 'Static' and 'Iteration 0' are omitted. 173

7.5 Converged $Q_{Ii,\text{field}}^{\text{total,pc}}$ for atoms in (a) NAPOAC01, and (b) FITZEO01 at different r_{scrp} values. Each line/color represents a different atom. $r_{\text{scrp}} = 0$ Å indicates 'Static' conditions. Test parameters are given in Table 7.2. 174

7.6 The maximum values of $\Delta Q_{Ii,t}^{\text{total,pc}}$ during the single-point SCERP evaluation of structures from the CE–755 dataset. The dashed line indicates $\varepsilon_{\text{scrp}} = 1 \times 10^{-3}$ a.u. 179

7.7 Predicted U_{ind} for the PCM- and SCERP-embedded models in (a) NHB, and (b) HB structures. The colors of each point reflect the Avg. $|V_{\text{field}}|$ values. The solid lines represent a $\pm 2\sigma_{\text{std.dev}}$ bound from the best fit line in each quantile. Points outside these bounds are flagged as outliers ($\diamond$), with the top-5 outliers labeled by their CSD Refcodes. 181

7.8 (a) and (c): ESP from the SCERP-field on the 0.8x vdW surface of KUXPUP and KUXPUP01 respectively; arrows point to the areas of greatest ESP. (b) and (d): The HB-motif in KUXPUP and KUXPUP01 respectively; hydrogen bonds are given by the cyan dotted lines. 182

7.9 Predicted $U_{\text{elec+ind}}$ for $r_{\text{H}}^{\text{GDMA}} = 0.650$ Å and 0.325 Å. NHB structures: (a) without PCM or SCERP, (b) with PCM ($\varepsilon_{\text{PCM}} = 11$), and (c) with SCERP. HB structures: (d) without PCM or SCERP, (e) with PCM ($\varepsilon_{\text{PCM}} = 11$), and (f) with SCERP. . 184

7.10 Predicted U_{ind} when $r_{\text{H}}^{\text{GDMA}} = 0.325$ Å, for the PCM- and SCERP-embedded models in (a) NHB, and (b) HB structures. The colors of each point reflect the Avg. $|V_{\text{field}}|$ values. The solid lines represent a $\pm 2\sigma_{\text{std.dev}}$ bound from the best fit line in each quantile. Points outside these bounds are flagged as outliers ($\diamond$), with the top-5 outliers labeled by their CSD Refcodes. 185

7.11 Schematic of the CSO–RM protocol, when used with SCERP. Key stages are enumerated and detailed in the main text. 187

8.1 The Tang–Toennies damping function, $f_n(\beta = 1.6400\ \text{bohr}^{-1}, r)$, for $n = 2$ to 6. . . 199

8.2 The Tang–Toennies damping function, $f_n(\beta = 1.7308\ \text{bohr}^{-1}, r)$, for $n = 2$ to 6. . . 200

8.3 Sum of geometry (blue) and energy (red) residual scores for the (a) NHB47, and (b) HB53 training sets. The total objective function value is summed above each pair. 204

8.4 Sum of geometry (blue) and energy (red) residual scores for the TP100 training set. The total objective function value is summed above each pair. 208

8.5 TP442 errors for PCM_{11} and post-CSORM $SCRP_{atm,90}$: (a) NHB crystals $RMSD_{15}$, (b) NHB crystals dU_{inter}, (c) HB crystals $RMSD_{15}$, and (d) HB crystals dU_{inter}. Structures with $|\Delta RMSD_{15}| \geq 0.20$ Å and $|\Delta dU_{inter}| \geq 15.0$ kJ/mol are marked with □ and ◇ respectively. Green areas in (b) and (d) denote where the y-axis model has a lower $|dU_{inter}|$. Histograms of each model's errors are shown on the corresponding axes. 211

8.6 Top: convergence of the SCERP-embedded CSO–RM minimization ($\varepsilon_{scrpRM} = 1 \times 10^{-3}$ kJ/mol (dashed line) and $k_{max} = 10$) for the $SCRP_{atm,90}$ model + Scheme 7 fitted FF over the CE–755 dataset. Inset: $U^{*}_{inter,k}$ for BEMKAI, which did not converge. Bottom: the average number of QM computations (normalized by Z') to converge the SCERP process at iteration k. 213

8.7 CE–755 TRN set errors for PCM_{11} and post-CSORM $SCRP_{atm,90}$, using Scheme 5: (a) NHB crystals $RMSD_{15}$, (b) NHB crystals dU_{inter}, (c) HB crystals $RMSD_{15}$, and (d) HB crystals dU_{inter}. Structures with $\Delta RMSD_{15} \geq 0.20$ Å and $|\Delta dU_{inter}| \geq 15.0$ kJ/mol are marked with □ and ◇ respectively. Green areas in (b) and (d) denote where the y-axis model has a lower $|dU_{inter}|$. Histograms of each model's errors are shown on the corresponding axes. OCAPUE is omitted from the HB statistics, due to its fusion with $SCRP_{atm,90}$. 215

8.8 CE–755 TRN set errors for PCM_{11} and post-CSORM $SCRP_{atm,90}$, using Scheme 7: (a) NHB crystals $RMSD_{15}$, (b) NHB crystals dU_{inter}, (c) HB crystals $RMSD_{15}$, and (d) HB crystals dU_{inter}. Structures with $\Delta RMSD_{15} \geq 0.20$ Å and $|\Delta dU_{inter}| \geq 15.0$ kJ/mol are marked with □ and ◇ respectively. Green areas in (b) and (d) denote where the y-axis model has a lower $|dU_{inter}|$. Histograms of each model's errors are shown on the corresponding axes. 216

8.9 CE–755 TST set errors for PCM_{11} and post-CSORM $SCRP_{atm,90}$, using Scheme 5: (a) NHB crystals $RMSD_{15}$, (b) NHB crystals dU_{inter}, (c) HB crystals $RMSD_{15}$, and (d) HB crystals dU_{inter}. Structures with $\Delta RMSD_{15} \geq 0.20$ Å and $|\Delta dU_{inter}| \geq 15.0$ kJ/mol are marked with □ and ◇ respectively. Green areas in (b) and (d) denote where the y-axis model has a lower $|dU_{inter}|$. Histograms of each model's errors are shown on the corresponding axes. 218

8.10 CE–755 TST set errors for PCM_{11} and post-CSORM $SCRP_{atm,90}$, using Scheme 7: (a) NHB crystals $RMSD_{15}$, (b) NHB crystals dU_{inter}, (c) HB crystals $RMSD_{15}$, and (d) HB crystals dU_{inter}. Structures with $\Delta RMSD_{15} \geq 0.20$ Å and $|\Delta dU_{inter}| \geq 15.0$ kJ/mol are marked with □ and ◇ respectively. Green areas in (b) and (d) denote where the y-axis model has a lower $|dU_{inter}|$. Histograms of each model's errors are shown on the corresponding axes. 219

8.11 ΔU_{poly} (kJ/mol) when using the $\text{SCRP}_{\text{atm,90}}$ multipole model. Each polymorphic pair is labeled with their CSD Refcodes ordered as 'More Stable/Less Stable'. $\Delta U_{\text{poly}}^{\text{ref}}$ values are given in black, while $\Delta U_{\text{poly}}^{\text{pred}}$ obtained using the Scheme 5 and 7 models are in cyan and magenta respectively. Hatched bars indicate that a member of the pair is from the test set. NHB and HB systems are on the left and right of the dashed line respectively. The KUXPUP/KUXPUP01 pair is omitted as KUXPUP01 presents with erroneous behavior with SCERP. 220

9.1 MH-DAP ($C_4H_6N_4O + H_2O$, 15 + 3 atoms). Both molecules are treated as rigid. 227

9.2 Summary of rank (top) and $\Delta U_{\text{latt}}^{\text{GM}}$ (bottom) for the MH-DAP matches using the refinement models in Table 9.2. MH-I, II, and III are given by ◇, □, and ⬠ respectively, while the marker fill colors reflect their RMSD_{20} scores (in Å). . . . 229

9.3 Urea (CH_4N_2O, 8 atoms). Four flexible torsions are defined: two proper (green and magenta), and two improper torsions (red and cyan). 232

9.4 Refined landscape of urea using Models 1 (top-left), 2 (top-right), 3 (bottom-left), and 4 (bottom-right). The lowest-energy match to F-I, F-III, and F-IV are labeled as ◇, □, and ⬠ respectively, with their corresponding RMSD_{20} scores. . . 234

9.5 Comparison of (a) $\Delta U_{\text{latt}}^{\text{F-III}}$, and (b) ranks, predicted by the HAIEFF models versus periodic DFT-D. The bottom 10 structures + structure 124 from Model 4 are denoted by blue ◇, while their corresponding Model 3 results are denoted by red ○. 236

9.6 Refined landscape of urea using Model 4, with Mode 2 (left) and Mode 3 (right) FD gradient approximations. The lowest-energy match to F-I, F-III, and F-IV are labeled as ◇, □, and ⬠ respectively, with their corresponding RMSD_{20} scores. 240

9.7 Glycine ($C_2H_5NO_2$, 10 atoms). Three flexible dofs are defined: two proper torsions (green and cyan), and one bond angle (magenta). 241

9.8 Refined landscape of glycine using Models 1 (top-left), 2 (top-right), 3 (bottom-left), and 4 (bottom-right). The lowest-energy match to α, β, γ, δ, and ε are labeled as ◇, □, ⬠, ✣, and ☆ respectively, with their corresponding RMSD_{20} scores. In the Model 4 plot, crystals belonging to polar space groups are denoted by •. 243

9.9 ESP exerted on atom i of α- (top), β- (middle), and γ-glycine (bottom) at different real-space cutoffs. All ESPs are reported as a ratio with respect to the ESP at r = 20 Å. 245

9.10 Comparison of (a) $\Delta U_{\text{latt}}^{\alpha}$, and (b) ranks, predicted by the HAIEFF models versus periodic DFT-D. The bottom 6 structures + structures 21 and 29 from Model 4 are denoted by blue ◇, while their corresponding Model 3 results are denoted by red ○. 246

9.11 Refined landscape of glycine using Model 4, with Mode 2 (left) and Mode 3 (right) FD gradient approximations. The lowest-energy match to α, β, γ, δ, and ε are labeled as $\diamond$, $\square$, $\pentagon$, $\clubsuit$, and $\star$ respectively, with their corresponding $RMSD_{20}$ scores. Crystals belonging to polar space groups are denoted by $\bullet$. 248

9.12 L-Proline ($C_5H_9NO_2$, 17 atoms). Three flexible proper torsions are defined (green, cyan, and magenta). 249

9.13 Refined landscape of proline using Models 1 (top-left), 2 (top-right), 3 (bottom-left), and 4 (bottom-right). The lowest-energy match to L-I, L-II, and DL-I are labeled as $\diamond$, $\square$, and $\pentagon$ respectively, with their corresponding $RMSD_{20}$ scores. In the Model 4 plot, crystals belonging to polar space groups are denoted by $\bullet$. . . . 251

9.14 Comparison of (a) $\Delta U_{\text{latt}}^{\text{L-I}}$, and (b) ranks, predicted by the HAIEFF models versus periodic DFT-D. The bottom 18 structures + structures 86 from Model 4 are denoted by blue $\diamond$, while their corresponding Model 3 results are denoted by red $\bigcirc$. 253

9.15 Refined landscape of proline using Model 4, with Mode 2 (left) and Mode 3 (right) FD gradient approximations. The lowest-energy match to L-I, L-II, and DL-I are labeled as $\diamond$, $\square$, and $\pentagon$ respectively, with their corresponding $RMSD_{20}$ scores. Crystals belonging to polar space groups are denoted by $\bullet$. 255

9.16 CAFF-mNBA ($C_8H_{10}N_4O_2 + C_7H_5NO_4$, 24 + 17 atoms). Three flexible torsions (green, cyan, and magenta) are defined for mNBA (left), while CAFF (right) is treated as rigid. 257

9.17 Crystal packing of CAFF-mNBA F-I: (a) face-on view of a single packing plane, (b) side-on view along the packing plane. HB contacts are given by the black dotted lines. 258

9.18 Refined landscape of CAFF-mNBA using Models 1 (top-left), 2 (top-right), 3 (bottom-left), and 4 (bottom-right). The lowest-energy match to F-I and F-II are labeled as $\diamond$ and $\square$ respectively, with their corresponding $RMSD_{20}$ scores. 259

D.1 Cartesian coordinates of the first 3 atoms in a molecule where (a) atom 3 is bonded to atom 1, or (b) atom 3 is bonded to atom 2. 312

D.2 The key problem we hope to solve is determine the position of atom 4 given knowledge of the connectivity relationship between 1, 2, 3, and 4. 313

D.3 A local coordinate system is defined around atom 2. 313

D.4 A spherical coordinate system is defined around atom 3. 314

E.1 A simplified representation of the clashing molecule problem. The red and blue structure represent distinct molecules in some arbitrary unit cell. The arrow points from the first-atom of each molecule to their respective 'clash sites'. . . . 325

E.2 An example in 2-dimensions where solving the linear equations results in a contracting solution. (a) The initial clash site distance (distance between arrowheads) is 1.35 units. (b) The final distance reaches the target of 2.20 units. 332

E.3 The program structure leading up to a call to the Shift Algorithm subroutine. . . 376

E.4 The program structure within the Shift Algorithm subroutine. 377

E.5 The Eps_3 subroutine, which solves the expansion when there are no space group constraints. 378

E.6 The Eps_2 subroutine, which solves the expansion when $a = b \neq c$. 379

E.7 The Eps_1 subroutine, which solves the expansion when $a = b = c$. 379

E.8 The select_sol subroutine, which carries out the solution selection procedure. . . . 380

I.1 Exemplars of the FIT-analogue determined using a Mie potential for (a) the C/H interaction, (b) the O/H_N interaction, and (c) the H_N/H_O interaction. The FIT reference (—) and the derived Mie potential (dashed —) are overlaid. 412

K.1 DCM model for the central molecule in (a) CYCYPR, and (b) KUXPUP01. DCM-determined satellite-charges are shown in magenta, while all other elements adhere to their default colors in Mercury. 448

K.2 Distribution of RMSEs in the model-predicted ESPs for structures in the CE–755 dataset: (a) the long-range RMSE, (b) the mid-range RMSE, (c) the short-range RMSE, and (d) the total RMSE. ADAMAN08 and FURANE10 have been omitted due to highly erroneous results with DCM. Over the remaining 753 structures, the bounds of each box give the 1st and 3rd quartile, while whiskers represent the ±1.5x interquartile range. The median is given by the orange — in each box, while means are given by the green ▲ (and at the top of the plot). Outliers are represented with ○. 451

K.3 Distribution of ΔRMSEs in the model-predicted ESPs for structures in the CE–755 dataset: (a) the long-range RMSE, (b) the mid-range RMSE, (c) the short-range RMSE, and (d) the total RMSE. ADAMAN08 and FURANE10 have been omitted due to highly erroneous results with DCM. Over the remaining 753 structures, the bounds of each box give the 1st and 3rd quartile, while whiskers represent the ±1.5x interquartile range. The median is given by the orange — in each box, while means are given by the green ▲ (and at the top of the plot). Outliers are represented with ○. 452

K.4 The 40 Å SCERP embedding-sphere around the central CYCYPR molecule (top) using: atomic-charges (represented with molecules) and DCM-charges (represented with magenta spheres). The charges associated with the first 100 molecules of the embedding-sphere are shown. The DCM-charges overlay exactly with the atomic-positions of their corresponding molecular periodic images, indicating that the transformation of DCM-charge coordinates to their SCERP-field positions has been correctly implemented. 454

K.5 The maximum values of $\Delta Q_{Ii,t}^{\text{total,pc}}$ during the single-point DCM-SCERP evaluation of structures from the CE–755 dataset. The dashed line indicates $\varepsilon_{\text{scrp}} = 1 \times 10^{-3}$ a.u. 455

K.6 Predicted U_{ind} for the PCM- and SCERP-embedded (with DCM-charges) models in (a) NHB, and (b) HB structures. The colors of each point reflect the Avg. $|V_{\text{field}}|$ values, defined by Equations 7.3.7 and 7.3.8. ΔU_{ind} is given by Equation 7.3.6. The solid lines represent a $\pm 2\sigma_{\text{std.dev}}$ bound from the best fit line in each quantile. Points outside these bounds are flagged as outliers ($\diamond$), with the top-5 outliers labeled by their CSD Refcodes. 456

K.7 Parity in U_{ind} for the SCERP-embedded models using atomic-charges (SCERP) and DCM-charges (DCM), in (a) NHB, and (b) HB structures. ΔU_{ind} is given by $\Delta U_{\text{ind}} = U_{\text{ind}}^{\text{DCM}} - U_{\text{ind}}^{\text{SCERP}}$ 457

M.1 Sum of geometry (blue) and energy (red) residual scores for the TP442 training set. The total objective function value is summed above each pair. 473

M.2 Post-CSORM TP442 errors for SCRP and $\text{SCRP}_{\text{atm,90}}$: (a) NHB crystals RMSD_{15}, (b) NHB crystals dU_{inter}, (c) HB crystals RMSD_{15}, and (d) HB crystals dU_{inter}. Structures with $|\Delta\text{RMSD}_{15}| \geq 0.20$ Å are marked with $\square$. Green areas in (b) and (d) denote where the y-axis model has a lower $|dU_{\text{inter}}|$. Histograms of each model's errors are shown on the corresponding axes. 475

M.3 CE–755 TRN set dU_{inter} for PCM_{11} and post-CSORM $\text{SCRP}_{\text{atm,90}}$: (a) NHB crystals with Scheme 5 FF, (b) HB crystals with Scheme 5 FF, (c) NHB crystals with Scheme 7 FF, and (d) HB crystals with Scheme 7 FF. In (b), OCAPUE is omitted from the HB statistics, due to its fusion with $\text{SCRP}_{\text{atm,90}}$. The outlier structures from Figures 8.7 and 8.8 (CYCYPR, KUXPUP01, CLTRZL, and GEYSEI) have been omitted from these plots. Green areas denote where the y-axis model has a lower $|dU_{\text{inter}}|$. Histograms of each model's errors are shown on the corresponding axes. 480

M.4 CE–755 TST set dU_{inter} for PCM_{11} and post-CSORM $\text{SCRP}_{\text{atm,90}}$: (a) HB crystals with Scheme 5 FF, and (b) HB crystals with Scheme 7 FF. The outlier structures from Figures 8.9d and 8.10d (EREFEP) have been omitted from these plots. Green areas denote where the y-axis model has a lower $|dU_{\text{inter}}|$. Histograms of each model's errors are shown on the corresponding axes. 481

N.1 Global Search landscape for MH-DAP. Matches to MH-I, MH-II, and MH-III are indicated with □, and labeled with their $RMSD_{20}$ scores. 484
N.2 Refined landscape for Static + FIT. Crystals post-clustering: 773. 485
N.3 Refined landscape for PCM_{11} + FIT. Crystals post-clustering: 772. 485
N.4 Refined landscape for Static + $DB1_{mp}$. Crystals post-clustering: 736. 486
N.5 Refined landscape for PCM_{11} + $DB14_{mp}$. Crystals post-clustering: 762. 486
N.6 Refined landscape for Static + $BT^{Static}_{dBhp,mp}$. Crystals post-clustering: 812. 487
N.7 Refined landscape for PCM_{11} + $BT^{PCM}_{dBhp,mp}$. Crystals post-clustering: 770. 487
N.8 Refined landscape for SCRP + $BT^{SCRP}_{dBhp,mp}$. Crystals post-clustering: 828. 488
N.9 Refined landscape for Static + $BT^{Static}_{Mie,mp}$. Crystals post-clustering: 793. 488
N.10 Refined landscape for PCM_{11} + $BT^{PCM}_{Mie,mp}$. Crystals post-clustering: 801. 489
N.11 Refined landscape for SCRP + $BT^{SCRP}_{Mie,mp}$. Crystals post-clustering: 836. 489
N.12 Top: convergence of the SCERP-embedded CSO–RM minimization (ε_{scrpRM} = 1×10^{-3} kJ/mol (dashed line) and k_{max} = 10) for the SCRP + $BT^{SCRP}_{Mie,mp}$ model. Bottom: the average number of QM computations (normalized by Z' = 2) to converge the SCERP process at iteration k. 490
N.13 The difference in $U^{*}_{inter,k}$ relative to $U^{*}_{inter,10}$ for structures that failed to converge in Figure N.12, using the SCRP + $BT^{SCRP}_{Mie,mp}$ model. Structures with oscillatory non-convergence are shown in color, while other failures are shown in gray. 490
N.14 F-IV's crystal structure over the periodic DFT-D minimization. (a) Initial structure, UREAXX26. (b) UREAXX26 (pink) overlaid with its counterpart after one round of periodic DFT-D optimization. (c) The first periodic DFT-D optimization structure (pink) overlaid with the final converged DFT-D structure. (d) The converged DFT-D structure overlaid with F-I, UREAXX25. 491
N.15 Converged LAM landscape for the Global Search of urea (2,588 LAMs). When evaluating DihX/DihY, all other dihedrals are kept at the midpoint of their respective search ranges shown in Table N.5. Magenta points show the LAM point placements. 493
N.16 Global Search landscape for urea. Matches to F-I, F-III, and F-IV are indicated in magenta, cyan, and green respectively, and labeled with their $RMSD_{20}$ scores. 495
N.17 Local minimization trajectories of urea F-I using Mode 1 (red), Mode 2 (blue), Mode 3 (orange), and Mode 4 (black) FD gradients: (a) the function value and gradient norm, and (b) the dof values. The yellow region in (b) indicate ±5° from the final Mode 4 dof value. 499
N.18 Local minimization trajectories of urea F-III using Mode 1 (red), Mode 2 (blue), Mode 3 (orange), and Mode 4 (black) FD gradients: (a) the function value and gradient norm, and (b) the dof values. The yellow region in (b) indicate ±5° from the final Mode 4 dof value. 500

N.19 Local minimization trajectories of urea F-IV using Mode 1 (red), Mode 2 (blue), Mode 3 (orange), and Mode 4 (black) FD gradients: (a) the function value and gradient norm, and (b) the dof values. The yellow region in (b) indicate ±5° from the final Mode 4 dof value. 501

N.20 Minimization trajectories of urea Refinement II by FD Mode 2: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $\|\boldsymbol{g}(x)\|$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★. 502

N.21 Minimization trajectories of urea Refinement II by FD Mode 3: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $\|\boldsymbol{g}(x)\|$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★. 502

N.22 Minimization trajectories of urea Refinement II by FD Mode 4: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $\|\boldsymbol{g}(x)\|$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★. 503

N.23 Comparison of optimized urea U_{latt} obtained using (a) Mode 2 and (b) Mode 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$. All comparisons are on a structure-by-structure basis. The yellow region indicates $|\Delta U_{\text{latt}}| \leq 0.5$ kJ/mol. 503

N.24 Converged LAM landscape for the Global Search of glycine (548 LAMs). Magenta points show the LAM point placements. 505

N.25 Global Search landscape for glycine. Matches to α, β, γ, δ and ε are indicated in magenta, cyan, green, orange, and blue respectively, and labeled with their RMSD_{20} scores. 507

N.26 ESP exerted on atom i of δ- (top) and ε-glycine (bottom) at different real-space cutoffs. All ESPs are reported as a ratio with respect to the ESP at $r = 20$ Å. . . 510

N.27 ESP exerted on atom i of KUXPUP (top), and KUXPUP01 (bottom) at different real-space cutoffs. All ESPs are reported as a ratio with respect to the ESP at $r = 20$ Å. 510

N.28 Minimization trajectories of glycine Refinement II by FD Mode 2: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $\|\boldsymbol{g}(x)\|$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★. 511

N.29 Minimization trajectories of glycine Refinement II by FD Mode 3: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $\|\boldsymbol{g}(x)\|$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★. 512

N.30 Minimization trajectories of glycine Refinement II by FD Mode 4: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $\|\boldsymbol{g}(x)\|$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★. 512

N.31 Comparison of optimized glycine U_{latt} obtained using (a) Mode 2 and (b) Mode 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$. All comparisons are on a structure-by-structure basis. The yellow region indicates $|\Delta U_{\text{latt}}| \leq 0.5$ kJ/mol. 513

N.32 Converged LAM landscape for the Global Search of L-proline (226 LAMs). Magenta points show the LAM point placements. 515

N.33 Global Search landscape for proline. Matches to L-I and DL-I are indicated in magenta and cyan respectively, and labeled with their RMSD_{20} scores. 517

N.34 Minimization trajectories of proline Refinement II by FD Mode 2: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $\|\boldsymbol{g}(x)\|$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★. 521

N.35 Minimization trajectories of proline Refinement II by FD Mode 3: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $\|\boldsymbol{g}(x)\|$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★. 521

N.36 Minimization trajectories of proline Refinement II by FD Mode 4: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $\|\boldsymbol{g}(x)\|$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★. 522

N.37 Comparison of optimized proline U_{latt} obtained using (a) Mode 2 and (b) Mode 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$. All comparisons are on a structure-by-structure basis. The yellow region indicates $|\Delta U_{\text{latt}}| \leq 0.5$ kJ/mol. 522

N.38 Converged mNBA LAM landscape (1,578 LAMs) for the Global Search of CAFF-mNBA. When evaluating DihX/DihY, the third dihedral is kept at its search range's midpoint, as given in Table N.28. Magenta points show the LAM point placements. 524

N.39 Global Search landscape for CAFF-mNBA. Matches to F-I and F-II are indicated in magenta and cyan respectively, and labeled with their RMSD_{20} scores. 526

N.40 Minimization trajectories of CAFF-mNBA Refinement II by FD Mode 2: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $\|\boldsymbol{g}(x)\|$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★. 529

N.41 Minimization trajectories of CAFF-mNBA Refinement II by FD Mode 3: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $\|\boldsymbol{g}(x)\|$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★. 529
N.42 Minimization trajectories of CAFF-mNBA Refinement II by FD Mode 4: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $\|\boldsymbol{g}(x)\|$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★. 530
N.43 Comparison of optimized CAFF-mNBA U_{latt} obtained using (a) Mode 2 and (b) Mode 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$. All comparisons are on a structure-by-structure basis. The yellow region indicates $|\Delta U_{\text{latt}}| \leq 0.5$ kJ/mol. 531
N.44 Refined landscape of CAFF-mNBA using Model 4, with Mode 2 (left) and Mode 3 (right) FD gradient approximations. The lowest-energy match to F-I and F-II are labeled as ◇ and □ respectively, with their corresponding RMSD_{20} scores. . 531

List of Tables

2.1 Summary of some transferable HAIEFF FFs and their parametrization strategies. All FFs use the Buckingham pair potential as shown in Equation 2.4.10. 32

3.1 Seven lattice types and their corresponding constraints. 49

4.1 E04UFF gradient test results for CPv2.4.4 and CPv2.5.2.1. The number of satisfactory $\frac{\partial F}{\partial \theta^{\text{ind}}}$ elements, determined by the E04UFF subroutine, are reported. . . 88

4.2 Average Global Searches statistics for RUTTEH and CANNOL: number of failed $\boldsymbol{x}$ to reach a valid candidate, time spent generating the candidate, time spent minimizing the candidate, overall production rate, and proportion of minimizations returning stable minima. 89

5.1 Description of the atomic environments (AEs) used for diversity analysis. 108

5.2 List of atom environment reference datasets obtained from DFT-D computations. 111

5.3 Total occurrences of each hydrogen bonding motif in the CE–755. A single motif can occur multiple times in a given structure. 116

5.4 Comparison of experimentally-derived lattice energies ($U_{\text{latt}}^{\text{exp}} \approx -\Delta H_{\text{sub}}^{\text{exp}} - 2RT$) with the DFT-D computed lattice energies ($U_{\text{latt}}^{\text{DFT}}$) for non-hydrogen bonding structures. The number of measurements associated with each structure is given (#), and if multiple sources give differing experimental values, both $U_{\text{latt}}^{\text{exp}}$ and the deviations in U_{latt} ($\Delta U_{\text{latt}} = U_{\text{latt}}^{\text{exp}} - U_{\text{latt}}^{\text{DFT}}$) are reported as ranges. Positive values of ΔU_{latt} indicate overbinding. 118

5.5 Comparison of experimentally-derived lattice energies ($U_{\text{latt}}^{\text{exp}} \approx -\Delta H_{\text{sub}}^{\text{exp}} - 2RT$) with the DFT-D computed lattice energies ($U_{\text{latt}}^{\text{DFT}}$) for hydrogen bonding structures. The number of measurements associated with each structure is given (#), and if multiple sources give differing experimental values, both $U_{\text{latt}}^{\text{exp}}$ and the deviations in U_{latt} ($\Delta U_{\text{latt}} = U_{\text{latt}}^{\text{exp}} - U_{\text{latt}}^{\text{DFT}}$) are reported as ranges. Positive values of ΔU_{latt} indicate overbinding. 119

6.1 Summary of CrystalEstimator (CE) program settings utilized. 129

6.2 Summary of the parameter estimation schemes investigated. 134

6.3 p-values for the Mann–Whitney U test ('MANWHIT'), the 'standard' t-test (HOMO-T), and Welch's t-test (HETERO-T), comparing the distributions of RMSD_{15} and dU_{inter} between the training and test sets. At a 5% significance level, the H_0 is accepted if $p > 0.05$. 148

6.4 Summary of the key performance metrics in the prediction of ΔU_{poly} for fitting Schemes 5 and 7, over the different models. These include the rank accuracy ('Rank Acc.', i.e., number of correctly ranked pairs), number of pairs with $|d\Delta U_{\mathrm{poly},i}| \leq 5$ kJ/mol ('Within 5 kJ/mol'), the AD in ΔU_{poly}, AAD in ΔU_{poly}, and the minimum/maximum $d\Delta U_{\mathrm{poly}}$ (Min/Max D respectively). 154

6.5 Summary of the key error metrics over the test set structures, for the different models tested. Metrics include the number of failed geometries, average RMSD_{15} ($\overline{\mathrm{RMSD}_{15}}$), AD in U_{inter}, AAD in U_{inter}, and SD in dU_{inter}. All errors metrics are relative to the CE–755 reference data. Scores are split as "NHB/HB". The first row contains the statistics for the reference model, which evaluates the test set using the same *ab initio* model applied for training (PBE0/6-311G(d,p), rank 4 multipoles, MIP, PCM with $\varepsilon_{\mathrm{PCM}} = 11.0$). 163

7.1 SCERP process convergence test specifications. 171

7.2 SCERP-field convergence test specifications. 174

7.3 Errors for the SCERP-field convergence test, using the 50 Å embedding as a reference. 175

7.4 Induction-coupled Electrostatic Ewald summation test specifications. 177

7.5 U_{real}, U_{rec}, and U_{corr} determined from single-point evaluations of PYRENE03 with a point-charge model, using a complete Ewald sum. U_{sum} is the sum of these three contributions. 178

7.6 FD modes for SCERP-embedded CSO–FM. For all modes, $\boldsymbol{\theta}^{\mathrm{dep}}$ and $\boldsymbol{Q}$ are updated by their respective LAMs during the FD computation (as they would be without induction). 192

8.1 Training sets for SCERP-compatible FF fitting. The number of training structures (N_{TRN}), atom types present, number of fitted parameters, and number of Sobol' seeds are shown. Values for Schemes 5 and 7 (only in the final fitting) are given before and after '/', respectively. 201

8.2 Summary of CrystalEstimator (CE) program settings used in SCERP-compatible FF parametrization. 202

8.3 Bondi atomic radii [3] used for each atom type, t. 202

8.4 Scheme 5 FF parameter estimations attempted in the trial FF fitting problems. . 203

8.5 Error metrics for the models trained on NHB47 and HB53. The metrics are reported for the minimized structures post-CE, and (for SCERP-embedded models) post-CSORM. The latter values are given after the '/'. Root-mean-square-deviations (between CE and CSORM) in $RMSD_{15}$ and dU_{inter} scores are also reported for the SCERP-embedded models. The numbers of structures with $|\Delta(RMSD_{15})_{CE\rightarrow CSORM}| \geq 0.2$ Å and $|\Delta(dU_{inter})_{CE\rightarrow CSORM}| \geq 1$ kJ/mol are given in parenthesis. $dU_{inter} = U_{inter}^{calc} - U_{inter}^{ref}$, such that $dU_{inter} > 0$ indicates underbinding. 206

8.6 Error metrics for the models trained on TP100. The metrics are reported for the minimized structures post-CE, and (for SCERP-embedded models) post-CSORM. The latter values are given after the '/'. Root-mean-square-deviations (between CE and CSORM) in $RMSD_{15}$ and dU_{inter} scores are also reported for the SCERP-embedded models. The numbers of structures with $|\Delta(RMSD_{15})_{CE\rightarrow CSORM}| \geq 0.2$ Å and $|\Delta(dU_{inter})_{CE\rightarrow CSORM}| \geq 1$ kJ/mol are given in parenthesis. $dU_{inter} = U_{inter}^{calc} - U_{inter}^{ref}$, such that $dU_{inter} > 0$ indicates underbinding. 209

8.7 dU_{inter} AAD and SD using PCM_{11} and $SCRP_{atm,90}$ for the CE–755 training set, omitting the outliers: CYCYPR, KUXPUP01, CLTRZL, and GEYSEI. 217

8.8 AAD and SD metrics for dU_{inter} in the CE–755 training and test set, using $SCRP_{atm,90}$. Outliers have been omitted from the corresponding training (CYCYPR, KUXPUP01, CLTRZL, and GEYSEI) and test set (EREFEP) statistics. $\Delta X = X_{TST} - X_{TRN}$ are given in parenthesis. 220

8.9 Summary of the key performance metrics in the prediction of ΔU_{poly} for fitting Schemes 5 and 7 over the different multipole models. These include the rank accuracy ('Rank Acc.', i.e., number of correctly ranked pairs), number of pairs with $|d\Delta U_{poly,i}| \leq 5$ kJ/mol ('Within 5 kJ/mol'), the AD in ΔU_{poly}, AAD in ΔU_{poly}, and the minimum/maximum $d\Delta U_{poly}$ (Min/Max D respectively). 221

9.1 Polymorphs of MH-DAP. *Z'/Z* include the active ingredient and water molecule. $RMSD_{15}$ values are obtained by comparing the starting and optimized crystals, and ΔU_{latt}^{GM} values are in kJ/mol. 228

9.2 Models used in the refinement of MH-DAP. All QM calculations are done at the PBE0/6-311G(d,p) level of theory. 228

9.3 Total CPU time (in secs) at each stage of a lattice energy minimization, averaged over all crystals minimized (3 experimental structures for periodic DFT-D, and 890 otherwise). 231

9.4 Polymorphs of Urea. $RMSD_{20}$ values are obtained by comparing the starting and optimized crystals, and ΔU_{latt}^{GM} values are in kJ/mol. 233

9.5 Models used in the refinement of urea. All QM calculations are done at the PBE0/6-311G(d,p) level of theory. 233

9.6 COMPACK comparison of refined crystals from Model 3 and 4, against F-I and F-IV. 235

9.7 Minimization[a] of urea by FD gradient Modes 1, 2, 3, and 4. Converged runs are denoted by '*'; other runs terminate from walltime or E04UFF failing to progress. All values are from the last E04UFF major iteration, except $||\boldsymbol{g}(\boldsymbol{x})||_0$ which is from the zeroth iteration. 237

9.8 Cost metrics of the urea local minimization by FD gradient Modes 1, 2, 3, and 4. Definitions for each metric are given in the main text. 238

9.9 Convergence behavior during urea Refinement II, using FD Modes 2–4. 239

9.10 Optimized U_{latt} and geometries for urea predicted by Modes 2 and 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$, while RMSD_{20} scores are generated by the COMPACK algorithm. All comparisons are done on a structure-by-structure basis. 239

9.11 Polymorphs of Glycine. RMSD_{20} values are obtained by comparing the starting and optimized crystals, and $\Delta U_{\text{latt}}^{\text{GM}}$ values are in kJ/mol. 241

9.12 Models used in the refinement of glycine. All QM calculations are done at the PBE0/6-311G(d,p) level of theory. 242

9.13 Polar unit cell classification of several crystal structures. 244

9.14 Convergence behavior during glycine Refinement II, using FD Modes 2–4. 247

9.15 Optimized U_{latt} and geometries for glycine predicted by Modes 2 and 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$, while RMSD_{20} scores are generated by the COMPACK algorithm. All comparisons are done on a structure-by-structure basis. 247

9.16 Polymorphs of Proline. RMSD_{20} values are obtained by comparing the starting and optimized crystals, and $\Delta U_{\text{latt}}^{\text{GM}}$ values are in kJ/mol. 'L' denotes the enantiopure Laevus form; 'DL' denotes the racemic form. 249

9.17 Models used in the refinement of proline. All QM calculations are done at the PBE0/6-311G(d,p) level of theory. 250

9.18 Convergence behavior during proline Refinement II, using FD Modes 2–4. 254

9.19 Optimized U_{latt} and geometries for proline predicted by Modes 2 and 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$, while RMSD_{20} scores are generated by the COMPACK algorithm. All comparisons are done on a structure-by-structure basis. 254

9.20 Polymorphs of the 1:1 CAFF + mNBA cocrystal. *Z'/Z* include both molecular species. RMSD_{20} values are obtained by comparing the starting and optimized crystals, and $\Delta U_{\text{latt}}^{\text{GM}}$ values are in kJ/mol. 256

9.21 Models used in the refinement of CAFF-mNBA. All QM calculations are done at the PBE0/6-311G(d,p) level of theory. 258

9.22 Summary of the case study results, ignoring forms that present with the "polar cell problem". Models 3 and 4 correspond to the the HAIEFF models employing the $\text{BT}_{\text{Mie,mp}}^{\text{PCM}}$ and $\text{BT}_{\text{Mie,mp}}^{\text{SCRP}}$ FF respectively. 261

A.1 Analytical solutions for $\Gamma(z)$ at different values of $l_1 + l_2$. 300

C.1 Analytical expressions for $\nabla_X \boldsymbol{R}$. 304
C.2 Analytical expressions for $\nabla_X \boldsymbol{R}^{-1}$. 305
C.3 Analytical expressions for $\nabla_X \nabla_{X'} \boldsymbol{R} \; \forall \; X, X' \in \{a, b, c, \alpha, \beta, \gamma\}$. Note that $\nabla_X \nabla_{X'} \boldsymbol{R} = \nabla_{X'} \nabla_X \boldsymbol{R}$. 306
C.4 Analytical expressions for $\nabla_X \nabla_{X'} \boldsymbol{R}^{-1} \; \forall \; X, X' \in \{a, b, c, \alpha, \beta, \gamma\}$. Note that $\nabla_X \nabla_{X'} \boldsymbol{R}^{-1} = \nabla_{X'} \nabla_X \boldsymbol{R}^{-1}$. 308

E.1 Summary of the relevant analytical equations (given by their number) for each problem formulation. The solutions are split based on: their space group constraints (SG-Con), whether any expansion factor is constrained (Con-x, y, z), as well as if p, q, r are zero-valued. Up to two sets of solutions are possible. Within these solutions, constrained values are indicated as 'Con. 1'. This is distinguished from cases where the expansion value being 1 arises from the solution itself (indicated as '= 1') 374

F.1 Comparison of experimentally-derived lattice energies ($U_{\text{latt}}^{\text{exp}} \approx -\Delta H_{\text{sub}}^{\text{exp}} - 2RT$) with the DFT-D computed lattice energies ($U_{\text{latt}}^{\text{DFT}}$). The errors are computed as ($\Delta U_{\text{latt}} = U_{\text{latt}}^{\text{exp}} - U_{\text{latt}}^{\text{DFT}}$). In some cases, the same structure is referenced using different Refcodes in the CE–755 database and in the experimental data. 382

G.1 Training set structures partitioned from the CE–755. This comprises 604 crystal structures: 356 non-hydrogen bonding (NHB) and 248 hydrogen bonding (HB) structures. No multicomponent crystals are included in the dataset, hence any structure with $Z' > 1$ has repeated units of the same molecule. 386
G.2 Test set structures partitioned from the CE–755. This comprises 151 crystal structures: 85 non-hydrogen bonding (NHB) and 66 hydrogen bonding (HB) structures. No multicomponent crystals are included in the dataset, hence any structure with $Z' > 1$ has repeated units of the same molecule. 402

I.1 FIT pair potential [4] parameters used in this work. Note that three hydrogen types are defined: a 'general' hydrogen (H), a nitrogen-bonded hydrogen (H_N), and an oxygen-bonded hydrogen (H_O). If only polarized hydrogens (H_p) are used, these assume the values of H_N. 410
I.2 Mie-analogue of the FIT pair potential [4] used in this work. Three hydrogen types are defined: a 'general' hydrogen (H), a nitrogen-bonded hydrogen (H_N), and an oxygen-bonded hydrogen (H_O). 412
I.3 Values of $\sigma_{\text{lbd}}^{t,t'}$ determined from $\sigma^{\text{FIT},t,t'}$ and $\sigma^{\text{CE755},t,t'}$. 414

J.1 Optimized FF for point-charge electrostatics without PCM, by Scheme 1. 416
J.2 Optimized FF for point-charge electrostatics without PCM, by Scheme 2. 417

J.3 Optimized FF for point-charge electrostatics without PCM, by Scheme 4. 418
J.4 Optimized FF for point-charge electrostatics without PCM, by Scheme 5. 419
J.5 Optimized FF for point-charge electrostatics without PCM, by Scheme 6. 420
J.6 Optimized FF for point-charge electrostatics without PCM, by Scheme 7. 421
J.7 Optimized FF for point-charge electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 1.423
J.8 Optimized FF for point-charge electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 2.424
J.9 Optimized FF for point-charge electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 3.425
J.10 Optimized FF for point-charge electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 4.426
J.11 Optimized FF for point-charge electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 5.427
J.12 Optimized FF for point-charge electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 6.428
J.13 Optimized FF for point-charge electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 7.429
J.14 Optimized FF for multipole electrostatics without PCM, by Scheme 1. 431
J.15 Optimized FF for multipole electrostatics without PCM, by Scheme 2. 432
J.16 Optimized FF for multipole electrostatics without PCM, by Scheme 3. 433
J.17 Optimized FF for multipole electrostatics without PCM, by Scheme 4. 434
J.18 Optimized FF for multipole electrostatics without PCM, by Scheme 5. 435
J.19 Optimized FF for multipole electrostatics without PCM, by Scheme 6. 436
J.20 Optimized FF for multipole electrostatics without PCM, by Scheme 7. 437
J.21 Optimized FF for multipole electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 1. . 439
J.22 Optimized FF for multipole electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 2. . 440
J.23 Optimized FF for multipole electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 3. . 441
J.24 Optimized FF for multipole electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 4. . 442
J.25 Optimized FF for multipole electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 5. . 443
J.26 Optimized FF for multipole electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 6. . 444
J.27 Optimized FF for multipole electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 7. . 445

L.1 Atom-pair specific Tang–Toennies damping parameters. All damping parameters are reported in units of bohr^{-1} 458

M.1 Training set structures partitioned from the CE–755 to form TP442. This comprises 442 crystal structures: 317 non-hydrogen bonding (NHB) and 125 hydrogen bonding (HB) structures. No multicomponent crystals are included in the dataset, hence any structure with $Z' > 1$ has repeated units of the same molecule. The NHB47, HB53, and TP100 datasets are subsets of this larger set. 460

M.2 Error metrics for the models trained on TP442. The metrics are reported for the minimized structures post-CE, and (for SCERP-embedded models) post-CSORM. The latter values are given after the '/'. Root-mean-square-deviations (between CE and CSORM) in $RMSD_{15}$ and dU_{inter} scores are also reported for the SCERP-embedded models. The numbers of structures with $|\Delta(RMSD_{15})_{CE\rightarrow CSORM}| \geq 0.2$ Å and $|\Delta(dU_{inter})_{CE\rightarrow CSORM}| \geq 1$ kJ/mol are given in parenthesis. $dU_{inter} = U_{inter}^{calc} - U_{inter}^{ref}$, such that $dU_{inter} > 0$ indicates underbinding. 474
M.3 Optimized FF for multipole electrostatics with $SCRP_{atm,90}$, by Scheme 5. 476
M.4 Optimized FF for multipole electrostatics with $SCRP_{atm,90}$, by Scheme 7. 478

N.1 Gas-phase conformational energies for the MH-DAP system. All computations are done at the PBE0/6-311G(d,p) level of theory, with $\varepsilon_{PCM} = 11$, in G09-d01. . 482
N.2 Local minimization of the MH-DAP experimental structures with the Global Search energy model, using CrystalPredictorv2.5.2.1. 483
N.3 MH-DAP Global Search, conducted with CrystalPredictorv2.5.2.1 and G09-d01. 484
N.4 Gas-phase conformational energies for urea. All computations are done at the PBE0/6-311G(d,p) level of theory in G09-d01. 492
N.5 Flexible definition of the urea molecule used in Global Search LAM construction. Site labels correspond to the labeling shown in Figure 9.3. 492
N.6 Adaptive LAM generation settings applied in the Global Search LAMs for urea. 493
N.7 Urea Global Search, conducted with CrystalPredictorv2.5.2.0 and G16-c01-avx2. 494
N.8 Refinement I of urea, conducted with CSO–FMv1.3.1, G09-d01, and GDMAv2.2. Model numbering corresponds to those shown in Table 9.5. 496
N.9 Refinement II of urea, conducted with CSO–FMv1.4.0.2, G09-d01, and GDMAv2.2. These settings correspond to Model 4 in Table 9.5. 497
N.10 COMPACK comparison between the Model 4 structures before and after local optimization by periodic DFT-D. Comparisons of the optimized structures to the experimental forms of urea are also presented. Blank entries indicate no valid $RMSD_{20}$ score could be generated. 498
N.11 Gas-phase conformational energies for glycine. All computations are done at the PBE0/6-311G(d,p) level of theory in G16-c01-avx2. 504
N.12 Flexible definition of the glycine molecule used in Global Search LAM construction. Site labels correspond to the labeling shown in Figure 9.7. 504
N.13 Adaptive LAM generation settings applied in the Global Search LAMs for glycine. 505
N.14 Local minimization of the glycine experimental structures with the Global Search energy model, using CrystalPredictorv2.5.2.1. 506
N.15 Glycine Global Search, conducted with CrystalPredictorv2.5.2.1 and G16-c01-avx2. 507
N.16 Refinement I of glycine, conducted with CSO–FMv1.3.1, G09-d01, and GDMAv2.2. Model numbering corresponds to those shown in Table 9.12. 508

N.17 Refinement II of glycine, conducted with CSO–FMv1.4.0.2, G09-d01, and GDMAv2.2. These settings correspond to Model 4 in Table 9.12. 509
N.18 COMPACK comparison between the Model 4 structures before and after local optimization by periodic DFT-D. Comparisons of the optimized structures to the experimental forms of glycine are also presented. Blank entries indicate no valid $RMSD_{20}$ score could be generated. 511
N.19 Gas-phase conformational energies for L-proline. All computations are done at the PBE0/6-311G(d,p) level of theory in G16-c01-avx2. 514
N.20 Flexible definition of the L-proline molecule used in Global Search LAM construction. Site labels correspond to the labeling shown in Figure 9.12. 514
N.21 Adaptive LAM generation settings applied in the Global Search LAMs for L-proline. 515
N.22 Local minimization of the L-/DL-proline experimental structures with the Global Search energy model, using CrystalPredictorv2.5.2.1. 516
N.23 Proline Global Search, conducted with CrystalPredictorv2.5.2.1 and G16-c01-avx2. 517
N.24 Refinement I of proline, conducted with CSO–FMv1.3.1, G09-d01, and GDMAv2.2. Model numbering corresponds to those shown in Table 9.17. 518
N.25 Refinement II of proline, conducted with CSO–FMv1.4.0.2, G09-d01, and GDMAv2.2. These settings correspond to Model 4 in Table 9.17. 519
N.26 COMPACK comparison between the Model 4 structures before and after local optimization by periodic DFT-D. Comparisons of the optimized structures to the experimental forms of proline are also presented. Blank entries indicate no valid $RMSD_{20}$ score could be generated. 520
N.27 Gas-phase conformational energies for CAFF and mNBA. All computations are done at the PBE0/6-311G(d,p) level of theory in G16-c01-avx2. 523
N.28 Flexible definition of the mNBA molecule used in Global Search LAM construction. Site labels correspond to the labeling shown in Figure 9.16. 523
N.29 Adaptive LAM generation settings applied in the Global Search LAMs for mNBA. 524
N.30 Local minimization of the CAFF-mNBA experimental structures with the Global Search energy model, using CrystalPredictorv2.5.2.1. 525
N.31 CAFF-mNBA Global Search with CrystalPredictorv2.5.2.1 and G16-c01-avx2. . 526
N.32 Refinement I of CAFF-mNBA, conducted with CSO–FMv1.3.2, G09-d01, and GDMAv2.2. Model numbering corresponds to those shown in Table 9.21. 527
N.33 Refinement II of CAFF-mNBA, conducted with CSO–FMv1.4.0.2, G09-d01, and GDMAv2.2. These settings correspond to Model 4 in Table 9.21. 528
N.34 Optimised U_{latt} and geometries for CAFF-mNBA predicted by Modes 2 and 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\mathrm{latt}}^{\mathrm{Mode\,2/3}} - U_{\mathrm{latt}}^{\mathrm{Mode\,4}}$, while $RMSD_{20}$ scores are generated by the COMPACK algorithm. All comparisons are on a structure-by-structure basis. 530

Chapter 1

Introduction

1.1 Background and Motivation

1.1.1 Organic Molecular Crystals

Organic crystalline materials are produced in many major chemical industries, as manufacturing intermediates and marketed products. In the pharmaceutical industry, it has been estimated that > 90% of small molecular drugs are sold as crystalline products [5]. Similarly, in agrochemicals, active ingredients for crop protection are most commonly marketed as solid formulations such as suspension concentrates and wettable powders [6]. Organic pigment colorants used to color paints, cosmetics, and plastics, come in the form of insoluble particulates that are applied via a liquid dispersing agent [7, 8]. Aside from these industries, organic crystals are also used in explosive materials [9, 10], and semiconductors [11]. Clearly, organic crystals are already immensely important, but with growing interest towards multicomponent crystals, their significance is likely to grow. Multicomponent crystals encompass solids containing more than one chemically distinct entity, such as salts, cocrystals, solvates, and hydrates. These systems present novel opportunities to tune the useful traits of organic crystals, whilst suppressing undesirable properties. For instance, the pharmaceutical industry has historically used salt formulations to bolster drug solubility and bioavailability [12]. More recently, cocrystals have gained attention since not all molecules easily ionize. Cocrystals have been shown to enhance drug analgesic effects [13, 14], tune bioavailability [12, 13], and improve drug stability towards moisture [15]. Lastly, hydrates and solvates are typically simple by-products of crystallization processes. However, understanding their formation and properties is still critical, especially for hydrates which can inadvertently form from moisture in the environment [16]. Overall, organic crystals are undoubtedly an important class of material, and their significance is likely to only grow as industries continue to explore the expanding chemical space of multicomponent crystals.

1.1.2 Polymorphism in Organic Crystals

As alluded to, the usefulness of organic crystals is closely tied to their physicochemical properties. In drugs, bioavailability and water solubility control their release rate and efficacy once administered. In pigments, other than color, insolubility of the crystalline material is also im-

portant to maintain the desired coloration. Crystal flow behavior, which is influenced by crystal morphology, can be important for both drug and agrochemical applications since this determines the necessary solid delivery method. Finally, in drugs, the mechanical properties of the crystal dictate the ability to compact the solid and form tablets. All of these properties are inherently linked to the crystal structure of the solid (i.e., the three-dimensional arrangement of atoms in the crystalline lattice), and for this reason it is vexing that many organic crystals exhibit polymorphism. Polymorphism is defined as the ability for a given solid material to crystallize in a multiplicity of packing arrangements, termed polymorphs [17]. It was once thought that polymorphic tendencies were a rarity. However today, we have a large (and ever-growing) body of evidence suggesting the contrary: polymorphism is the rule and not the exception [18–22]. This near-ubiquity is captured by the famous words of Walter McCrone [23]:

"*It is at least this author's opinion that every compound has different polymorphic forms and that, in general, the number of forms known for a given compound is proportional to the time and energy spent in research on that compound.*"

Due to their distinct crystal structures, polymorphs of the same material can have very different properties, despite sharing the same chemical identities. This is not inherently problematic, if the desired polymorphic form could be reliably crystallized and maintained in that state. Unfortunately, this is more challenging than it sounds. Over time, all metastable polymorphs will spontaneously transform into thermodynamically more-stable forms that are lower in free energy, until they reach the most-stable form. Determining the relative stability of different forms is therefore vital, but this is complicated by several factors.

Firstly, the relative stability of polymorphs can depend on the conditions of the system. For instance, polymorphs can exhibit enantiotropic/temperature-dependent stability relationships [24–26], and/or pressure-dependent polymorphic phase-transitions [27, 28]. Additionally, multicomponent crystals must also consider if there is sufficient thermodynamic driving force to favor multicomponent crystallization, over neat crystallization [29–32]. Even if the correct thermodynamic conditions are known, wherein the desired polymorph is most-stable, there is a delicate interplay with kinetic factors, such that less-stable/metastable forms tend to crystallize out first, sometimes referred to as 'Ostwald's rule of stages' [33]. These metastable states can be kinetically trapped for prolonged durations, before spontaneously (and unpredictably) transforming into more-stable forms (termed 'late-appearing polymorphs').

Failing to characterize the risk of polymorphism can result in dramatic repercussions for all stakeholders in the chemical supply chain [34–40]. Perhaps the most notorious example of this was the case of ritonavir, a HIV-protease inhibitor that was brought to market in 1996. In 1998, a majority of the marketed drug began displaying a significant loss in efficacy. After thorough investigation, this was attributed to the sudden transformation of the formulated solid form (denoted Form I) into a more stable, and significantly less soluble, polymorph (denoted Form II) [35]. The emergence of Form II also led to Form I 'disappearing': manufacturing sites could no longer reliably produce Form I. If any remnants of Form II were present on-site, they would

act as seeds for crystallization directly to the more-stable form, bypassing Form I altogether [34]. Ritonavir had to be withdrawn from the market and could only be reintroduced after the development of new manufacturing routes to suppress Form II formation. For the manufacturer, the financial cost of this debacle was estimated at about US$250 million [34], yet this neglects the social cost borne by consumers who relied on the drug. Since then, three other forms of ritonavir have been found [37]. In response to the ritonavir case, systematic changes were implemented in hopes of addressing the polymorph-problem. Yet, in 2009 a similar case of late-appearing polymorphism occurred for rotigotine [38, 39], a Parkinson's disease therapeutic, reminding us that polymorphism is still a lurking problem.

Late-appearing and unforeseen polymorphic transformations are only one facet of polymorphism. When properly characterized and managed, polymorphism can also offer opportunity for the design and engineering of more-effective crystalline products. For instance, if kinetic inertness can be elicited, metastable polymorphs with superior physicochemical properties could be employed in formulations [40]. With this, legislative and intellectual property issues related to the patenting of different polymorphic modifications will also become pertinent [34].

1.1.3 Crystal Structure Prediction

For better or for worse, polymorphism is here to stay. Therefore, there is a need for methods to determine a system's propensity towards polymorphism, and the relative stability of its polymorphs. From a statistical analysis of the Cambridge Crystallographic Data Centre's (CCDC's) Crystal Structure Database (CSD), Cruz-Cabeza *et al.* suggested that chemical/molecular features are poor indicators for polymorphic behavior [21], precluding the development of screening heuristics. As such, experimental polymorph screens remain the primary mode of characterization. This entails subjecting the system of interest to a barrage of crystallization experiments, spanning various design parameters: temperature, pressure, solvent/antisolvent choice, composition, crystallization method, etc. Since it is unknown *a priori* how many (if any) polymorphs can be observed and under what conditions, a 'brute force' approach is the only option [41]. This experimental effort is further compounded in the context of multicomponent crystals, where this search must be repeated across different coformers/ion-pairs/solvents. In practice, experimental polymorph screens are severely limited by the availability of time and resource, and one rarely obtains a definitive answer to the question of polymorphism in this way. These constraints can be particularly acute early in the product design cycle, where material available for polymorph screening may be scarce, and there is time pressure to advance development [16]. Yet, it is also at these initial stages, when investment into the project is lowest, that one hopes to identify any threats of polymorphism and 'pull-the-plug' if necessary. It is these shortcomings that have inspired the development of computational methods, like Crystal Structure Prediction (CSP).

As outlined by Nyman [41], early concepts of CSP were attempted long before the repercussions of polymorphism had been fully appreciated, likely driven by scientific interest more so than necessity. Today, CSP primarily seeks to predict the structure and relative stability

of all observable crystal structures, from only a description of their molecular connectivity and stoichiometry [42]. This modern goal of CSP was likely formalized around the time of the first CCDC Blind Test [43]. During the Blind Test, participants are tasked with predicting (and ranking) the crystal structures of known polymorphs for several systems, starting from only their molecular diagrams. No additional experimental data is provided, hence the 'blind' nature of the test. Between 2001–2020, six more Blind Tests have been conducted [44–49].

The Blind Tests provide invaluable insight on the current state of CSP methods and their capabilities. Over the seven tests, the target systems have progressed from small, rigid molecules to increasingly large, flexible (sometimes multicomponent) systems, indicating the progress made in CSP methods. Part of this growing success stems from the collective recognition that a staged CSP workflow promotes better CSP outcomes [50, 51]. Along those lines, a more realistic definition of 'success' has also emerged from the Blind Tests. The stated goal of CSP alludes to a definitive answer to the polymorph problem, replacing the need for experimental screens. However, most CSP practitioners today acknowledge the extreme difficulty of this feat, and instead tout CSP as a *complementary* tool, which can provide evidence in support of the completeness (or otherwise) of an experimental screen. Moreover, CSP-generated results are increasingly finding value beyond just evaluating *known* forms. Assessing entire landscapes of theoretical crystal structures can offer guidance into solid state behavior, and open opportunities for the discovery and engineering of crystals with more-desirable properties. These trends are reflected in the Blind Tests. Earlier submissions required participants to submit a finalized list of only three most-likely candidates for the experimental forms. In the Sixth and Seventh tests, there were separate submission stages (reflecting the stages of a typical CSP workflow), each comprising much larger lists of structures than would be realistically observed.

While the Blind Tests do suggest a positive trajectory in CSP capabilities, they also indicate that CSP still does not reliably fulfill its intended goals. Nevertheless, with suitable expertise, useful insights can still be derived from imperfect CSP results [41, 52]. Multiple examples from Academia and Industry [16, 41, 51, 52] illustrate how CSP can be applied to support polymorph screens. In studies of anhydrous creatine [53] and 4-aminoquinaldine [54], CSP proposed novel crystals that were more-stable than the known forms at the time. These new forms were subsequently crystallized and shown to have stability relationships in agreement with theory. Similarly, an early CSP conducted on the ROY molecule produced a solid form landscape containing all 10 known forms [55], three of which had not been experimentally observed at the time. ROY was later revisited [56], from which close alignment with experiments was achieved by a more-sophisticated CSP model. CSP has also been used to validate that the known form of isoniazid is the most stable [57], and experimental efforts could potentially have concluded earlier. These examples demonstrate how CSP can help rationalize if more experimental screens are necessary. Moreover, judicious analysis of the solid form landscape can also inform how novel forms may be accessed experimentally, for example through extreme conditions (e.g., high-pressure [54, 56, 57]) or non-standard techniques (e.g., isostructural-templating [58]).

Beyond its role in polymorph screens, CSP has also demonstrated value for material discov-

ery. Where crystal structures of observed polymorphs cannot be resolved by experiment (usually due to difficulty in isolating suitable single-crystals), the theoretical solid form landscape can be used for indirect structure determination via other experimental data like powder X-ray diffraction (pXRD) [53, 59, 60]. In multicomponent systems, *in silico* coformer screens are possible by assessment of the thermodynamic competition between neat and cocrystal CSP landscapes [29–32]. For material design, CSP landscapes provide a rich source of theoretical structures that can be analyzed through structure–property maps to guide decision-making between different candidates [61, 62]. Unlike other crystal engineering approaches (e.g., supramolecular synthons [63, 64]), theoretical structures predicted by CSP have the added consideration of thermodynamic viability. Finally, if conducted *en masse*, it has been proposed that CSP studies can offer large theoretical datasets for enhancing our fundamental understanding of packing behavior and to train general machine-learning models [65].

From demonstrations like these, there is growing belief in CSP as a useful technology today. In fact, some industries have begun integrating CSP methods into their workflows [16, 41], and a market for CSP-services has emerged through commercial providers like Avant-garde Materials Simulations and XtalPi. With that said, as an applied technology, CSP is still in its infancy and continued effort is required to address its outstanding problems. First and foremost, is the accuracy attainable with the current suite of CSP methods. It is widely recognized that energy differences between observable polymorphs are in the range of only a few kJ/mol, or even sub-kJ/mol [21, 24]. However, even the best CSP models today have uncertainties in excess of this accuracy [66], precluding truly reliable ranking of relative stabilities. Moreover, most CSP practiced today still falls under what Price terms 'CSP_0' [52], wherein the effects of the thermodynamic state (temperature in particular) are neglected in the prediction of relative stability. As such, there is an inherent disconnect between the CSP_0 solid form landscape and a landscape at 'real-world' conditions. The neglect of temperature may not only lead to the incorrect assessment of relative stability [24, 67], but also manifests in the gross overprediction of metastable forms on the theoretical landscape [68, 69]. In addressing these issues, one must also be mindful of the trade-off between computational costs and model accuracy. Since the structures of interest are unknown *a priori*, exploration of the solid form landscape must necessarily be exhaustive (analogous to a brute-force experimental screen), and even small increases in computational cost can compound over an entire landscape. Without satisfactory resolution to these problems, the usefulness of CSP will undoubtedly be limited.

1.2 Thesis Scope and Outline

While each problem highlighted above is important in its own right, in this work I focus on developing the energy models used in the domain of CSP_0, specifically the class of hybrid *ab initio*/empirical force-field (HAIEFF) models [2, 70]. As its name suggests, HAIEFF models combine *ab initio* quantum mechanical (QM) calculations with empirical force-fields to describe the interactions in the crystal. These models have a long history in CSP [42], and their presence

in modern CSP is a testament to their effectiveness [47–49]. The continued use of HAIEFF models is partly due to their versatility in modulating computational cost and accuracy, making them a natural fit for the modern CSP workflow. While these models fall short of the accuracy required to qualify as 'final answers' in the CSP workflow, their continued improvement is still desirable, since the success of the CSP workflow is cumulative over all stages. By raising the 'accuracy-floor' of the lower-fidelity models used early in the workflow, more-stringent criteria can be applied to the structures advanced to later stages. Ultimately, the high-accuracy models necessary for useful CSP will be practicable only if a solid form landscape can be reliably reduced to a handful of relevant structures.

Previous work in our group has independently considered different components of the HAIEFF model [2, 71], and herein I aim to unify these under a consistent framework. This pertains primarily to the intermolecular components of the HAIEFF model: the *ab initio*-derived electrostatic interactions, and the empirical force-field. For each interaction, I will study different physical features related to that component. In so doing, I hope to derive insight into the underlying physical nature of the interaction and maximize any improvements in model accuracy. In my study of the electrostatic component, I will also revisit the algorithmic implementation proposed by Zhang [71], so as to bolster computational performance. Overall, an underlying theme of this work will be the importance of congruity between the force-field and *ab initio* components.

The rest of this thesis is structured as follows. In Chapter 2, I will review the state of CSP, with focus on the energy models that are currently employed in the field. In Chapter 3, I will introduce the mathematical framework used in the description of crystal geometry and energy. In Chapter 4, I will present the implementation of these formulations within the CSP programs developed by our group (which I have termed the 'Imperial MSE CSP Suite'). Chapters 5–6 cover my investigation into the empirical force-field. I begin by building on work from Bowskill [2], in order to have a comprehensive and balanced reference dataset for subsequent parametrization work. Following this, I will apply our group's previously developed parameter estimation methodology to derive effective transferable force-fields for different electrostatic models, and demonstrate the importance of HAIEFF model congruity. Chapters 7–8 encompass my foray into the electrostatic interaction, specifically towards incorporating polarization/induction into the HAIEFF model. The theoretical and computational considerations of our implementation are discussed in Chapter 7, after which I derive a complementary force-field for the modified electrostatic model in Chapter 8. In Chapter 9, I will apply the unified HAIEFF model on several case studies, in order to evaluate model improvements and shortcomings. Finally, I offer some concluding remarks and potentially avenues for future work in Chapter 10.

‘

Chapter 2

Literature Review

In this chapter, I review the the current state of CSP. This begins with an explanation of the thermodynamics governing the stability of crystalline materials. Following this, I give a brief overview of the hierarchical CSP framework, before discussing the two main operations within the CSP framework, (1) candidate generation and (2) lattice energy minimization. Finally, I will hone in on the two intermolecular interactions explored in this work.

2.1 Solid State Thermodynamics

At a given temperature (T) and pressure (P), metastable polymorphs are free energy local minima, while the most-stable polymorph is the global minimum. Since all systems are driven to minimize their free energy, the polymorphic transformation of metastable forms to more-stable forms (and ultimately the most-stable form) occurs spontaneously. The kinetics of this transition are influenced by the energy barriers between minima, wherein kinetically-trapped metastable forms are isolated from other forms by large barriers. A core tenet of CSP is that observable polymorphs exist near the global minimum, while kinetic effects are neglected, such that any crystal structure low enough in free energy should also be accessible by experiment.

For a perfect molecular crystal, the Gibbs free energy of the crystal (ΔG) is given by

$$\Delta G(T,P) = \Delta U(T) + PV(T,P) - TS(T,P), \tag{2.1.1}$$

where ΔU is the internal energy of the crystal lattice, V is the lattice's volume, and S is the entropy of the system. ΔG and ΔU are typically defined relative to the ground-state energy of the crystal's molecular constituents, *in vacuo*, and at *static* conditions ($U^{\mathrm{gp}}_{\mathrm{intra}}$). The '*in vacuo*' (or 'gas-phase') environment refers to a molecule at infinite separation, while 'static' indicates the absence of any vibrational contributions to the energy. Formally, V and S are also defined relative to the reference. However, for the chosen reference V is negligibly small while $S = 0$, hence for brevity Δ is omitted from these terms. Here, TS comprises only vibrational entropy, but in disordered systems this includes contributions from configurational disorder [24, 50]; such systems are beyond the scope of this work. The internal energy can be further expressed as

$$\Delta U(T) = U_{\mathrm{latt}} + U_{\mathrm{ZPE}} + U_{\mathrm{thermal}}(T), \tag{2.1.2}$$

where U_{ZPE} is the zero-point energy (ZPE) of the crystal, $U_{\mathrm{thermal}}(T)$ is the energy from thermal vibrations at $T > 0$ K, and U_{latt} is the *static* lattice energy, which again neglects all vibrations. For consistency with nomenclature commonly used in the CSP community, I denote U_{latt} without Δ, although the static lattice energy is defined relative to $U_{\mathrm{intra}}^{\mathrm{gp}}$. It is important to recognize that while static conditions imply T = 0 K, 0 K conditions are not synonymous with static conditions due to zero-point vibrations.

Regrouping the terms in Equations 2.1.1 and 2.1.2, the free energy can be re-expressed as

$$\Delta G(T,P) = U_{\mathrm{latt}} + PV + F_{\mathrm{vib}}(T), \tag{2.1.3}$$

where F_{vib} is the energy contribution from intramolecular and intermolecular vibrational modes: zero-point vibrations, thermal energy, and vibrational entropy [24]. Evaluating F_{vib} requires assessment of phonon frequencies [24, 67], which is too costly for exhaustive CSP searches. As such, for CSP_0, it is assumed that ΔG is adequately approximated by the static lattice enthalpy:

$$\Delta G(T,P) \approx H_{\mathrm{latt}}(P) = U_{\mathrm{latt}} + PV, \tag{2.1.4}$$

or more commonly just the static lattice energy:

$$\Delta G(T,P) \approx U_{\mathrm{latt}}. \tag{2.1.5}$$

Leveraging error cancellation between structures, CSP_0 asserts that the neglect of F_{vib} should not significantly impact predictions of *relative* crystal stabilities. This assumption is obviously imperfect since enantiotropic transformations occur. In their testing, Nyman & Day found that near room temperature, relative stabilities in ~9% of polymorphic pairs re-ordered upon introduction of F_{vib} [24], rising to ~21% when considering temperatures up to the melting point [72]. Nevertheless, even proponents of temperature-dependent CSP (which Price terms 'CSP_thd' [52]) will acknowledge the practical limitations of applying these calculations at scale. Whilst assessing static enthalpy is inexpensive, crystals are extremely dense states with small molar volumes. At the near-ambient conditions of interest to most CSP studies, PV might be in the order of several J/mol, while U_{latt} is in the order of kJ/mol. Between crystal structures, differences in V are even smaller [73], hence ignoring PV rarely affects relative stabilities. If high-pressure (several GPa) behavior is studied, PV is trivial to re-introduce. Herein, I will operate under the paradigm of CSP_0 that employs Equation 2.1.5 to approximate ΔG.

2.2 The Hierarchical CSP Framework

Even with this simplified thermodynamic view, reliable CSP_0 is still extremely challenging. In principle, CSP only involves two major operations: (1) generate a trial crystal structure, and then (2) minimize U_{latt} with respect to the crystal geometry. I refer to these operations as 'candidate generation' and 'lattice energy minimization' respectively. A naïve implementation of CSP could simply repeat these actions until the solid form landscape has been adequately

explored. Unfortunately, such an approach is unlikely to succeed since the demands of these operations are directly at odds with one another.

In candidate generation, exhaustive exploration is vital to ensure all thermodynamically relevant structures have been considered. However, this search space suffers acutely from the curse of dimensionality. Briefly, a crystal structure is described by: the unit cell parameters (cell lengths and angles), the symmetry of the unit cell (230 possible space groups), the number of symmetrically unrelated molecules in the unit cell (denoted Z'), the position and orientation of the Z' distinct molecules, and (in the case of flexible molecules) their conformations. Sampling this space uniformly is usually intractable, thus CSP methodologies must make concessions to limit the search space and employ strategies to explore the constrained search space more-efficiently. Even with the most-sophisticated candidate generation strategies, it is not uncommon to evaluate several millions of crystal structures.

Orthogonal to the candidate generation is the task of lattice energy minimization. It has been suggested that ΔU_{latt} between observable polymorphs very rarely exceed 10 kJ/mol and in most cases will be $\leq$ 4 kJ/mol [21, 24, 65]. Crystal lattice energies are the result of a delicate interplay between intra- and intermolecular effects including: conformational changes, electrostatic interactions, polarization, hydrogen bonding, and van der Waals interactions. Unfortunately, models capable of describing all aspects of U_{latt} to the level of accuracy required to resolve the purportedly small ΔU_{latt} cannot be applied to the previously quoted quantity of crystal structures, within a reasonable computing time.

To balance the need for exploration and accuracy, a multi-fidelity hierarchical/multistage framework [50, 51] has become the norm for almost all CSP workflows (Figure 2.1). This begins with an input of basic molecular and stoichiometric information, and any additional constraints on the search space. The first stage, denoted the 'Global Search', involves candidate generation and an initial round of lattice energy minimization, prioritizing the extensive exploration of the search space. To that end, the lattice energy model employed is relatively coarse, enabling 'cheap-and-dirty' evaluation of 10^5–10^6 crystal structures. As mentioned earlier, many modern candidate generation strategies also incorporate some protocol to more-efficiently explore the space. Whilst the solid form landscape is vast, not all structures will be thermodynamically relevant. This can be exploited, with feedback from the energy model/lattice energy minimization, to guide the candidate generation away from regions of the search space which are unlikely to generate thermodynamically competitive crystal structures.

Energy models compatible with the Global Search stage are rarely accurate enough for meaningful CSP results, therefore the Global Search must be followed by a Refinement stage. During refinement, a more-sophisticated energy model is used to re-minimize and re-rank the crystal structures obtained from a preceding stage, bolstering the credibility of the CSP results. Refinement can be repeated as many times as desired, with increasingly accurate (but costlier) energy models, until a satisfactory accuracy is achieved. While this multistage approach was formulated in the CSP_0 paradigm, a natural evolution could be introducing CSP_thd techniques/models as late-stage refinements of the CSP_0 workflow [51].

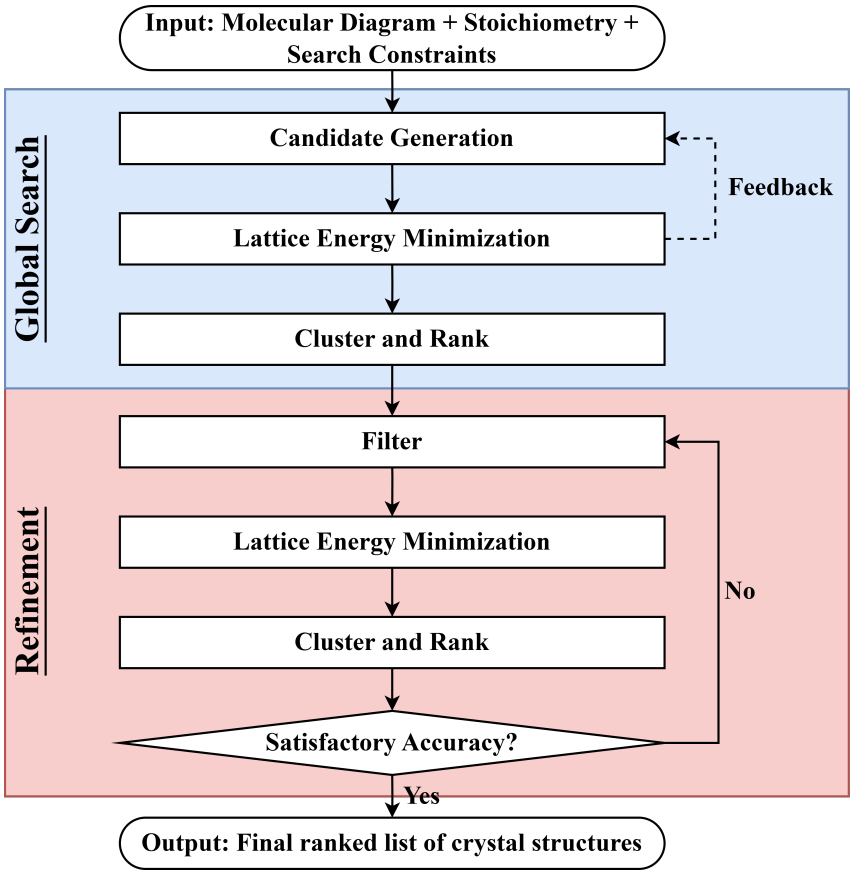


Figure 2.1: The multistage, hierarchical CSP workflow.

Critically, this multistage workflow balances model fidelity and computational costs in two ways. Firstly, each lattice energy minimization is initialized from the minimized structures of the preceding stage. If the lower-fidelity model is not too crude, the higher-fidelity energy minimizations start already near a minimum, thereby reducing costs. If the crystal geometry is not expected to change significantly, or if the higher-fidelity model is too costly for minimizations, single-point evaluations with the higher-fidelity model can be considered. Secondly, and more importantly, the number of structures advanced for refinement is significantly reduced via intermediate clustering and filtering steps. Clustering reduces the number of crystal structures on the solid form landscape by grouping related minima together by some similarity criteria, usually crystal geometry. It has been suggested that the static CSP landscape has a 'bumpy' energy surface that gives rise to many energetically- and geometrically-related structures. If subject to thermal and vibrational effects, these families should represent a single finite-temperature crystal structure [68, 69], justifying their coalescence on the static landscape. The unique structures remaining after clustering are ranked by U_{latt} and thermodynamically un-competitive structures beyond some energetic cutoff are filtered out. The choice of cutoff is critical here: ΔU_{latt} are not expected to exceed 5 kJ/mol [21, 24, 65], hence at the highest fidelities, one may employ a cutoff window around this value. However, at lower fidelities, a larger energy window (e.g., ~20 kJ/mol at the Global Search) is necessary to accommodate model inaccuracies.

It is important to highlight that the decision-making involved at each stage of the work-

flow is not prescriptive, and other factors often take precedence. For instance, when deciding appropriate cutoffs between stages, one should consider that these widely touted U_{latt} cutoffs are based only on theoretical work [21, 24, 31], and thus their validity is caveated by their relevant model assumptions and uncertainties. More often, computational limitations still dominate the decision-making here and practical cutoffs may be defined solely based on the number of structures which can be processed by the available resources. In deciding how many refinement stages are necessary, practitioners should not treat the question of 'satisfactory accuracy' solely as an academic question of physical-model fidelity. As noted by Beran [51], for CSP to become more useful, 'reasonably reliable' CSP must become the standard, wherein the computational effort expended is commensurate with the accuracy necessary for the intended application. As an example, in product discovery and development, CSP studies of different quality may be appropriate at different stages of the development pipeline, depending on the available time and the certainty required for practical decision-making.

As evidenced from the most recent Blind Tests [47–49], the multistage CSP workflow is almost ubiquitous. Where individual CSP methodologies differ is in their approach towards candidate generation and lattice energy minimization. In the next two sections, I will describe the popular strategies/models that are used for these operations in modern CSP.

2.3 Candidate Generation Strategies

Candidate generation is the first major operation in the CSP workflow and the central problem can be distilled down to "*How can we most-efficiently generate starting crystal structures that lead to thermodynamically competitive minima?*". Strategies that are too conservative will frequently explore superfluous regions of the search space that only generate high energy minima, wasting computational effort. Conversely, strategies that aggressively exclude regions of the search space reduce costs, but are more susceptible to missing relevant structures.

2.3.1 Constraining the Search Space

2.3.1.1 Space Group Constraints

Efficient candidate generation starts with imposing sensible constraints to the crystal structure search space. Most CSP codes search each space group independently, i.e., the space group is selected during the candidate generation and fixed throughout a crystal's lifetime in the workflow. As such, restricting the choice of space groups is a common and important strategy. In the CSD, the 13 most-common space groups already account for > 90% of all observed crystal structures, while the top 60 space groups account for ~98% [74]. Among a given set of space groups, the frequency at which they are sampled may be uniform or proportional to their observed frequencies. Taylor *et al.* [65] performed 1,000 CSP studies and in each the top 26 space groups were sampled uniformly. Across all CSPs, they found that the space group distribution for stable structures was consistent with experimentally observed statistics, justifying frequency-

biased sampling. However, they also noted that within a reasonable ΔU_{latt} cutoff, this agreement in space group frequency starts to break down, suggesting that observable metastable forms may not adhere strictly to experimentally observed frequencies. Our group has also observed this: Molecule XXXI in the Seventh Blind Test [49] had a metastable de-hydrated form which we failed to find, in part because the polymorph occurs in the $R\bar{3}$ space group which has a relatively low observation frequency (~0.8%). Overall, depending on the CSP goals, the space group sampling strategy may need to be adapted: frequency-biased sampling if only highly-stable forms are of interest (e.g., polymorph screens), while uniform sampling may preferred for materials discovery, where superior properties are accessible only through more-metastable forms.

2.3.1.2 Z' Constraints

Another common search constraint is on the number of symmetrically-independent molecules in the unit cell, Z'. Each independent molecule is described by its position in the asymmetric unit cell, its orientation, and its conformation, hence additional molecules expand the search space dramatically. Moreover, there are other practical complications associated with the placement of independent molecules with respect to one another [71]. Altogether, search complexity rises considerably for each additional distinct molecule, and it is preferable to keep Z' low. Fortunately, $Z' > 1$ crystal structures are less frequently observed (~20%) and tend to be metastable relative to their $Z' = 1$ counterparts [21], hence most modern CSPs default to only $Z' = 1$ searches. Obviously, this risks excluding relevant high Z' structures, like the $Z' = 3$ form of Molecule XXIX in the Seventh Blind Test [49]. Moreover, multicomponent crystals are by definition $Z' > 1$, hence $Z' > 1$ searches will undoubtedly become more routine as these systems gain relevance.

2.3.1.3 Molecular Flexibility Constraints

The final major constraint relates to the treatment of molecular flexibility within the candidate generation. In a crystal, the thermodynamic driving force is to lower U_{latt} as a whole. This gives reason for molecules to deform from their *in vacuo* conformation to a crystalline conformation ($U_{\text{intra}}^{\text{crys}}$). While this deformation penalizes the conformational energy ($U_{\text{intra}}^{\text{crys}} > U_{\text{intra}}^{\text{gp}}$), it can be favorable for U_{latt} if it enhances intermolecular interactions that counterbalance the penalty. This explains the occurrence of both packing and *conformational* polymorphism [19, 75]. Depending on the chemical moieties present, molecular deformation can be accommodated to different extents: ~20 kJ/mol for molecules without intramolecular hydrogen bonds [76], ~35 kJ/mol in hydrates/solvates [77], and ≥ 50 kJ/mol for molecules with intramolecular hydrogen bonds [75]. The inclusion of molecular flexibility steeply increases costs since each flexible feature (bond lengths, bond angles, and dihedral angles) adds a search dimension. To manage this problem, several strategies may be employed to deal with flexible molecules [77].

In one approach, only conformer-related flexibility is considered. Multiple gas-phase (local) minimizations of the molecule are initialized to generate a collection of conformers. Conformers deemed too high in energy (relative to $U_{\text{intra}}^{\text{gp}}$) may be removed, leaving only those that are likely

in the crystal. Independent, rigid CSP searches are then performed for each conformer. This strategy was successfully employed to search for polymorphs of piracetam [78], and is justified on the basis that in most systems exhibiting conformational polymorphism, only a handful of conformers are observed (≤ 5) [75]. However, this does not diminish the importance of conformational adjustments in polymorphism. When comparing crystalline conformations between polymorphs, Cruz-Cabeza & Bernstein [75] observed a continuous distribution in conformational differences that peaked at low difference scores (Figure 5 in Ref [75]). If conformers alone explained polymorphism, this distribution should instead be multi-nodal over larger difference scores. To accommodate such adjustments, rigid candidate generation can be based on low-energy conformations identified by scanning conformational degrees of freedom (dofs). This was applied with success on phenobarbital [79].

While approaches that discretise molecular flexibility may be suitable for small molecules with limited flexibility, they fall victim to some critical issues [77]. Firstly, the energy cutoff for relevant conformations necessitates finding the true ground-state conformation. If a molecule is extremely flexible, this can entail many trial minimizations. Moreover, the cutoff requires that conformational energies are correctly computed. Density Functional Theory (DFT) [80, 81] is commonly used for evaluating conformational energies, but DFT has well-known issues such as lack of dispersion [82] and delocalization error [83, 84]. While dispersion corrections [85–89] have helped address the former, delocalization error has been shown to cause seriously erroneous conformational energies [56, 90, 91]. Most importantly, a discrete treatment of flexibility is suboptimal, since even small adjustments in molecular conformation can lead to large changes in U_{latt} [73, 92]. Rigid candidate generation, even based on several low-energy conformations, have higher risk of eliminating crystal structures which could have otherwise been energetically competitive if its conformation was allowed to change slightly. Whilst molecular flexibility can be reintroduced later in the CSP workflow, the use rigid candidate generation invariably biases certain regions of conformational space. Overall, a continuous description of flexibility in the candidate generation (and lattice energy minimization) are critical for flexible molecule CSP.

In light of this, several strategies have been developed for the continuous description of molecular flexibility. These methods permit either partial molecular flexibility in specific conformational dofs (usually dihedral angles) [93–98], or atomistic flexibility (bonds, angles, and torsions) [99–101]. In both approaches, the continuous description is achieved by formulating a surrogate model for the intramolecular energy, as a function of the flexible dofs. These models can be used not only for sampling the conformational space during candidate generation, but are also employed during the lattice energy minimization. Because the details of these methods are closely tied to the lattice energy model, they will be described in Section 2.4.

2.3.2 Landscape Exploration Strategies

Once the candidate search space has been constrained, starting structures for lattice energy minimization may be generated. To that end, there are two main classes of production method.

2.3.2.1 Unbiased Exploration

In unbiased search methods, there is minimal feedback from the lattice energy minimization and the candidate generation is mostly ignorant of the performance of preceding candidates. In effect, similar effort is expended on regions of the solid form landscape that result in high and low U_{latt}, such that exploration is prioritized over exploitation. The simplest implementation of this is a uniform grid search, but as alluded to previously, this does not scale well for the systems of interest in modern CSP (higher Z', more flexibility). In the Sixth Blind Test, only three groups out of 21 utilized a grid search [48]. Pseudo-random structure generation using uniform or Gaussian distributions [102] is a popular alternative, with seven groups employing this strategy in the Sixth Blind Test [48]. However, pseudo-random structure generation is susceptible to uneven coverage of the search space, even when using a uniform distribution. Finally, quasi-random, low-discrepancy sequence methods, like Sobol' sequences [103], are employed in CSP programs like GLEE [104] and CrystalPredictor [93, 94, 96–98]. For a given number of samples, Sobol' sequences attempt to achieve the most even coverage of a search space, by maximizing the distance between each new sample and all previous points. The deterministic nature of this process allows samples to be added retroactively, without compromising the quality of the search coverage. In the Sixth Blind Test, four groups utilized this method [48].

By themselves, unbiased candidate generation may favor excessive exploration, hence auxiliary screening can be added to discard inferior candidates. This includes unphysical candidates (e.g., crystals with extremely close/overlapping atomic sites), candidates that are likely far from a local minimum and would take too long to minimize (e.g., unreasonably low-density crystals), and candidates that are unlikely to be relevant upon minimization (e.g., crystals that are very high in energy, relative to the incumbent global minimum lattice energy).

2.3.2.2 Biased Exploration

Using various global optimization techniques, biased search methods actively feedback information from the lattice energy minimization, to inform the production of new candidates. These strategies begin with an initial random population of candidate structures. Following lattice energy minimization, the performance of a population is evaluated through U_{latt}. A new population of candidates is then stochastically generated, conditioned on the performance of preceding populations. Exploitation is achieved by assigning higher probabilities to features that encourage superior performance, while this is balanced with exploration by permitting low probability changes. This iterative process continues until there is convergence. A wide range of strategies emerge from this framework, differing in how the conditional sampling probabilities are applied. In the Sixth Blind Test, only six out of 21 groups used biased candidate generation strategies, but this encompassed four distinct methods [48]. Some notable CSP implementations that utilize biased methods include: the Generation Ranking and Characterization Engine (GRACE) [99, 100, 105] which employs Monte Carlo parallel tempering, GAtor [106, 107] which uses a genetic algorithm, and XTalPi [108, 109] which combines particle swarm optimization

and a Monte Carlo algorithm. More recently, the CCDC has released the CSD Landscape Generator [110], which can also be used for biased candidate generation. Given a molecule (or molecules), the Landscape Generator encodes the system-of-interest as a set of molecular fingerprints, and searches for crystals within the CSD with similar molecules. The target molecules are embedded into the experimentally observed crystal structure, substituting its corresponding molecular-analogue [110]. In effect, the Landscape Generator acts as a machine learning based search algorithm, that is biased by existing empirical knowledge.

2.4 Lattice Energy Models

Lattice energy minimization is the second major operation in the CSP workflow. This task requires energy models that can describe the various interactions in the crystal, for which different classes of lattice energy model have been proposed at varying levels of accuracy and computational cost [50]. In line with the multistage workflow, no singular model is appropriate at all stages and it is necessary to employ multiple fidelities of energy model for successful CSP.

2.4.1 Electronic Structure Methods

Electronic structure methods are so-called as they aim to describe the distribution of electrons in a system through Schrödinger's equation. From the electronic structure, other physical quantities (e.g., energies) may be derived. These methods are based wholly on first-principles, but their complexity precludes solving them analytically. Instead iterative numerical methods must be used, such that they are also termed self-consistent field (SCF) methods. Because electronic structure methods treat systems atomistically, there is no explicit distinction between intra- and intermolecular interactions. Atomistic flexibility is also inherent to these methods. This class of energy model generally represents the upper bound of cost and accuracy for CSP.

2.4.1.1 Electron Wavefunction Methods: Hartree–Fock and Post-Hartree–Fock

Hartree–Fock (HF) and post-Hartree–Fock (post-HF) methods, directly seek solutions to the electron wavefunction (Ψ_{elec}) [111, 112]. To that end, HF theory does describe many key quantum mechanical effects, like electron-exchange and Pauli exclusion [112], but fails to represent electron-correlation beyond a mean-field. This limits HF's efficacy in molecular crystals, where dispersion interactions can be significant. Post-HF correlated-wavefunction methods, which improve the description of electron-correlation via multi-determinant wavefunctions [111], are thus the preferred choice for molecular crystals. Notably, the MP2 Møller & Plesset many-body perturbation theory [113] (MPx, where x is the perturbation order) has been demonstrated on several small molecular crystals [114–116], but its widespread use is hindered by several factors.

Firstly, while MP2 is considered one of the cheapest post-HF methods, it is still relatively cost prohibitive. For canonical MP2, computational costs scale by ~$\mathcal{O}(N^5)$, where N is the number of basis functions [117, 118]. Paired with MP2's poor basis set convergence [117, 118],

costs can quickly balloon in large periodic systems like crystals. Beyond its high cost, MP2 theory is also prone to inaccurate predictions of dispersions, depending on the polarizability and/or degree of electron delocalization of the system [118, 119]. For instance, using MP2, benzene crystals were overbound by ~10 kJ/mol compared to more-sophisticated wavefunction theories (e.g., MP3, or coupled-cluster methods) [118]. While these higher levels of theory can mitigate MP2's inaccuracies, their computational costs are even less tenable. Instead, empirical corrections may be a more-viable means to fix the underlying issues with MP2, whilst preserving computational feasibility in periodic systems [118, 119].

In its current state, direct application of wavefunction methods to large periodic systems faces several challenges before it can achieve a competitive balance in cost and accuracy. A key benefit of wavefunction methods is that their accuracy improves systematically with level of theory and basis set size [120]. However, it is unlikely that this can be leveraged in a cost effective manner, without more innovative application of these theories.

2.4.1.2 Electron Density Methods: Periodic DFT

In lieu of directly determining Ψ_{elec}, DFT solves for the electron density (ρ_{elec}), and then determines Ψ_{elec} as a functional of ρ_{elec} [80, 81]. This strategy makes DFT less costly than post-HF methods, scaling by only ~$\mathcal{O}(N^3)$–$\mathcal{O}(N^4)$ [121]. For its (relatively) low cost, DFT achieves respectable accuracy that is not a far cry from post-HF methods. Overall, periodic DFT offers a good balance of cost and accuracy, and is routinely employed in molecular CSP.

The efficacy of periodic DFT is predicted upon several modeling decisions, the first of which is the choice of exchange-correlation (XC) functional. Based on the approximations employed during their formulation, XC functionals are classified into a hierarchy, known as the 'Jacob's Ladder' [122]. On the lowest 'rung', local density approximation (LDA) functionals are generally too simplistic to yield meaningful CSP results. Conversely, at the top of the ladder are hybrid functionals (e.g., the hybrid Perdew–Burke–Ernzerhof (PBE0) functional [123–125], and the 3-parameter Becke exchange + Lee–Yang–Parr correlation (B3LYP) functional [126–128]), which incorporate a proportion of exact HF exchange into the DFT formulation. Hybrid functionals do help mitigate DFT delocalization error (discussed below), but computational costs rise steeply when they are employed [122, 129]. This precludes their use for minimization in large periodic systems. As such, molecular CSP predominantly relies on the middle rungs: the generalized gradient approximations (GGAs) (e.g., the Perdew–Burke–Ernzerhof (PBE) functional [123]) and meta-GGAs (e.g., the Tao-–Perdew—Staroverov—Scuseria (TPSS) functional [130]).

All of the stated functionals are local/semilocal in nature, thus DFT utilizing these cannot describe long-range electron-correlation effects, like dispersion. While some semi-empirical functionals have shown dispersion-like behavior (e.g., B3LYP), this is likely an artefact from their empirical fitting [82]. Dispersion is important in many organic crystals [118, 131, 132], thus dispersion-corrected DFT (DFT-D) is the standard for CSP. Popular corrections include: Grimme's D3 method [85], Tkatchenko & Scheffler's dispersion correction (TS) [87], exchange-

hole dipole moment (XDM) theory [88, 89, 133], and van der Waals density functionals (vdW-DFs) [134]. D3, TS, and XDM 'append' the classical pairwise r^{-2n} London dispersions to the DFT energy, but differ in their truncation of the dispersion series, choice of damping function, atom types, and derivation of dispersion coefficients [135]. Many-body dispersions (atomic many-body effects [136], or Dobson-B effects [137]) can be important to molecular crystals [135]. As such, D3 [85] and XDM [133] include the leading Axilrod–Teller–Muto (ATM) [138, 139] three-body term, while TS was extended to a many-body dispersion (MBD) variant [140, 141]. Unlike the previous methods, vdW-DFs [134] are non-local functionals that natively embed long-range correlation into DFT. vdW-DFs are generally less popular than the classical methods, as they tend to be more costly but do not necessarily provide commensurate gain in accuracy.

Multiple authors [131, 132, 142, 143] have performed benchmarking of the DFT-D methods. DFT-XDM, DFT-TS, DFT-D2 (an earlier version of D3), and vdW-DF were compared using the C21 test set, with the PBE and B86b functionals [131]. DFT-XDM predicted lattice energies most accurately and was least prone to overbinding, while DFT-TS was the least accurate. All models gave good predictions of crystal geometry. Two studies [142, 143] used the X23 test set to compare DFT-TS, DFT-MBD, and DFT-D3, over a range of functionals. Their results corroborate that crystal lattice energies and geometries are accurately predicted for most DFT-D schemes. DFT-TS generally performed poorly, but results improved greatly upon progressing to DFT-MBD [142]. Conversely, the accuracy of DFT-D3 was found to be relatively insensitive to the inclusion of the ATM term [143], and overall the performance of DFT-D3 was on par with DFT-MBD with the same functional [142, 143]. The mean absolute error and mean absolute relative error in the C21 test set using PBE-D3 are 5.47 kJ/mol and 7.14% respectively (using the predicted values from Ref [143] and reference energies from Ref [131]). This places the accuracy of PBE-D3 in a similar ballpark to PBE-XDM [131]. Finally, using the POLY59 test set, it was demonstrated that PBE-MBD gives comparable results to TPSS-D3, with both approaches successful predicting relative lattice energies amongst several polymorphic families [132].

DFT-D mitigates only one of DFT's flaws, but it has other shortcomings. Firstly, DFT is prone to delocalization error [83, 84], which is the proclivity for LDA, GGA, and meta-GGA functionals to overstabilize geometries with delocalized electron density. This can lead to spurious errors in conformational energy, significant enough to affect relative stabilities. Hybrid functionals which incorporate a portion of exact HF exchange do slightly mitigate this issue [51, 84], but not reliably. A more promising route to resolve this might be *post hoc* monomer-correction of the erroneous conformational energies with wavefunction methods (that are free of delocalization error) [91, 144]. However, these strategies are currently not routinely applied in CSP. Secondly, unlike Ψ_{elec} methods, DFT does not exhibit a systematic improvement in accuracy with level of theory (e.g., basis set size, higher rungs on the 'Jacob's Ladder'). This can result in a degree of arbitrariness in modeling quality across different systems.

The best performing periodic DFT-D methods in the previously discussed benchmarking tests achieve average absolute deviations (AADs) in lattice energies ~3–6 kJ/mol and standard deviations (SDs) in error ~5 kJ/mol. This is still inadequate to resolve ΔU_{latt} in extreme cases.

However, periodic DFT-D still offers the most tenable balance of accuracy and computation cost (among electronic structure methods), hence it remains the predominant choice for high-fidelity calculations in CSP (i.e., final refinement in Figure 2.1). In the Sixth Blind Test, almost half (11 out of 25) of the participating groups had employed periodic DFT at some stage of their workflow, usually for final rankings [48]. DFT-D in the form of a GGA/meta-GGA functional + Grimme/TS/XDM/MBD dispersion-correction was employed by 8 groups, while 2 groups used non-local vdW-DFs; the last group (Group 14) used an in-house developed DFT-D model.

2.4.1.3 Fragmentation Methods

Thus far, I have discussed the application of Ψ_{elec} and ρ_{elec} methods to treat an *entire periodic system*. Fragmentation methods propose an 'indirect' approach for applying electronic structure methods. In lieu of evaluating the entire periodic crystal, the system is expressed as a many-body expansion involving 'fragments' of the system (e.g., the molecules in a crystal) [118, 145]:

$$U_{\text{latt}} = \sum_{i} U_i + \sum_{i>j} \Delta^2 U_{ij} + \sum_{i>j>k} \Delta^3 U_{ijk} + \cdots, \tag{2.4.1}$$

where i, j, k are fragment indices. U_i is the monomer contribution to the lattice energy, $\Delta^2 U$ is the dimer contribution, $\Delta^3 U$ is the trimer contribution, and so on. 'Δ^N' terms are N-body *perturbations* of the lower-body terms, and thus exclude contributions from lower-body terms. By fragmenting a periodic crystal into smaller systems, expensive electronic structure methods become accessible, especially in the lower-body terms ($\sim\mathcal{O}(nN^m)$ scales better, if $n > 1$ and N decreases). Leveraging the fact that the N-body terms are independent of one another, high-fidelity electronic structure theory (e.g., MP2, or other post-HF methods) can be used to describe the lower-body terms, while cheaper models are applied to the rest of the expansion. Since the bulk of U_{latt} is attributable to the lower-body terms ($N \leq 2$ in particular), overall model accuracy can be insensitive to such a multi-fidelity construction.

Formally, the expansion of a periodic system is an infinite series, for which several treatments have been proposed [145]. In the simplest approach, the higher-body terms are truncated and approximated via electronic embedding of the lower-body terms. While this can capture long-range electrostatic effects, it neglects long-range dispersion. As such, it is preferable to substitute the higher-body terms with a periodic evaluation of the entire system at low-fidelity, being careful to avoid double-counting lower-body contributions. The need to conduct this periodic evaluation restricts the low-fidelity model to HF, and certain classes of periodic DFT. In the hybrid many-body interaction (HMBI) approach, Beran & co-workers propose that the low-fidelity model can even be a polarizable, classical force-field [118, 145], giving a quantum mechanical/molecular mechanics (QM/MM) approach analogous to the ONIOM method [146].

When tested on several small-molecular (≤ 12 atoms) systems [147, 148], the HMBI method (employing an AMOEBA force-field) predicted lattice energies within 4 kJ/mol of experimental estimates [147]. This error was further reduced to ~2 kJ/mol by using a bespoke, semi-empirical polarizable classical force-field [148]. HMBI was later applied to two forms of aspirin and pre-

dicted relative lattice energies that agreed better with experiment than periodic DFT [149]. Recently, HMBI was tested on three moderately sized molecules (ROY, acetamidobenzamide, and oxalyl dihydrazide) using a D3-corrected variant of MP2 [90], and in all cases HMBI gave superior predictions over periodic DFT-D. For ROY and acetamidobenzamide, this was due to periodic DFT-D exhibiting large delocalization errors, whereas HMBI predicted monomer energies accurately. While HMBI has demonstrated commendable accuracy, competitive with full periodic Ψ_{elec} methods [148], HMBI is still more costly than periodic DFT-D [90, 147]. As such, it is unlikely that fragmentation methods will supersede periodic DFT-D for routine solid state calculations, unless better cost efficiency is achieved.

2.4.1.4 Symmetry-Adapted Perturbation Theory with DFT Monomers

The final electronic structure method I briefly discuss is the DFT formulation of Symmetry-Adapted Perturbation Theory (SAPT(DFT)) [150, 151]. By design, SAPT, and SAPT(DFT), probe the interaction energy between *dimers*, hence they cannot be used directly in the evaluation of crystals. However, SAPT-based models provide nuanced information about interaction energies, which can be used to parametrize highly accurate intermolecular force-fields. Given the growing popularity of this strategy (four out of 25 groups in the Sixth Blind Test [48]), I would be remiss not to provide at least a cursory explanation of its underlying theory.

The description given here summarizes key points from Refs [151, 152]. In a supermolecular approach, total dimer *interaction* energy is determined by subtracting individual monomer energies from the total dimer energy. In SAPT, this dimer interaction energy ($U_{\text{int}}^{\text{SAPT}}$) can be directly evaluated from the electronic wavefunction as a perturbation series,

$$U_{\text{int}}^{\text{SAPT}} = \sum_{ij} (U_{\text{int}}^{ij} + U_{\text{exch}}^{ij}), \tag{2.4.2}$$

where U_{int}^{ij} and U_{exch}^{ij} are the ij^{th}-order interaction and exchange energy contribution respectively. The exchange counterpart to each interaction term is a result of enforcing antisymmetry [152], and can be interpreted as a manifestation of the Pauli exclusion principle. SAPT(DFT) applies SAPT theory but with wavefunctions derived by DFT, instead of HF/post-HF methods, exploiting the computational efficiency of DFT without significant loss in accuracy. Critically, the individual terms in the SAPT perturbation series can be correlated to known physical interactions: electrostatics, polarization/induction, dispersion, and their exchange-repulsion counterparts [151]. Through this neat decomposition of the *ab initio*-determined total interaction energy, one has access to reference energies for each physical force, that are amenable towards targeted-fitting of force-field parameters; this will be discussed later.

2.4.2 Force-field Methods

Force-field (FF) methods lie on the other end of the QM/MM spectrum relative to electronic structure methods. In these models, U_{latt} is described classically through independent contributions to the lattice energy. A general FF formulation for the lattice energy is given as,

$$U_{\text{latt}} = U_{\text{bonded}} + U_{\text{nonbonded}}, \tag{2.4.3}$$

$$U_{\text{bonded}} = U_{\text{stretch}} + U_{\text{bend}} + U_{\text{tors}}, \tag{2.4.4}$$

$$U_{\text{nonbonded}} = U_{\text{vdW}} + U_{\text{elec}}, \tag{2.4.5}$$

where U_{bonded} and $U_{\text{nonbonded}}$ are the bonded and nonbonded interaction energies respectively. The distinction between these terms is based on bonded neighbor separation, rather than 'intra-' and 'intermolecular', and atoms on the *same molecule* can have nonbonded interactions if they are adequately separated. Bonded interactions typically include bond stretching (U_{stretch}), bond angle bending (U_{bend}), and torsion angle rotations (U_{tors}), while the nonbonded interactions comprise van der Waals (U_{vdW}) and Coulomb (U_{elec}) interactions. Other contributions may be appended to this general formulation, such as out-of-plane bending and hydrogen-bonding interactions [99, 111]. Each FF contribution is given by a closed-form analytical expression, involving sets of participating atoms. For example, the electrostatic interaction between two atoms, i and j, is usually described by Coulomb's law,

$$U_{ij}^{\text{elec}} = \frac{1}{4\pi\varepsilon_0} \cdot \frac{q_i q_j}{r_{ij}}, \tag{2.4.6}$$

where ε_0 is the vacuum electric permittivity, r_{ij} is the interatomic distance, and $q_{i/j}$ is the partial-charge of the corresponding atomic site. To evaluate these expressions, each atom must be assigned an atom type. This distinguishes atoms minimally by their proton number, but can also include other chemical features which separate distinct atomic environments, such as: bonded neighbors, hybridization, valency, etc. Each atom type has its own set of FF parameters that quantify the strength of each interaction (e.g., $q_{i\,\text{or}\,j}$ in Equation 2.4.6). Due to their closed-form, FF models are extremely cheap to evaluate and are amenable to atomistic flexibility. As described below, the simplicity of FF models does not preclude excellent model accuracy, depending on the strategy used to define atom types and parametrize the FF.

2.4.2.1 Transferable Molecular Mechanics Force-fields

Transferable molecular mechanics force-fields (MMFFs) have been developed to describe a variety of states-of-matter, molecular chemistries, and physicochemical properties. These general FFs employ predefined atom types (and FF parameters), which are assumed to be universally valid across all molecular systems. To attain this transferability, MMFFs are parametrized on extensive datasets comprising different molecules (and conformations), so as to achieve accurate predictions (on aggregate) over the entire dataset. General MMFFs were popular in the earlier Blind Tests [43–46], with a significant number of participants utilizing MMFFs like: DREIDING [153], CVFF [154], COMPASS [155], OPLS [156], and GAFF [157]. However, this was likely motivated by necessity, more so than model accuracy. Recently, Robinson *et al.* [158]

benchmarked the performance of six MMFFs on 210 crystal structures. Among these, only the COMPASSII FF [159] presented with reasonable accuracy (AAD ~9–10 kJ/mol), while all other MMFFs tested had AADs > 20 kJ/mol. In the X23 test set, their COMPASSII protocol had an AAD of 10.72 kJ/mol [158], which is competitive with some hybrid *ab initio*/empirical force-field models (Section 2.4.3), but 2–3 times larger than the best periodic DFT-D methods [66, 143].

The results of Robinson *et al.* [158] indicate that, in most cases, transferable MMFFs are too inaccurate for reliable CSP. This aligns with most modern CSP workflows, as observed during the Sixth Blind Test [48]. While several groups did employ MMFFs, this was only under specific circumstances. Some groups (e.g., 3 and 21) used MMFFs solely to identify conformers for use in their CSP search. Groups who used an MMFF for lattice energy minimization and ranking, did so almost exclusively in the Global Search stage, before re-ranking the landscape by periodic DFT-D. Since MMFFs cannot modulate their accuracy in any meaningful way, they are ill-suited for progressive refinement in the multistage workflow. Only Group 1 used a COMPASS FF in their final ranking step [48], but they failed to find the target structures in their submissions.

2.4.2.2 Tailor-made Force-fields

The limited accuracy of MMFFs stems from the general strategy applied in their parametrization. These FFs are not only expected to work on a wide range of molecules, but they are also not exclusively used to describe crystal energies/geometries. Tailor-made Force-Fields (TMFFs) employ a diametrically opposed strategy, by parametrizing molecule-specific FFs, exclusively for the solid state. Below, I give an overview of two methodologies that use the TMFF model.

The TMFF concept was first introduced by Neumann & co-workers [99, 100], and was motivated by accurate CSP results obtained using a custom variant of DFT-D, denoted 'DFT-NP' [160]. While a DFT-D model is too expensive for large-scale lattice energy minimization throughout a CSP workflow, it was proposed that this energy model could instead provide high-quality reference data for fitting an accurate bespoke FF [99, 100]. In their TMFF workflow, molecule-specific atom types are first defined to capture unique atomic environments. A variety of (hypothetical) crystal structures and molecular conformations are assembled from the target molecule, or representative fragments of the molecule if it is too large. Importantly, these geometries probe different interatomic distances between atom types. The DFT-NP model is used to evaluate the energy, forces, and electrostatic potential for these geometries, forming reference data for FF fitting. Non-equilibrium data (i.e., geometries that have not been energy minimized) are also employed: this is more economical than synthesizing only equilibrium data, and offers information about the topology of the energy surface. The parametrized TMFF is similar to Equations 2.4.3–2.4.5, except that the TMFF has additional coupling terms and a hydrogen bonding interaction term. Unlike MMFFs, which usually assume conformation-independent FF parameters, the TMFFs incorporate some conformation-dependence into the atomic-charges used in the Coulomb interaction. Through this rigorous parametrization strategy, the resulting TMFF model can achieve excellent reproduction of the DFT-NP energy surface [100].

Neumann & co-worker's TMFF framework is currently implemented in the GRACE software package, developed by Avant-garde Materials Simulations [105]. In the Fourth to Sixth Blind Tests [46–48], TMFFs were applied at the Global Search stage, followed by final Refinement and ranking with the DFT-NP model. This was met with good success, in most cases accurately predicting the relative rankings of the target molecules. However, the accuracy achieved by the TMFF's molecule-specific nature comes at the expense of relatively high computational cost, not from its deployment but from generating new reference data for each system. In all of the aforementioned Blind Tests, the TMFF approach frequently ranked as one of the most computationally demanding methods [46–48]. Recent work has explored several developments to this framework [101]. Firstly, the TMFF model was augmented to allow the use of dipoles and quadrupoles in the Coulomb term. Secondly, it was demonstrated that the use of representative molecular-fragments confers some FF-transferability among related families of molecules. From this, the authors [101] proposed a workflow that begins with a 'transferable' TMFF for the Global Search, followed by intermediate Refinement with a backfitted molecule-specific TMFF, and concluding with final ranking by the DFT-NP model. In effect, the modified framework introduces some facility for modulating the specificity and accuracy of the TMFF for intermediate Refinements.

The success of GRACE has motivated similar TMFF approaches, for instance by the industrial service-provider XTalPi [109]. Only scant details regarding the formulation of XTalPi's TMFF have been published. Unlike GRACE's TMFF, and other MMFFs, the XTalPi TMFF has distinct FF parameters for intramolecular and intermolecular nonbonded interactions, but no information was found regarding the atom typing scheme used. Similar to GRACE, XTalPi fits their TMFF to large-scale DFT-generated synthetic data, including energy data for monomer, dimer, and periodic systems. Following the Global Search using their TMFF, the XTalPi workflow conducts final energy ranking with periodic DFT-D. XTalPi did not participate in the Sixth Blind Test [48], but their methodology has been demonstrated across a range of pharmaceutical systems [161–163], and for coformer screening [164]. As noted by some authors [163, 165], an appealing aspect of using FF models in CSP is the potential to integrate this with other FF-based molecular mechanics modeling approaches, like molecular dynamics. From these auxiliary techniques, other finite-temperature properties about the crystal may be predicted.

2.4.3 Hybrid Ab Initio/Empirical Force-Field Methods

Hybrid *ab initio*/empirical force-field methods are a class of lattice energy model that combine both QM and MM approaches to describe U_{latt}. HAIEFF models partition the lattice energy in the crystal into an intramolecular (ΔU_{intra}) and intermolecular contribution (U_{inter}),

$$U_{\text{latt}} = \Delta U_{\text{intra}} + U_{\text{inter}}. \tag{2.4.7}$$

The intramolecular contribution is given by

$$\Delta U_{\text{intra}} = U_{\text{intra}}^{\text{crys}} - U_{\text{intra}}^{\text{gp}}, \tag{2.4.8}$$

where $U_{\text{intra}}^{\text{crys}}$ is the conformational energy of the molecule in its crystalline conformation and $U_{\text{intra}}^{\text{gp}}$ is the gas-phase conformational energy. The intermolecular energy is given by

$$U_{\text{inter}} = U_{\text{elec}} + U_{\text{FF}}, \tag{2.4.9}$$

where U_{elec} is the electrostatic contribution from Coulombic interactions and U_{FF} is the empirical FF contribution. U_{FF} primarily describes van der Waals-type interactions like dispersion and exchange-repulsion, but it implicitly contains other residual intermolecular interactions.

In HAIEFF models, ΔU_{intra} and U_{elec} use QM-derived quantities, determined under the assumption that QM calculations on an isolated-molecule provide information that is applicable to the crystal. Since these calculations involve only monomers, they are extremely cost efficient compared to the previous methods employing QM calculations with ≥ 2 molecules. Through these relative cheap QM computations, HAIEFF models benefit in a few ways. Firstly, these isolated-molecule computations are practicable even with relatively high levels of electronic structure theory (e.g., hybrid DFT functionals, large basis sets). Secondly, by judicious sampling of the conformational landscape, surrogate models of the QM-derived quantities as a function of conformation can be constructed at QM-accuracy, which enables flexible molecular descriptions in the HAIEFF model. These benefits are conditioned on the validity of the isolated-molecule assumption, which for ΔU_{intra} and U_{elec} can give adequate results. Conversely, the residual intermolecular interactions mostly relate to many-body effects (e.g., electron-correlation, polarization), which cannot be represented with an isolated-molecule. This precludes a similar *ab initio* treatment of the residual interactions and explains their relegation to the FF term. Below I review the approaches used to treat each HAIEFF component.

2.4.3.1 Intramolecular Contributions

In principle, ΔU_{intra} is trivial to evaluate for a given molecular geometry. However, more-nuanced strategies may be necessary depending on: (1) the chosen QM method, and (2) the treatment of molecular flexibility. Regarding (1), DFT (usually with hybrid functionals) and MP2 are widely used for these monomer calculations. The pros and cons of these electronic structure methods have been detailed in Section 2.4.1, so here I only reiterate that delocalization error is a pertinent problem in DFT-predicted conformational energies (even when using hybrid functionals) [56, 90]. Like with periodic DFT, *post hoc* correction of DFT conformational energies with MP2 energies has been shown to be highly effective in HAIEFF models [56].

Addressing (2), there are a few common treatments of molecular flexibility in HAIEFF models. In the simplest case, the molecule is assumed to be rigid, i.e., its crystalline conformation is fixed to the gas-phase reference, and $\Delta U_{\text{intra}} \approx 0$. This is rarely applicable for industrially-relevant systems, and molecular flexibility must be accommodated in some way. As discussed in Section 2.3.1.3, flexibility can be restricted to several discrete rigid conformations during

candidate generation, for which ΔU_{intra} can be evaluated for each geometry. While this might be adequate for simple molecules, less-restrictive (and continuous) descriptions of molecular flexibility are desirable for more-complex molecules.

To that end, intramolecular models based on partial flexibility of conformational dofs have been applied to the HAIEFF framework [30, 93, 95–98]. To describe ΔU_{intra} as a continuous function of the flexible conformational dofs, it is too costly to continually evaluate $U_{\text{intra}}^{\text{crys}}$. Instead, discrete QM sampling of the conformational space is performed, with flexible dofs acting as independent variables. From the sampled data, surrogate models that describe a set of dependent variables as functions of the independent variables are constructed using derivative information. Here, the dependent variables are $U_{\text{intra}}^{\text{crys}}$ and the values of the other conformational dofs. An early implementation based on restricted cubic Hermite interpolants had been proposed [30, 93], but this suffered from several shortcomings [96]. Firstly, flexible dofs had to be sampled on uniform grids and, secondly, the quality of the interpolants deteriorated with more flexible dofs. Altogether, this approach had limited usefulness with more flexible dofs, and was ultimately superseded by the Local Approximate Model (LAM) [95] framework. Using LAMs, a second-order Taylor expansion approximates the relationship between flexible dofs and $U_{\text{intra}}^{\text{crys}}$, while a first-order expansion is applied to the conformational relationship [95]. LAMs have persistent quadratic accuracy and are not restricted to grid-sampling. This enables the adaptive placement of sample points, to focus computational effort on energetically relevant parts of conformational space [97]. Finally, a smoothing algorithm can be applied over the LAMs, to enhance the numerical continuity of the surrogate model [98]. Programs developed in our group that utilize the LAM model will be detailed in Chapter 4.

Permitting partial flexibility bolsters model fidelity, but introduces other challenges. Firstly, the cost of the energy model increases with the need to assess QM sample points. In larger molecules, this may restrict the QM method to lower levels of theory. Secondly, the selection of flexible dofs impacts the model predictions, yet this is often a subjective decision based on chemical intuition [77]. In my own experience, incorrect treatment of flexible dofs as dependent dofs can also have unpredictable numerical consequences on the LAM model.

2.4.3.2 Electrostatic Contributions

Unlike ΔU_{intra}, U_{elec} is not directly determined by QM calculations. Instead, from a QM calculation, molecular electrostatic features (e.g., partial-charges, dipoles, and higher-rank multipoles) are derived. These are then used to evaluate the electrostatic interaction according to Coulomb's law. Using distributed multipoles is critical for accurate descriptions of crystal energy, particularly in hydrogen bonded structures [166–168]. However, multipole models add considerable cost to lattice energy calculations (~10x [167, 169]), relative to atomic-charge models. Not only does each multipole rank introduce more pairwise terms to the Coulomb sum, higher-rank multipoles also employ a more-complex, anisotropic formulation of Coulomb's law. Near multipole-accuracy could be achieved using multiple point-charges per atomic site (e.g., off-atom satellite-charges

[93, 94]), but this approach is rare in modern CSP since determining the placement of the auxiliary charges is non-trivial. Considering their relative costs, in the multistage CSP workflow atomic-charge models are commonly used at the Global Search stage, while Refinement stages almost exclusively use distributed multipoles.

Beyond the electrostatic feature representation, other key modeling decisions include: (1) the choice of QM method, (2) the treatment of molecular flexibility, and (3) the method to derive the electrostatic features. I will not review (1) as it reiterates the previous discussion on the use of electronic structure methods for monomers. Regarding (2), if a rigid molecule representation is used, then electrostatic features need only be derived at the target conformation(s). If partial molecular flexibility is used, the LAMs framework permits similar surrogate modeling of charges/multipole moments from QM sample points [95–98]. Below, I discuss only (3).

Approaches to derive electrostatic features from the electronic structure fall into two broad categories: electrostatic potential (ESP) fitting methods, and atom-in-molecule (AIM) methods [170]. In the former, electrostatic features are empirically parametrized to the *ab initio* ESP determined from the electronic structure. In the latter, the electronic structure of the monomer is apportioned to individual atomic sites, using an AIM-scheme. From the distributed electronic structure, electrostatic features are then derived analytically. In HAIEFF models, both classes are employed to derive electrostatic features for the isolated-molecule. Analogous methods for periodic systems do exist, and interested readers are directed to Refs [170–172].

To determine atomic-charges, molecular ESP-fitting methods, like the Hu–Lu–Yang (HLY) [173] and CHelpG [174] algorithms, are popular choices for HAIEFF models. The HLY method is particularly suited for studying the conformational dependence of atomic-charges. Methods like CHelpG [174] and Merz–Singh–Kollman [175, 176], which use discrete fitting grids, are susceptible to 'bumpy' progression in their derived atomic-charges. The HLY method mitigates this problem by using numerical integration instead [173]. A well-known issue with empirical ESP-fitting methods is that they can lead to unphysical atomic-charges, especially for atomic sites far from the molecular surface [177]. Since fitting points within the molecular cavity are usually excluded, these 'buried atoms' have few data points associated with their fitting, leading to ill-defined partial-charge values. Nevertheless, ESP-fitting is convenient and provides adequate accuracy for the Global Search stage. AIM methods can provide atomic-charges, but these tend to perform poorly without higher-rank multipoles [177].

To derive distributed multipoles, ESP-fitting methods are atypical. ESP-fitting can already give rise to unphysical solutions when deriving atomic-charges, and this would be exacerbated for multipoles which have more decision variables. Instead, the multipoles used in CSP are most often determined by AIM methods. While AIM methods are more costly to execute, they tend to give more-reasonable multipole values. A popular choice for this is the Gaussian Distributed Multipole Analysis (GDMA) method of Stone [178–181]. In GDMA, overlap charge densities are defined at the intersection between pairs of primitive basis functions. From these charge densities, multipoles can be analytically determined, and then 'distributed' to appropriate atomic sites. In effect, GDMA apportions the basis functions of the molecule. Originally,

GDMA employed a nearest-neighbor distribution scheme [180], but this was found to give poor multipole convergence with basis set size, especially with diffuse basis functions [181]. As such, Stone proposed an alternative grid-based quadrature distribution scheme [181, 182], wherein each multipole is distributed over all atoms, weighted by its distance to each site.

GDMA has been suggested to give suboptimal partitioning of charge density, which leads to slow convergence of the multipole expansion [2, 183]. As such, truncating the GDMA multipoles at low ranks (≤ 1) tends to give poor predictions of the molecular ESP and electrostatic energy (hence why GDMA atomic-charges should not be used in isolation) [183, 184]. In lieu of GDMA, other AIM methods based on the real-space partitioning of charge density have been suggested, such as iterative-Hirshfeld analysis [185], and the iterated stockholder atom approach [186]. In these approaches, a molecule's electron density is distributed to atomic sites using a weighted average of proto-atom densities. Currently, these methods are as not widely used in HAIEFF models, but they offer potential for more-accurate predictions of electrostatic energy [2].

2.4.3.3 Validity of the Isolated-Molecule Assumption

Before I review the U_{FF} contribution, it is pertinent I discuss the implications of the isolated-molecule assumption on the preceding QM-derived quantities. This boils down to the question, "*Are the conformations, conformational energies, and charge densities derived in vacuo representative of the same molecule in a crystal?*". Unfortunately, the answer to this is often "*No*". Importantly, in crystals, a significant electric field can develop across the periodic lattice, particularly if the crystal contains molecules with strong charges (e.g., salts/zwitterions), highly polar moieties (e.g., hydrogen bonding systems), and/or large polarizable electron densities (e.g., large conjugated systems). Isolated-molecule QM calculations done in the gas-phase completely neglect this, which can have profound impact on the model predictions.

In the intramolecular model, neglect of the crystal's electric field can lead to catastrophically incorrect predictions of molecular conformation. For example, *in vacuo*, amino acids are most stable in their neutral form, yet in the crystal the electric field of the crystalline environment stabilizes the zwitterionic form instead [27, 187]. Without constraining specific bond lengths [187, 188], monomer calculations done *in vacuo* will not predict the desired zwitterionic form. In the electrostatic model, *in vacuo* evaluation of the monomer provides only the charge density inherent to the molecular geometry. I refer to this unperturbed state as the 'static' charge density (not to be confused with the 'static' definition used for U_{latt}). In the crystalline electric field, this charge density would be polarized from its static state. It has been shown that this polarization effect (also termed induction) can contribute significant electrostatic stabilization to the crystal and have large impact on relative crystal stabilities [183, 189].

Explicit modeling of the crystalline electric field is non-trivial and not routinely done in most HAIEFF models. A convenient implicit approximation is to conduct the QM calculation with the Polarizable Continuum Model (PCM) [190]. When using PCM, the isolated-molecule is subject to an external (polarizable) electric field, which is intended to provide a mean representation

of the crystalline environment. Several authors have noted that using PCM improves HAIEFF model predictions [48, 188, 191], due to the implicit embedding of some induction effects into the electrostatic model. In my experience, PCM can also help stabilize conformations that would otherwise be unstable *in vacuo*. However, as will be discussed in Section 2.5.2.1, there is little consensus on how the PCM dielectric constant should be determined, which somewhat hinders the robustness of the model. This, among other issues discussed in Section 2.5.2, make it desirable to explore explicit models for polarization for high-accuracy CSP.

2.4.3.4 Empirical Force-Field Contribution

The final contribution in the conventional HAIEFF framework is the empirical force-field term. This is sometimes referred to as the 'RD interaction', since it is expected to predominantly capture van der Waals-type interactions, like close-range repulsions from Pauli Exclusion and attractive long-range London dispersions. Like the vdW term in FF methods, U_{FF} is formulated as the pairwise sum of a pair potential. To reflect the leading RD interactions, HAIEFF models most commonly adopt the Buckingham pair potential (BHP),

$$U_{ij}^{\mathrm{FF,BHP}} = A_{ij} \exp\left(-\frac{r_{ij}}{B_{ij}}\right) - \frac{C_{6,ij}}{r_{ij}^{6}}, \tag{2.4.10}$$

where $U_{ij}^{\mathrm{FF,BHP}}$ is the energy contribution from the interaction between atom sites i and j separated by distance r_{ij}. A_{ij}, B_{ij}, and $C_{6,ij}$ are the Buckingham potential model parameters. The preponderance of the Buckingham potential in HAIEFF FFs stems from the good physical basis of its individual terms. Firstly, atomic orbitals described by Slater determinants are exponential in form, thus it is reasonable to assume a similar form for repulsions arising from orbital overlap [192, 193]. Secondly, the attractive contribution contains the leading r^{-6} term of the dispersion series. Higher-order r^{-2n} terms are truncated, since they decay more-rapidly at long-range and should have diminishing significance to U_{FF} [192, 193]. However, a well-known deficiency with the BHP is that the attractive r^{-6} term diverges faster than the repulsive exponential term at short-range, leading to unphysical attractions at these distances. Nevertheless, with careful selection of A_{ij}, B_{ij}, and $C_{6,ij}$, this is a non-issue as long as the divergence occurs at inaccessibly short range and is 'insulated' by the repulsive wall.

Like the MMFF models discussed earlier, HAIEFF FFs assign values to their model parameters on the basis of atom types and are usually parametrized for transferable-use. As such, most HAIEFF FFs employ predefined atom types with model parameters empirically fit to diverse datasets. Unlike MMFFs, HAIEFF FFs are trained specifically on crystal energies and geometries/forces, limiting their use to only CSP. Some popular transferable HAIEFF FFs include the FIT potential [166, 194–198], the William's potential (W99) [192, 193, 199], and the revised William's potential (W99rev) [169]. There is no restriction that HAIEFF FFs must be transferable, and molecule-specific HAIEFF FFs have been attempted in the past [70]. Molecule-specific HAIEFF FFs generally offer superior accuracy for the target system, but they are less common due to the added effort required to parametrize the FFs.

As a focal point to the current work, I discuss HAIEFF FF parametrization in more detail in Section 2.5. However, a critical consequence of the FF's empirical fitting is that it implicitly contains contributions beyond the RD interactions. These include residual intermolecular interactions such as: higher-order two-body dispersions (e.g., r^{-8}, r^{-10}), and many-body dispersions (e.g., the three-body ATM term). As electrostatic contributions are usually not decoupled from the FF training data, errors from the electrostatic model are also absorbed into U_{FF} (e.g., induction effects). Finally, if finite-temperature reference data is used, thermal effects are embedded into the FF. Using the empirical FF as an outlet for model errors is a double-edged sword. On one hand, it can serve as a band aid solution for complex, but important, interactions that would otherwise be absent in the model. On the other hand, if too much 'junk' is lumped into the FF, the chosen pair potential may struggle to reflect the gamut of contributions and FF parameter estimation can become extremely challenging. This can ultimately become detrimental to the FF's overall accuracy and transferability.

2.4.3.5 HAIEFF Models in CSP

Thus far, I have discussed the most common form of the HAIEFF model, specifically with transferable empirical FFs. Five out of 21 groups in the Sixth Blind Test [48] employed this type of HAIEFF model, demonstrating its continued relevance. Clearly, HAIEFF models are extremely versatile, offering a myriad of modeling options for each contribution. The 'modular' construction of the HAIEFF model makes it well-suited for CSP: depending on the system of interest and/or stage of the CSP workflow, there are numerous avenues to modulate individual components and tune the accuracy/cost of the lattice energy model. With that said, it must be acknowledged that these HAIEFF models are predominantly used for the Global Search and intermediate Refinement stages, but they are generally not accurate enough for the final ranking stage of CSP. Nyman *et al.* [66] benchmarked some HAIEFF models against popular periodic DFT-D approaches using the X23 test set. They found that the best performing HAIEFF models have AAD in predicted lattice energies ~9 kJ/mol, which is about 2–3 times larger than the periodic DFT-D methods [66]. In terms of standard deviations (SDs) in lattice energy error, the HAIEFF models were more competitive (~8.6 kJ/mol) with the periodic DFT-D models (~4–6 kJ/mol). Overall, HAIEFF models with transferable FFs give inferior accuracy to periodic DFT-D, thus it is still preferred to conduct final refinements with the latter.

2.4.4 SAPT(DFT)-Parametrized Intermolecular Force-Fields

SAPT(DFT)-parametrized intermolecular force-fields are a class of energy model that is seeing increasing popularity. Developers of this method aim to construct FFs that are theoretically robust and rely minimally on empirical fitting. The distributed intermolecular force-field (DIFF) model [92, 165, 200–202] is an example of this approach. The DIFF model shares many features with the HAIEFF intermolecular model: both use Born–Mayer repulsions, r^{-6} dispersions, and electrostatics described by a Coulomb sum. However, the DIFF model is further supplemented

with several other features, like anisotropic repulsions, higher-order two-body dispersions, induction, and damping for the relevant dispersion and electrostatic terms. In the DIFF model, many of the FF parameters (e.g., dispersion coefficients, multipole moments, polarizabilities) are determined wholly from first-principle calculations, while the remaining FF parameters (e.g., damping and Born–Mayer parameters) are obtained by empirical fitting to SAPT(DFT) dimer data. Since SAPT(DFT) decomposes total energy into contributions from the various physical interactions, this parameter estimation can be done in a highly targeted fashion. Variants of this FF have been tested on medium-sized rigid molecules, and were found to give sensible predictions of relative polymorphic stability [92, 165].

Podeszwa and co-workers demonstrated that a similar approach could be used to parametrize a simpler FF, which only comprises a Born–Mayer term, point-charge electrostatics, and an r^{-6} dispersion term [203]. Atomic-charges and r^{-6} coefficients are fit to long-range electrostatic and dispersion + induction energies, obtained from an *ab initio*-derived model. Damping and Born–Mayer parameters are then fit to SAPT(DFT) dimer data. This FF has been applied to CSP studies of medium-sized target molecules [204, 205], and in the Sixth Blind Test all four groups using SAPT(DFT)-parametrized intermolecular force-fields employed this approach [48].

Whilst such theoretically-driven models are appealing, I believe their applicability to modern CSP is currently limited. Incorporating molecular flexibility alongside the SAPT(DFT)-FF is non-trivial, since the *ab initio*-derived FF parameters need to be re-determined for large conformational changes [9]. If the empirically-fit parameters also need to be re-estimated following this update, this would further prohibit effective use of the model. Moreover, the reliance on *dimer*-based SAPT(DFT) data complicates FF parametrization in multicomponent systems, especially if there are distinct atom types defined for each molecule. Consider a multicomponent crystal with molecules A and B. To fit all pairwise interactions between the atom types of both molecules, data must be generated for all unique dimer pairs (i.e., A + A, A + B, and B + B). This is illustrated by Molecule XXIV in the Sixth Blind Test, a three-component salt hydrate [48]. Group 24 performed six distinct parameter estimations, each with ~900 SAPT(DFT) dimer evaluations [48]. Although SAPT(DFT) is a relatively cheap electronic structure method, costs can quickly balloon in this manner. Finally, like the TMFFs, these SAPT(DFT)-parametrized FFs are ultimately system-specific, limiting their use to other systems.

2.4.5 Machine Learning Methods

The lattice energy models discussed above can be classified as physics-based models, which describe physical phenomenon through explicitly defined semi-empirical/theoretical models. While physics-based models are appealing since they have explainable theoretical underpinnings, they also have several flaws. Precise formulation of a theoretical model is a time-consuming and complex process. Even after conceptualization, a theoretical model may not always be implemented faithfully to its formulation, and more often approximations/simplifications must be made for practical reasons. As such, machine learning (ML) models are an appealing option for CSP.

ML models rely on highly parametrized (1,000s of model parameters) general mathematical frameworks, which are 'flexible' enough to describe arbitrary relationships between given inputs and outputs. By leveraging large sets of reference data, these models can be parameterized to accurately reproduce their training data, under relatively short development times. Implicitly, the ML model 'learns' the underlying physics within the data, and embeds it in its general formulation. A detailed discussion of ML techniques is beyond the scope of this work. Below, I summarize some recent applications of ML to CSP model development

To enable their extensive parametrization, ML methods are extremely data-hungry. This is a problem for organic CSP, since experimental data is often scarce and of poor quality [206]. Synthetic data is the only viable option for ML, but it is challenging to generate sufficient high-quality data, whilst balancing costs. Organic crystals can be stabilized by extremely long-ranged intermolecular interactions (e.g., electrostatics). Unfortunately, the atomic descriptors used in ML force-fields (MLFFs) are often restricted to local atomic environments [207], so as to limit the demand for reference data. Training MLFFs to describe all crystal interactions, including these long-ranged effects, will exacerbate the data scarcity problem. As such, direct parametrization of MLFFs is rare in CSP, and instead Δ-ML models are a more popular strategy [65, 208–211]. In lieu of describing the lattice energy solely by ML, the Δ-ML strategy requires the ML model to learn only the correction between an existing physics-based energy model (e.g., a HAIEFF model) and a higher-fidelity reference model (e.g., periodic DFT-D). Since the physics-based model describes most of the underlying physics, the ML model does not need to 're-learn' this, reducing the data requirement for model training [210].

Various Δ-ML fitting schemes have been tested in CSP, which explore:

- Using data from monomers/clusters [208, 210], versus periodic systems [65, 209, 211].
- Gaussian Process Regression potentials [208–210], versus Neural Network potentials [65, 211].
- Fitting a single ML model to the total interaction energy [208, 209, 211], versus separate ML models for the intra- and intermolecular interactions [65, 210].

In all cases, the Δ-ML model accurately corrected between the low- (usually a HAIEFF model) and high-fidelity models, significantly improving CSP results. Going further, Δ-ML models can be 'stacked', such that additional ML models correct the error between the cumulative Δ-ML model and progressively higher-fidelity reference models [209]. To reach extremely high levels of accuracy, this cumulative multi-fidelity approach is more cost effective, since not all data needs to be generated at the highest fidelity [209]. Fully ML-based models could also be built this way. An MLFF trained on low-fidelity data acts as a baseline that is corrected by Δ-ML at higher-fidelity [212]. ML methods were used by a few groups in the Seventh Blind Test [49].

2.5 Current State of HAIEFF Intermolecular Models

Thus far, I have reviewed the different methodologies used for CSP. I will now focus on areas relevant to this thesis. In this work, I consider the HAIEFF class of energy model. Its versatility

makes it applicable to many stages of the CSP workflow and there is evidence that it will remain relevant to CSP for the foreseeable future. Unfortunately, there are several deficiencies in its intermolecular components, U_{FF} and U_{elec}, which hamper the accuracy of the HAIEFF model. Below, I review the issues pertinent to each component.

2.5.1 Empirical Force-Field parametrization

The reliability and accuracy of a FF is directly tied to the rigor employed in its parametrization. Nyman and co-workers [66] suggested that the accuracy gap between HAIEFF models and periodic DFT-D methods is partially contributed to by the HAIEFF FFs, and better HAIEFF FFs are needed. In my view, HAIEFF FF development has largely stagnated and in Table 2.1 I summarize some notable HAIEFF FFs. Other than the FIT [166, 194–198], W99 [192, 193, 199], and W99rev [169] FFs, that are widely used in CSP, I have included parametrizations by Gatsiou *et al.* [213] and Bowskill (denoted 'DB' in this work) [2]. The process of fitting these FFs comprises five key steps: (1) reference data curation, (2) definition of the FF model, (3) definition of the electrostatic model, (4) FF parameter estimation, and lastly (5) deployment in CSP.

2.5.1.1 Reference Data Curation

The first step of performing any parameter fitting is to assemble an adequate collection of reference data, which for CSP predominantly comprises crystal energies and geometries. Most parametrizations (all except DB in Table 2.1) have relied on experimentally reported crystal structures and sublimation energies, but this is problematic for a few reasons.

Firstly, there is a scarcity of reliable experimental data. While good-quality crystal structure data is available from the CSD, sublimation enthalpy (ΔH_{sub}) data tends to be much rarer. Recently, new databases of sublimation energies have been collated [214], but nowhere near the scale of the CSD. Even when reported, ΔH_{sub} data is often unreliable, due to different measurement methodologies, impure samples, and human error. Large variations in reported ΔH_{sub} are common [214] with one estimate suggesting measurement uncertainties ~4–5 kJ/mol [206], which is of similar magnitude to expected polymorphic energy differences. This lack of reliable experimental data can be especially severe in systems with uncommon elements like sulphur or chlorine [213]. As such, FF parametrization employing experimental data often have highly imbalanced datasets that bias geometry data (*cf.* Table 2.1).

Another issue with experimental data is its implicit inclusion of thermal effects. In principle, F_{vib} can be evaluated so as to remove thermal effects from the reference data. However, doing this at high accuracy is prohibitively expensive and impractical for FF parameter estimation. Instead, previous FF parametrizations have opted to either ignore the thermal contribution (e.g., FIT and W99) or apply a simple Dulong–Petit correction (e.g., W99rev and Gatsiou *et al.*):

$$\Delta H_{\mathrm{sub}} = -U_{\mathrm{latt}}^{\mathrm{DLP}} - 2RT, \tag{2.5.1}$$

where R is the universal gas constant. It has been noted that this correction is approximate

Table 2.1: Summary of some transferable HAIEFF FFs and their parametrization strategies. All FFs use the Buckingham pair potential as shown in Equation 2.4.10.

Name	**Fitting Data**[a]	**Data Size**[b]	**Atom Typing**[c]	**Heavy Elements**	**Unlike?**	***Ab Initio* Model**	**Electrostatic Model**[d]	**Reference(s)**
FIT	ΔH_{sub} + Geom.	49 (49)[e]	Element + H neighbor	C, O, N, F, Cl, S[f]	✗	MP2/6-31G**	MP	[166, 194–198]
W99	ΔH_{sub} + Geom.	334 (26)	Element + CN + H neighbor	C, O, N	✗	HF/6-31G**	PC	[192, 193, 199]
W99rev	$U_{\text{latt}}^{\text{DLP}}$ + Geom.	186 (53)	Element + CN + H neighbor	C, O, N	✓	B3LYP/6-31G** or 6-311G** with and without PCM	MP	[169]
Gatsiou *et al.*	$U_{\text{latt}}^{\text{DLP}}$ + Geom.	106 (53)	Element + H neighbor	C, O, N, F, Cl, S	✓	M06/6-31G(d,p)	MP	[213]
Bowskill (DB)	$U_{\text{inter}}^{\text{ref}}$ + Geom.$_{\text{DFT-D}}$	445 (445)[g]	Element + H neighbor	C, O, N, F, Cl, S	✓	Multiple models	PC and MP	[2]

[a] All reference data are from experimental sources, unless otherwise stated. Experimental estimates of U_{latt} may be derived from ΔH_{sub} with the Dulong–Petit correction (denoted 'DLP'). Bowskill determines $U_{\text{inter}}^{\text{ref}}$ (and the corresponding geometry) from periodic DFT-D, with care to remove intramolecular contributions from the reference energy.

[b] Geometry information is available for all data points, but not all training set crystals have energy data. The values in parenthesis denote frequency of energy data.

[c] Atom typing usually involves the element and bonded neighbors (for hydrogen), but coordination number (CN) is sometimes used as well.

[d] The electrostatic description used during the parametrization include point-charges (PC) and multipole (MP) models.

[e] The parameters for FIT are taken from various sources, hence the training data used in the final set is slightly unclear [166, 198].

[f] The parameters for F, Cl, and S were not parametrized in conjunction with the the other elements.

[g] All 445 structures had both geometry and energy data available, but not all energy data was used.

and may introduce more errors than it solves [131, 142]. In terms of geometry, crystal unit cells usually expand by 2–15% upon heating [215, 216], but reversing this expansion is non-trivial since it is often anisotropic [216]. Ultimately, there is no convenient and accurate way to remove thermal effects from experimental data, hence FFs parametrized to this type of data will implicitly incorporate these effects [66]. Pressure also affects experimental data, but its effects are likely negligible at typical measurement conditions.

In lieu of experimental data, members of our group have used synthetic data, from periodic DFT-D, to fit HAIEF FFs [2, 70]. This follows the lead of TMFFs, MLFFs, and SAPT(DFT)-parametrized FFs, which all rely on data generated with electronic structure methods. Using synthetic data eludes the issues associated with experimental data: QM calculated data is unambiguous at static conditions, and any data shortage can be remedied with more computational effort. From a 'philosophical' perspective, fitting to periodic DFT-D data is also more consistent with the overarching CSP workflow. Under 'blind' conditions, periodic DFT-D is the 'ground truth' and the preceding steps in the CSP workflow should aim to reproduce its results.

Whether experimental or synthetic data is used, *rigid* molecules are preferably selected, since an accurate description of molecular flexibility and ΔU_{intra} is challenging to incorporate into the parameter estimation. By using rigid molecules, one can approximate $\Delta U_{\text{intra}} \approx 0$ and assume the reference energy is a result of only the intermolecular components. Unfortunately, this is not always possible since some atomic environments are inherently flexible (e.g., sulfonyl groups). In these cases, one could decouple the intra- and intermolecular energies by (for example) calculating ΔU_{intra} via QM calculations and deducting it from the total reference energy. However, to my knowledge, none of the parametrizations using experimental data have done this. Where synthetic data has been used [2], the reference intermolecular energies ($U_{\text{inter}}^{\text{ref}}$) have been appropriately isolated from any intramolecular contributions.

2.5.1.2 Force-field Model Definition

Once the reference data has been curated, the FF model must be defined. This entails several modeling decisions, beginning with the choice of atom types. Among HAIEFF FFs, most do not use particularly detailed atom types: heavy atoms are distinguished by only proton number, while separate hydrogens may be defined according to bonded-neighbor (polarizing versus non-polarizing species). The exception to this are the W99 [192, 193, 199] and W99rev [169] FFs, which also consider the coordination number (CN) of heavy atoms and the hydrogen's chemical moiety. Unfortunately, these FFs only span a limited range of elements. Intuitively, an atom's local environment (e.g., bonded neighbors, hybridization) will influence the nature of its interactions [217] and the atom types should reflect this. For instance, models like TMFFs [99], MMFFs [153–157], and DFT dispersion-corrections [85] all employ more-extensive atom typing schemes. In short, to achieve high accuracy, the atom types currently employed in most HAIEFF FFs is inadequate. The choice to limit atom types may have (in the past) been driven by data scarcity, which hinders robust determination of many parameters.

The formulation of the pair potential is another important decision. Although all existing HAIEFF FFs default to the Buckingham potential (Equation 2.4.10), it is unclear that this is necessarily the best option. While the BHP has some sound physical traits, it is deficient in other aspects. This is illustrated by the DIFF model [92, 165, 200–202], which employs a more-complex formulation to describe the same interactions as U_{FF}:

$$U_{ij}^{\mathrm{FF,DIFF}} = A_{ij}\exp\left(-\frac{r_{ij}}{B_{ij}}\right) - f_6(\beta_{\mathrm{disp}}, r_{ij})\cdot\frac{C_{6,ij}}{r_{ij}^6} - f_8(\beta_{\mathrm{disp}}, r_{ij})\cdot\frac{C_{8,ij}}{r_{ij}^8}, \tag{2.5.2}$$

where f_n is a damping function for r^{-n} dispersions, and β_{disp} is the associated damping parameter. While the r^{-n}-dependency of dispersions is correct at long-range, dispersion interactions should converge to a constant, non-zero finite value at $r = 0$, due to charge overlap [86]. Damping functions are (semi-)empirical approaches to 'fix' this erroneous behavior at short-range, and are also employed in all DFT dispersion-corrections based on the classical r^{-2n} model of dispersions [85, 87–89]. The absence of higher-order dispersion terms (r^{-8} and r^{-10}) is another deficiency of the BHP. While the leading r^{-6} term should be the dominant dispersion contribution, studies have shown that the higher-order terms (particularly r^{-8}) can contribute non-negligible stabilization to crystals [136]. More generally, these higher-order terms are important to the fine-details of medium-range dispersions. Dispersion damping and higher-order dispersions are just two examples of the BHP's physical inadequacies.

Going further, there are also questions about the fundamental form of the pair potential. While the BHP is a sound formulation for RD interactions, it is known that U_{FF} contains other implicit contributions. To that end, a different pair potential may offer greater flexibility to model the collective interactions in U_{FF}. For instance, the Mie potential could be considered

$$U_{ij}^{\mathrm{FF,Mie}} = C_{ij}^{\mathrm{Mie}}\cdot\varepsilon_{ij}\cdot\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{\lambda_{ij}^{\mathrm{r}}} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{\lambda_{ij}^{\mathrm{a}}}\right], \tag{2.5.3}$$

$$C_{ij}^{\mathrm{Mie}} = \frac{\lambda_{ij}^{\mathrm{r}}}{\lambda_{ij}^{\mathrm{r}} - \lambda_{ij}^{\mathrm{a}}}\cdot\left(\frac{\lambda_{ij}^{\mathrm{r}}}{\lambda_{ij}^{\mathrm{a}}}\right)^{\frac{\lambda_{ij}^{\mathrm{a}}}{\lambda_{ij}^{\mathrm{r}} - \lambda_{ij}^{\mathrm{a}}}}, \tag{2.5.4}$$

where ε_{ij}, σ_{ij}, $\lambda_{ij}^{\mathrm{r}}$, and $\lambda_{ij}^{\mathrm{a}}$ are the energy parameter, radius parameter, repulsive exponent, and attractive exponent respectively. The 12–6 ($\lambda_{ij}^{\mathrm{r}}$ and $\lambda_{ij}^{\mathrm{a}}$ respectively) Lennard–Jones potential is a popular fixed-form variant of the Mie potential used in MMFFs (e.g., GAFF [157] and OPLS [156]). In TMFFs, 12–6 potentials [109] and 9–6 potentials [101] have also been used to describe vdW interactions, while 10–6 [109] and 12–9 [99] potentials have been used in hydrogen bonding terms. In all of these models, good predictions are achieved without using the BHP.

Numerically, FFs based on the BHP are handicapped as well. Due to its exponential formulation, the BHP is extremely sensitive towards B_{ij}'s value [169, 213], and there exists extremely strong correlation between the A_{ij}/B_{ij} parameters [192, 193, 199]. As such, simultaneous fitting of all three parameters often leads to numerical failure, and is almost never attempted. In effect, the BHP has only two effective parameters per pairwise interaction (A_{ij}, and $C_{6,ij}$), rendering it

a less flexible model. Moreover, the short-range divergent behavior of the BHP can be a massive source of grief during parameter estimation. While normally insulated by the repulsive wall, the unphysical region of the BHP may inadvertently be exposed as parameter values are varied, leading to undesirable numerical behavior. This is especially so when extremely attractive interactions (like hydrogen bonds) are implicitly embedded into U_{FF} [2].

The final decision related to the FF model definition is the treatment of homoatomic (like, $i = j$) and heteroatomic (unlike, $i \neq j$) interactions. While like-interactions are always parametrized, unlike-interaction parameters are commonly determined by combining rules, like the geometric mean rule (Equation 2.5.5) and the arithmetic mean rule (Equation 2.5.6). Although these rules lack a robust physical justification, limiting the number of fitted parameters in this way was still frequently done, most likely due to data scarcity. Recently, Pyzer-Knapp *et al.* [169], Gatsiou *et al.* [213], and Bowskill [2] have all performed FF fitting with parametrization of the unlike-interactions. As expected, this generally led to improved model performance.

$$A_{ij} = \sqrt{A_{ii}A_{jj}}, \qquad C_{6,ij} = \sqrt{C_{6,ii}C_{6,jj}}, \tag{2.5.5}$$

$$\frac{1}{B_{ij}} = \frac{\frac{1}{B_{ii}} + \frac{1}{B_{jj}}}{2}. \tag{2.5.6}$$

2.5.1.3 Electrostatic Model Definition

During HAIEFF FF fitting, evaluation of the electrostatic component is still necessary to complete the (rigid) lattice energy model. HAIEFF FFs can be parametrized for any electrostatic model, but only the DB [2] and W99rev FFs [169] have been fit with more-modern strategies for determining electrostatic features (e.g., PCM). Since reference data used in the FF parametrization is not decomposed into separate HAIEFF contributions, errors in the electrostatic model will also be absorbed into the FF during parameter estimation. This in itself is not generally an issue, but becomes relevant during FF deployment (Section 2.5.1.5).

2.5.1.4 Force-field Parameter Estimation

The preceding steps have led up to the FF parameter estimation. Some key methodological decisions have been explored by previous authors. First, is the construction of the objective function (OF). Since only stable/meta-stable crystals are used as reference data, the HAIEFF model should predict these structures as energy minima. To enforce this optimality criteria, two approaches have been proposed. For the FIT and W99 FFs, crystals are held completely rigid at their reference geometries during the parameter estimation, and the OF is formulated as

$$\mathrm{OF}(\bar{\boldsymbol{p}}) = \sum_{s=1}^{N_{\mathrm{struc}}} \omega_{e,s} \cdot \left[U_{\mathrm{inter},s}^{\mathrm{HAIEFF}}(\bar{\boldsymbol{p}}; \boldsymbol{\Omega}^{\mathrm{ref}}, \boldsymbol{\theta}^{\mathrm{ref}}) - U_s^{\mathrm{ref}}\right]^2 + \sum_{s=1}^{N_{\mathrm{struc}}} \sum_{v=1}^{N_{\mathrm{DV},s}} \omega_{f,v,s} \cdot F_{v,s}^2(\bar{\boldsymbol{p}}; \boldsymbol{\Omega}^{\mathrm{ref}}, \boldsymbol{\theta}^{\mathrm{ref}}), \tag{2.5.7}$$

where $\bar{\boldsymbol{p}}$ are the FF parameters being estimated. The first term involves minimizing the absolute error between the predicted intermolecular energy ($U_{\text{inter},s}^{\text{HAIEFF}}$) and the reference energy (U_s^{ref}) over N_{struc} structures in the training data. Crucially, the predicted energies are determined at the reference conformations ($\boldsymbol{\theta}^{\text{ref}}$) and reference crystal geometries ($\boldsymbol{\Omega}^{\text{ref}}$), i.e., only *single-point* energy evaluations need to be done at the reference geometry. Values after the ';' are input parameters into the energy evaluation, and are not varied. To enforce optimality, the second term minimizes the forces ($F_{v,s}$) acting on each structure s, with respect to their rigid-body crystal structure variables (indexed v). The energy and force terms are weighted respectively by $\omega_{e,s}$ and $\omega_{f,v,s}$. In principle, if the forces are adequately minimized, then during a (rigid) energy minimization, the reference geometry would rightfully be an energy minimum. This framework is appealing for its relative simplicity and low computational cost, since only single-point energy evaluations are done. However, the optimality criteria at the reference geometry is enforced only as a 'soft' constraint, by embedding it as a penalty within the OF. The weight of this criteria may be further swayed by the chosen values of ω. Ultimately, the optimum parameter estimates may fail to predict the reference geometry as an energy minimum.

Perhaps facilitated by greater computing power, more-recent FF parametrizations have opted to explicitly embed the (rigid) energy minimization into the parameter estimation [2, 169, 213]. The corresponding OF that can be generalized as

$$\text{OF}(\bar{\boldsymbol{p}}) = \sum_{s=1}^{N_{\text{struc}}} \omega_{e,s} \cdot \Delta U_{\text{inter},s}^{\text{HAIEFF}}(\bar{\boldsymbol{p}}, \boldsymbol{\Omega}^*(\bar{\boldsymbol{p}}); \boldsymbol{\theta}^{\text{ref}}) + \sum_{s=1}^{N_{\text{struc}}} \omega_{g,v,s} \cdot \Delta G_s^{\text{HAIEFF}}(\bar{\boldsymbol{p}}, \boldsymbol{\Omega}^*(\bar{\boldsymbol{p}}); \boldsymbol{\theta}^{\text{ref}}), \quad (2.5.8)$$

where $\Delta U_{\text{inter},s}^{\text{HAIEFF}}$ and $\Delta G_{\text{inter},s}^{\text{HAIEFF}}$ are respectively measures of energy and geometry deviations between the HAIEFF-predicted quantities and the reference data. As before, each of these terms are weighted by $\omega_{e,s}$ and $\omega_{g,v,s}$ respectively. Unlike the earlier framework, all OF quantities are determined at explicit energy minima, denoted by their dependency on the optimum crystal structure at a given set of parameter estimates ($\boldsymbol{\Omega}^*(\bar{\boldsymbol{p}})$). In short, during the parameter estimation, all reference structures must be energy minimized at a given $\bar{\boldsymbol{p}}$, before optimality of the parameter estimation can be assessed. In this way, this framework strictly enforces the optimality criteria at the reference geometry, but at higher cost.

Other than the chosen OF-framework, another important decision in FF parametrization is whether to estimate parameter values simultaneously, or sequentially. During sequential fitting, the set $\bar{\boldsymbol{p}}$ is partitioned into smaller subsets, most commonly by atom types [213]. The subsets of $\bar{\boldsymbol{p}}$ are fit in succession, preserving any parameter values determined earlier in the sequence. This approach is attractive for several reasons, and has been more widely adopted (e.g., FIT, W99, W99rev, and Gatsiou *et al.*). By fitting smaller sets of $\bar{\boldsymbol{p}}$, fewer decision variables and less reference data are involved per estimation. This can make each parameter estimation more affordable, compared to mass parametrization of all $\bar{\boldsymbol{p}}$. Moreover, sequential fitting can reduce correlation between FF parameters, as unrelated parameters are not fit together and cannot assimilate each other's errors. However, sequential fitting can be more time-consuming, since several rounds of parameter estimation must be performed to obtain a complete FF. Moreover,

sequential fitting is not immune to unphysical parameter values. In fact, errors in the parameter estimates can accumulate and be inadvertently embedded into parameters that are fit later in the sequence. In light of this, our group has favored developing tools for large-scale parameter estimation, allowing the treatment of several hundred structures in the dataset and simultaneous parametrization of > 60 FF parameters [1, 2]. In our view, this permits more-efficient FF fitting and avoids error-propagation, leading to higher quality parameter estimates.

2.5.1.5 Force-field Deployment

The deployment of the HAIEFF FF into real CSP studies is not strictly part of the 'parameter estimation', but there are issues here that are deeply entwined with how parameter estimation is viewed and approached. As mentioned before, the empirical nature of HAIEFF FFs mean that they are an outlet for errors from other components of the HAIEFF model [70, 169]. In principle, this means that FFs parametrized with a given *ab initio* model should be used only with that same model during CSP, lest they introduce indeterminate errors from mismatched *ab initio*/FF models. While most practitioners of CSP would concur with this best practice, in reality CSP studies are rife with *ad hoc* deployment of FFs. For instance, during the Sixth Blind Test [48], several groups that used HAIEFF models would pair the FIT FF with atomic-charges during the Global Search, despite it being parametrized for use with multipole models [166, 198]. Groups that employed PCM when deriving multipole moments were sometimes careful to use the W99rev FFs, which were parametrized for use with a PCM dielectric constant ($\varepsilon_{\mathrm{PCM}}$) of 3.0 [169]. However the same FF was also used at higher values of $\varepsilon_{\mathrm{PCM}}$. An egregious practice with molecules containing more 'exotic' atoms, like sulphur or chlorine, has been retroactively 'tacking on' parameters from other parametrizations, without re-fitting the rest of the empirical FF for consistency. In some cases, the level of theory and basis set used in the CSP are also incommensurate with the parameter fitting *ab initio* model. Some might attribute these improvisations to the time constraints of the Blind Tests, but even outside the Blind Tests, these practices are commonly observed [54, 57, 96]. Ultimately, I believe that CSP practitioners have grown accustomed to overlooking this bad practice, since the extent to which these modeling choices might affect CSP outcomes has not been clearly quantified.

2.5.2 Induction in HAIEFF

Moving away from the FF term, I now evaluate the HAIEFF electrostatic term. In my view, the biggest flaw with U_{elec} is its reliance on the isolated-molecule assumption, which neglects all effects of induction/polarization. Induction has been suggested to contribute non-negligibly to the stabilization of many crystals, with estimates of as much as 20–40% of the total electrostatic energy [189]. Strong induction effects are not restricted to only hydrogen bonding compounds, and even non-hydrogen bonding compounds can have significant induction contributions (e.g., polyaromatic systems) [183, 189]. Unfortunately, induction is by definition a many-body effect, hence its accurate modeling is challenging to accomplish at reasonable cost. Several approaches

to model induction have been explored within the HAIEFF framework, and below I review some of these strategies.

2.5.2.1 Implicit Models for Induction

As described in Section 2.4.3.3, only the PCM model developed by Tomasi *et al.* [190] is routinely applied today to approximate polarization in HAIEFF models. When PCM is used during a molecular QM calculation, the isolated-molecule is placed in a polarizable continuum environment, instead of a vacuum. As its name suggests, the continuum polarizes in response to the molecular charge density, and develops a charge at its interface with the molecule. The sensitivity of the continuum towards the molecular charge density is characterized by its dielectric constant, $\varepsilon_{\mathrm{PCM}}$, where larger $\varepsilon_{\mathrm{PCM}}$ values reflects greater continuum polarizability. The continuum and molecule continue to polarize each other until convergence is reached in a self-consistent reaction field (SCRF) calculation. In principle, the PCM continuum provides an implicit approximation of the polarizable crystalline environment, such that the SCRF molecular charge density has induction embedded in it. The corresponding electrostatic features derived from this polarized charge density can be used in the evaluation of U_{elec}. Multiple CSP studies using HAIEFF models report that the addition of PCM leads to improved CSP results [2, 48, 169, 188, 191]. However, there are reasons to believe that this implicit model is still inadequate.

Firstly, evidence suggests that there is a failure in the PCM model, where it pertains to choosing appropriate values of $\varepsilon_{\mathrm{PCM}}$. While $\varepsilon_{\mathrm{PCM}} \approx 3.0$ is purported to be representative of organic crystals [48, 191, 218, 219], there is significant anecdotal evidence that using larger $\varepsilon_{\mathrm{PCM}}$ may be better for CSP. For instance, when modeling amino acids, Cooper *et al.* found increasing $\varepsilon_{\mathrm{PCM}}$ to as much as 11.0 improved CSP predictions [188]. Likewise, in the Sixth Blind Test [48], Price's group used $\varepsilon_{\mathrm{PCM}} = 11.0$ for molecule XXIV, a salt system, and Day's group commented that $\varepsilon_{\mathrm{PCM}} > 3.0$ is likely beneficial for salt and hydrate systems. The enhanced outcomes obtained as $\varepsilon_{\mathrm{PCM}}$ is increased to 'unphysical' values is a warning that the underlying model is flawed. A theoretical understanding of induction in the crystal supports this proposition, and discredit the validity of the PCM model. PCM polarization occurs only at the locality of the molecular cavity, with a 'mean-field' representation. Highly directional, close-range polarizing effects like hydrogen bonding cannot be captured in this way, and some have suggested that the use of larger $\varepsilon_{\mathrm{PCM}}$ could be compensating for this [188]. Concurrently, induction can also arise from long-range propagation of electrostatic effects through the periodic lattice. A continuum model like PCM cannot reflect this effect, since it exhibits no periodicity.

Overall, PCM is a useful first-approximation of induction, appropriate for early-stage CSP where poorer accuracy is tolerated for better computational cost. However, PCM simply cannot provide the nuance required to faithfully reproduce differences in polarization that arise between polymorphs. Describing these effects more accurately will likely aid in the resolution of small polymorphic energy differences.

2.5.2.2 Explicit Models for Induction

Induction is a complex many-body effect that emerges from the interaction between the charge density of molecules in the crystal [220, 221]. Various models of induction have been proposed [220], but I focus my review solely on explicit models developed specifically for CSP. Common to these approaches is a reliance on the classical model of induction energy, where the pairwise induction contribution ($U_{ij}^{\mathrm{ind},l_i m_i l_j m_j}$) is given by the interaction between an *induced* multipole on atom site i with rank l_i and moment m_i ($\Delta Q_i^{l_i m_i}$) interacting with a *static* multipole on atom site j with rank l_j and moment m_j ($Q_j^{l_j m_j}$) [221, 222],

$$U_{ij}^{\mathrm{ind},l_i m_i l_j m_j} = \frac{1}{2} \cdot f_{l_i m_i l_j m_j}(\beta_{\mathrm{ind}}, r_{ij}) \cdot \Delta Q_i^{l_i m_i} T_{ij}^{l_i m_i l_j m_j} Q_j^{l_j m_j}, \tag{2.5.9}$$

where $f_{l_i m_i l_j m_j}$ is an induction damping function with damping factor β_{ind}, and $T_{ij}^{l_i m_i l_j m_j}$ is the corresponding interaction tensor element describing the distance and orientation dependency of the multipole interaction. The $\frac{1}{2}$ prefactor prevents double-counting when the pairwise sum is performed over the infinite lattice. This formulation is an analogue of the static charge–charge Coulomb interaction shown in Equation 2.4.6. The classical model is convenient as it has a simple closed-form expression, but it only approximates the full induction energy. Firstly, the classical model assumes pairwise additivity of the induction contribution, but this is not the case and induction can have a significant many-body component [220, 223]. Nevertheless, the two-body contribution is the leading term and should describe the majority of induction stabilization. Moreover, while the classical model is asymptotically correct at long-range, it does not correctly reflect phenomena that occur at short-range when the charge densities of two molecules overlap (e.g., charge penetration and transfer) [221, 224]. The damping function is a coarse empirical tool to fix this (and provide better numerical behavior), yet the form of the damping function is not always well justified. To employ this formulation, two methods for determining $\Delta Q_i^{l_i m_i}$ have been proposed, based on either 'distributed polarizabilities' [222–225], or 'electronic-embedding' [189, 226].

In the distributed polarizabilities framework, $\Delta Q_i^{l_i m_i}$ are determined by [189, 223, 224]

$$\Delta Q_i^{l_i m_i} = -\sum_{J \neq I} \sum_{j \in J} \sum_{l_i' m_i' l_j m_j} \alpha_i^{l_i m_i l_i' m_i'} \cdot f_{l_i' m_i' l_j m_j}(\beta_{\mathrm{ind}}, r_{ij}) \cdot T_{ij}^{l_i' m_i' l_j m_j} (Q_j^{l_j m_j} + \Delta Q_j^{l_j m_j}), \tag{2.5.10}$$

where $\alpha_i^{l_i m_i l_i' m_i'}$ is the distributed polarizability of site i. $\alpha_i^{l_i m_i l_i' m_i'}$ quantifies the linear-response of the $l_i m_i$-multipole to the $l_i' m_i'$-component of an electric field at i [223]. The summation of $T_{ij}^{l_i' m_i' l_j m_j}(Q_j^{l_j m_j} + \Delta Q_j^{l_j m_j})$ over all sites j on J molecules (excluding sites on molecule I, which contains i) is the intermolecular electric field at site i from all other (static and induced) multipoles. This follows an intuitive view of the induced multipole as a response to perturbations from the surrounding electric field. Since each induced multipole contributes to the electric field of the other induced multipoles, Equation 2.5.10 must be solved simultaneously and self-consistently for all $\Delta Q_i^{l_i m_i}$ [189, 223, 224]. The iterative nature of this process reflects the higher-order propagation of induction throughout the periodic crystal [223, 224]. As in Equation 2.5.9,

the inclusion of the damping function serves both a physical (capturing effects of charge-overlap) and a numerical (preventing the 'polarization catastrophe') role [189].

The challenge in the distributed polarizabilities induction model is determining reliable values of $\alpha_i^{l_i m_i l'_i m'_i}$. To do this, the frequency-dependent density susceptibility (FDDS) is first evaluated. Using a density-fitting scheme, Misquitta & Stone proposed a method to determine non-local distributed polarizabilities from the FDDS [222]. Thereafter, the Williams–Stone–Misquitta (WSM) localization scheme [223, 224] can be applied to eliminate non-local elements of $\alpha_i^{l_i m_i l'_i m'_i}$. The localized $\alpha_i^{l_i m_i l'_i m'_i}$ can then be used in Equation 2.5.10. More recently, Misquitta & Stone suggested an alternative method (ISA-Pol) for obtaining non-local distributed polarizabilities using the iterated stockholder atoms AIM partitioning scheme [225], that can also be localized by the WSM method if necessary. They suggest that this real-space partitioning scheme gives more physically sensible distributed polarizabilities [225].

The second method for determining $\Delta Q_i^{l_i m_i}$ is based on electronic-embedding [226], and is termed the self-consistent electronic response to point-charge (SCERP) model [189]. To initialize the SCERP procedure, all symmetrically-unique molecules in the asymmetric unit cell are subject to an *in vacuo* SCF calculation, from which atomic-charges are derived from the SCF charge density. Thereafter, each molecule is embedded in a field of point-charges. Notably, these point-charges are positioned at the crystallographic positions of the atoms in the lattice and each point-charge assumes the value corresponding to its respective atom's atomic-charge. In effect, a cluster forms wherein only the target molecule is represented by real atoms, while all other crystallographic molecules are represented by only their QM-derived charges. Monomer SCF calculations are repeated to update the target molecule's charge density and atomic-charges, in the presence of the charge field. This is done until self-consistency is reached, mimicking the higher-order propagation of polarization through the crystal, similar to the WSM method. Once the derived point-charges converge, a proxy for evaluating convergence of the charge density, multipoles can be evaluated through a chosen multipole derivation scheme (e.g., GDMA) [71]. The multipoles derived from the perturbed charge density reflect the 'total' multipoles ($Q_{\mathrm{total,i}}^{l_i m_i}$), and the induced multipoles are trivially obtained by

$$\Delta Q_i^{l_i m_i} = Q_{\mathrm{total,i}}^{l_i m_i} - Q_i^{l_i m_i}. \tag{2.5.11}$$

In terms of cost expended on QM calculations, distributed polarizability methods only perform one monomer calculation to obtain non-local distributed polarizabilities, while the SCERP method must iterate through multiple QM calculations. However, it must be emphasized that SCERP uses only *ground-state* SCF calculations which are fairly economical. In contrast, the FDDS required by distributed polarizability methods are obtained by *linear-response* DFT calculations, which are significantly more expensive than 'regular' DFT. To illustrate, when modeling carbamazepine (30 atoms) [189], SCERP QM calculations could be directly performed on the target molecule, while application of the distributed polarizability method required the molecule to be partitioned into computationally more-manageable fragments, during QM calcu-

lations. Moreover, SCERP's self-consistent point-charge framework has strong parallels with the PCM approach, except that the former uses 'polarizable' point-charges at discrete sites instead of a continuum. In commercial implementations like Gaussian [227], using PCM add little to no cost over *in vacuo* calculations, suggesting that efficient implementations of SCERP could also be extremely cost effective.

Integrating either of these methods into a lattice energy minimization requires re-evaluation of their relevant QM-derived quantities. For the distributed polarizability method, $\Delta Q_i^{l_i m_i}$ can be determined by Equation 2.5.10 at different packing geometries, but $\alpha_i^{l_i m_i l_i' m_i'}$ must be re-derived if the conformation changes. As such, molecular flexibility cannot be trivially introduced in this approach, and indeed all demonstrations of the WSM-method have been on rigid molecules [92, 165, 189]. On the other hand, the SCERP model is dependent on both the molecular conformation *and* the packing geometry. Consequently, over the course of a flexible lattice energy minimization, it is likely that the SCERP method will need to conduct considerably more QM calculations than a distributed polarizability method.

In terms of accuracy of the derived $\Delta Q_i^{l_i m_i}$, solving Equation 2.5.10 iteratively and self-consistently can become challenging with higher multipole ranks, thus use of the WSM induction model has been largely restricted to rank 2 (up to induced quadrupoles) [165, 189]. Conversely, the SCERP methodology is agnostic to the desired multipole rank and it is trivial to obtain $\Delta Q_i^{l_i m_i}$ up to any rank. However, the fidelity of SCERP suffers for other reasons. Firstly, its electric field is currently limited to point-charges [71, 189], which permits only a crude representation of the crystalline electric field. Secondly, SCERP is able to reflect only short- to medium-range polarization, since it is restricted to embedding-clusters of finite-size. Extremely long-range polarization effects which manifest only in the limit of an infinite lattice may not be well described by this finite-range electronic-embedding approach [226, 228–230]. Distributed polarizability methods do not have these limitations since they can explicitly use higher-ranked multipoles in Equation 2.5.10, while techniques like Ewald summations [231–233] facilitate assessment of the electric field over the infinite lattice (although Welch [183] suggests that this may not be needed). Finally, neither $\Delta Q_i^{l_i m_i}$ model has a particularly robust description of short-range charge-overlap effects. WSM approximates this by damping functions, which arguably lack strong theoretical justification, while SCERP captures only some penetration effects due to charges in close proximity to the central molecule's charge density [189].

2.5.2.3 Previous Attempts to Incorporate Induction into HAIEFF

Ultimately, frameworks to incorporate induction explicitly within HAIEFF models exist, but there has been an incommensurate effort to develop them as part of the CSP workflow. Within CSP, the distributed polarizability model has been tested with HAIEFF models in the past [189], but further development of this approach has since been mostly limited to SAPT(DFT)-parametrized FFs [92, 165]. Conversely, electronic-embedding methods like SCERP have not seen as much use in CSP, but they do have precedence in related fields of solid state chemistry

[149, 226]. It should be noted that electronic-embedding strategies have been proposed in conjunction with fragmentation approaches for evaluating crystal energies [234], however this has not been widely adopted in modern CSP.

In my view, an important criteria that will determine which approach (distributed polarizability or electronic-embedding) is more 'useful' will be their capability to handle large, flexible molecules that are industrially-relevant. Unfortunately, distributed polarizability methods seem to be disadvantaged in this aspect, due to their reliance on linear-response DFT complicating the treatment of larger molecules. Thus far, no studies have demonstrated the use of the distributed polarizability model in flexible CSP. In contrast, initial integration of SCERP into the HAIEFF workflow has been investigated by a previous member of our group, with an implementation for flexible CSP as well [71]. From this early proof of concept, SCERP shows potential for use in flexible HAIEFF CSPs, however further development is needed to bolster performance and reliability, and validate the physical robustness of the methodology. One additional shortcoming of this previous investigation [71] was the absence of empirical FFs compatible with the new electrostatic description.

2.6 Summary

To summarize this review of the current state of CSP, there is a growing sentiment that CSP, or at least CSP_0, is mature enough to find practical use in some industrial applications [41, 51, 52]. Part of this success stems from the establishment of the hierarchical CSP workflow, which has formalized an effective strategy to balance the two main demands of CSP: breadth and accuracy. To enable this workflow, numerous energy models have been proposed across a range of classes, each with their respective strengths and weaknesses. Amongst these, the HAIEFF model is a versatile class that can be used across several stages of the workflow, making it worthy of further development. To that end, in this work, I attempt to address shortcomings of the current suite of HAIEFF models, related to the components of the intermolecular contribution.

The issues with the HAIEFF model's empirical FF term can be distilled into three interconnected areas. This begins with accessibility to resources for parameter estimation, which pertains to having effective tools for parameter estimation, and having adequate reference data to train FFs and benchmark their performance. Previous work in our group has helped alleviate some of these problems through the development of the CrystalEstimator program [1, 2], but I believe the available reference data should be enriched to represent more-diverse atomic environments. Poor accessibility discourages routine parameter estimation and detailed investigation of HAIEFF FF modeling decisions, which leads to subpar FFs that cannot describe all crystal properties effectively [2]. Ultimately, when new fit-for-purpose FFs are prohibitive to parametrize, *ad hoc* utilization of existing FFs is encouraged instead.

The issues with the HAIEFF model's electrostatic term are more fundamental. Induction is an important contribution to the lattice energy, but methods for its explicit description have not been rigorously developed and studied in HAIEFF models. While some groundwork for the

utilization of the SCERP model has already been laid [71], more must be done to demonstrate that the methodology is not only physically robust, but is computationally efficient. Moreover, in the spirit of the work done for the empirical FF, it will also be vital to parametrize corresponding FFs that are compatible with this development in the electrostatic component.

Chapter 3

HAIEFF Model Formulation

In this chapter I lay out the overarching theory and mathematical formulations underpinning all HAIEFF model calculations. This encompasses two main aspects. In Section 3.1, I expound on the formalism used to describe crystal geometries. This forms the basis for the calculation of crystal energy for a given crystal geometry, explained in Section 3.2. I will approach the HAIEFF lattice energy formulation first from the intermolecular contribution (Section 3.2.1), followed by the intramolecular contribution (Section 3.2.2). The mathematical framework detailed in this chapter underpins all programs used in this work, which will be described in Chapter 4.

3.1 Describing Crystal Geometries

Describing the geometry of a crystal is a fundamental task for CSP. The crystal geometry underlies all properties of the crystal, including its lattice energy. Formally, the crystal geometry is defined as the positions of all atoms in the crystal, but this geometric description can be simplified by exploiting concepts of the crystal's unit cell and asymmetric unit cell.

3.1.1 Unit Cells

The unit cell is a three-dimensional subset of the crystal's infinite lattice. More precisely, the unit cell represents the smallest possible repeating subunit which can recover the full crystalline lattice solely through translational operations. In short, with the unit cell definition (i.e., the geometry of the three-dimensional volume and the position of all atomic sites within it) one can recover the entire crystal lattice.

The unit cell is a parallelepiped structure constructed within a Cartesian coordinate system (Figure 3.1). I refer to this coordinate system as the space-fixed or the global coordinate system. The unit cell is determined by three lattice lengths (a, b, and c), as well as three lattice angles (α, β, and γ). a is the cell edge lying along the x-axis of the coordinate system, b is the edge lying on the xy-plane, while c is the last edge within the three-dimensional coordinate system. α, β, and γ are the angles between the b/c, a/c, and a/b edges respectively. I refer to these six variables (a, b, c, α, β, and γ) collectively as the unit cell or lattice parameters.

Further manipulation of this geometry is best handled in vector notation. The unit cell matrix ($\boldsymbol{R}$), uniquely defines each unit cell, where its columns are defined by the vectors $\boldsymbol{a}$, $\boldsymbol{b}$,

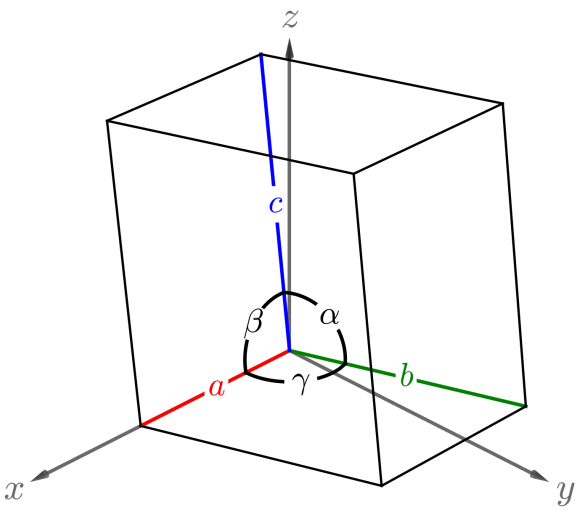


Figure 3.1: Illustration of an arbitrary unit cell of a crystal in the space-fixed coordinate system.

and $\boldsymbol{c}$ according to

$$\boldsymbol{R} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \begin{pmatrix} a & b\cos(\gamma) & c\cos(\beta) \\ 0 & b\sin(\gamma) & c\mu \\ 0 & 0 & c\omega \end{pmatrix}, \tag{3.1.1}$$

where μ and ω are given by

$$\mu = \frac{\cos(\alpha) - \cos(\beta)\cos(\gamma)}{\sin(\gamma)}, \tag{3.1.2}$$

and

$$\omega = \frac{\sqrt{W}}{\sin(\gamma)}, \tag{3.1.3}$$

respectively. Finally, W is defined as the cell deformation parameter:

$$W = 1 - \sum_{\zeta=\alpha,\beta,\gamma} \cos^2(\zeta) + 2 \prod_{\zeta=\alpha,\beta,\gamma} \cos(\zeta). \tag{3.1.4}$$

Related to $\boldsymbol{R}$ is its inverse, the reciprocal unit cell matrix $\boldsymbol{R}^{-1}$. $\boldsymbol{R}^{-1}$ has rows defined by the inverse of the previous three column vectors:

$$\boldsymbol{a}^{-1} = \frac{\boldsymbol{b} \times \boldsymbol{c}}{\boldsymbol{a} \cdot (\boldsymbol{b} \times \boldsymbol{c})}, \tag{3.1.5}$$

$$\boldsymbol{b}^{-1} = \frac{\boldsymbol{a} \times \boldsymbol{c}}{\boldsymbol{b} \cdot (\boldsymbol{a} \times \boldsymbol{c})}, \tag{3.1.6}$$

$$\boldsymbol{c}^{-1} = \frac{\boldsymbol{a} \times \boldsymbol{b}}{\boldsymbol{c} \cdot (\boldsymbol{a} \times \boldsymbol{b})}, \tag{3.1.7}$$

such that $\boldsymbol{R}^{-1}$ can be defined as

$$\boldsymbol{R}^{-1} = \begin{pmatrix} (\boldsymbol{a}^{-1})^T \\ (\boldsymbol{b}^{-1})^T \\ (\boldsymbol{c}^{-1})^T \end{pmatrix} = \begin{pmatrix} \frac{1}{a} & -\frac{\cos(\gamma)}{a\sin(\gamma)} & -\frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{a\sin(\gamma)\sqrt{W}} \\ 0 & \frac{1}{b\sin(\gamma)} & -\frac{\mu}{b\sqrt{W}} \\ 0 & 0 & \frac{1}{c\omega} \end{pmatrix}. \tag{3.1.8}$$

Finally, the unit cell volume (V) can be determined simply as the determinant of $\boldsymbol{R}$:

$$V = \det|\boldsymbol{R}| = abc\sqrt{W}. \tag{3.1.9}$$

3.1.2 Atomic Coordinate Positions

To recover the entire crystal lattice, the position of atoms within the unit cell are also required. To illustrate how this is described, consider a unit cell containing Z molecules, where the Ith molecule comprises η_I atoms. In this work, I denote the Cartesian coordinates (in the global coordinate system) of an atom i on molecule I by the vector $\boldsymbol{r}_{Ii}$. Its corresponding *fractional* coordinates within the unit cell are denoted by the vector $\hat{\boldsymbol{r}}_{Ii}$. These vectors are given by $\boldsymbol{r}_{Ii} = (x_{Ii}, y_{Ii}, z_{Ii})^T$ and $\hat{\boldsymbol{r}}_{Ii} = (\hat{x}_{Ii}, \hat{y}_{Ii}, \hat{z}_{Ii})^T$ respectively. The fractional and absolute coordinates are related via the unit cell matrix according to

$$\boldsymbol{r}_{Ii} = \boldsymbol{R}\hat{\boldsymbol{r}}_{Ii}. \tag{3.1.10}$$

For all atomic positions to be defined within the convex hull of unit cell, each element of $\hat{\boldsymbol{r}}_{Ii}$ is constrained within [0,1]. Finally, from the individual atomic positions, the position of the molecular centre-of-mass (COM) ($\boldsymbol{r}_{0,I}$) is given by

$$\boldsymbol{r}_{0,I} = \frac{\sum_{i=1}^{\eta_I} m_{Ii} \cdot \boldsymbol{r}_{Ii}}{\sum_{i=1}^{\eta_I} m_{Ii}}, \tag{3.1.11}$$

where m_{Ii} is the atomic mass of atom i in molecule I. COM positions can be equivalently expressed in fractional coordinates and converted to absolute coordinates by Equation 3.1.10.

3.1.3 Body-Fixed Coordinates and Molecular Orientation

It is often desirable to treat the molecular conformation independently of the unit cell and global coordinate system. To that end, an alternate Cartesian coordinate system, the body-fixed coordinates, is defined. In principle, the definition of this system is arbitrary, but I follow the approach of Bowskill [2]. In the body-fixed coordinate system, the origin is defined as the molecular COM, $\boldsymbol{r}_{0,I}$. The x-dimension is given by the unit vector from this origin to the first atom of the molecule. The y-dimension is a unit vector orthogonal to the x-dimension, and lies on the plane containing the origin, the first atom, and the next atom which is not collinear to atom 1. Finally, the z-dimension is given by the unit vector orthogonal to the x- and y-dimensions. The body-fixed coordinates of each atom i on molecule I is then given by $\boldsymbol{r}_{Ii}^{\text{bf}} = (x_{Ii}^{\text{bf}}, y_{Ii}^{\text{bf}}, z_{Ii}^{\text{bf}})^T$. By definition, the body-fixed coordinate system is independent of any translations or rotations of the molecule in the space-fixed system. Moreover, the body-fixed coordinates of all atoms in

a molecule can be obtained from details of its Zmatrix definition and associated values of the conformational degrees of freedom ($\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ include bond, angle, and dihedral values).

To interconvert between the space-fixed and body-fixed coordinate systems, the orientation of a molecule I within the unit cell is described by three Euler angles: ϕ_I, θ_I, and ψ_I. To relate the two coordinate systems with these angles, I follow the convention employed by Karamertzanis [233] and Bowskill [2] for intrinsic rotations $z_1 - x' - z_2''$. The body-fixed coordinate system can be obtained from the space-fixed coordinate system through a series of operations. First, a rotation by ϕ_I around the space-fixed z-axis is executed. Second, a rotation by θ_I about the reorientated x-axis. Finally, a rotation by ψ_I is performed about the reorientated z-axis (which now corresponds to the body-fixed z-axis). For a molecule I, these individual rotation operations are expressed in the matrices $\boldsymbol{P}_{\phi,I}$, $\boldsymbol{P}_{\theta,I}$, and $\boldsymbol{P}_{\psi,I}$, defined by

$$\boldsymbol{P}_{\phi,I} = \begin{pmatrix} \cos(\phi_I) & \sin(\phi_I) & 0 \\ -\sin(\phi_I) & \cos(\phi_I) & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{3.1.12}$$

$$\boldsymbol{P}_{\theta,I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_I) & \sin(\theta_I) \\ 0 & -\sin(\theta_I) & \cos(\theta_I) \end{pmatrix}, \tag{3.1.13}$$

$$\boldsymbol{P}_{\psi,I} = \begin{pmatrix} \cos(\psi_I) & \sin(\psi_I) & 0 \\ -\sin(\psi_I) & \cos(\psi_I) & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{3.1.14}$$

respectively. To accommodate all possible rotational relationships between the two systems, it is sufficient to limit the Euler angles to $\phi_I, \psi_I \in (0, 2\pi]$, and $\theta_I \in (0, \pi]$. The full rotation matrix for molecule I ($\boldsymbol{A}_{\mathrm{rot},I}$) is given by the multiplication of the previous three rotation matrices:

$$\boldsymbol{A}_{\mathrm{rot},I} = \boldsymbol{P}_{\psi,I}\boldsymbol{P}_{\theta,I}\boldsymbol{P}_{\phi,I} =$$

$$\begin{pmatrix} \cos(\phi_I)\cos(\psi_I) & \sin(\phi_I)\cos(\psi_I) & \sin(\theta_I)\sin(\psi_I) \\ -\sin(\phi_I)\cos(\theta_I)\sin(\psi_I) & +\cos(\phi_I)\cos(\theta_I)\sin(\psi_I) & \\ -\cos(\phi_I)\sin(\psi_I) & -\sin(\phi_I)\sin(\psi_I) & \sin(\theta_I)\cos(\psi_I) \\ -\sin(\phi_I)\cos(\theta_I)\cos(\psi_I) & +\cos(\phi_I)\cos(\theta_I)\cos(\psi_I) & \\ \sin(\phi_I)\sin(\theta_I) & -\cos(\phi_I)\sin(\theta_I) & \cos(\theta_I) \end{pmatrix}. \tag{3.1.15}$$

Due to its orthonormal nature, the full rotation matrix must satisfy $\boldsymbol{A}_{\mathrm{rot},I}^{-1} = \boldsymbol{A}_{\mathrm{rot},I}^{T}$. With this, the interconversion between coordinate systems can be done via

$$\boldsymbol{r}_{Ii}^{\mathrm{bf}} = \boldsymbol{A}_{\mathrm{rot},I}(\boldsymbol{r}_{Ii} - \boldsymbol{r}_{0,I}), \tag{3.1.16}$$

to obtain body-fixed atomic positions from space-fixed (e.g., from an input crystal structure where all atomic coordinates have been defined in the space-fixed frame), and conversely

$$\boldsymbol{r}_{Ii} = \boldsymbol{A}_{\mathrm{rot},I}^{T} \boldsymbol{r}_{Ii}^{\mathrm{bf}} + \boldsymbol{r}_{0,I}, \tag{3.1.17}$$

to obtain space-fixed atomic positions from body-fixed (e.g., to insert a molecule altered in the body-fixed frame back into the unit cell).

Finally, there are a few important characteristics of the rotation that should be discussed. Firstly, by definition, the (space-fixed) vectors representing the three Cartesian dimensions of the body-fixed coordinate system will correspond to the columns of $\boldsymbol{A}_{\mathrm{rot},I}^{T}$. When defining the z-dimension of the body-fixed coordinates, there is a choice of 'direction' since it need only be orthogonal to the other two dimensions. To avoid ambiguity in this definition, $\boldsymbol{A}_{\mathrm{rot},I}^{T}$ must form a right-handed coordinate system (i.e., $\det|\boldsymbol{A}_{\mathrm{rot},I}^{T}| = 1$). The handedness of the coordinate system can be changed by flipping the direction of the body-fixed z-dimension.

3.1.4 Space Groups and Asymmetric Unit Cells

The formulation discussed above is sufficient to obtain a complete description of the entire infinite lattice, but treating each molecule in a unit cell independently is inefficient for the purposes of describing a crystal structure. In addition to the periodicity of the crystal exhibited by the unit cell, crystalline solids must also exhibit specific internal symmetries within their unit cells. This means that not all Z molecules within the unit cell are independent of one another.

The required crystalline symmetries are obtained through various symmetry operations like translations, inversions, rotations, and/or combinations of these operations [235]. There are a total of 230 sets of symmetry operations, termed space groups, that lead to valid unit cells [236]. Leveraging space group symmetry, one can define an asymmetric unit cell containing only Z' molecules that are symmetrically-independent of each other. For a space group that comprises σ_s symmetry operations, $Z' = \frac{Z}{\sigma_s}$. By applying σ_s symmetry operations on the asymmetric unit cell, the positions of all other symmetrically-related molecules in the unit cell ($\hat{\boldsymbol{r}}_{(s-1)Z'+N,i}$) may be recovered from the positions of molecules in the asymmetric unit cell ($\hat{\boldsymbol{r}}_{Ni}$):

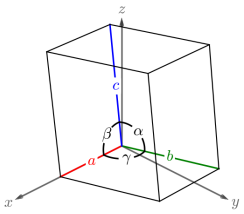
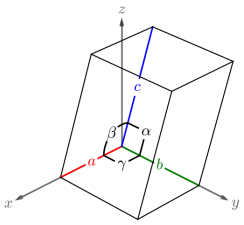
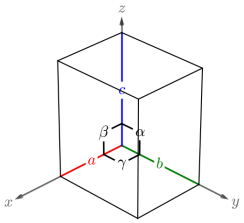
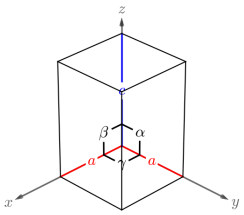
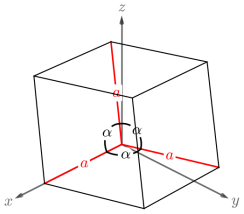
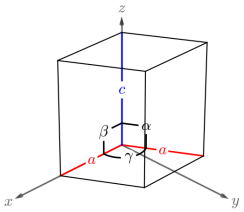
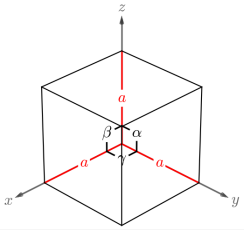
$$\hat{\boldsymbol{r}}_{(s-1)Z'+N,i} = \boldsymbol{M}_s \hat{\boldsymbol{r}}_{Ni} + \boldsymbol{T}_s, \qquad s = 1,\cdots,\sigma_s, \quad i = 1,\cdots,\eta_N, \quad N = 1,\cdots,Z', \tag{3.1.18}$$

where $\boldsymbol{M}_s$ is a transformation matrix representing various inversions, rotations, and reflections, while $\boldsymbol{T}_s$ is a translation vector. When $s = 1$, $\boldsymbol{M}_s$ and $\boldsymbol{T}_s$ are given by

$$\boldsymbol{M}_s = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \boldsymbol{T}_s = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \tag{3.1.19}$$

such that the $s = 1$ set of molecules placed into the unit cell correspond exactly to the molecules from the asymmetric unit cell (i.e., no transformation or translation). Fractional values of Z' can occur in molecules which exhibit internal symmetries in their molecular conformation (e.g.,

Table 3.1: Seven lattice types and their corresponding constraints.

Lattice System	**Schematic**	**Constraints**	**Unconstrained Lattice Parameters**
Triclinic		None	$a, b, c, \alpha, \beta, \gamma$
Monoclinic		$a \neq b \neq c$ $\alpha = \gamma = 90°$ $\beta \neq 90°$	a, b, c, β
Orthorhombic		$a \neq b \neq c$ $\alpha = \beta = \gamma = 90°$	a, b, c
Tetragonal		$a = b \neq c$ $\alpha = \beta = \gamma = 90°$	a, c
Rhombohedral		$a = b = c$ $\alpha = \beta = \gamma \neq 90°$	a, α
Hexagonal		$a = b \neq c$ $\alpha = \beta = 90°$ $\gamma = 120°$	a, c
Cubic		$a = b = c$ $\alpha = \beta = \gamma = 90°$	a

benzene). If the point group symmetry of the molecule coincides with special positions in the unit cell, fractional Z' can emerge. Currently, all codes developed in our group permit only integer values of Z'. If fractional Z' are encountered, tools available in the CCDC program Mercury [237] are used to convert between space group definitions to one with an integer Z'.

Formulating the crystal geometry using space group symmetries is more succinct for two reasons. First, as has already been suggested, only Z' molecules need to be treated as independent geometric variables in the crystal structure. Second, each space group belongs to one of seven possible lattice systems and depending on the system, additional constraints may be imposed on the unit cell parameters (Table 3.1). This in turn further reduces the number of free variables necessary to describe the crystal structure. If no space group symmetries are exploited, this is equivalent to using the *P1* space group, wherein all molecules in the unit cell and all unit cell parameters are treated as independent variables to the crystal geometry.

To complete this discussion of crystal geometries, the equations from Section 3.1.1–3.1.4 can be combined to obtain an expression for the atomic coordinates of all molecules in a unit cell, whilst leveraging space group symmetry. Starting from Equation 3.1.18, it is trivial to convert this fractional coordinate expression to one in absolute coordinates using Equation 3.1.10:

$$\begin{aligned}\boldsymbol{r}_{(s-1)Z'+N,i} &= \boldsymbol{R}\hat{\boldsymbol{r}}_{(s-1)Z'+N,i} \\ &= \boldsymbol{R}\boldsymbol{M}_s\boldsymbol{R}^{-1}\boldsymbol{r}_{Ni} + \boldsymbol{R}\boldsymbol{T}_s, \qquad s = 1,\cdots,\sigma_s, \quad i = 1,\cdots,\eta_N, \quad N = 1,\cdots,Z'.\end{aligned} \tag{3.1.20}$$

From here, Equation 3.1.17 can be used to obtain an equivalent expression in terms of the body-fixed coordinates of the molecule,

$$\begin{aligned}\boldsymbol{r}_{(s-1)Z'+N,i} =& \boldsymbol{R}\boldsymbol{M}_s\boldsymbol{R}^{-1}(\boldsymbol{A}^T_{\mathrm{rot},N}\boldsymbol{r}^{\mathrm{bf}}_{Ni} + \boldsymbol{r}_{0,N}) \\ &+ \boldsymbol{R}\boldsymbol{T}_s, \qquad s = 1,\cdots,\sigma_s, \quad i = 1,\cdots,\eta_N, \quad N = 1,\cdots,Z'.\end{aligned} \tag{3.1.21}$$

Finally, the COM of every symmetry-related molecule is normalized such that they are contained in the central unit cell, by adding a translation vector ($\boldsymbol{t}_{N,s}$) to the previous expression,

$$\begin{aligned}\boldsymbol{r}_{(s-1)Z'+N,i} =& \boldsymbol{R}\boldsymbol{M}_s\boldsymbol{R}^{-1}(\boldsymbol{A}^T_{\mathrm{rot},N}\boldsymbol{r}^{\mathrm{bf}}_{Ni} + \boldsymbol{r}_{0,N}) \\ &+ \boldsymbol{R}(\boldsymbol{T}_s + \boldsymbol{t}_{N,s}), \qquad s = 1,\cdots,\sigma_s, \quad i = 1,\cdots,\eta_N, \quad N = 1,\cdots,Z'\end{aligned} \tag{3.1.22}$$

It is trivial to express the COM of each symmetry-related molecule by setting $\boldsymbol{r}^{\mathrm{bf}}_{Ni}$ to the corresponding body-fixed origin:

$$\boldsymbol{r}_{0,(s-1)Z'+N} = \boldsymbol{R}\boldsymbol{M}_s\boldsymbol{R}^{-1}\boldsymbol{r}_{0,N} + \boldsymbol{R}(\boldsymbol{T}_s + \boldsymbol{t}_{N,s}), \qquad s = 1,\cdots,\sigma_s, \quad N = 1,\cdots,Z', \tag{3.1.23}$$

and the values of $\boldsymbol{t}_{N,s}$ may be found as the unique set of integer values which satisfy the following (in fractional terms):

$$\mathbf{0} \leq \boldsymbol{M}_s \boldsymbol{R}^{-1} \boldsymbol{r}_{0,N} + (\boldsymbol{T}_s + \boldsymbol{t}_{N,s}) \leq \mathbf{1}, \qquad s = 1, \cdots, \sigma_s, \quad N = 1, \cdots, Z'. \tag{3.1.24}$$

From Equation 3.1.21, it can be inferred that since symmetry-related molecules have the same set of body-fixed coordinates (i.e., identical conformations, $\boldsymbol{r}^{\mathrm{bf}}_{((s-1)Z'+N)i} = \boldsymbol{r}^{\mathrm{bf}}_{Ni}$), effects of the symmetry-operations are instead contained within the rotation matrix

$$\boldsymbol{A}^T_{\mathrm{rot},(s-1)Z'+N} = \boldsymbol{R}\boldsymbol{M}_s\boldsymbol{R}^{-1}\boldsymbol{A}^T_{\mathrm{rot},N}, \qquad s = 1, \cdots, \sigma_s, \quad N = 1, \cdots, Z'. \tag{3.1.25}$$

With that, all equations necessary to describe the crystal lattice from the asymmetric unit cell and unit cell parameters have been detailed.

3.2 HAIEFF Lattice Energy Formulation

With the crystal structure mathematically defined, the corresponding lattice energy can be evaluated. HAIEFF models are constructed by a partitioning of the lattice energy into independent intra- and intermolecular contributions. The dependencies of the energy terms shown in Equations 2.4.7–2.4.9 may be further expanded as

$$U_{\mathrm{latt}}(\boldsymbol{\theta}, \boldsymbol{\Omega}; \boldsymbol{p}) = \Delta U_{\mathrm{intra}}(\boldsymbol{\theta}) + U_{\mathrm{inter}}(\boldsymbol{\theta}, \boldsymbol{\Omega}; \boldsymbol{p}), \tag{3.2.1}$$

$$\Delta U_{\mathrm{intra}}(\boldsymbol{\theta}) = U^{\mathrm{crys}}_{\mathrm{intra}}(\boldsymbol{\theta}) - U^{\mathrm{gp}}_{\mathrm{intra}}, \tag{3.2.2}$$

$$U_{\mathrm{inter}}(\boldsymbol{\theta}, \boldsymbol{\Omega}; \boldsymbol{p}) = U_{\mathrm{elec}}(\boldsymbol{\theta}, \boldsymbol{\Omega}) + U_{\mathrm{FF}}(\boldsymbol{\theta}, \boldsymbol{\Omega}; \boldsymbol{p}). \tag{3.2.3}$$

$\boldsymbol{\theta}$ is the vector of conformational degrees of freedom (bond lengths, bond angles, and torsion values) defining the geometry of each molecule in the asymmetric unit cell. $\boldsymbol{\Omega}$ are all other variables necessary to define the crystal geometry: lattice lengths and angles, the (fractional) COM position of all molecules in the asymmetric unit cell ($\hat{\boldsymbol{r}}_{0,N}$ for $N = 1, \cdots, Z'$), and the Euler angles defining the orientation of each asymmetric molecule (ϕ_N, θ_N, and ψ_N for $N = 1, \cdots, Z'$). $\boldsymbol{\Omega}$ will also be referred to as 'rigid-body variables' since they are the crystal geometric variables for a given molecular conformation, $\boldsymbol{\theta}$. Lastly, $\boldsymbol{p}$ is the vector of FF parameters used during the evaluation of U_{FF}. During a lattice energy evaluation, $\boldsymbol{p}$ are treated as fixed parameters. For clarity, I denote such parameters after the semi-colon. These energetic contributions have additional implicit dependencies (e.g., space group constraints, QM-derived quantities, etc.), but these have been omitted for brevity. In the subsequent sections, I detail the calculations involved in the assessment of each individual contribution ($\Delta U_{\mathrm{intra}}$, U_{elec}, and U_{FF}).

3.2.1 Intermolecular Interactions

To begin, I consider the intermolecular contributions. The calculation of the electrostatic and FF contributions involves evaluating functions based on interatomic Euclidean distances. Expressing

these functions explicitly in terms of $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$ is cumbersome, hence following the approach of Bowskill [2], I re-express the energetic contributions in terms of more-convenient variables,

$$U_{\text{inter}}(\boldsymbol{r}, \boldsymbol{X}) = U_{\text{elec}}(\boldsymbol{r}, \boldsymbol{X}) + U_{\text{FF}}(\boldsymbol{r}, \boldsymbol{X}), \tag{3.2.4}$$

where $\boldsymbol{r}$ are the Cartesian coordinates of all atoms in the unit cell ($\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{\theta}, \boldsymbol{\Omega})$), and $\boldsymbol{X}$ are the lattice parameters ($\boldsymbol{X} \in \{a, b, c, \alpha, \beta, \gamma\}$). The dependence on the FF parameters, $\boldsymbol{p}$, has been omitted for brevity. From here, a general expression for the intermolecular term is given by

$$U_{\text{inter}} = \frac{1}{2Z} \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} \sum_{\boldsymbol{l}=-\boldsymbol{\infty}}^{\boldsymbol{\infty}} \underbrace{\sum_{J=1}^{Z}}_{J \neq I \text{ if } \boldsymbol{l}=\boldsymbol{0}} \sum_{j=1}^{\eta_J} \left[U^{\text{elec}}_{\boldsymbol{0}Ii,\boldsymbol{l}Jj}(r_{\boldsymbol{0}Ii,\boldsymbol{l}Jj}) + U^{\text{FF}}_{\boldsymbol{0}Ii,\boldsymbol{l}Jj}(r_{\boldsymbol{0}Ii,\boldsymbol{l}Jj}) \right], \tag{3.2.5}$$

where $U^{\text{elec}}_{\boldsymbol{0}Ii,\boldsymbol{l}Jj}$ and $U^{\text{FF}}_{\boldsymbol{0}Ii,\boldsymbol{l}Jj}$ are pairwise interaction functions for the electrostatic and FF contribution respectively. Both depend on the interatomic distance between two interacting sites ($r_{\boldsymbol{0}Ii,\boldsymbol{l}Jj}$). Equation 3.2.5 can be interpreted as the sum of all pairwise interactions involving atoms i belonging to molecules I in the central unit cell (indexed as $\boldsymbol{l} = \boldsymbol{0}$), with all other atoms j belonging to molecules J in the periodic images of the central unit cell (indexed as $\boldsymbol{l}$). The summation runs over all atoms in the corresponding molecules (up to η_I and η_J for molecules I and J respectively). Importantly, molecular self-interaction must be avoided within the central unit cell, hence the condition $J \neq I$ if $\boldsymbol{l} = \boldsymbol{0}$. This self-interaction is undesirable since: (1) there is already an independent intramolecular term (ΔU_{intra}) in the HAIEFF formulation, and (2) the intermolecular interaction functions are not formulated with extremely short-range interactions in mind and are liable to unphysical results when used for *intramolecular* calculations. The $\frac{1}{2}$ prefactor prevents double-counting of the pairwise interactions, while normalization by Z converts the energy per unit cell to an energy per molecule. This ensures a common basis of comparison between crystals with different Z. Finally, the interatomic distance is simply the Euclidean norm between the two interacting sites

$$r_{\boldsymbol{0}Ii,\boldsymbol{l}Jj}(\boldsymbol{\theta}, \boldsymbol{\Omega}) = \| \boldsymbol{r}_{\boldsymbol{0}Jj}(\boldsymbol{\theta}, \boldsymbol{\Omega}) + \boldsymbol{R}(\boldsymbol{\Omega}) \cdot \boldsymbol{l} - \boldsymbol{r}_{\boldsymbol{0}Ii}(\boldsymbol{\theta}, \boldsymbol{\Omega}) \|, \tag{3.2.6}$$

where $\boldsymbol{r}_{\boldsymbol{0}Jj}(\boldsymbol{\theta}, \boldsymbol{X}) + \boldsymbol{R}(\boldsymbol{X}) \cdot \boldsymbol{l}$ gives the Cartesian coordinate of site $\boldsymbol{l}Jj$ in periodic unit cell $\boldsymbol{l}$.

3.2.1.1 Electrostatic Interactions

The Generalized Coulomb Sum

The first interaction function in Equation 3.2.5 corresponds to the electrostatic contribution to the lattice energy, which arise from interactions between the localized charge densities of molecules. At medium- and long-range, where there is negligible charge density overlap between the interacting sites, the two-body electrostatic contribution adheres exactly to Coulomb's law, and in the case of two interacting point-charges this is given by a relatively simple expression (Equation 2.4.6). Atomic-charges are often used in the Global Search stage. However, utilization of higher-ranked multipoles is critical for accurate energy ranking [166], thus distributed

multipole models are almost ubiquitous in Refinement stages. In the workflows used by our group, it is common to consider interactions between charges, dipoles, quadrupoles, octupoles, and hexadecapoles during refinement calculations.

Given this fact, it is desirable to consider a more general form of the Coulomb sum that is applicable to all ranks of multipole interactions. The formulation outlined here follows the derivation given by Stone [221]. In his approach, the pairwise electrostatic interaction function can be expressed with as a generalized Coulomb interaction,

$$U^{\text{elec}}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \frac{1}{4\pi\varepsilon_0}\sum_{l_1,l_2}\sum_{m_1,m_2}(-1)^{l_1}\left[\frac{(2l_1+2l_2+1)!}{(2l_1)!(2l_2)!}\right]^{\frac{1}{2}} Q^{l_1m_1}_{Ii}Q^{l_2m_2}_{Jj}\frac{S^{m_1,m_2}_{l_1,l_2,l_1+l_2}(\boldsymbol{A}^T_{\text{rot},I},\boldsymbol{A}^T_{\text{rot},J},\hat{\boldsymbol{r}}_{\mathbf{0}Ii,\boldsymbol{l}Jj})}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^{l_1+l_2+1}}, \tag{3.2.7}$$

where Q^{lm}_{Ii} is the mth multipole moment with rank l on site i of molecule I with respect to the body-fixed frame of that molecule. Multipole ranks $l = 0, 1, 2, 3$, and 4 correspond to charges, dipoles, quadrupoles, octupoles, and hexadecapoles respectively. At each rank, there exists $2l+1$ moments, which are indexed $m = 0, 1c, 1s, \cdots, lc, ls$. The added complexity of this expression relative to the charge–charge Coulomb's law stems from the need to account for anisotropy in the interaction between higher-ranked multipoles. This is accomplished by the S-function, $S^{m_1,m_2}_{l_1,l_2,l_1+l_2}$, which describes the relative orientation between two interacting molecules, hence its dependence on the inverse rotation matrices, $\boldsymbol{A}^T_{\text{rot},I}$ and $\boldsymbol{A}^T_{\text{rot},J}$. $\hat{\boldsymbol{r}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ is the unit vector from site $\mathbf{0}Ii$ to site $\boldsymbol{l}Jj$. The methods used to derive Q^{lm}_{Ii} will be discussed in Section 4.1.4.

By grouping terms, Equation 3.2.7 can be simplified to

$$U^{\text{elec}}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \frac{1}{4\pi\varepsilon_0}\sum_{l_1,l_2}\sum_{m_1,m_2} Q^{l_1m_1}_{Ii}Q^{l_2m_2}_{Jj}\frac{\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}(\boldsymbol{A}^T_{\text{rot},I},\boldsymbol{A}^T_{\text{rot},J},\hat{\boldsymbol{r}}_{\mathbf{0}Ii,\boldsymbol{l}Jj})}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^{l_1+l_2+1}}, \tag{3.2.8}$$

where all anisotropic/angular effects of the interaction are contained in the interaction function $\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}$. The formulation of these interaction functions has been tabulated by Stone [221] for all interactions with $l_1 + l_2 \leq 5$. Note that in the special case of $l_1 = l_2 = 0$, this constrains $m_1 = m_2 = 0$ and $\widetilde{T}^{0,0}_{0,0,0}$ is just 1. This is the specific case of charge–charge interactions wherein Equation 3.2.8 reduces exactly to Equation 2.4.6.

The Generalized Ewald Summation

An important characteristic of the pairwise Coulomb contribution shown in Equation 3.2.8 is that multipole interactions decay with interatomic distance according to $r^{-l_1-l_2-1}$. This introduces a problem when $U^{\text{elec}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ is embedded into the infinite-sum shown in Equation 3.2.5. For multipole interactions where $l_1 + l_2 \geq 3$, the interactions decay sufficiently quickly at long-range that the infinite-sum can be truncated to a finite-range sum (within some distance cutoff) without significant loss of accuracy. Unfortunately, this is untrue of interactions where $l_1 + l_2 \leq 2$ (i.e., charge–charge, charge–dipole, dipole–dipole, and charge–quadrupole interactions). In these low-ranked interactions, the summation exhibits poor convergence with distance and its value depends on the order in which the terms are summed. A consequence of this conditionally convergent behavior is that the infinite-sum cannot be simply truncated, and a direct summation

in real-space must be evaluated to infinity.

A strategy to mitigate conditional convergence is the use of an Ewald sum [231, 238]. Below, I first outline the core concepts underpinning the Ewald sum, before detailing the implemented formulation. This follows the description of the Ewald sum as described by Nijboer & De Wette [231], and as implemented by Vasileiadis [232] and Bowskill [2]. Nijboer & De Wette [231] consider a conditionally convergent infinite-sum in a periodic system, involving a slow-decaying function, $f(x)$, where $f(x) \to 0$ slowly as $x \to \infty$. This summation can be re-expressed as

$$\sigma = \sum_{n=1}^{\infty} f(n) = \sum_{n=1}^{\infty} f(n)\mathcal{F}(n) + \sum_{n=1}^{\infty} f(n)\left[1 - \mathcal{F}(n)\right], \tag{3.2.9}$$

where $\mathcal{F}(n)$ is an auxiliary function. $\mathcal{F}(n)$ can be selected such that it is finite at $\mathcal{F}(0)$, but decays quickly to 0 as $x \to \infty$. From this, the first sum, $\sum_{n=1}^{\infty} f(n)\mathcal{F}(n)$, is no longer conditionally convergent, and can be directly assessed. However, the second summation, $\sum_{n=1}^{\infty} f(n)\left[1 - \mathcal{F}(n)\right]$, remains conditionally convergent. Critically, Nijboer & De Wette [231] propose that instead of direct assessment of this second sum, the periodicity of the system can be exploited to assess the sum in reciprocal-space using a Fourier transformation. Fourier transforms of smooth functions will approach 0 rapidly in reciprocal-space [231], hence as long as $\mathcal{F}(n)$ is carefully selected such that $f(n)\left[1 - \mathcal{F}(n)\right]$ is smooth and 'flat', the second sum is convergent in reciprocal-space. A consequence of this reciprocal-space summation is that a 'self-interaction' contribution invariably emerges and must be corrected for. In the original work by Nijboer & De Wette [231], they only considered interactions between discrete point-charges in a periodic lattice. Vasileiadis [232] later generalized this to interacting *multipoles* between *molecules*. This required extending the self-interaction correction to exclude all *intramolecular* interactions. The Ewald sum is fundamentally a mathematical 'trick' to converge the Coulomb sum, but a different interpretation has been discussed elsewhere [233, 238].

Nijboer & De Wette [231] suggest that an appropriate form of $\mathcal{F}(n)$ can be obtained from the gamma function ($\Gamma(z)$),

$$\Gamma(z) = \int_0^{\infty} t^{z-1}e^{-t}dt = \Gamma(z,x) + \gamma(z,x), \qquad \forall\ z, \tag{3.2.10}$$

where $\Gamma(z,x)$ and $\gamma(z,x)$ are the upper and lower incomplete gamma functions respectively. These are given by

$$\Gamma(z,x) = \int_x^{\infty} t^{z-1}e^{-t}dt, \qquad \text{and} \qquad \gamma(z,x) = \int_0^{x} t^{z-1}e^{-t}dt. \tag{3.2.11}$$

From here, an appropriate auxiliary function can be defined as

$$\mathcal{F}(r_{\mathbf{0}Ii,\mathbf{l}Jj}) = \frac{\Gamma(\frac{1}{2} + l_1 + l_2, \alpha_{l_1+l_2} \cdot \pi \cdot (r_{\mathbf{0}Ii,\mathbf{l}Jj})^2)}{\Gamma(\frac{1}{2} + l_1 + l_2)}, \tag{3.2.12}$$

with

$$1 - \mathcal{F}(r_{\mathbf{0}Ii,\mathbf{l}Jj}) = \frac{\gamma(\frac{1}{2} + l_1 + l_2, \alpha_{l_1+l_2} \cdot \pi \cdot (r_{\mathbf{0}Ii,\mathbf{l}Jj})^2)}{\Gamma(\frac{1}{2} + l_1 + l_2)}, \tag{3.2.13}$$

where $\alpha_{l_1+l_2}$ is an interaction-rank dependent convergence parameter. Since the z arguments of $\Gamma(z)$, $\Gamma(z,x)$, and $\gamma(z,x)$ involve only discrete values of l_1+l_2, analytical expressions for all three functions can be derived, based on recursive gamma function relationships, for the interaction ranks of interest. This is shown in Appendix A.

Applying the Ewald sum framework, the total electrostatic contribution to U_{inter} in Equation 3.2.5 can be reformulated as

$$\begin{aligned} U_{\text{elec}} =& \frac{1}{2Z}\sum_{I=1}^{Z}\sum_{i=1}^{\eta_I}\sum_{\boldsymbol{l}=-\infty}^{\infty}\underbrace{\sum_{J=1}^{Z}}_{J\neq I \text{ if } \boldsymbol{l}=\boldsymbol{0}}\sum_{j=1}^{\eta_J} U^{\text{elec}}_{\boldsymbol{0}Ii,\boldsymbol{l}Jj} \\ =& \frac{1}{2Z}\sum_{I=1}^{Z}\sum_{i=1}^{\eta_I}\sum_{\boldsymbol{l}=-\infty}^{\infty}\underbrace{\sum_{J=1}^{Z}}_{J\neq I \text{ if } \boldsymbol{l}=\boldsymbol{0}}\sum_{j=1}^{\eta_J} U^{\text{real}}_{\boldsymbol{0}Ii,\boldsymbol{l}Jj} + \frac{1}{2Z}\frac{1}{\varepsilon_0 V}\underbrace{\sum_{\boldsymbol{l}_k=-\infty}^{\infty}}_{\boldsymbol{l}_k\neq\boldsymbol{0}} U^{\text{rec}}_{\boldsymbol{l}_k} + U_{\text{corr}} \\ =& U_{\text{real}} + U_{\text{rec}} + U_{\text{corr}}, \end{aligned} \tag{3.2.14}$$

where $U^{\text{real}}_{\boldsymbol{0}Ii,\boldsymbol{l}Jj}$, $U^{\text{rec}}_{\boldsymbol{l}_k}$, and U_{corr} are the real-space (or direct-space) contribution, reciprocal-space contribution, and self-interaction correction contribution respectively (V is the unit cell volume, defined earlier). Each contribution is detailed below.

The pairwise real-space contribution to U_{elec} follows the pairwise Coulomb term shown in Equation 3.2.8, but augmented with the upper incomplete gamma function:

$$\begin{aligned} U^{\text{real}}_{\boldsymbol{0}Ii,\boldsymbol{l}Jj} =& \frac{1}{4\pi\varepsilon_0}\sum_{l_1,l_2}\frac{\Gamma(\frac{1}{2}+l_1+l_2,\alpha_{l_1+l_2}\cdot\pi\cdot(r_{\boldsymbol{0}Ii,\boldsymbol{l}Jj})^2)}{\Gamma(\frac{1}{2}+l_1+l_2)} \\ &\cdot\sum_{m_1,m_2} Q^{l_1m_1}_{Ii}Q^{l_2m_2}_{Jj}\frac{\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}(\boldsymbol{A}^T_{\text{rot},I},\boldsymbol{A}^T_{\text{rot},J},\hat{\boldsymbol{r}}_{\boldsymbol{0}Ii,\boldsymbol{l}Jj})}{(r_{\boldsymbol{0}Ii,\boldsymbol{l}Jj})^{l_1+l_2+1}}, \end{aligned} \tag{3.2.15}$$

and the total real-space contribution (U_{real}) is simply given by

$$U_{\text{real}} = \frac{1}{2Z}\sum_{I=1}^{Z}\sum_{i=1}^{\eta_I}\sum_{\boldsymbol{l}=-\infty}^{\infty}\underbrace{\sum_{J=1}^{Z}}_{J\neq I \text{ if } \boldsymbol{l}=\boldsymbol{0}}\sum_{j=1}^{\eta_J} U^{\text{real}}_{\boldsymbol{0}Ii,\boldsymbol{l}Jj}. \tag{3.2.16}$$

The derivation of the reciprocal-space term is significantly more involved. Applying a Fourier transformation to the generalized Coulomb sum gives

$$\begin{aligned} U^{\text{rec}}_{\boldsymbol{l}_k} =& \sum_{l_1,l_2}\frac{2^{-l_1-l_2}\sqrt{\pi}}{\Gamma(\frac{1}{2}+l_1+l_2)}\frac{1}{\|\boldsymbol{k}\|^{2-l_1-l_2}}\exp\left(\frac{-\|\boldsymbol{k}\|^2}{4\cdot\pi\cdot\alpha_{l_1+l_2}}\right) \\ &\cdot\sum_{m_1,m_2}\sum_{I=1}^{Z}\sum_{J=1}^{Z}\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}(\boldsymbol{A}^T_{\text{rot},I},\boldsymbol{A}^T_{\text{rot},J},\hat{\boldsymbol{k}}) \\ &\cdot\sum_{i=1}^{\eta_I}\sum_{j=1}^{\eta_J} Q^{l_1m_1}_{Ii}Q^{l_2m_2}_{Jj}\{i^{l_1+l_2}\cdot e^{-i[\boldsymbol{k}\cdot(\boldsymbol{r}_{\boldsymbol{0}Jj}-\boldsymbol{r}_{\boldsymbol{0}Ii})]}\}, \end{aligned} \tag{3.2.17}$$

where $U^{\text{rec}}_{\boldsymbol{l}_k}$ is the reciprocal-space contribution from a reciprocal lattice vector $\boldsymbol{k}$,

$$\boldsymbol{k} = 2\pi(\boldsymbol{R}^{-1})\boldsymbol{l}_k. \tag{3.2.18}$$

Applying Euler's formula, the imaginary components in the above expression can be recast as

$$\begin{aligned} U_{\boldsymbol{l}_k}^{\text{rec}} = & \sum_{l_1,l_2} \frac{2^{-l_1-l_2}\sqrt{\pi}}{\Gamma(\frac{1}{2}+l_1+l_2)} \frac{1}{\|\boldsymbol{k}\|^{2-l_1-l_2}} \exp\left(\frac{-\|\boldsymbol{k}\|^2}{4\cdot\pi\cdot\alpha_{l_1+l_2}}\right) \\ & \cdot \sum_{m_1,m_2} \sum_{I=1}^{Z} \sum_{J=1}^{Z} \widetilde{T}_{l_1,l_2,l_1+l_2}^{m_1,m_2}(\boldsymbol{A}_{\text{rot},I}^T, \boldsymbol{A}_{\text{rot},J}^T, \hat{\boldsymbol{k}}) \\ & \cdot \sum_{i=1}^{\eta_I} \sum_{j=1}^{\eta_J} Q_{Ii}^{l_1m_1} Q_{Jj}^{l_2m_2} \{ i^{l_1+l_2} \cdot \cos[\boldsymbol{k}\cdot(\boldsymbol{r}_{\boldsymbol{0}Jj} - \boldsymbol{r}_{\boldsymbol{0}Ii})] - i^{l_1+l_2+1} \cdot \sin[\boldsymbol{k}\cdot(\boldsymbol{r}_{\boldsymbol{0}Jj} - \boldsymbol{r}_{\boldsymbol{0}Ii})] \}. \end{aligned} \tag{3.2.19}$$

In Equation 3.2.19, it is desirable to avoid evaluating any imaginary terms. When $l_1 + l_2$ is even (i.e., $l_1 + l_2 = 0, 2, 4, \cdots$), only the sine term is imaginary. Bowskill [2] demonstrated that in this case, the imaginary terms cancel out due to the summation over all $\boldsymbol{k}$ vectors. By trigonometric manipulation (Appendix B), the remaining (non-imaginary) cosine term can be re-expressed as

$$\begin{aligned} U_{\boldsymbol{l}_k}^{\text{rec}} = & \sum_{l_1,l_2} \frac{2^{-l_1-l_2}\sqrt{\pi}}{\Gamma(\frac{1}{2}+l_1+l_2)} \frac{1}{\|\boldsymbol{k}\|^{2-l_1-l_2}} \exp\left(\frac{-\|\boldsymbol{k}\|^2}{4\cdot\pi\cdot\alpha_{l_1+l_2}}\right) \\ & \cdot \sum_{m_1,m_2} \sum_{I=1}^{Z} \sum_{J=1}^{Z} \widetilde{T}_{l_1,l_2,l_1+l_2}^{m_1,m_2}(\boldsymbol{A}_{\text{rot},I}^T, \boldsymbol{A}_{\text{rot},J}^T, \hat{\boldsymbol{k}}) (-1)^{\frac{l_1+l_2}{2}} \\ & \cdot [\rho_{c,I}^{l_1m_1}(\boldsymbol{k})\rho_{c,J}^{l_2m_2}(\boldsymbol{k}) + \rho_{s,I}^{l_1m_1}(\boldsymbol{k})\rho_{s,J}^{l_2m_2}(\boldsymbol{k})], \qquad l_1 + l_2 = 0, 2, 4, \cdots, \end{aligned} \tag{3.2.20}$$

where

$$\rho_{c,I}^{lm}(\boldsymbol{k}) = \sum_{i=1}^{\eta_I} Q_{Ii}^{lm} \cos(\boldsymbol{k}\cdot\boldsymbol{r}_{\boldsymbol{0}Ii}), \qquad I = 1, 2, \cdots, Z, \tag{3.2.21}$$

and

$$\rho_{s,I}^{lm}(\boldsymbol{k}) = \sum_{i=1}^{\eta_I} Q_{Ii}^{lm} \sin(\boldsymbol{k}\cdot\boldsymbol{r}_{\boldsymbol{0}Ii}), \qquad I = 1, 2, \cdots, Z, \tag{3.2.22}$$

respectively. This re-formulation of Equation 3.2.19 is preferable as it decouples the double summation over the atoms in each interacting molecule (η_I and η_J), permitting $\mathcal{O}(N)$ rather than $\mathcal{O}(N^2)$ scaling (where N is the number of atoms in the unit cell).

When $l_1 + l_2$ is odd (i.e., $l_1 + l_2 = 1, 3, 5, \cdots$), only the cosine term in Equation 3.2.19 is imaginary. Following a similar negation of imaginary terms and trigonometric reformulation, Bowskill [2] proposed

$$\begin{aligned} U_{\boldsymbol{l}_k}^{\text{rec}} = & \sum_{l_1,l_2} \frac{2^{-l_1-l_2}\sqrt{\pi}}{\Gamma(\frac{1}{2}+l_1+l_2)} \frac{1}{\|\boldsymbol{k}\|^{2-l_1-l_2}} \exp\left(\frac{-\|\boldsymbol{k}\|^2}{4\cdot\pi\cdot\alpha_{l_1+l_2}}\right) \\ & \cdot \sum_{m_1,m_2} \sum_{I=1}^{Z} \sum_{J=1}^{Z} \widetilde{T}_{l_1,l_2,l_1+l_2}^{m_1,m_2}(\boldsymbol{A}_{\text{rot},I}^T, \boldsymbol{A}_{\text{rot},J}^T, \hat{\boldsymbol{k}}) (-1)^{\frac{l_1+l_2-1}{2}} \\ & \cdot [\rho_{c,I}^{l_1m_1}(\boldsymbol{k})\rho_{s,J}^{l_2m_2}(\boldsymbol{k}) - \rho_{s,I}^{l_1m_1}(\boldsymbol{k})\rho_{c,J}^{l_2m_2}(\boldsymbol{k})], \qquad l_1 + l_2 = 1, 3, 5, \cdots, \end{aligned} \tag{3.2.23}$$

using the same definitions of $\rho_{s,I}^{lm}$ and $\rho_{c,I}^{lm}$ as before.

Overall, combining Equation 3.2.18 with Equations 3.2.20–3.2.23, the total reciprocal-space contribution (U_{rec}) can be evaluated as

$$U_{\text{rec}} = \frac{1}{2Z}\frac{1}{\varepsilon_0 V}\underbrace{\sum_{\boldsymbol{l}_k=-\infty}^{\infty}}_{\boldsymbol{l}_k\neq\mathbf{0}} U_{\boldsymbol{l}_k}^{\text{rec}}. \tag{3.2.24}$$

The Γ-point term ($U_{\boldsymbol{l}_k}^{\text{rec}}$ when $\boldsymbol{l}_k = \mathbf{0}$) is commonly referred to as the surface-correction term [232]. In systems with unit cells of non-zero dipole moment, a charge can emerge at the surface of the crystal, leading to this contribution. Several authors have discussed the importance of this term [228–230], but as pointed out by Van Eijck & Kroon [229], it is most practical to assume 'tin-foil boundary conditions' (i.e., the surface charge is annulled by accumulation of external charges from the surrounding environment) and neglect the surface contribution (as done here). Without this assumption, evaluation of U_{rec} requires knowledge of the crystal morphology to describe the shape of the crystal surface, which is impractical for CSP.

The final term in Equation 3.2.14 is the self-interaction correction term, U_{corr}. This term corrects for the self-interaction artefact that arises from the Fourier transform in the reciprocal-space term. By definition, the self-correction occurs only once per molecule in the central unit cell and can be evaluated in real-space. The total correction over the unit cell is given by

$$\begin{aligned} U_{\text{corr}} = &-\frac{1}{Z}\frac{1}{4\pi\varepsilon_0}\sqrt{\alpha_{0+0}}\sum_{I=1}^{Z}\sum_{i=1}^{\eta_I}(Q_{Ii}^{0,0})^2 \\ &-\frac{1}{2Z}\frac{1}{4\pi\varepsilon_0}\sum_{I=1}^{Z}\sum_{i=1}^{\eta_I}\underbrace{\sum_{i'=1}^{\eta_I}}_{i'\neq i}\sum_{l_1,l_2}\frac{\gamma(\frac{1}{2}+l_1+l_2,\alpha_{l_1+l_2}\cdot\pi\cdot(r_{\mathbf{0}Ii,\mathbf{0}Ii'})^2)}{\Gamma(\frac{1}{2}+l_1+l_2)} \\ &\cdot\sum_{m_1,m_2}Q_{Ii}^{l_1m_1}Q_{Ii'}^{l_2m_2}\frac{\widetilde{T}_{l_1,l_2,l_1+l_2}^{m_1,m_2}(\boldsymbol{A}_{\text{rot},I}^{T},\boldsymbol{A}_{\text{rot},I}^{T},\hat{\boldsymbol{r}}_{\mathbf{0}Ii,\mathbf{0}Ii'})}{(r_{\mathbf{0}Ii,\mathbf{0}Ii'})^{l_1+l_2+1}}. \end{aligned} \tag{3.2.25}$$

The first sum using $Q_{Ii}^{0,0}$ is the correction for pointwise self-interaction derived by Nijboer & De Wette [231] (i.e., an atom on a molecule interacting with itself). For interactions with $l_1 + l_2 > 0$ this term disappears, hence only the $l_1 = m_1 = 0$ charge–charge interaction is evaluated. There is no double sum over atoms and no possibility of double-counting, thus the prefactor $\frac{1}{2}$ is unnecessary. The second sum is proposed by Vasileiadis [232] for intramolecular self-interaction. This sum is an analogue to the intermolecular real-space interaction (Equations 3.2.15 and 3.2.16), except that this sum involves only atoms in molecule I, and the $i' = i$ self-interaction is excluded since it is in the first sum. Additionally, the lower incomplete gamma function substitutes the upper incomplete gamma function used in Equation 3.2.15. This is intuitive since this correction term emerges from the reciprocal term, which is multiplied by $1-\mathcal{F}$ instead of $\mathcal{F}$. Finally, note that both sums are *deducted* (minus sign before sum) from U_{elec} since these are corrections for double-counting in the reciprocal-space sum.

To summarize, I have detailed the formulation for assessing the total electrostatic contribution (U_{elec}) for models involving multipole interactions of arbitrary rank. Crucially, to deal with conditional convergence, an Ewald sum is employed (Equation 3.2.14). This splits U_{elec} into independent contributions from U_{real} (Equations 3.2.15 and 3.2.16), U_{rec} (Equations 3.2.18 and 3.2.20–3.2.24), and U_{corr} (Equation 3.2.25). While only interactions with $l_1 + l_2 \leq 2$ are conditionally convergent, in the codes developed by our group the Ewald sum is applied to *all* values of $l_1 + l_2$. This is slightly more expensive, but should give better accuracy [2].

Returning to Point-Charges

For completeness, I also give the electrostatic formulation for models based on atomic-charges, that are used in the Global Search. When $l = m = 0$, the general pairwise Coulomb contribution (Equation 3.2.8) reduces to the pairwise charge–charge Coulomb term,

$$U^{\text{elec}}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \frac{1}{4\pi\varepsilon_0}\frac{Q_{Ii}Q_{Jj}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}, \tag{3.2.26}$$

where $\widetilde{T}^{0,0}_{0,0,0} = 1$. Superscripts for l and m have been omitted for brevity (and obviously there are no sums over these indices). Charge–charge interactions are conditionally convergent, hence the Ewald summation technique must be applied to obtain the total electrostatic energy (identically to Equation 3.2.14). The following expressions are obtained using the analytical solutions for $\Gamma(\frac{1}{2})$, $\Gamma(\frac{1}{2}, x)$, and $\gamma(\frac{1}{2}, x)$ in Appendix A.

The pairwise real-space contribution to U_{elec} (analogue to Equation 3.2.15) is given by

$$\begin{aligned} U^{\text{real}}_{\mathbf{0}Ii,\boldsymbol{l}Jj} =& \frac{1}{4\pi\varepsilon_0}\frac{\Gamma(\frac{1}{2}, \alpha_0\cdot\pi\cdot(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2)}{\Gamma(\frac{1}{2})}\frac{Q_{Ii}Q_{Jj}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \\ =& \frac{1}{4\pi\varepsilon_0}\frac{Q_{Ii}Q_{Jj}\text{erfc}(r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\sqrt{\alpha_0\cdot\pi})}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}, \end{aligned} \tag{3.2.27}$$

and can be applied in Equation 3.2.16 to get the total real-space contribution. Similarly, the reciprocal-space contribution to U_{elec} (analogues to Equations 3.2.20–3.2.22) is given by

$$\begin{aligned} U^{\text{rec}}_{\boldsymbol{l}_k} =& \frac{\sqrt{\pi}}{\Gamma(\frac{1}{2})}\frac{1}{\|\boldsymbol{k}\|^2}\exp\left(\frac{-\|\boldsymbol{k}\|^2}{4\cdot\pi\cdot\alpha_0}\right)\cdot\sum_{I=1}^{Z}\sum_{J=1}^{Z}[\rho_{c,I}(\boldsymbol{k})\rho_{c,J}(\boldsymbol{k}) + \rho_{s,I}(\boldsymbol{k})\rho_{s,J}(\boldsymbol{k})] \\ =& \frac{1}{\|\boldsymbol{k}\|^2}\exp\left(\frac{-\|\boldsymbol{k}\|^2}{4\cdot\pi\cdot\alpha_0}\right)\cdot\sum_{I=1}^{Z}\sum_{J=1}^{Z}[\rho_{c,I}(\boldsymbol{k})\rho_{c,J}(\boldsymbol{k}) + \rho_{s,I}(\boldsymbol{k})\rho_{s,J}(\boldsymbol{k})], \end{aligned} \tag{3.2.28}$$

$$\rho_{c,I}(\boldsymbol{k}) = \sum_{i=1}^{\eta_I} Q_{Ii}\cos(\boldsymbol{k}\cdot\boldsymbol{r}_{\mathbf{0}Ii}), \qquad I = 1, 2, \cdots, Z, \tag{3.2.29}$$

$$\rho_{s,I}(\boldsymbol{k}) = \sum_{i=1}^{\eta_I} Q_{Ii}\sin(\boldsymbol{k}\cdot\boldsymbol{r}_{\mathbf{0}Ii}), \qquad I = 1, 2, \cdots, Z. \tag{3.2.30}$$

Recognizing that I and J run over the same indices in Equation 3.2.28, I and J can be treated equivalently, and by extension so can their constituent atoms i and j. This could not be done in the more general case (Equations 3.2.20 and 3.2.23) due to $\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}(\boldsymbol{A}^T_{\text{rot},I}, \boldsymbol{A}^T_{\text{rot},J}, \hat{\boldsymbol{k}})$, which considers the cross-interaction of I and J. The above expressions simplify as

$$U_{\boldsymbol{l}_k}^{\mathrm{rec}} = \frac{1}{\|\boldsymbol{k}\|^2} \exp\left(\frac{-\|\boldsymbol{k}\|^2}{4 \cdot \pi \cdot \alpha_0}\right) [\rho_c^2(\boldsymbol{k}) + \rho_s^2(\boldsymbol{k})], \tag{3.2.31}$$

$$\rho_c(\boldsymbol{k}) = \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} Q_{Ii} \cos(\boldsymbol{k} \cdot \boldsymbol{r}_{\mathbf{0}Ii}), \tag{3.2.32}$$

$$\rho_s(\boldsymbol{k}) = \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} Q_{Ii} \sin(\boldsymbol{k} \cdot \boldsymbol{r}_{\mathbf{0}Ii}). \tag{3.2.33}$$

These equations are applied in Equation 3.2.24 to obtain the total reciprocal-space contribution. Finally, the self-correction analogue to Equation 3.2.25 is straightforwardly obtained as

$$\begin{aligned} U_{\mathrm{corr}} = & -\frac{1}{Z} \frac{1}{4\pi\varepsilon_0} \sqrt{\alpha_0} \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} (Q_{Ii})^2 \\ & - \frac{1}{2Z} \frac{1}{4\pi\varepsilon_0} \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} \underbrace{\sum_{i'=1}^{\eta_I}}_{i' \neq i} \frac{\gamma(\frac{1}{2}, \alpha_0 \cdot \pi \cdot (r_{\mathbf{0}Ii,\mathbf{0}Ii'})^2)}{\Gamma(\frac{1}{2})} \frac{Q_{Ii} Q_{Ii'}}{r_{\mathbf{0}Ii,\mathbf{0}Ii'}} \\ = & -\frac{1}{Z} \frac{1}{4\pi\varepsilon_0} \sqrt{\alpha_0} \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} (Q_{Ii})^2 - \frac{1}{2Z} \frac{1}{4\pi\varepsilon_0} \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} \underbrace{\sum_{i'=1}^{\eta_I}}_{i' \neq i} \frac{Q_{Ii} Q_{Ii'} \mathrm{erf}(r_{\mathbf{0}Ii,\mathbf{0}Ii'} \sqrt{\alpha_0 \cdot \pi})}{r_{\mathbf{0}Ii,\mathbf{0}Ii'}}. \end{aligned} \tag{3.2.34}$$

With this, the mathematical formulations used in our group's CSP methods to evaluate both point-charge and multipole electrostatic energies have been explained.

3.2.1.2 Repulsion–Dispersion Interactions

The empirical FF term in Equation 3.2.5 is a residual term that describes all intermolecular interactions not in U_{elec}. In principle, this includes contributions from effects like induction, many-body effects, etc, but a common view is that the predominant contribution to U_{FF} are van der Waals-type interactions: repulsions from Pauli exclusion and London dispersions. On this basis, semi-empirical pair potentials formulated for van der Waals-type interactions are commonly used to describe $U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\mathrm{FF}}$. In this work, two functional forms are studied.

First, is the Buckingham pair potential (BHP),

$$U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\mathrm{FF,BHP}}(r_{\mathbf{0}Ii,\boldsymbol{l}Jj}) = A^{t,t'} \exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right) - \frac{C^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}^6}, \qquad t = t(Ii), \quad t' = t'(Jj). \tag{3.2.35}$$

$A^{t,t'}$, $B^{t,t'}$, and $C^{t,t'}$ are the empirical parameters of the BHP. $A^{t,t'}$ and $B^{t,t'}$ affect the repulsive term that decays exponentially with interatomic distance. Conversely, $C^{t,t'}$ influences the attractive term given by the leading pairwise London dispersion term that decays by r^{-6}. t and t' denote the atom types of the participating atoms (Ii and Jj respectively). The typing rules (represented by $t(Ii)$ and $t'(Jj)$) are predefined and can be applied transferably across all

molecular systems. Note that the empirical parameters (and by extension the interaction) are symmetric (i.e., $X^{t,t'} = X^{t',t}$ for $X \in \{A, B, C\}$).

The BHP is probably the most popular functional form for U_{FF} used in HAIEFF models. However, in this work, I also consider a second functional form, the Mie pair potential (MIP):

$$U^{\mathrm{FF,MIP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}(r_{\mathbf{0}Ii,\boldsymbol{l}Jj}) = C^{t,t'}_{\mathrm{Mie}} \cdot \varepsilon^{t,t'} \cdot \left[\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \right)^{\lambda_{\mathrm{r}}^{t,t'}} - \left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \right)^{\lambda_{\mathrm{a}}^{t,t'}} \right], \qquad t = t(Ii), \quad t' = t'(Jj), \tag{3.2.36}$$

where

$$C^{t,t'}_{\mathrm{Mie}} = \frac{\lambda_{\mathrm{r}}^{t,t'}}{\lambda_{\mathrm{r}}^{t,t'} - \lambda_{\mathrm{a}}^{t,t'}} \cdot \left(\frac{\lambda_{\mathrm{r}}^{t,t'}}{\lambda_{\mathrm{a}}^{t,t'}} \right)^{\frac{\lambda_{\mathrm{a}}^{t,t'}}{\lambda_{\mathrm{r}}^{t,t'} - \lambda_{\mathrm{a}}^{t,t'}}}, \qquad t = t(Ii), \quad t' = t'(Jj). \tag{3.2.37}$$

$\varepsilon^{t,t'}$, $\sigma^{t,t'}$, $\lambda_{\mathrm{r}}^{t,t'}$, and $\lambda_{\mathrm{a}}^{t,t'}$ are the model parameters for the MIP, while $C^{t,t'}_{\mathrm{Mie}}$ is a prefactor determined from $\lambda_{\mathrm{r}}^{t,t'}$ and $\lambda_{\mathrm{a}}^{t,t'}$. By using $C^{t,t'}_{\mathrm{Mie}}$ in its formulation, the MIP exhibits the convenient traits that $\varepsilon^{t,t'}$ is exactly the depth of the attractive well in the pair potential, while $\sigma^{t,t'}$ is the interatomic distance where the pair potential crosses the $r_{\mathbf{0}Ii,\boldsymbol{l}Jj}$-axis. $\lambda_{\mathrm{r}}^{t,t'}$ and $\lambda_{\mathrm{a}}^{t,t'}$ are commonly referred to as the repulsive and attractive exponent, and they reflect the r^{-n} distance-dependency of the repulsive and attractive interactions respectively. In principle, both $\lambda_{\mathrm{r}}^{t,t'}$ and $\lambda_{\mathrm{a}}^{t,t'}$ can be varied, but I have chosen to fix $\lambda_{\mathrm{a}}^{t,t'} = 6$ to reflect the leading London dispersion term. Like the BHP, atom typing and interaction symmetry applies to the MIP.

For both the BHP and MIP, the repulsive and attractive terms decay sufficiently fast with interatomic distance that conditional convergence is not a concern. As such, no special summation techniques (e.g., the Ewald sum) need to be considered. The total FF contribution to U_{latt} (U_{FF}) can be evaluated by substituting either the BHP or MIP into Equation 3.2.5,

$$U_{\mathrm{FF}} = \frac{1}{2Z} \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} \sum_{\boldsymbol{l}=-\boldsymbol{\infty}}^{\boldsymbol{\infty}} \underbrace{\sum_{J=1}^{Z}}_{J \neq I \text{ if } \boldsymbol{l}=\mathbf{0}} \sum_{j=1}^{\eta_J} U^{\mathrm{FF},P}_{\mathbf{0}Ii,\boldsymbol{l}Jj}(r_{\mathbf{0}Ii,\boldsymbol{l}Jj}), \qquad P = \text{BHP or MIP}. \tag{3.2.38}$$

3.2.1.3 Gradients of Isotropic Intermolecular Interactions

The formulation discussed above is sufficient to conduct single-point evaluations of the intermolecular energy, but energy *minimizations* are necessary for CSP. To that end, our group's CSP programs rely on gradient-based optimization techniques. Below, I describe the analytical formulation of gradient quantities relevant to the intermolecular energy functions, as derived by Bowskill [2]. I address the *isotropic* intermolecular interactions here, while anisotropic gradients are tackled Section 3.2.1.4. I first seek gradient expressions for the $U_{\mathrm{inter}}(\boldsymbol{r}, \boldsymbol{X})$ functions with respect to the Cartesian coordinates, $\boldsymbol{r}$, and the lattice parameters, $\boldsymbol{X}$, i.e., $\nabla_{\boldsymbol{r}} U_{\mathrm{inter}}$ and $\nabla_{\boldsymbol{X}} U_{\mathrm{inter}}$. These are the 'natural' variables used to evaluate the intermolecular functions, but they are not explicit decision variables in the minimization. As such, once gradients with respect

to $\boldsymbol{r}$ and $\boldsymbol{X}$ have been formulated, I relate them back to the rigid-body decision variable vector, $\boldsymbol{\Omega}$ (*cf.* Equation 3.2.3). Readers are reminded that bolded variables are vector/matrix quantities while non-bolded variables are scalar quantities.

First-Derivatives in Real-Space

For an arbitrary function, f, dependent on the Euclidean distance between two points connected by the vector $\boldsymbol{r}$, the gradient vector of f with respect to $\boldsymbol{r}$ can be obtained via chain-rule as

$$\nabla_{\boldsymbol{r}} f = \frac{\partial f(r)}{\partial r} \nabla_{\boldsymbol{r}} r, \quad \text{where} \quad \nabla_{\boldsymbol{r}} r = \frac{\boldsymbol{r}}{r}. \tag{3.2.39}$$

This applies to the real-space intermolecular functions, which are explicit functions of $r_{\mathbf{0}Ii,\boldsymbol{l}Jj}$,

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \frac{\partial U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \frac{(\boldsymbol{r}_{\mathbf{0}Jj} + \boldsymbol{R}\cdot\boldsymbol{l} - \boldsymbol{r}_{\mathbf{0}Ii})}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}, \qquad \kappa \in \{\text{real}, \text{FF}\}. \tag{3.2.40}$$

Given that $\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \boldsymbol{r}_{\mathbf{0}Jj} + \boldsymbol{R}\cdot\boldsymbol{l} - \boldsymbol{r}_{\mathbf{0}Ii}$, it follows that

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}} U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = -\nabla_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}, \qquad \kappa \in \{\text{real}, \text{FF}\}, \tag{3.2.41}$$

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Jj}} U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \nabla_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}, \qquad \kappa \in \{\text{real}, \text{FF}\}. \tag{3.2.42}$$

Since $\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ is a function of $\boldsymbol{X}$ (i.e., $\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}(\boldsymbol{X})$), chain-rule can be applied to obtain

$$\frac{\partial U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial X} = (\nabla_X \boldsymbol{R}\cdot\boldsymbol{l})^T \nabla_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}, \qquad \kappa \in \{\text{real}, \text{FF}\}, \quad X \in \{a, b, c, \alpha, \beta, \gamma\}. \tag{3.2.43}$$

Analytical expressions for $\nabla_X \boldsymbol{R}$ have been tabulated in Table C.1. A special case for Equations 3.2.41 and 3.2.42 arises when an atom interacts with its own periodic image (i.e., $Ii = Jj$). In this case, it follows that

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}} U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \nabla_{\boldsymbol{r}_{\mathbf{0}Jj}} U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = 0, \qquad \kappa \in \{\text{real}, \text{FF}\}, \quad Ii = Jj. \tag{3.2.44}$$

Analytical expressions for $\frac{\partial U^{\text{FF}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}$ are determined from the corresponding expressions for the BHP and MIP interaction functions (Equations 3.2.35, and 3.2.36), respectively:

$$\frac{\partial U^{\text{FF,BHP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} = -\frac{A^{t,t'}}{B^{t,t'}} \exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right) + \frac{6C^{t,t'}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^7}, \tag{3.2.45}$$

$$\frac{\partial U^{\text{FF,MIP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} = C^{t,t'}_{\text{Mie}} \cdot \varepsilon^{t,t'} \cdot \left[-\lambda^{t,t'}_{\text{r}} \frac{(\sigma^{t,t'})^{\lambda^{t,t'}_{\text{r}}}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^{\lambda^{t,t'}_{\text{r}}+1}} + \lambda^{t,t'}_{\text{a}} \frac{(\sigma^{t,t'})^{\lambda^{t,t'}_{\text{a}}}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^{\lambda^{t,t'}_{\text{a}}+1}} \right]. \tag{3.2.46}$$

Similarly, an expression for the real-space electrostatic contribution can be derived for point-charge electrostatics (Equation 3.2.27):

$$\frac{\partial U^{\text{real}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} = \frac{1}{4\pi\varepsilon_0} \frac{Q_{Ii}Q_{Jj}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \left[-2\sqrt{\alpha_0} \exp\left(-\alpha_0 \cdot \pi \cdot (r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2\right) - \frac{\text{erfc}(r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\sqrt{\alpha_0 \cdot \pi})}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \right]. \tag{3.2.47}$$

Substituting these pairwise gradient expressions into their appropriate summations over the entire lattice will give the total derivative with respect to $\boldsymbol{r}$ and $\boldsymbol{X}$. It should be noted that no derivatives need to be determined for the self-interaction correction term, U_{corr}. Since U_{corr} is an *intramolecular* (correction) term, it is constant for a rigid molecular geometry and has no dependencies on the intermolecular rigid-body variables.

First-Derivatives in Reciprocal-Space

For the reciprocal-space contribution for point-charges (Equation 3.2.31), a similar derivation is performed. I consider the derivative of $U_{\boldsymbol{l}_k}^{\text{rec}}$ with respect to $\boldsymbol{r}_{\boldsymbol{0}Ii}$,

$$\nabla_{\boldsymbol{r}_{\boldsymbol{0}Ii}} U_{\boldsymbol{l}_k}^{\text{rec}} = \frac{1}{\|\boldsymbol{k}\|^2} \exp\left(\frac{-\|\boldsymbol{k}\|^2}{4 \cdot \pi \cdot \alpha_0}\right) [2\rho_c(\boldsymbol{k}) \nabla_{r_{\boldsymbol{0}Ii}} \rho_c(\boldsymbol{k}) + 2\rho_s(\boldsymbol{k}) \nabla_{r_{\boldsymbol{0}Ii}} \rho_s(\boldsymbol{k})], \tag{3.2.48}$$

where

$$\nabla_{\boldsymbol{r}_{\boldsymbol{0}Ii}} \rho_c(\boldsymbol{k}) = -\boldsymbol{k} Q_{Ii} \sin(\boldsymbol{k} \cdot \boldsymbol{r}_{\boldsymbol{0}Ii}), \tag{3.2.49}$$

$$\nabla_{\boldsymbol{r}_{\boldsymbol{0}Ii}} \rho_s(\boldsymbol{k}) = \boldsymbol{k} Q_{Ii} \cos(\boldsymbol{k} \cdot \boldsymbol{r}_{\boldsymbol{0}Ii}). \tag{3.2.50}$$

Note that from Equation 3.2.18, $\boldsymbol{k}$ is independent of $\boldsymbol{r}_{\boldsymbol{0}Ii}$. Also, in the derivatives of ρ_c and ρ_s the differentiation is with respect to only one index of Ii, hence the summations are eliminated. Derivatives with respect to the lattice parameters are more involved. From Equation 3.2.18, the reciprocal lattice vectors, $\boldsymbol{k}$, are dependent on the reciprocal unit cell, $\boldsymbol{R}^{-1}$, which is a function of $\boldsymbol{X}$. By extension,

$$\nabla_X \boldsymbol{k} = 2\pi(\nabla_X \boldsymbol{R}^{-1}) \boldsymbol{l}_k, \qquad X \in \{a, b, c, \alpha, \beta, \gamma\}, \tag{3.2.51}$$

and by chain-rule

$$\frac{\partial \|\boldsymbol{k}\|}{\partial X} = \frac{\boldsymbol{k}^T}{\|\boldsymbol{k}\|} \nabla_X \boldsymbol{k}, \qquad X \in \{a, b, c, \alpha, \beta, \gamma\}. \tag{3.2.52}$$

First-derivatives of the reciprocal unit cell matrix are necessary to evaluate $\nabla_X \boldsymbol{k}$; these have been tabulated in Table C.2. Accounting for the implicit dependencies on X, differentiation of Equations 3.2.31–3.2.33 gives

$$\begin{aligned} \frac{\partial U_{\boldsymbol{l}_k}^{\text{rec}}}{\partial X} =& \frac{1}{\|\boldsymbol{k}\|^2} \exp\left(\frac{-\|\boldsymbol{k}\|^2}{4 \cdot \pi \cdot \alpha_0}\right) (\nabla_X \boldsymbol{k})^T \Bigg\{ -\boldsymbol{k} \left(\frac{1}{\|\boldsymbol{k}\|^2} + \frac{1}{2 \cdot \pi \cdot \alpha_0}\right) \left[\rho_c^2(\boldsymbol{k}) + \rho_s^2(\boldsymbol{k})\right] \\ &+ [2\rho_c(\boldsymbol{k}) \nabla_{\boldsymbol{k}} \rho_c(\boldsymbol{k}) + 2\rho_s(\boldsymbol{k}) \nabla_{\boldsymbol{k}} \rho_s(\boldsymbol{k})] \Bigg\}, \qquad X \in \{a, b, c, \alpha, \beta, \gamma\}, \end{aligned} \tag{3.2.53}$$

$$\nabla_{\boldsymbol{k}} \rho_c(\boldsymbol{k}) = \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} -Q_{Ii} \boldsymbol{r}_{\boldsymbol{0}Ii} \sin(\boldsymbol{k} \cdot \boldsymbol{r}_{\boldsymbol{0}Ii}), \tag{3.2.54}$$

$$\nabla_{\boldsymbol{k}} \rho_s(\boldsymbol{k}) = \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} Q_{Ii} \boldsymbol{r}_{\boldsymbol{0}Ii} \cos(\boldsymbol{k} \cdot \boldsymbol{r}_{\boldsymbol{0}Ii}). \tag{3.2.55}$$

Equations 3.2.48–3.2.50 can be used directly within Equation 3.2.24 to evaluate the *total* derivative with respect to $\boldsymbol{r}_{\mathbf{0}Ii}$ over the entire reciprocal-space summation,

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}} U_{\text{rec}} = \frac{1}{2Z}\frac{1}{\varepsilon_0 V} \underbrace{\sum_{\boldsymbol{l}_k=-\infty}^{\infty}}_{\boldsymbol{l}_k \neq \mathbf{0}} \nabla_{\boldsymbol{r}_{\mathbf{0}Ii}} U^{\text{rec}}_{\boldsymbol{l}_k}. \tag{3.2.56}$$

However, the corresponding total derivative for $\boldsymbol{X}$ must also account for the dependency of the cell volume, V, on $\boldsymbol{X}$. Using chain-rule, this results in

$$\frac{\partial U_{\text{rec}}}{\partial X} = \frac{1}{2Z}\frac{1}{\varepsilon_0 V} \underbrace{\sum_{\boldsymbol{l}_k=-\infty}^{\infty}}_{\boldsymbol{l}_k \neq \mathbf{0}} \frac{\partial U^{\text{rec}}_{\boldsymbol{l}_k}}{\partial X} - \frac{1}{V}\frac{\partial V}{\partial X} U_{\text{rec}}, \qquad X \in \{a, b, c, \alpha, \beta, \gamma\}, \tag{3.2.57}$$

wherein $\frac{\partial V}{\partial X}$ can be easily obtained from Equation 3.1.9. Derivatives with respect to a cell length, L, are given by

$$\frac{\partial V}{\partial L} = \sqrt{W} \prod_{L'=a,b,c,L' \neq L} L', \qquad \forall\ L \in \{a, b, c\}, \tag{3.2.58}$$

while derivatives with respect to a cell angle, ζ, are given by

$$\frac{\partial V}{\partial \zeta} = \frac{abc}{2\sqrt{W}}\frac{\partial W}{\partial \zeta}, \qquad \forall\ \zeta \in \{\alpha, \beta, \gamma\}. \tag{3.2.59}$$

$\frac{\partial W}{\partial \zeta}$ has been defined in Appendix C.

Second-Derivatives in Real-Space

The CSP programs developed in our group employ quasi-Newton optimization techniques and require only first-derivatives to operate. However, as will be described in Chapter 4, second-derivatives are also employed in certain parts of our programs. Below, I outline the analytical formulations of these derivatives with respect to Cartesian coordinates ($\nabla^2_{\boldsymbol{r}} U_{\text{inter}}$), lattice parameters ($\nabla^2_{\boldsymbol{X}} U_{\text{inter}}$), and their cross-derivatives ($\nabla_{\boldsymbol{r}} \nabla^T_{\boldsymbol{X}} U_{\text{inter}}$).

Starting from the general, first-derivative formulation (Equation 3.2.40), a second differentiation with respect to $\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ is done. This gives

$$\begin{aligned}\nabla^2_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj} =& \frac{1}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \left\{ \left[\frac{1}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \frac{\partial^2 U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial (r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2} - \frac{1}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2} \frac{\partial U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \right] \right. \\ & \left. \cdot (\boldsymbol{r}_{\mathbf{0}Jj} + \boldsymbol{R}\cdot\boldsymbol{l} - \boldsymbol{r}_{\mathbf{0}Ii})(\boldsymbol{r}_{\mathbf{0}Jj} + \boldsymbol{R}\cdot\boldsymbol{l} - \boldsymbol{r}_{\mathbf{0}Ii})^T + \frac{\partial U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \boldsymbol{I} \right\}, \\ & \kappa \in \{\text{real}, \text{FF}\},\end{aligned} \tag{3.2.60}$$

where $\boldsymbol{I}$ is the identity matrix. As before, Equation 3.2.60 can be used to determine other double-derivatives with respect to $\boldsymbol{r}_{\mathbf{0}Ii}$, $\boldsymbol{r}_{\mathbf{0}Ii}$, and $\boldsymbol{X}$:

$$\nabla^2_{\boldsymbol{r}_{\mathbf{0}Ii}} U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \nabla^2_{\boldsymbol{r}_{\mathbf{0}Jj}} U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \nabla^2_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}, \qquad \kappa \in \{\text{real}, \text{FF}\}, \tag{3.2.61}$$

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}}\nabla^T_{\boldsymbol{r}_{\mathbf{0}Jj}}U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \nabla_{\boldsymbol{r}_{\mathbf{0}Jj}}\nabla^T_{\boldsymbol{r}_{\mathbf{0}Ii}}U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = -\nabla^2_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}, \qquad \kappa \in \{\text{real}, \text{FF}\}, \tag{3.2.62}$$

$$\begin{aligned}\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}}\frac{\partial U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial X} = -\nabla_{\boldsymbol{r}_{\mathbf{0}Jj}}\frac{\partial U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial X} = -\nabla^2_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}(\nabla_X\boldsymbol{R}\cdot\boldsymbol{l}),\\ \kappa \in \{\text{real}, \text{FF}\}, \quad X \in \{a, b, c, \alpha, \beta, \gamma\},\end{aligned} \tag{3.2.63}$$

$$\begin{aligned}\frac{\partial^2 U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial X \partial X'} =&(\nabla_X\boldsymbol{R}\cdot\boldsymbol{l})^T\nabla^2_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}(\nabla_{X'}\boldsymbol{R}\cdot\boldsymbol{l}) + (\nabla_X\nabla_{X'}\boldsymbol{R}\cdot\boldsymbol{l})^T\nabla_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj},\\ &\kappa \in \{\text{real}, \text{FF}\}, \quad X, X' \in \{a, b, c, \alpha, \beta, \gamma\},\end{aligned} \tag{3.2.64}$$

with the special case of an atom interacting with its periodic image (i.e., $Ii = Jj$):

$$\begin{aligned}\nabla^2_{\boldsymbol{r}_{\mathbf{0}Ii}}U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj} =&\nabla^2_{\boldsymbol{r}_{\mathbf{0}Jj}}U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \nabla_{\boldsymbol{r}_{\mathbf{0}Ii}}\nabla^T_{\boldsymbol{r}_{\mathbf{0}Jj}}U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \nabla_{\boldsymbol{r}_{\mathbf{0}Jj}}\nabla^T_{\boldsymbol{r}_{\mathbf{0}Ii}}U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \mathbf{0},\\ &\kappa \in \{\text{real}, \text{FF}\}, \quad Ii = Jj.\end{aligned} \tag{3.2.65}$$

All quantities required to evaluate these real-space second-derivatives have been previously defined except for $\frac{\partial^2 U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2}$ and $\nabla_X\nabla_{X'}\boldsymbol{R}$. The double-derivatives of $\boldsymbol{R}$ with respect to X and X' are tabulated in Appendix C, while $\frac{\partial^2 U^{\kappa}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2}$ for $\kappa \in \{\text{FF}, \text{real}\}$ can be formulated by respective differentiation of Equations 3.2.45–3.2.47:

$$\frac{\partial^2 U^{\text{FF,BHP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2} = \frac{A^{t,t'}}{(B^{t,t'})^2}\exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right) - \frac{42C^{t,t'}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^8}, \tag{3.2.66}$$

$$\frac{\partial^2 U^{\text{FF,MIP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2} = C^{t,t'}_{\text{Mie}}\cdot\varepsilon^{t,t'}\cdot\left[\frac{\lambda^{t,t'}_{\text{r}}(\lambda^{t,t'}_{\text{r}}+1)(\sigma^{t,t'})^{\lambda^{t,t'}_{\text{r}}}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^{\lambda^{t,t'}_{\text{r}}+2}} + \frac{\lambda^{t,t'}_{\text{a}}(\lambda^{t,t'}_{\text{a}}+1)(\sigma^{t,t'})^{\lambda^{t,t'}_{\text{a}}}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^{\lambda^{t,t'}_{\text{a}}+2}}\right], \tag{3.2.67}$$

$$\begin{aligned}\frac{\partial^2 U^{\text{real}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2} = \frac{Q_{Ii}Q_{Jj}}{4\pi\varepsilon_0}\Bigg\{&\left[4\pi\alpha_0^{3/2} + \frac{4\sqrt{\alpha_0}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2}\right]\exp\left(-\alpha_0\cdot\pi\cdot(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2\right)\\ &+ \frac{2\text{erfc}(r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\sqrt{\alpha_0\cdot\pi})}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^3}\Bigg\}.\end{aligned} \tag{3.2.68}$$

Substituting these pairwise gradient expressions into their appropriate summations over the entire lattice gives the total second-derivative with respect to $\boldsymbol{r}$ and $\boldsymbol{X}$. As before, no derivatives with respect to U_{corr} are needed.

Second-Derivatives in Reciprocal-Space

Progressing to the reciprocal-space contributions, the second-derivatives with respect to Cartesian coordinates can be obtained by differentiation of Equations 3.2.48–3.2.50 by $\boldsymbol{r}_{\mathbf{0}Ii}$ (recognizing $\boldsymbol{k}$ has no dependency on $\boldsymbol{r}_{\mathbf{0}Ii}$):

$$\nabla^2_{\boldsymbol{r}_{\mathbf{0}Ii}} U^{\mathrm{rec}}_{\boldsymbol{l}_k} = \frac{1}{\|\boldsymbol{k}\|^2} \exp\left(\frac{-\|\boldsymbol{k}\|^2}{4\cdot\pi\cdot\alpha_0}\right)\Bigg[2(\nabla_{r_{\mathbf{0}Ii}}\rho_c(\boldsymbol{k}))(\nabla^T_{r_{\mathbf{0}Ii}}\rho_c(\boldsymbol{k})) + 2\rho_c(\boldsymbol{k})\nabla^2_{r_{\mathbf{0}Ii}}\rho_c(\boldsymbol{k}) + 2(\nabla_{r_{\mathbf{0}Ii}}\rho_s(\boldsymbol{k}))(\nabla^T_{r_{\mathbf{0}Ii}}\rho_s(\boldsymbol{k})) + 2\rho_s^2(\boldsymbol{k})\nabla^2_{r_{\mathbf{0}Ii}}\rho_s(\boldsymbol{k})\Bigg], \tag{3.2.69}$$

where

$$\nabla^2_{\boldsymbol{r}_{\mathbf{0}Ii}}\rho_c(\boldsymbol{k}) = -\boldsymbol{k}\boldsymbol{k}^T Q_{Ii}\cos(\boldsymbol{k}\cdot\boldsymbol{r}_{\mathbf{0}Ii}), \tag{3.2.70}$$

$$\nabla^2_{\boldsymbol{r}_{\mathbf{0}Ii}}\rho_s(\boldsymbol{k}) = -\boldsymbol{k}\boldsymbol{k}^T Q_{Ii}\sin(\boldsymbol{k}\cdot\boldsymbol{r}_{\mathbf{0}Ii}). \tag{3.2.71}$$

In the case of cross-derivatives between Cartesian coordinates, this simplifies to

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}}\nabla^T_{\boldsymbol{r}_{\mathbf{0}Jj}} U^{\mathrm{rec}}_{\boldsymbol{l}_k} = \frac{1}{\|\boldsymbol{k}\|^2} \exp\left(\frac{-\|\boldsymbol{k}\|^2}{4\cdot\pi\cdot\alpha_0}\right)\Bigg[2(\nabla_{r_{\mathbf{0}Ii}}\rho_c(\boldsymbol{k}))(\nabla^T_{r_{\mathbf{0}Jj}}\rho_c(\boldsymbol{k})) + 2(\nabla_{r_{\mathbf{0}Ii}}\rho_s(\boldsymbol{k}))(\nabla^T_{r_{\mathbf{0}Jj}}\rho_s(\boldsymbol{k}))\Bigg]. \tag{3.2.72}$$

Cross-derivatives of Cartesian coordinates and lattice parameters can be derived by differentiation of Equations 3.2.53–3.2.55 by $\boldsymbol{r}_{\mathbf{0}Ii}$,

$$\begin{aligned}\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}}\frac{\partial U^{\mathrm{rec}}_{\boldsymbol{l}_k}}{\partial X} =& \frac{1}{\|\boldsymbol{k}\|^2}\exp\left(\frac{-\|\boldsymbol{k}\|^2}{4\cdot\pi\cdot\alpha_0}\right)(\nabla^T_X\boldsymbol{k}) \\ &\Bigg\{-\left(\frac{1}{\|\boldsymbol{k}\|^2}+\frac{1}{2\cdot\pi\cdot\alpha_0}\right)\left[2\rho_c(\boldsymbol{k})\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}}\rho_c(\boldsymbol{k}) + 2\rho_s(\boldsymbol{k})\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}}\rho_s(\boldsymbol{k})\right]\boldsymbol{k}^T \\ &+ 2(\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}}\rho_c(\boldsymbol{k}))(\nabla^T_{\boldsymbol{k}}\rho_c(\boldsymbol{k})) + 2\rho_c(\boldsymbol{k})\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}}\nabla^T_{\boldsymbol{k}}\rho_c(\boldsymbol{k}) \\ &+ 2(\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}}\rho_s(\boldsymbol{k}))(\nabla^T_{\boldsymbol{k}}\rho_s(\boldsymbol{k})) + 2\rho_s(\boldsymbol{k})\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}}\nabla^T_{\boldsymbol{k}}\rho_s(\boldsymbol{k})\Bigg\}, \qquad X\in\{a,b,c,\alpha,\beta,\gamma\},\end{aligned} \tag{3.2.73}$$

where

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}}\nabla^T_{\boldsymbol{k}}\rho_c(\boldsymbol{k}) = -Q_{Ii}\left[\boldsymbol{I}\sin(\boldsymbol{k}\cdot\boldsymbol{r}_{\mathbf{0}Ii}) + \boldsymbol{k}\boldsymbol{r}^T_{\mathbf{0}Ii}\cos(\boldsymbol{k}\cdot\boldsymbol{r}_{\mathbf{0}Ii})\right], \tag{3.2.74}$$

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}}\nabla^T_{\boldsymbol{k}}\rho_s(\boldsymbol{k}) = Q_{Ii}\left[\boldsymbol{I}\cos(\boldsymbol{k}\cdot\boldsymbol{r}_{\mathbf{0}Ii}) - \boldsymbol{k}\boldsymbol{r}^T_{\mathbf{0}Ii}\sin(\boldsymbol{k}\cdot\boldsymbol{r}_{\mathbf{0}Ii})\right]. \tag{3.2.75}$$

Finally, progressing to double-derivatives of $U^{\mathrm{rec}}_{\boldsymbol{l}_k}$ with respect to X. By judicious application of chain-rule on Equations 3.2.53–3.2.55, one can derive

$$
\begin{aligned}
\frac{\partial^2 U_{\boldsymbol{l}_k}^{\mathrm{rec}}}{\partial X \partial X'} =& \frac{1}{\|\boldsymbol{k}\|^2} \exp\left(\frac{-\|\boldsymbol{k}\|^2}{4\cdot\pi\cdot\alpha_0}\right)\Bigg\{\left(-\frac{1}{\|\boldsymbol{k}\|^2}-\frac{1}{2\pi\alpha_0}\right)\boldsymbol{k}^T\nabla_{X'}\boldsymbol{k} \\
&\cdot\Bigg[\left(-\frac{1}{\|\boldsymbol{k}\|^2}-\frac{1}{2\pi\alpha_0}\right)\cdot\left[\rho_c^2(\boldsymbol{k})+\rho_s^2(\boldsymbol{k})\right]\boldsymbol{k}^T \\
&+ 2\rho_c(\boldsymbol{k})\nabla_{\boldsymbol{k}}^T\rho_c(\boldsymbol{k}) + 2\rho_s(\boldsymbol{k})\nabla_{\boldsymbol{k}}^T\rho_s(\boldsymbol{k}) \\
&+\left(-\frac{1}{\|\boldsymbol{k}\|^2}-\frac{1}{2\pi\alpha_0}\right)\left[\rho_c^2(\boldsymbol{k})+\rho_s^2(\boldsymbol{k})\right]\nabla_{X'}\boldsymbol{k}^T + \frac{2\boldsymbol{k}^T}{\|\boldsymbol{k}\|^4}\nabla_{X'}\boldsymbol{k}\left[\rho_c^2(\boldsymbol{k})+\rho_s^2(\boldsymbol{k})\right]\boldsymbol{k}^T \\
&+\left(-\frac{1}{\|\boldsymbol{k}\|^2}-\frac{1}{2\pi\alpha_0}\right)\left[2\rho_c(\boldsymbol{k})\nabla_{\boldsymbol{k}}^T\rho_c(\boldsymbol{k}) + 2\rho_s(\boldsymbol{k})\nabla_{\boldsymbol{k}}^T\rho_s(\boldsymbol{k})\right]\left[\nabla_{X'}\boldsymbol{k}\right]\boldsymbol{k}^T \\
&+ 2\nabla_{\boldsymbol{k}}^T\rho_c(\boldsymbol{k})\nabla_{X'}\boldsymbol{k}\nabla_{\boldsymbol{k}}^T\rho_c(\boldsymbol{k}) + 2\nabla_{\boldsymbol{k}}^T\rho_s(\boldsymbol{k})\nabla_{X'}\boldsymbol{k}\nabla_{\boldsymbol{k}}^T\rho_s(\boldsymbol{k}) \\
&+ 2\rho_c(\boldsymbol{k})\nabla_{X'}\boldsymbol{k}^T\nabla_{\boldsymbol{k}}^2\rho_c(\boldsymbol{k}) + 2\rho_s(\boldsymbol{k})\nabla_{X'}\boldsymbol{k}^T\nabla_{\boldsymbol{k}}^2\rho_s(\boldsymbol{k})\Bigg]\nabla_X\boldsymbol{k} \\
&+\Bigg[\left(-\frac{1}{\|\boldsymbol{k}\|^2}-\frac{1}{2\pi\alpha_0}\right)\left[\rho_c^2(\boldsymbol{k})+\rho_s^2(\boldsymbol{k})\right]\boldsymbol{k}^T \\
&+ 2\rho_c(\boldsymbol{k})\nabla_{\boldsymbol{k}}^T\rho_c(\boldsymbol{k}) + 2\rho_s(\boldsymbol{k})\nabla_{\boldsymbol{k}}^T\rho_s(\boldsymbol{k})\Bigg]\nabla_X\nabla_{X'}\boldsymbol{k}\Bigg\}, \qquad X, X' \in \{a,b,c,\alpha,\beta,\gamma\},
\end{aligned}
\tag{3.2.76}
$$

where

$$
\nabla_{\boldsymbol{k}}^2\rho_c(\boldsymbol{k}) = \sum_{I=1}^{Z}\sum_{i=1}^{\eta_I} -Q_{Ii}\boldsymbol{r}_{\boldsymbol{0}Ii}\boldsymbol{r}_{\boldsymbol{0}Ii}^T\cos\left(\boldsymbol{k}\cdot\boldsymbol{r}_{\boldsymbol{0}Ii}\right), \tag{3.2.77}
$$

$$
\nabla_{\boldsymbol{k}}^2\rho_s(\boldsymbol{k}) = \sum_{I=1}^{Z}\sum_{i=1}^{\eta_I} -Q_{Ii}\boldsymbol{r}_{\boldsymbol{0}Ii}\boldsymbol{r}_{\boldsymbol{0}Ii}^T\sin\left(\boldsymbol{k}\cdot\boldsymbol{r}_{\boldsymbol{0}Ii}\right), \tag{3.2.78}
$$

and

$$
\nabla_X\nabla_{X'}\boldsymbol{k} = 2\pi(\nabla_X\nabla_{X'}\boldsymbol{R}^{-1})^T\boldsymbol{l}_k \qquad X, X' \in \{a,b,c,\alpha,\beta,\gamma\}. \tag{3.2.79}
$$

Expressions for $\nabla_X\nabla_{X'}\boldsymbol{R}^{-1}$ are tabulated in Appendix C. Substituting these expressions into their corresponding summations, the second-derivatives of the total reciprocal-space contribution are given by

$$
\nabla_{\boldsymbol{r}_{\boldsymbol{0}Ii}}^2 U_{\mathrm{rec}} = \frac{1}{2Z}\frac{1}{\varepsilon_0 V}\underbrace{\sum_{\boldsymbol{l}_k=-\infty}^{\infty}}_{\boldsymbol{l}_k\neq\boldsymbol{0}}\nabla_{\boldsymbol{r}_{\boldsymbol{0}Ii}}^2 U_{\boldsymbol{l}_k}^{\mathrm{rec}}, \tag{3.2.80}
$$

$$
\nabla_{\boldsymbol{r}_{\boldsymbol{0}Ii}}\nabla_{\boldsymbol{r}_{\boldsymbol{0}Jj}}^T U_{\mathrm{rec}} = \frac{1}{2Z}\frac{1}{\varepsilon_0 V}\underbrace{\sum_{\boldsymbol{l}_k=-\infty}^{\infty}}_{\boldsymbol{l}_k\neq\boldsymbol{0}}\nabla_{\boldsymbol{r}_{\boldsymbol{0}Ii}}\nabla_{\boldsymbol{r}_{\boldsymbol{0}Jj}}^T U_{\boldsymbol{l}_k}^{\mathrm{rec}}, \tag{3.2.81}
$$

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}} \frac{\partial U_{\text{rec}}}{\partial X} = \frac{1}{2Z}\frac{1}{\varepsilon_0 V} \underbrace{\sum_{\boldsymbol{l}_k=-\infty}^{\infty}}_{\boldsymbol{l}_k \neq \mathbf{0}} \nabla_{\boldsymbol{r}_{\mathbf{0}Ii}} \frac{\partial U_{\boldsymbol{l}_k}^{\text{rec}}}{\partial X} - \frac{1}{V}\frac{\partial V}{\partial X} \nabla_{\boldsymbol{r}_{\mathbf{0}Ii}} U_{\text{rec}}, \qquad X \in \{a, b, c, \alpha, \beta, \gamma\}, \tag{3.2.82}$$

$$\begin{aligned} \frac{\partial^2 U_{\text{rec}}}{\partial X \partial X'} =& \frac{1}{2Z}\frac{1}{\varepsilon_0 V} \underbrace{\sum_{\boldsymbol{l}_k=-\infty}^{\infty}}_{\boldsymbol{l}_k \neq \mathbf{0}} \frac{\partial^2 U_{\boldsymbol{l}_k}^{\text{rec}}}{\partial X \partial X'} - \frac{1}{V}\left[\frac{\partial V}{\partial X}\frac{\partial U_{\text{rec}}}{\partial X'} + \frac{\partial V}{\partial X'}\frac{\partial U_{\text{rec}}}{\partial X} + \frac{\partial^2 V}{\partial X \partial X'} U_{\text{rec}}\right], \\ & X, X' \in \{a, b, c, \alpha, \beta, \gamma\}. \end{aligned} \tag{3.2.83}$$

Derivatives with Respect to $\boldsymbol{\Omega}$

Analyzing derivatives with respect to $\boldsymbol{r}$ and $\boldsymbol{X}$ is natural given the formulation of the intermolecular interaction functions. However, lattice energy minimizations are performed in terms of $\boldsymbol{\Omega}$. As such, the already-derived gradient expressions must be related to the corresponding gradients required by the optimization routines. To recap, $\boldsymbol{\Omega}$ includes: lattice parameters, and the corresponding fractional COM positions and orientation (i.e., Euler angles) for each molecule in the asymmetric unit cell. Following Bowskill [2], I partition these variables as

$$\begin{aligned} &\boldsymbol{\Omega} \in \{\boldsymbol{\mu}, \boldsymbol{\nu}\}, \\ &\boldsymbol{\mu} \in \{\hat{\boldsymbol{r}}_{0,N}, \phi_N, \theta_N, \psi_N\}, \qquad N = 1, 2, \cdots, Z', \\ &\boldsymbol{\nu} \in \{a, b, c, \alpha, \beta, \gamma\}, \qquad s.t. \quad \text{Space group constraints.} \end{aligned} \tag{3.2.84}$$

It is clear that $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{\nu})$. By embedding these into the expression for U_{inter},

$$U_{\text{inter}} = U_{\text{inter}}(\boldsymbol{r}, \boldsymbol{X}) = U_{\text{inter}}(\boldsymbol{r}(\boldsymbol{\mu}, \boldsymbol{\nu}), \boldsymbol{X}(\boldsymbol{\nu})), \tag{3.2.85}$$

and by application of chain-rule, one obtains

$$\frac{\partial U_{\text{inter}}}{\partial \mu_i} = \nabla_{\mu_i} \boldsymbol{r}^T \nabla_{\boldsymbol{r}} U_{\text{inter}}, \qquad \mu_i \in \boldsymbol{\mu}, \tag{3.2.86}$$

$$\frac{\partial U_{\text{inter}}}{\partial \nu_i} = \nabla_{\nu_i} \boldsymbol{r}^T \nabla_{\boldsymbol{r}} U_{\text{inter}} + \nabla_{\nu_i} \boldsymbol{X}^T \nabla_{\boldsymbol{X}} U_{\text{inter}}, \qquad \nu_i \in \boldsymbol{\nu}. \tag{3.2.87}$$

To evaluate these equations, expressions for $\nabla_{\mu_i}^T \boldsymbol{r}$, $\nabla_{\nu_i}^T \boldsymbol{r}$, and $\nabla_{\nu_i} \boldsymbol{X}^T$ are needed (derivatives with respect to $\boldsymbol{r}$ and $\boldsymbol{X}$ were already derived). By definition, $\boldsymbol{X} \equiv \boldsymbol{\nu}$, hence it follows that

$$\frac{\partial X_i}{\partial \nu_j} = \begin{cases} 1, & \text{if } X_i = \nu_j, \\ 0, & \text{otherwise,} \end{cases} \tag{3.2.88}$$

although care must be exercised in the case of space groups which have constraints on their lattice parameters. Variables in Equation 3.1.22 can be re-grouped to express the Cartesian coordinates of an atom in terms of its molecular fractional COM position and orientation:

$$\begin{aligned}\boldsymbol{r}_{(s-1)Z'+N,i} =&(\boldsymbol{R}\boldsymbol{M}_s\boldsymbol{R}^{-1})\boldsymbol{A}^T_{\mathrm{rot},N}\boldsymbol{r}^{\mathrm{bf}}_{Ni} + (\boldsymbol{R}\boldsymbol{M}_s)\hat{\boldsymbol{r}}_{0,N}\\ &+\boldsymbol{R}(\boldsymbol{T}_s+\boldsymbol{t}_{N,s}), \qquad s=1,\cdots,\sigma_s, \quad i=1,\cdots,\eta_N, \quad N=1,\cdots,Z'.\end{aligned} \tag{3.2.89}$$

From here, it is trivial to show

$$\begin{aligned}&\left(\nabla_{\hat{x}_{0,N}}\boldsymbol{r}_{(s-1)Z'+N,i} \quad \nabla_{\hat{y}_{0,N}}\boldsymbol{r}_{(s-1)Z'+N,i} \quad \nabla_{\hat{z}_{0,N}}\boldsymbol{r}_{(s-1)Z'+N,i}\right) = \boldsymbol{R}\boldsymbol{M}_s,\\ &s=1,\cdots,\sigma_s, \quad i=1,\cdots,\eta_N, \quad N=1,\cdots,Z',\end{aligned} \tag{3.2.90}$$

and

$$\begin{aligned}&\nabla_{\omega_N}\boldsymbol{r}_{(s-1)Z'+N,i} = (\boldsymbol{R}\boldsymbol{M}_s\boldsymbol{R}^{-1})\nabla_{\omega_N}\boldsymbol{A}^T_{\mathrm{rot},N}\boldsymbol{r}^{\mathrm{bf}}_{Ni},\\ &\omega_N\in\{\phi_N,\theta_N,\psi_n\}, \quad s=1,\cdots,\sigma_s, \quad i=1,\cdots,\eta_N, \quad N=1,\cdots,Z'.\end{aligned} \tag{3.2.91}$$

Expressions for $\nabla_{\omega_N}\boldsymbol{A}^T_{\mathrm{rot},N}$ can be derived by differentiation of the matrix shown in Equation 3.1.15. Finally, the derivatives of Cartesian coordinates with respect to $\boldsymbol{\nu}$ can be obtained as

$$\begin{aligned}\nabla_\nu\boldsymbol{r}_{(s-1)Z'+N,i} = \sum_{X=a,b,c,\alpha,\beta,\gamma}\frac{\partial X}{\partial\nu}\Bigg[&\left(\nabla_X\boldsymbol{R}\boldsymbol{M}_s\boldsymbol{R}^{-1}+\boldsymbol{R}\boldsymbol{M}_s\nabla_X\boldsymbol{R}^{-1}\right)\boldsymbol{A}^T_{\mathrm{rot},N}\boldsymbol{r}^{\mathrm{bf}}_{Ni}\\ &+\nabla_X\boldsymbol{R}(\boldsymbol{T}_s+\boldsymbol{t}_{N,s})\Bigg],\\ &\nu\in\{a,b,c,\alpha,\beta,\gamma\}, \quad s=1,\cdots,\sigma_s, \quad i=1,\cdots,\eta_N, \quad N=1,\cdots,Z'.\end{aligned} \tag{3.2.92}$$

By an identical approach, second-derivatives with respect to $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ can be expressed as

$$\frac{\partial^2 U_{\mathrm{inter}}}{\partial\mu_i\partial\mu_j} = \nabla_{\mu_i}\nabla_{\mu_j}\boldsymbol{r}^T\nabla_{\boldsymbol{r}}U_{\mathrm{inter}} + \nabla_{\mu_i}\boldsymbol{r}^T\nabla^2_{\boldsymbol{r}}U_{\mathrm{inter}}\nabla_{\mu_j}\boldsymbol{r}, \qquad \mu_i,\mu_j\in\boldsymbol{\mu}, \tag{3.2.93}$$

$$\begin{aligned}\frac{\partial^2 U_{\mathrm{inter}}}{\partial\mu_i\partial\nu_j} =&\nabla_{\mu_i}\nabla_{\nu_j}\boldsymbol{r}^T\nabla_{\boldsymbol{r}}U_{\mathrm{inter}} + \nabla_{\mu_i}\boldsymbol{r}^T\nabla^2_{\boldsymbol{r}}U_{\mathrm{inter}}\nabla_{\nu_j}\boldsymbol{r}\\ &+\nabla_{\mu_i}\boldsymbol{r}^T\nabla_{\boldsymbol{r}}\nabla^T_{\boldsymbol{X}}U_{\mathrm{inter}}\nabla_{\nu_j}\boldsymbol{X}, \qquad \mu_i\in\boldsymbol{\mu}, \quad \nu_i\in\boldsymbol{\nu},\end{aligned} \tag{3.2.94}$$

$$\begin{aligned}\frac{\partial^2 U_{\mathrm{inter}}}{\partial\nu_i\partial\nu_j} =&\nabla_{\nu_i}\nabla_{\nu_j}\boldsymbol{r}^T\nabla_{\boldsymbol{r}}U_{\mathrm{inter}} + \nabla_{\nu_i}\boldsymbol{r}^T\nabla^2_{\boldsymbol{r}}U_{\mathrm{inter}}\nabla_{\nu_j}\boldsymbol{r}\\ &+\nabla_{\nu_i}\boldsymbol{r}^T\nabla_{\boldsymbol{r}}\nabla^T_{\boldsymbol{X}}U_{\mathrm{inter}}\nabla_{\nu_j}\boldsymbol{X} + \nabla_{\nu_j}\boldsymbol{r}^T\nabla_{\boldsymbol{r}}\nabla^T_{\boldsymbol{X}}U_{\mathrm{inter}}\nabla_{\nu_i}\boldsymbol{X}\\ &+\nabla_{\nu_i}\boldsymbol{X}^T\nabla^2_{\boldsymbol{X}}U_{\mathrm{inter}}\nabla_{\nu_j}\boldsymbol{X}, \qquad \nu_i,\nu_j\in\boldsymbol{\nu}.\end{aligned} \tag{3.2.95}$$

Like the first-derivatives, double-derivatives with respect to $\boldsymbol{r}$ and $\boldsymbol{X}$ have been previously defined, and the only undetermined quantities are the second-derivatives with respect to $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$. Double-derivatives with respect to $\boldsymbol{\mu}$ can be derived for the Cartesian coordinates. From Equation 3.2.90, the first-derivatives of $\boldsymbol{r}$ with respect to the fractional COM coordinates are independent of the fractional COM coordinates themselves. Therefore, all $\nabla_{\mu_i}\nabla_{\mu_j}\boldsymbol{r}^T = \boldsymbol{0}$ when

both μ_i and μ_j are fractional COM coordinates. Applying the same logic to Equation 3.2.91, all $\nabla_{\mu_i}\nabla_{\mu_j}\boldsymbol{r}^T = \mathbf{0}$ when μ_i and μ_j are a pair of fractional COM coordinates and Euler angles. Double-derivatives of two Euler angles can be obtained by differentiation of Equation 3.2.91,

$$\begin{aligned}
&\nabla_{\omega_N}\nabla_{\omega'_{N'}}\boldsymbol{r}_{(s-1)Z'+N,i} = \delta_{N,N'}(\boldsymbol{R}\boldsymbol{M}_s\boldsymbol{R}^{-1})\nabla_{\omega_N}\nabla_{\omega'_{N'}}\boldsymbol{A}^T_{\mathrm{rot},N}\boldsymbol{r}^{\mathrm{bf}}_{Ni},\\
&\omega_N \in \{\phi_N,\theta_N,\psi_N\}, \quad \omega'_{N'} \in \{\phi_{N'},\theta_{N'},\psi_{N'}\}, \quad s = 1,\cdots,\sigma_s, \quad i = 1,\cdots,\eta_N,\\
&N = 1,\cdots,Z', \quad N' = 1,\cdots,Z',
\end{aligned} \tag{3.2.96}$$

where $\delta_{N,N'}$ is the Kronecker delta. Second-derivatives of $\boldsymbol{A}^T_{\mathrm{rot},N}$ can be found from Equation 3.1.15. Next, cross-derivatives of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ can be derived for the Cartesian coordinates, from Equation 3.2.90. These are

$$\begin{aligned}
&(\nabla_{\hat{x}_{0,N}}\nabla_\nu\boldsymbol{r}_{(s-1)Z'+N,i} \quad \nabla_{\hat{y}_{0,N}}\nabla_\nu\boldsymbol{r}_{(s-1)Z'+N,i} \quad \nabla_{\hat{z}_{0,N}}\nabla_\nu\boldsymbol{r}_{(s-1)Z'+N,i})\\
&= \sum_{X=a,b,c,\alpha,\beta,\gamma}\frac{\partial X}{\partial\nu}\nabla_X\boldsymbol{R}\boldsymbol{M}_s,\\
&\nu \in \{a,b,c,\alpha,\beta,\gamma\}, \quad s = 1,\cdots,\sigma_s, \quad i = 1,\cdots,\eta_N, \quad N = 1,\cdots,Z',
\end{aligned} \tag{3.2.97}$$

and

$$\begin{aligned}
\nabla_{\omega_N}\nabla_\nu\boldsymbol{r}_{(s-1)Z'+N,i} = &\sum_{X=a,b,c,\alpha,\beta,\gamma}\frac{\partial X}{\partial\nu}(\nabla_X\boldsymbol{R}\boldsymbol{M}_s\boldsymbol{R}^{-1} + \boldsymbol{R}\boldsymbol{M}_s\nabla_X\boldsymbol{R}^{-1})\nabla_{\omega_N}\boldsymbol{A}^T_{\mathrm{rot},N}\boldsymbol{r}^{\mathrm{bf}}_{Ni},\\
&\omega_N \in \{\phi_N,\theta_N,\psi_n\}, \quad \nu \in \{a,b,c,\alpha,\beta,\gamma\},\\
&s = 1,\cdots,\sigma_s, \quad i = 1,\cdots,\eta_N, \quad N = 1,\cdots,Z',
\end{aligned} \tag{3.2.98}$$

for the derivatives with respect to fractional COM coordinates and the Euler angles respectively. Finally, double-derivatives with respect to $\boldsymbol{\nu}$ can be derived for the Cartesian coordinates,

$$\begin{aligned}
\nabla_\nu\nabla_{\nu'}\boldsymbol{r}_{(s-1)Z'+N,i} = &\sum_{X,X'=a,b,c,\alpha,\beta,\gamma}\frac{\partial X}{\partial\nu}\frac{\partial X'}{\partial\nu'}\Bigg[\Big(\nabla_X\nabla_{X'}\boldsymbol{R}\boldsymbol{M}_s\boldsymbol{R}^{-1} + \nabla_X\boldsymbol{R}\boldsymbol{M}_s\nabla_{X'}\boldsymbol{R}^{-1}\\
&+ \nabla_{X'}\boldsymbol{R}\boldsymbol{M}_s\nabla_X\boldsymbol{R}^{-1} + \boldsymbol{R}\boldsymbol{M}_s\nabla_X\nabla_{X'}\boldsymbol{R}^{-1}\Big)\boldsymbol{A}^T_{\mathrm{rot},N}\boldsymbol{r}^{\mathrm{bf}}_{Ni}\\
&+ \nabla_X\nabla_{X'}\boldsymbol{R}(\boldsymbol{T}_s + \boldsymbol{t}_{N,s})\Bigg],\\
&\nu,\nu' \in \{a,b,c,\alpha,\beta,\gamma\}, \quad s = 1,\cdots,\sigma_s, \quad i = 1,\cdots,\eta_N, \quad N = 1,\cdots,Z'.
\end{aligned} \tag{3.2.99}$$

Double-derivatives of X with respect to $\boldsymbol{\nu}$ are not required, and they are all zero in any case. With this, all first- and second-derivatives related to isotropic intermolecular interactions are defined in terms of the rigid-body optimization variables. These equations are adequate for the evaluation and minimization of intermolecular energies described by point-charge electrostatics.

3.2.1.4 Gradients of Anisotropic Intermolecular Interactions

The use of multipole models in the CSP workflow demands a corresponding set of gradients for anisotropic intermolecular interactions. In the interest of brevity, I will not detail the full derivation of the anisotropic gradients, since this adhere closely to the previous section. I discuss only the salient changes necessary to accommodate the introduction of the orientation-dependent interaction function, $\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}$, in the anisotropic gradients.

Cartesian Coordinate Dependence

In real-space, the interaction function has implicit dependency on the Cartesian coordinates (due to $\hat{\boldsymbol{r}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$) and derivatives involving $\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ must account for this by chain-rule. The derivatives of $\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}$ with respect to $\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ can be obtained by differentiation of the relevant relations tabulated by Stone [221]. In general, these derivatives can be expressed as

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2} = \nabla_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}(\boldsymbol{A}^T_{\mathrm{rot,I}}, \boldsymbol{A}^T_{\mathrm{rot,J}}, \hat{\boldsymbol{r}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}, \nabla_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \hat{\boldsymbol{r}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}), \tag{3.2.100}$$

where

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \hat{\boldsymbol{r}}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \frac{\begin{pmatrix} (y_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2 + (z_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2 & -(x_{\mathbf{0}Ii,\boldsymbol{l}Jj})(y_{\mathbf{0}Ii,\boldsymbol{l}Jj}) & -(x_{\mathbf{0}Ii,\boldsymbol{l}Jj})(z_{\mathbf{0}Ii,\boldsymbol{l}Jj}) \\ -(x_{\mathbf{0}Ii,\boldsymbol{l}Jj})(y_{\mathbf{0}Ii,\boldsymbol{l}Jj}) & (x_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2 + (z_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2 & -(y_{\mathbf{0}Ii,\boldsymbol{l}Jj})(z_{\mathbf{0}Ii,\boldsymbol{l}Jj}) \\ -(x_{\mathbf{0}Ii,\boldsymbol{l}Jj})(z_{\mathbf{0}Ii,\boldsymbol{l}Jj}) & -(y_{\mathbf{0}Ii,\boldsymbol{l}Jj})(z_{\mathbf{0}Ii,\boldsymbol{l}Jj}) & (x_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2 + (y_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2 \end{pmatrix}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^3}. \tag{3.2.101}$$

$x_{\mathbf{0}Ii,\boldsymbol{l}Jj}$, $y_{\mathbf{0}Ii,\boldsymbol{l}Jj}$, and $z_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ are the 1st, 2nd, and 3rd element of the vector $\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ respectively. Similarly, in the reciprocal-space interaction, the interaction function has implicit dependency on the reciprocal lattice vector (due to $\hat{\boldsymbol{k}}$). The derivatives of $\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}$ with respect to $\boldsymbol{k}$ can be written generally as

$$\nabla_{\boldsymbol{k}} \widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2} = \nabla_{\boldsymbol{k}} \widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}(\boldsymbol{A}^T_{\mathrm{rot,I}}, \boldsymbol{A}^T_{\mathrm{rot,J}}, \hat{\boldsymbol{k}}, \nabla_{\boldsymbol{k}} \hat{\boldsymbol{k}}). \tag{3.2.102}$$

The formulation of $\nabla_{\boldsymbol{k}} \hat{\boldsymbol{k}}$ is identical to Equation 3.2.101, except $\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ (and its respective elements) are replaced with the corresponding terms from $\boldsymbol{k}$. Further differentiation of these first-derivative terms gives the generalized second-derivative expressions,

$$\nabla^2_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2} = \nabla^2_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}(\boldsymbol{A}^T_{\mathrm{rot,I}}, \boldsymbol{A}^T_{\mathrm{rot,J}}, \hat{\boldsymbol{r}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}, \nabla_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \hat{\boldsymbol{r}}_{\mathbf{0}Ii,\boldsymbol{l}Jj} \nabla^2_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \hat{\boldsymbol{r}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}), \tag{3.2.103}$$

and

$$\nabla^2_{\boldsymbol{k}} \widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2} = \nabla^2_{\boldsymbol{k}} \widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}(\boldsymbol{A}^T_{\mathrm{rot,I}}, \boldsymbol{A}^T_{\mathrm{rot,J}}, \hat{\boldsymbol{k}}, \nabla_{\boldsymbol{k}} \hat{\boldsymbol{k}}, \nabla^2_{\boldsymbol{k}} \hat{\boldsymbol{k}}), \tag{3.2.104}$$

for the real-space and reciprocal-space interaction functions respectively. Second-derivatives of U_{inter} with respect to $\boldsymbol{r}$, $\boldsymbol{X}$, and their cross-derivatives have been previously defined.

Molecular Orientation Dependence

The addition of $\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}$ in the anisotropic energy function also introduces an orientation dependence from $\boldsymbol{A}_{\mathrm{rot},I}$. To reflect this, I consider the derivatives of $U^{\mathrm{real}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ and $U^{\mathrm{rec}}_{\boldsymbol{l}_k}$ with respect to the Euler angles. From Equation 3.2.15, the (pairwise) real-space gradient is given by

$$\begin{aligned}\frac{\partial U^{\mathrm{real}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial\omega} =& \frac{1}{4\pi\varepsilon_0}\sum_{l_1,l_2}\frac{\Gamma(\frac{1}{2}+l_1+l_2,\alpha_{l_1+l_2}\cdot\pi\cdot(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2)}{\Gamma(\frac{1}{2}+l_1+l_2)}\\ &\cdot\sum_{m_1,m_2}\frac{\partial\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}}{\partial\omega}\frac{Q^{l_1m_1}_{Ii}Q^{l_2m_2}_{Jj}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^{l_1+l_2+1}}, \qquad \omega\in\{\phi_N,\theta_N,\psi_N\},\quad N=1,\cdots,Z',\end{aligned} \tag{3.2.105}$$

where

$$\begin{aligned}\frac{\partial\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}}{\partial\omega} =& \frac{\partial\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}}{\partial\omega}(\boldsymbol{A}^T_{\mathrm{rot},I},\boldsymbol{A}^T_{\mathrm{rot},J},\hat{\boldsymbol{r}}_{\mathbf{0}Ii,\boldsymbol{l}Jj},\nabla_\omega\boldsymbol{A}^T_{\mathrm{rot},I},\nabla_\omega\boldsymbol{A}^T_{\mathrm{rot},J},),\\ &\omega\in\{\phi_N,\theta_N,\psi_N\},\quad N=1,\cdots,Z'.\end{aligned} \tag{3.2.106}$$

Total derivatives are obtained by performing a real-space summation over the infinite lattice. The same derivation applies to Equations 3.2.20 and 3.2.23 to obtain the reciprocal-space terms,

$$\begin{aligned}\frac{\partial U^{\mathrm{rec}}_{\boldsymbol{l}_k}}{\partial\omega} =& \sum_{l_1,l_2}\frac{2^{-l_1-l_2}\sqrt{\pi}}{\Gamma(\frac{1}{2}+l_1+l_2)}\frac{1}{\|\boldsymbol{k}\|^{2-l_1-l_2}}\exp\left(\frac{-\|\boldsymbol{k}\|^2}{4\cdot\pi\cdot\alpha_{l_1+l_2}}\right)\\ &\cdot\sum_{m_1,m_2}\sum_{I=1}^{Z}\sum_{J=1}^{Z}\frac{\partial\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}}{\partial\omega}(-1)^{\frac{l_1+l_2}{2}}\\ &\cdot[\rho^{l_1m_1}_{c,I}(\boldsymbol{k})\rho^{l_2m_2}_{c,J}(\boldsymbol{k})+\rho^{l_1m_1}_{s,I}(\boldsymbol{k})\rho^{l_2m_2}_{s,J}(\boldsymbol{k})],\\ &\omega\in\{\phi_N,\theta_N,\psi_N\},\quad N=1,\cdots,Z',\quad l_1+l_2=0,2,4,\cdots,\end{aligned} \tag{3.2.107}$$

and

$$\begin{aligned}\frac{\partial U^{\mathrm{rec}}_{\boldsymbol{l}_k}}{\partial\omega} =& \sum_{l_1,l_2}\frac{2^{-l_1-l_2}\sqrt{\pi}}{\Gamma(\frac{1}{2}+l_1+l_2)}\frac{1}{\|\boldsymbol{k}\|^{2-l_1-l_2}}\exp\left(\frac{-\|\boldsymbol{k}\|^2}{4\cdot\pi\cdot\alpha_{l_1+l_2}}\right)\\ &\cdot\sum_{m_1,m_2}\sum_{I=1}^{Z}\sum_{J=1}^{Z}\frac{\partial\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}}{\partial\omega}(-1)^{\frac{l_1+l_2-1}{2}}\\ &\cdot[\rho^{l_1m_1}_{c,I}(\boldsymbol{k})\rho^{l_2m_2}_{s,J}(\boldsymbol{k})-\rho^{l_1m_1}_{s,I}(\boldsymbol{k})\rho^{l_2m_2}_{c,J}(\boldsymbol{k})],\\ &\omega\in\{\phi_N,\theta_N,\psi_N\},\quad N=1,\cdots,Z',\quad l_1+l_2=1,3,5,\cdots,\end{aligned} \tag{3.2.108}$$

where

$$\begin{aligned}\frac{\partial\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}}{\partial\omega} =& \frac{\partial\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}}{\partial\omega}(\boldsymbol{A}^T_{\mathrm{rot},I},\boldsymbol{A}^T_{\mathrm{rot},J},\hat{\boldsymbol{k}},\nabla_\omega\boldsymbol{A}^T_{\mathrm{rot},I},\nabla_\omega\boldsymbol{A}^T_{\mathrm{rot},J},),\\ &\omega\in\{\phi_N,\theta_N,\psi_N\},\quad N=1,\cdots,Z'.\end{aligned} \tag{3.2.109}$$

Similar expressions can be derived for the corresponding second-derivative terms $\nabla^2_{\boldsymbol{\omega}} U_{\text{inter}}$, $\nabla_{\boldsymbol{r}} \nabla^T_{\boldsymbol{\omega}} U_{\text{inter}}$, and $\nabla_{\boldsymbol{X}} \nabla^T_{\boldsymbol{\omega}} U_{\text{inter}}$.

Derivatives with Respect to $\boldsymbol{\Omega}$

Lastly, I derive the relationship between the preceding gradient terms and the rigid-body optimization variables, $\boldsymbol{\Omega}$. Specifically, the 'natural' variables describing the intermolecular energy must be expanded to explicitly include the Euler angles ($\boldsymbol{\omega}$), i.e., $U_{\text{inter}} = U_{\text{inter}}(\boldsymbol{r}(\boldsymbol{\mu}, \boldsymbol{\nu}), \boldsymbol{X}(\boldsymbol{\nu}), \boldsymbol{\omega})$ (*cf.* Equation 3.2.85). Note that $\boldsymbol{\omega}$ are also contained in $\boldsymbol{\mu}$. This results in adaptation of some of the previously derived equations. For the first-derivatives, derivatives with respect to $\boldsymbol{\mu}$ (Equation 3.2.86) must append on the explicit dependence on $\boldsymbol{\omega}$ when $\mu = \omega$:

$$\frac{\partial U_{\text{inter}}}{\partial \mu_i} = \begin{cases} \nabla_{\mu_i} \boldsymbol{r}^T \nabla_{\boldsymbol{r}} U_{\text{inter}}, & \text{if } \mu_i \notin \boldsymbol{\omega}, \\ \nabla_{\mu_i} \boldsymbol{r}^T \nabla_{\boldsymbol{r}} U_{\text{inter}}, + \nabla_{\mu_i} U_{\text{inter}}, & \text{if } \mu_i \in \boldsymbol{\omega}. \end{cases} \tag{3.2.110}$$

The same must be done for any double-derivatives involving $\mu = \omega$ (Equations 3.2.93 and 3.2.94):

$$\begin{aligned} \frac{\partial^2 U_{\text{inter}}}{\partial \mu_i \partial \mu_j} =& \nabla_{\mu_i} \nabla_{\mu_j} \boldsymbol{r}^T \nabla_{\boldsymbol{r}} U_{\text{inter}} + \nabla_{\mu_i} \boldsymbol{r}^T \nabla^2_{\boldsymbol{r}} U_{\text{inter}} \nabla_{\mu_j} \boldsymbol{r} \\ &+ \nabla_{\mu_i} \boldsymbol{r}^T \nabla_{\boldsymbol{r}} \nabla_{\mu_j} U_{\text{inter}} + \nabla_{\mu_j} \boldsymbol{r}^T \nabla_{\boldsymbol{r}} \nabla_{\mu_i} U_{\text{inter}} + \nabla_{\mu_i} \nabla_{\mu_j} U_{\text{inter}}, \\ & \mu_i, \mu_j \in \boldsymbol{\omega}, \end{aligned} \tag{3.2.111}$$

$$\begin{aligned} \frac{\partial^2 U_{\text{inter}}}{\partial \mu_i \partial \mu_j} =& \nabla_{\mu_i} \nabla_{\mu_j} \boldsymbol{r}^T \nabla_{\boldsymbol{r}} U_{\text{inter}} + \nabla_{\mu_i} \boldsymbol{r}^T \nabla^2_{\boldsymbol{r}} U_{\text{inter}} \nabla_{\mu_j} \boldsymbol{r} + \nabla_{\mu_i} \boldsymbol{r}^T \nabla_{\boldsymbol{r}} \nabla_{\mu_j} U_{\text{inter}}, \\ & \mu_i, \mu_j \in \boldsymbol{\mu}, \quad \mu_i \in \{\hat{x}_{0,N}, \hat{y}_{0,N}, \hat{z}_{0,N}\}, \quad \mu_j \in \boldsymbol{\omega}, \end{aligned} \tag{3.2.112}$$

$$\begin{aligned} \frac{\partial^2 U_{\text{inter}}}{\partial \mu_i \partial \nu_j} =& \nabla_{\mu_i} \nabla_{\nu_j} \boldsymbol{r}^T \nabla_{\boldsymbol{r}} U_{\text{inter}} + \nabla_{\mu_i} \boldsymbol{r}^T \nabla^2_{\boldsymbol{r}} U_{\text{inter}} \nabla_{\nu_j} \boldsymbol{r} \\ &+ \nabla_{\mu_i} \boldsymbol{r}^T \nabla_{\boldsymbol{r}} \nabla^T_{\boldsymbol{X}} U_{\text{inter}} \nabla_{\nu_j} \boldsymbol{X} + \nabla_{\nu_j} \boldsymbol{r}^T \nabla_{\boldsymbol{r}} \nabla_{\mu_i} U_{\text{inter}} + \nabla_{\nu_j} \boldsymbol{X}^T \nabla_{\boldsymbol{X}} \nabla_{\mu_i} U_{\text{inter}}, \\ & \mu_i \in \boldsymbol{\omega}, \quad \nu_j \in \boldsymbol{\nu}. \end{aligned} \tag{3.2.113}$$

With that, the first- and second-derivative expressions can be used with anisotropic intermolecular interactions. These formulations are necessary for the evaluation and minimization of intermolecular energies described by multipole electrostatic models.

3.2.1.5 Truncation of Infinite-Sums

The formulations for intermolecular energies/gradients formally involve *infinite-sums*, in either the real- or reciprocal-space (except U_{corr}). Fortunately, all of the intermolecular contributions either decay quickly with interatomic distance (e.g., U_{FF}), or have been subject to appropriate manipulation (i.e., the Ewald sum) to incite rapid decay in the real- (e.g., U_{real}) and reciprocal-space (e.g., U_{rec}). As such, each infinite-sum can be truncated at an appropriate distance (or $\|\boldsymbol{k}\|$

for the reciprocal sum) without introducing significant errors in the summation result. In effect, when evaluating an interaction function, the (pairwise) interaction energy ($U^{\mathrm{FF}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$, $U^{\mathrm{real}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$, or $U^{\mathrm{rec}}_{\boldsymbol{l}_k}$) is added to the total sum only if the participating interaction sites are within the relevant cutoffs for that contribution. In this work, $U^{\mathrm{FF}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ is subject to radial distance cutoffs in real-space. Assuming predominantly van der Waals-type interactions, appropriate cutoff distances are typically between 15–25 Å [2]; interactions beyond this range are usually negligible for most systems. The electrostatic term is slightly less straightforward. $U^{\mathrm{real}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ is also subject to radial distance cutoffs in real-space, while $U^{\mathrm{rec}}_{\boldsymbol{l}_k}$ decays with increasing magnitude of the reciprocal lattice vector and has a cutoff defined in terms of $\|\boldsymbol{k}\|$. $U^{\mathrm{real}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ and $U^{\mathrm{rec}}_{\boldsymbol{l}_k}$ synergistically determine U_{elec}, hence selection of appropriate values for their respective cutoffs must consider how U_{elec} has been partitioned between the two contributions, based on $\alpha_{l_1+l_2}$. Bowskill [2] has outlined a methodology to determine $\alpha_{l_1+l_2}$ and the reciprocal-space cutoff from a given real-space cutoff and target truncation error. This approach has been adopted in this work.

As the crystal geometry is altered, truncation of the infinite-sums can lead to discontinuous behavior in the evaluated intermolecular energy. This occurs when participants of the intermolecular energy function 'wander' in/out of the cutoff limits, leading to non-smooth changes in the energy contribution. This is pertinent to energy minimization, since a discontinuous energy function can lead to poor numerical behavior during optimization. Instead, Bowskill [2] proposes using quintic polynomial splines to ensure each interaction decays smoothly to 0 when approaching its cutoff limit. This is an effective means to mitigate effects of transient-participation near the cutoff thresholds. The formulation of this spline function, and its derivatives, have been described by Bowskill [2] and is implemented in our group's programs accordingly.

Before concluding this section, I wish to clarify the language around the terms in the Ewald sum. In my view, reference to the real- and reciprocal-space contributions as 'short-range' and 'long-range' contributions (*cf.* Refs [2, 71]) is misleading. As mentioned before, the Ewald sum is just a mathematical 'trick', and should not be confused for a physically-meaningful partitioning of electrostatic contributions. Whilst it is practical to use relatively 'short-range' real-space cutoffs (~10–20 Å), there is no restriction that precludes selecting a 'long-range' real-space cutoff. Moreover, the reciprocal term lacks any strict delineation in terms of interaction distance since it is (in effect) a periodic evaluation over the *entire* lattice. As such, a portion of interaction energy within the real-space cutoff distance is in fact encompassed in U_{rec}. This is exemplified by the need for U_{corr}, which (by definition) is a '*short-range*' interaction. Overall, one should avoid relating individual Ewald sum terms to distance-dependent physical interactions.

Altogether, the mathematical framework in Sections 3.2.1.1–3.2.1.5 enables effective evaluation and minimization of the intermolecular energy for fixed molecular conformations.

3.2.2 Intramolecular Interactions and Flexible CSP

The intramolecular contribution is an equally important part of the HAIEFF energy formulation. In principle, this contribution is easy to compute as the difference in conformational energies

between the crystalline conformation ($U_{\text{intra}}^{\text{crys}}$) and the gas-phase reference conformation ($U_{\text{intra}}^{\text{gp}}$) (Equation 3.2.2). In the HAIEFF framework, these energies are simply determined by any appropriate QM method (e.g., DFT, post–HF methods). Note that determining the 'true' value of $U_{\text{intra}}^{\text{gp}}$ is important if *absolute* values of U_{latt} were of interest. This can be challenging for flexible molecules which can have many locally optimum conformations. However, if only *relative* values of U_{latt} are important (e.g., to rank polymorphic stability), $U_{\text{intra}}^{\text{gp}}$ acts as a fixed reference energy common to all predicted structures, hence its value is somewhat inconsequential.

Straightforward evaluation of ΔU_{intra} by two discrete conformational energy assessments is adequate if molecular flexibility is not varied continuously in the lattice energy minimization. However, if it is desired to treat the conformation as a continuous decision variable during the lattice energy minimization (and candidate generation), this introduces a few problems. Firstly, the previous (naïve) approach would no longer be practical since it demands re-assessment of $U_{\text{intra}}^{\text{crys}}$ at every $\boldsymbol{\theta}$. This potentially translates to a huge number of QM computations for a single lattice energy minimization. Moreover, assessment of U_{elec} requires accurate values of Q_{Ii}^{lm}, which reflect the localized regions of charge density that emerge from a molecule's electronic structure. As the molecular conformation changes, its electronic structure is also altered and by extension the values of Q_{Ii}^{lm} must respond to $\boldsymbol{\theta}$. Thus, additional effort must also be expended to re-evaluate Q_{Ii}^{lm} at every $\boldsymbol{\theta}$. While derivation of Q_{Ii}^{lm} (Section 4.1.4) is generally less expensive than the QM calculations for determining $U_{\text{intra}}^{\text{crys}}$, it can still add considerable cost to a minimization if performed at every $\boldsymbol{\theta}$.

To rigorously address molecular flexibility as part of the HAIEFF model, our group adopts the Local Approximate Model (LAM) approach. This framework was proposed by Kazantsev [95], and later incorporated into the various programs developed by our group [96–98]. In the rest of this section, I describe the core tenants of the LAM methodology described in Refs [95, 98], while program-specific details are discussed in Chapter 4. Henceforth, when referring to the conformational energy of a molecule in a crystal, I will use $U_{\text{intra}}^{\text{crys}}$ and U_{intra} interchangeably.

3.2.2.1 Defining Molecular Flexibility

The conformational variables in $\boldsymbol{\theta}$ are given by a Zmatrix describing the molecular conformation. A molecule I, containing η_I atoms, requires minimally $3\eta_I - 6$ conformational degrees of freedom (cdofs) to have a fully defined conformation ($\eta_I - 1$ bond lengths, $\eta_I - 2$ bond angles, and $\eta_I - 3$ torsion angles). Any deviation in $\boldsymbol{\theta}$ from the gas-phase conformation ($\boldsymbol{\theta}^{\text{gp}}$) incurs an energy penalty, i.e., $\Delta U_{\text{intra}} > 0$. This is only thermodynamically favorable if it drives a larger enhancement of U_{inter} that lowers U_{latt} as a whole. Considering the strength of intermolecular interactions in organic molecular crystals, this places an intuitive limit to ΔU_{intra} [75]. Consequently, only a subset of $\boldsymbol{\theta}$ may deviate significantly from their value in $\boldsymbol{\theta}^{\text{gp}}$.

With this knowledge, $\boldsymbol{\theta}$ can be partitioned as

$$\boldsymbol{\theta} = \boldsymbol{\theta}^{\text{ind}} \cup \boldsymbol{\theta}^{\text{dep}}, \tag{3.2.114}$$

where $\boldsymbol{\theta}^{\text{ind}}$ is the subset of independent cdofs, while $\boldsymbol{\theta}^{\text{dep}}$ is the subset of dependent cdofs. The former are dofs that can vary without incurring an unreasonably large energy penalty, whilst the latter cannot. As such, $\boldsymbol{\theta}^{\text{dep}}$ should exhibit only minor adjustments that minimize the conformational energy for a given $\boldsymbol{\theta}^{\text{ind}}$. In other words:

$$\begin{aligned} U_{\text{intra}}(\boldsymbol{\theta}) =& U_{\text{intra}}(\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{\theta}^{\text{dep}}) \\ =& U_{\text{intra}}(\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{\theta}^{\text{dep}}(\boldsymbol{\theta}^{\text{ind}})) \\ =& \min_{\boldsymbol{\theta}^{\text{dep}}} U_{\text{intra}}(\boldsymbol{\theta}^{\text{dep}}; \boldsymbol{\theta}^{\text{ind}}). \end{aligned} \tag{3.2.115}$$

It is important to stress that while the values of $\boldsymbol{\theta}^{\text{dep}}$ are not expected to change much, in this framework they are not truly 'rigid', but are instead functions of $\boldsymbol{\theta}^{\text{ind}}$:

$$\boldsymbol{\theta}^{\text{dep}}(\boldsymbol{\theta}^{\text{ind}}) = \arg\min_{\boldsymbol{\theta}^{\text{dep}}} U_{\text{intra}}(\boldsymbol{\theta}^{\text{dep}}; \boldsymbol{\theta}^{\text{ind}}). \tag{3.2.116}$$

As such, the 'flexible'/'rigid' labels are substituted by the more-precise 'independent'/'dependent' terminology. Often, chemical intuition is employed to partition these groups. $\boldsymbol{\theta}^{\text{ind}}$ is expected to contain torsions around single bonds, which can rotate without significant restriction, and some bond angles. Conversely, $\boldsymbol{\theta}^{\text{dep}}$ should comprise most bond lengths, bond angles, and some restricted torsions (e.g., torsions in ring systems and delocalized π-networks). Identification of $\boldsymbol{\theta}^{\text{ind}}$ can also be supported by scans of the molecular conformation's potential energy surface, although this must be done sparingly to limit cost. In my experience, a critical distinction between $\boldsymbol{\theta}^{\text{ind}}$ and $\boldsymbol{\theta}^{\text{dep}}$, is that $\boldsymbol{\theta}^{\text{dep}}$ should be *uniquely* defined by $\boldsymbol{\theta}^{\text{ind}}$. This is a subtle trait that is difficult to perceive, but becomes important deeper in the LAM framework.

Ultimately, effective partitioning of $\boldsymbol{\theta}$ is central to efficient descriptions of molecular flexibility since it lowers the dimensionality of the flexible problem. This benefits both the construction of the LAM model (explained below) as well as the lattice energy minimization. In the latter, only the $\boldsymbol{\theta}^{\text{ind}}$ are treated as optimization decision variables, reducing problem complexity. Dudek & Drużbicki [77] class this as a model with 'partial molecular flexibility'.

3.2.2.2 The LAM Surrogate Model

The central problem that the LAM methodology seeks to address is: given a set of $\boldsymbol{\theta}^{\text{ind}}$, all other quantities related to the molecular conformation ($\boldsymbol{\theta}^{\text{dep}}$, U_{intra}, and $\boldsymbol{Q}$) must be defined to evaluate U_{latt}. More precisely, we seek a *surrogate* model of these dependent quantities in terms of $\boldsymbol{\theta}^{\text{ind}}$, such that $\boldsymbol{\theta}^{\text{ind}}$ can be explored without continuous need for QM calculations, i.e.,

$$\begin{aligned} \boldsymbol{\theta}^{\text{dep}} =& \boldsymbol{\theta}^{\text{dep}}(\boldsymbol{\theta}^{\text{ind}}), \\ U_{\text{intra}}(\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{\theta}^{\text{dep}}) =& U_{\text{intra}}(\boldsymbol{\theta}^{\text{ind}}), \\ \boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{\theta}^{\text{dep}}) =& \boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}}). \end{aligned} \tag{3.2.117}$$

The LAM approach constructs such a surrogate model using local information about the energy surface's topology (i.e., gradient information), provided by QM calculations. From a QM cal-

culation performed at $\boldsymbol{\theta}_{\mathrm{LAM}}$, where $\boldsymbol{\theta}_{\mathrm{LAM}} = \boldsymbol{\theta}^{\mathrm{ind}}_{\mathrm{LAM}} \cup \boldsymbol{\theta}^{\mathrm{dep}}_{\mathrm{LAM}}(\boldsymbol{\theta}^{\mathrm{ind}}_{\mathrm{LAM}})$, the intramolecular energy at $(\boldsymbol{\theta}^{\mathrm{ind}}, \boldsymbol{\theta}^{\mathrm{dep}})$ can be approximated by a second-order Taylor expansion,

$$\begin{aligned} U_{\mathrm{intra}}(\boldsymbol{\theta}^{\mathrm{ind}}, \boldsymbol{\theta}^{\mathrm{dep}}) = U_{\mathrm{intra}}(\boldsymbol{\theta}_{\mathrm{LAM}}) \\ + (\nabla^T_{\boldsymbol{\theta}^{\mathrm{ind}}} U_{\mathrm{intra}})_{\boldsymbol{\theta}_{\mathrm{LAM}}}(\boldsymbol{\theta}^{\mathrm{ind}} - \boldsymbol{\theta}^{\mathrm{ind}}_{\mathrm{LAM}}) + (\nabla^T_{\boldsymbol{\theta}^{\mathrm{dep}}} U_{\mathrm{intra}})_{\boldsymbol{\theta}_{\mathrm{LAM}}}(\boldsymbol{\theta}^{\mathrm{dep}} - \boldsymbol{\theta}^{\mathrm{dep}}_{\mathrm{LAM}}) \\ + \frac{1}{2}(\boldsymbol{\theta}^{\mathrm{ind}} - \boldsymbol{\theta}^{\mathrm{ind}}_{\mathrm{LAM}})^T (\nabla^2_{\boldsymbol{\theta}^{\mathrm{ind}}} U_{\mathrm{intra}})_{\boldsymbol{\theta}_{\mathrm{LAM}}}(\boldsymbol{\theta}^{\mathrm{ind}} - \boldsymbol{\theta}^{\mathrm{ind}}_{\mathrm{LAM}}) \\ + \frac{1}{2}(\boldsymbol{\theta}^{\mathrm{dep}} - \boldsymbol{\theta}^{\mathrm{dep}}_{\mathrm{LAM}})^T (\nabla^2_{\boldsymbol{\theta}^{\mathrm{dep}}} U_{\mathrm{intra}})_{\boldsymbol{\theta}_{\mathrm{LAM}}}(\boldsymbol{\theta}^{\mathrm{dep}} - \boldsymbol{\theta}^{\mathrm{dep}}_{\mathrm{LAM}}) \\ + (\boldsymbol{\theta}^{\mathrm{dep}} - \boldsymbol{\theta}^{\mathrm{dep}}_{\mathrm{LAM}})^T (\nabla_{\boldsymbol{\theta}^{\mathrm{ind}}} \nabla_{\boldsymbol{\theta}^{\mathrm{dep}}} U_{\mathrm{intra}})_{\boldsymbol{\theta}_{\mathrm{LAM}}}(\boldsymbol{\theta}^{\mathrm{ind}} - \boldsymbol{\theta}^{\mathrm{ind}}_{\mathrm{LAM}}). \end{aligned} \tag{3.2.118}$$

Note that all derivative terms are evaluated at the LAM point (indicated by the subscript '$\boldsymbol{\theta}_{\mathrm{LAM}}$'). Since the QM calculation at the LAM is formulated according to Equations 3.2.115 and 3.2.116, all first-derivatives with respect to $\boldsymbol{\theta}^{\mathrm{dep}}$ must satisfy the optimality criterion:

$$(\nabla_{\boldsymbol{\theta}^{\mathrm{dep}}} U_{\mathrm{intra}})_{\boldsymbol{\theta}_{\mathrm{LAM}}} = \mathbf{0}. \tag{3.2.119}$$

Using a first-order Taylor expansion, the corresponding approximation of $\boldsymbol{\theta}^{\mathrm{dep}}$ is given by

$$\boldsymbol{\theta}^{\mathrm{dep}}(\boldsymbol{\theta}^{\mathrm{ind}}) = \boldsymbol{\theta}^{\mathrm{dep}}_{\mathrm{LAM}} + (\nabla_{\boldsymbol{\theta}^{\mathrm{ind}}} \boldsymbol{\theta}^{\mathrm{dep}})_{\boldsymbol{\theta}_{\mathrm{LAM}}}(\boldsymbol{\theta}^{\mathrm{ind}} - \boldsymbol{\theta}^{\mathrm{ind}}_{\mathrm{LAM}}). \tag{3.2.120}$$

It is desirable to express these equations wholly in terms of $\boldsymbol{\theta}^{\mathrm{ind}}$ and their gradients with respect to U_{intra}. To that end, we enforce the definition of $\boldsymbol{\theta}^{\mathrm{dep}}$ as a response to $\boldsymbol{\theta}^{\mathrm{ind}}$ that *minimizes* U_{intra}. Accordingly, the optimality criterion shown in Equation 3.2.119 must hold at the values of $\boldsymbol{\theta}^{\mathrm{dep}}$ predicted by the LAM. In short, for a perturbation $(\delta\boldsymbol{\theta})$ from $\boldsymbol{\theta}_{\mathrm{LAM}}$,

$$(\nabla_{\boldsymbol{\theta}^{\mathrm{dep}}} U_{\mathrm{intra}})_{\boldsymbol{\theta}_{\mathrm{LAM}} + \delta\boldsymbol{\theta}} = \mathbf{0}. \tag{3.2.121}$$

Taking a first-order Taylor expansion of $\nabla_{\boldsymbol{\theta}^{\mathrm{dep}}} U_{\mathrm{intra}}$ at the LAM point, one obtains

$$\begin{aligned} \nabla_{\boldsymbol{\theta}^{\mathrm{dep}}} U_{\mathrm{intra}}(\boldsymbol{\theta}^{\mathrm{ind}}, \boldsymbol{\theta}^{\mathrm{dep}}) = (\nabla_{\boldsymbol{\theta}^{\mathrm{dep}}} U_{\mathrm{intra}})_{\boldsymbol{\theta}_{\mathrm{LAM}}} \\ + (\nabla^2_{\boldsymbol{\theta}^{\mathrm{dep}}} U_{\mathrm{intra}})_{\boldsymbol{\theta}_{\mathrm{LAM}}}(\boldsymbol{\theta}^{\mathrm{dep}} - \boldsymbol{\theta}^{\mathrm{dep}}_{\mathrm{LAM}}) \\ + (\nabla_{\boldsymbol{\theta}^{\mathrm{ind}}} \nabla_{\boldsymbol{\theta}^{\mathrm{dep}}} U_{\mathrm{intra}})_{\boldsymbol{\theta}_{\mathrm{LAM}}}(\boldsymbol{\theta}^{\mathrm{ind}} - \boldsymbol{\theta}^{\mathrm{ind}}_{\mathrm{LAM}}). \end{aligned} \tag{3.2.122}$$

Applying Equations 3.2.119 and 3.2.121, Equation 3.2.122 can be rearranged to give

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{\theta}_{\mathrm{LAM}}) = (\nabla_{\boldsymbol{\theta}^{\mathrm{ind}}} \boldsymbol{\theta}^{\mathrm{dep}})_{\boldsymbol{\theta}_{\mathrm{LAM}}} \\ = -(\nabla^2_{\boldsymbol{\theta}^{\mathrm{dep}}} U_{\mathrm{intra}})^{-1}_{\boldsymbol{\theta}_{\mathrm{LAM}}} (\nabla_{\boldsymbol{\theta}^{\mathrm{ind}}} \nabla_{\boldsymbol{\theta}^{\mathrm{dep}}} U_{\mathrm{intra}})^T_{\boldsymbol{\theta}_{\mathrm{LAM}}}. \end{aligned} \tag{3.2.123}$$

Combining Equations 3.2.118–3.2.120, and 3.2.123, the surrogate models for U_{intra} and $\boldsymbol{\theta}^{\mathrm{dep}}$ can be re-expressed as

$$\boldsymbol{\theta}^{\mathrm{dep}}(\boldsymbol{\theta}^{\mathrm{ind}}) = \boldsymbol{\theta}^{\mathrm{dep}}_{\mathrm{LAM}} + \boldsymbol{A}(\boldsymbol{\theta}_{\mathrm{LAM}})(\boldsymbol{\theta}^{\mathrm{ind}} - \boldsymbol{\theta}^{\mathrm{ind}}_{\mathrm{LAM}}), \tag{3.2.124}$$

and

$$\begin{aligned} U_{\text{intra}}(\boldsymbol{\theta}^{\text{ind}}) =& U_{\text{intra}}(\boldsymbol{\theta}_{\text{LAM}}) + [\boldsymbol{b}(\boldsymbol{\theta}_{\text{LAM}})]^T(\boldsymbol{\theta}^{\text{ind}} - \boldsymbol{\theta}^{\text{ind}}_{\text{LAM}}) \\ &+ \frac{1}{2}(\boldsymbol{\theta}^{\text{ind}} - \boldsymbol{\theta}^{\text{ind}}_{\text{LAM}})^T\boldsymbol{C}(\boldsymbol{\theta}_{\text{LAM}})(\boldsymbol{\theta}^{\text{ind}} - \boldsymbol{\theta}^{\text{ind}}_{\text{LAM}}), \end{aligned} \tag{3.2.125}$$

where

$$\boldsymbol{b}(\boldsymbol{\theta}_{\text{LAM}}) = (\nabla_{\boldsymbol{\theta}^{\text{ind}}} U_{\text{intra}})_{\boldsymbol{\theta}_{\text{LAM}}}, \tag{3.2.126}$$

$$\begin{aligned} \boldsymbol{C}(\boldsymbol{\theta}_{\text{LAM}}) =& (\nabla^2_{\boldsymbol{\theta}^{\text{ind}}} U_{\text{intra}})_{\boldsymbol{\theta}_{\text{LAM}}} \\ &- (\nabla_{\boldsymbol{\theta}^{\text{ind}}}\nabla_{\boldsymbol{\theta}^{\text{dep}}} U_{\text{intra}})_{\boldsymbol{\theta}_{\text{LAM}}}(\nabla^2_{\boldsymbol{\theta}^{\text{dep}}} U_{\text{intra}})^{-1}_{\boldsymbol{\theta}_{\text{LAM}}}(\nabla_{\boldsymbol{\theta}^{\text{ind}}}\nabla_{\boldsymbol{\theta}^{\text{dep}}} U_{\text{intra}})^T_{\boldsymbol{\theta}_{\text{LAM}}}. \end{aligned} \tag{3.2.127}$$

To summarize, I have derived expressions that enable the approximation of $U_{\text{intra}}(\boldsymbol{\theta}^{\text{ind}})$ and $\boldsymbol{\theta}^{\text{dep}}(\boldsymbol{\theta}^{\text{ind}})$ (respectively) as second- and first-order Taylor expansions of $\boldsymbol{\theta}^{\text{ind}}$, around a given LAM/QM calculation point.

The ESP and AIM methods used to determine $\boldsymbol{Q}$ (point-charges or multipoles) do not provide gradient information relating $\boldsymbol{Q}$ to either U_{intra} or $\boldsymbol{\theta}$. To build an analogous Taylor expansion would necessitate using finite-differences to derive this information [95]. In lieu of this, it is assumed that in the vicinity of a LAM point, the conformational-dependence of the multipole moments can be captured by simply rotating $\boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}}_{\text{LAM}})$ about their local environments to match the predicted conformation at $(\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{\theta}^{\text{dep}}(\boldsymbol{\theta}^{\text{ind}}))$. In short,

$$\boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{\theta}^{\text{dep}}(\boldsymbol{\theta}^{\text{ind}})) = \mathcal{R}[\boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}}_{\text{LAM}}, \boldsymbol{\theta}^{\text{dep}}_{\text{LAM}}(\boldsymbol{\theta}^{\text{ind}}_{\text{LAM}}))], \tag{3.2.128}$$

where $\mathcal{R}$ is the rotation operator. For isotropic atomic-charges, the predicted atomic-charge values are unchanged from their values at the LAM point.

3.2.2.3 Overcoming LAM Local-Validity

From a QM/LAM sampling point, the basic LAM framework constructs surrogate models which provide estimates of $\boldsymbol{\theta}^{\text{dep}}$, U_{intra}, and $\boldsymbol{Q}$ as functions of $\boldsymbol{\theta}^{\text{ind}}$. However, as their name suggests, the LAM model is a *local approximation* valid only in the vicinity of the LAM point. As $\boldsymbol{\theta}^{\text{ind}}$ deviates further from its reference LAM, the gradient-information from $\boldsymbol{\theta}_{\text{LAM}}$ becomes increasingly divorced from the current conformation, precluding accurate predictions of $\boldsymbol{\theta}^{\text{dep}}$, U_{intra}, and $\boldsymbol{Q}$. Across a landscape of crystal structures, a myriad of crystalline conformations will be explored and to describe this breadth of molecular conformations, a single LAM point cannot suffice. A solution to this is to have multiple QM/LAM sampling points probing different parts of the conformational landscape (as defined by $\boldsymbol{\theta}^{\text{ind}}$). The surrogate model used to describe any conformation can be chosen as the nearest LAM to $\boldsymbol{\theta}^{\text{ind}}$, such that each LAM represents only the conformations in its locality. This nearest-LAM approach was adopted in early implementations of the framework [97]. However, 'jumping' between LAM reference points as $\boldsymbol{\theta}^{\text{ind}}$ varies can

introduce discontinuities into the values of $\boldsymbol{\theta}^{\text{dep}}$, U_{intra}, and $\boldsymbol{Q}$. This is detrimental to gradient-based optimization, especially if a lattice energy minimum lies near the boundary between two LAMs. To address this, Sugden *et al.* [98] proposed the concept of smoothed LAMs.

To illustrate this approach, consider a system where K LAM sampling points have been generated by QM calculations. The methodology to decide the placement of these LAM points will not be discussed here, since it is program-specific (see Chapter 4). The details of LAM placement are also irrelevant to the smoothing algorithm. For a conformation defined by $\boldsymbol{\theta}^{\text{ind}}$, each LAM in this collection can provide estimates for $\boldsymbol{\theta}^{\text{dep}}$, U_{intra}, and $\boldsymbol{Q}$ per Equations 3.2.124, 3.2.125, and 3.2.128 respectively. I denote these as

$$X_k = X_k(\boldsymbol{\theta}^{\text{ind}}), \qquad X \in \{\boldsymbol{\theta}^{\text{dep}}, U_{\text{intra}}, \boldsymbol{Q}\}, \quad k = 1, 2, \cdots, K. \tag{3.2.129}$$

Rather than using only a *single* LAM, the smoothed LAM framework proposes employing a weighted-average (denoted by the subscript 'wa') of *all* LAMs within the collection:

$$X_{\text{wa}}(\boldsymbol{\theta}^{\text{ind}}) = \frac{\sum_{k=1}^{K} w_k(\boldsymbol{\theta}^{\text{ind}}) X_k(\boldsymbol{\theta}^{\text{ind}})}{\sum_{k=1}^{K} w_k(\boldsymbol{\theta}^{\text{ind}})}, \qquad X \in \{\boldsymbol{\theta}^{\text{dep}}, U_{\text{intra}}, \boldsymbol{Q}\}, \quad k = 1, 2, \cdots, K, \tag{3.2.130}$$

where

$$w_k(\boldsymbol{\theta}^{\text{ind}}) = \exp\left(-\frac{\|\boldsymbol{\theta}^{\text{ind}} - \boldsymbol{\theta}^{\text{ind}}_{\text{LAM},k}\|^2}{s^2}\right), \qquad \forall\ k = 1, 2, \cdots, K. \tag{3.2.131}$$

$\|\boldsymbol{\theta}^{\text{ind}} - \boldsymbol{\theta}^{\text{ind}}_{\text{LAM},k}\|^2$ is the square of the Euclidean distance between the two sets of independent cdof, while s is a user-specified, positive-valued smoothing parameter. In effect, the above formulation weights the contribution of each LAM to X_{wa} based on the distance from the current conformation ($\boldsymbol{\theta}^{\text{ind}}$) to the LAM conformation ($\boldsymbol{\theta}^{\text{ind}}_{\text{LAM},k}$). This contribution decays exponentially with distance, respecting the local-validity of the LAM. The smaller the value of s, the faster the rate of this decay. In practice, values of s ~0.2–0.4 lead to negligible contributions to the weighted-average from all but the nearest LAM points.

From here, corresponding gradient expressions can be derived as

$$\frac{\partial X_{\text{wa}}}{\partial \theta^{\text{ind}}}(\boldsymbol{\theta}^{\text{ind}}) = \frac{\sum_{k=1}^{K} \left\{ w_k(\boldsymbol{\theta}^{\text{ind}}) \frac{\partial X_k}{\partial \theta^{\text{ind}}}(\boldsymbol{\theta}^{\text{ind}}) + \left[X_k(\boldsymbol{\theta}^{\text{ind}}) - X_{\text{wa}}(\boldsymbol{\theta}^{\text{ind}}) \right] \frac{\partial w_k}{\partial \theta^{\text{ind}}}(\boldsymbol{\theta}^{\text{ind}}) \right\}}{\sum_{k=1}^{K} w_k(\boldsymbol{\theta}^{\text{ind}})}, \tag{3.2.132}$$
$$X \in \{\boldsymbol{\theta}^{\text{dep}}, U_{\text{intra}}, \boldsymbol{Q}\}, \quad k = 1, 2, \cdots, K,$$

where

$$\frac{\partial w_k}{\partial \theta^{\text{ind}}}(\boldsymbol{\theta}^{\text{ind}}) = \frac{2 w_k(\boldsymbol{\theta}^{\text{ind}})}{s^2}(\theta^{\text{ind}} - \theta^{\text{ind}}_{\text{LAM}}), \qquad \forall\ k = 1, 2, \cdots, K. \tag{3.2.133}$$

Expressions for $\frac{\partial X_k}{\partial \theta^{\text{ind}}}$ can be obtained by differentiation of the relevant equations for the LAM-predicted quantities (Equations 3.2.124, 3.2.125, and 3.2.128).

Overall, the LAM framework enables efficient and accurate incorporation of molecular flexibility into HAIEFF models, suitable for lattice energy minimization (and candidate generation). The basic concept of a LAM enables *continuous* descriptions of conformation-dependent quantities like $\boldsymbol{\theta}^{\text{dep}}$, U_{intra}, and $\boldsymbol{Q}$ (at near QM-accuracy) without the incessant need for QM calculation. The local-validity of each LAM can be overcome by adequate sampling about the $\boldsymbol{\theta}^{\text{ind}}$ space, extending the LAM model over a range of molecular conformations. Finally, undesirable numerical behavior associated with using multiple LAMs can be suppressed by smoothing. An important feature of LAMs is once generated, they can be stored and re-used. Over a landscape containing crystal structures that explore similar 'pockets' of the conformational space, QM information can be shared further reducing cost.

With that said, there are some shortcomings to the LAM framework. Firstly, the number of LAMs required for an effective description of molecular flexibility scales steeply with the dimensionality of $\boldsymbol{\theta}^{\text{ind}}$, hence extremely flexible molecules ($\dim(\boldsymbol{\theta}^{\text{ind}}) \geq 7$ or 8) remain prohibitive under the current LAM framework (*cf.* Molecule XXXII from the Seventh Blind Test [49]). Secondly, the appropriate selection of $\boldsymbol{\theta}^{\text{ind}}$ can be non-trivial. If the value of θ^{dep} is not uniquely determined by $\boldsymbol{\theta}^{\text{ind}}$ (i.e., θ^{dep} has multiple distinct local-minima associated with Equation 3.2.116), then θ^{dep} should be included in $\boldsymbol{\theta}^{\text{ind}}$. Failing to do this can, in the worst case, result in spurious numerical behavior (even with smoothing) when LAMs in close proximity predict extremely distinct values of θ^{dep}. Unfortunately, it may not always be obvious when a cdof might exhibit this behavior, even from chemical intuition and conformational scans. In my experience, 'warning signs' include hydrogens that can form intramolecular hydrogen bonds in specific conformations, and strained rings that can exhibit puckering. When in doubt, it would be safest to include a cdof in $\boldsymbol{\theta}^{\text{ind}}$, but this adds to the dimensionality problem, leading to a catch-22 situation. Exploring remedies for these issues is beyond the scope of this work.

3.3 Summary

In this chapter I have laid out the mathematical foundations that underpin all calculations conducted in this work. I began by formulating the mathematical description of the crystal geometry, which then enables the rigorous computation of crystal lattice energies. Where necessary, analytical expressions for various derivative quantities were also derived. This facilitates more-efficient execution of the lattice energy minimization in our group's CSP codes. The HAIEFF framework explained in this chapter encompasses all levels of model fidelity that might be used in a real CSP study. This includes electrostatic models based on atomic-charges and multipoles, empirical FFs based on two popular pair potentials, and models that include continuous descriptions of molecular flexibility. In Chapter 4, I will outline the algorithmic implementation of these formalisms in the CSP programs that our research group develop.

Chapter 4

The Imperial MSE CSP Suite

In the preceding chapter, the mathematical formalism behind HAIEFF calculations was detailed. To apply this in practice, several CSP tools have been developed in our research group. These programs target distinct stages of the CSP workflow: Candidate Generation → Global Search → Refinement. In this chapter, I first outline some external packages/libraries that support the programs used by our group. Thereafter, I will introduce the collection of CSP programs developed in our group, which I have denoted the 'Imperial Molecular Systems Engineering (MSE) CSP Suite'. Separate to this primary set of programs, I will also discuss our auxiliary program, CrystalEstimator, which is specifically employed for HAIEFF FF parametrization. Over the course of this work, several augmentations have been made to these programs in support of the intended research goals. Here, I describe the stable versions of these codes, while other relevant developments will be discussed in subsequent chapters.

4.1 Supporting External Programs/Libraries

I begin by briefly introducing the supporting external packages and libraries which interface with our codes to support CSP calculations.

4.1.1 NAG Libraries

The Numerical Algorithms Group (NAG) [239, 240] provides Fortran subroutine libraries for performing various numerical operations. In our CSP codes, lattice energy minimizations employ the NAG E04UFF subroutine, which solves constrained non-linear optimizations of the form

$$\min_{\boldsymbol{x}\in\mathbb{R}} F(\boldsymbol{x}), \tag{4.1.1}$$

where $\boldsymbol{x}$ is the decision variable vector for the objective function $F(\boldsymbol{x})$. For HAIEFF FF parameter estimation, the NAG E04USA subroutine is employed, which solves constrained non-linear least-squares problems given by

$$\min_{\boldsymbol{x}\in\mathbb{R}} F(\boldsymbol{x}) = \frac{1}{2}\sum_i [y_i - f_i(\boldsymbol{x})]^2, \tag{4.1.2}$$

where y_i and $f_i(\boldsymbol{x})$ are respectively the reference and predicted value at the decision variable vector, $\boldsymbol{x}$, of the ith residual in the parameter estimation. To solve these optimizations, both subroutines employ gradient-based, quasi-Newton optimization techniques based on the Broyden–Fletcher–Goldfarb–Shanno (BFGS) approximation of the Hessian. Analytical first-derivatives, like those derived in Sections 3.2.1.3 and 3.2.1.4, are provided to the subroutines. Accurate gradients are crucial for effective minimization with these subroutines. In support of this, each subroutine has an in-built utility for validating user-provided gradients against finite-difference gradients. This utility has been employed in various parts of this work to verify that any implemented gradients are indeed correct. The NAG optimization subroutines consider a combination of convergence criteria, including: first-order Karush–Kuhn–Tucker (KKT) criteria (i.e., stationarity), step size, and constraint violation.

In this work, both Mark 27 and Mark 29 NAG libraries have been used across the different CSP codes. This mismatch is due to changes in our compute cluster's licensing. To my knowledge, there are no changes to either subroutine between the two releases.

4.1.2 CCDC Suite

The CCDC lead the development of several supporting tools that are invaluable for CSP. CCDC-developed programs like Mercury [237] offer utility for visualizing, analyzing, comparing, and manipulating crystal structures. Many of these functions have also been integrated into the CCDC Python API [241], facilitating more-convenient utilization from the command line. Both Mercury and the Python API have been used throughout this work.

An important utility within Mercury and the Python API is the CCDC's implementation of the COMPACK algorithm [242], which forms the basis for quantifying crystal similarity in this work. Given two crystal structures, the COMPACK algorithm attempts to generate a coordination sphere of N molecules that overlap between the two crystals, comprising a central molecule and $N-1$ neighbors. The limits within which molecules are considered to be 'overlapping' is defined by distance and angle tolerances. For N overlapping molecules, the algorithm optimizes the coordination sphere overlap by minimizing the root–mean–square–deviation in atomic positions over N molecules (RMSD_N). Generally, lower RMSD_N scores indicate greater similarity (less deviation) between two crystal structures. A few practical caveats are noted. Firstly, when comparing CSP-generated structures against experimental structures, it is typical to ignore hydrogen positions. This is because experimental determination of hydrogen positions tends to be less reliable than for heavy atom positions. Secondly, larger target values of N represent a more-stringent geometric test, since the overlap cluster probes a greater extent of the crystal packing. However, given two COMPACK comparisons yielding N and N' overlapping molecules relative to a reference, where $N > N'$, it is inconclusive that the structure with an N-molecule overlap is more similar to the reference. In some cases, loosening the COMPACK overlap tolerances to permit the previous $\text{RMSD}_{N'}$ score to become RMSD_N can reveal the opposite. In short, comparisons of RMSD_N are only meaningful when both scores have the same N. As noted by

Mayo *et al.* [243], loosening the COMPACK tolerances too much can degrade the algorithm's performance, hence this should be done judiciously.

In this work, unless otherwise stated, CSP-generated structures are considered matches to an experimental reference if they achieve $\text{RMSD}_{20} \leq 0.8$ Å. This aligns with the highest RMSD_N score permitted during the Sixth Blind Test [48]. Instead of the CCDC default settings (20%/20°), I use 40%/40° distance/angle COMPACK tolerances to facilitate generation of RMSD_{20} scores. If necessary, these tolerances may be further increased.

4.1.3 Ab Initio Packages

QM calculations are a central part of the HAIEFF model. These calculations are used to determine: conformational energies, stable molecular geometries, and other molecular electronic properties (e.g., point-charges and multipoles). To that end, the Gaussian quantum chemistry package is used to conduct non-periodic QM computations. Either Gaussian09 (Revision D.01) [227], or Gaussian16 (Revision C.01) [244] (abbreviated as G09 and G16 respectively) is employed. From my testing, G16 tends to be slightly faster than G09, but slight differences in their output file formats may prevent some of our CSP codes from interfacing correctly with G16 (testing this is beyond the scope of the current work). Periodic DFT computations are not required in the HAIEFF framework, but these calculations are an important part of the overarching CSP workflow. Periodic DFT-D remains the gold-standard for CSP, hence it is typical to refine HAIEFF-model results with periodic DFT-D at the end of a CSP study. Moreover, in this work, synthetic crystal data generated by periodic DFT-D is also employed during parametrization of HAIEFF FFs. In service of this, the Vienna Ab initio Simulation Package (VASP) is used to perform periodic QM computations, specifically VASPv5.4.4 [245–249].

4.1.4 Packages for Electrostatic Feature Determination

To evaluate U_{elec}, values for the electrostatic features must be determined. The method for deriving Q_{Ii}^{lm} depends on the desired electrostatic model (point-charge or multipole).

If multipole electrostatics are requested, Stone's GDMA methodology is used [178–181]. The GDMA algorithm considers overlap charge densities represented by the product of two primitive Gaussian wavefunctions (e.g., ϕ_t and ϕ_u). The product of these functions is a Gaussian function centred at $\boldsymbol{P}$, and the multipole associated with this overlap density is given by [180, 221]

$$Q^{lm}(\boldsymbol{P}) = -\int \boldsymbol{R}^{lm}(\boldsymbol{r}-\boldsymbol{P})\rho_{tu}\phi_t\phi_u d^3\boldsymbol{r}. \tag{4.1.3}$$

$\boldsymbol{R}^{lm}$ is the spherical harmonic associated with the mth moment of the rank l multipole, and ρ_{tu} is the corresponding element of the density matrix for ϕ_t and ϕ_u. The multipole expansion at $\boldsymbol{P}$ can be 'moved' to another site (e.g., $\boldsymbol{S}$) according to

$$Q^{l'm'}(\boldsymbol{S}) = \sum_{l=0}^{l'}\sum_{m=-l}^{l}\left[\binom{l'+m'}{l+m}\binom{l'-m'}{l-m}\right]^{\frac{1}{2}} \times Q^{lm}(\boldsymbol{P})\boldsymbol{R}^{(l'-l)(m'-m)}(\boldsymbol{S}-\boldsymbol{P}). \tag{4.1.4}$$

In this way, GDMA distributes multipole moments to atomic sites (i.e., $Q^{lm}(\boldsymbol{P}) \Rightarrow Q_{Ii}^{lm}(\boldsymbol{r}_{Ii})$). The shift of multipoles from one site to another changes the values of the multipole moments [180]. Different distribution schemes have been tested with GDMA. In this work, I use GDMAv2.2, which defaults to a quadrature grid distribution scheme [181]. The 'legacy' GDMA nearest-neighbor distribution scheme can be requested within GDMAv2.2, but this has not be used unless otherwise stated. GDMAv2.3 is freely available, but due to changes in its output formatting, it is inconvenient to interface this version of GDMA with our existing suite of programs. From the GDMAv2.3 change-logs [182], the two versions should not have any meaningful differences in their derived multipoles, so long as the same input settings are used; I have verified this from my own testing. Where relevant, details of the chosen GDMA settings will be discussed in later chapters.

Electrostatic features derived by AIM models prioritize preserving the physical fidelity of their calculated values. However, when the multipole expansion is truncated at low-ranks, e.g., in atomic-charge models, these methods do not have any facility to absorb the truncation error into the remaining multipoles. The severity of this error depends on how quickly the multipole expansion converges for the chosen AIM-method. In the case of GDMA, it is often reported that using GDMA-derived atomic-charges without any higher-ranked multipoles leads to extremely unsatisfactory modeling performance [177, 182, 183]. As such, for CSP calculations involving atomic-charge electrostatic models, I use ESP-fitting methods to determine the charge values. The empirical determination of atomic-charges intrinsically absorbs multipole truncation error into the fitted charge values, albeit at the expense of their physical fidelity.

Several ESP-fitting methods are offered in G09 and G16 [227, 244], but herein I opt to use Gaussian's implementation of the HLY [173] charge fitting method (denoted 'HLYGAt'). From a QM-derived charge density, the molecular ESP at a point $\boldsymbol{R}$ ($V_{\mathrm{QM}}(\boldsymbol{R})$) is defined as

$$V_{\mathrm{QM}}(\boldsymbol{R}) = \int \frac{\rho_{\mathrm{mol}}(\boldsymbol{r})}{|\boldsymbol{r}-\boldsymbol{R}|} d^3\boldsymbol{r}, \tag{4.1.5}$$

where ρ_{mol} is the molecular charge density. The corresponding ESP (at $\boldsymbol{R}$) for a collection of atomic-charges on molecule I is given by

$$V_{\mathrm{pc}}(\boldsymbol{R},\boldsymbol{q}) = \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{\eta_I} \frac{q_{Ii}}{|\boldsymbol{r}_{Ii}-\boldsymbol{R}|}, \tag{4.1.6}$$

where $\boldsymbol{q}$ is the vector of atomic-charge values associated with sites i on molecule I (q_{Ii}), and $\boldsymbol{r}_{Ii}$ is the corresponding position of each site. The HLY method treats the values of q_{Ii} as decision variables in the optimization,

$$\min_{\boldsymbol{q}} \int \omega_{\mathrm{HLY}}(\boldsymbol{R}) \big[V_{\mathrm{pc}}(\boldsymbol{R},\boldsymbol{q}) - V_{\mathrm{QM}}(\boldsymbol{R})\big]^2 d^3\boldsymbol{R}. \tag{4.1.7}$$

This problem attempts to minimize the absolute error between V_{pc} and V_{QM}, over relevant regions of space, by varying $\boldsymbol{q}$. To obtain more-physically meaningful values of $\boldsymbol{q}$, a smooth weighting function ($\omega_{\mathrm{HLY}}(\boldsymbol{R})$) is introduced, which assigns lower weight to regions of space 'near'

and 'far' from the molecule, while prioritizing the 'medium' range. The partitioning of these ranges depends on predefined protoatom electron densities, and in the HLYGAt implementation Gaussian has tabulated its own values for these reference densities. The use of integration in the HLY/HLYGAt charge-fitting algorithm is suggested to give smoother evolution in atomic-charge values as the molecular conformation varies [173], relative to methods using discrete sums in their fitting (e.g., CHelpG [174]). This is beneficial for our group's CSP methods, which often include molecular flexibility during the Global Search stage.

4.2 The Imperial MSE CSP Suite

In this section, I introduce the programs developed by our group for performing CSP with HAIEFF models. In alignment with the multistage CSP workflow, each program is relevant to different stages of the CSP pipeline. All of these programs are underpinned by the mathematical formalism described in Chapter 3. All Imperial MSE CSP programs are written in Fortran90.

4.2.1 CrystalPredictor II

CrystalPredictor II (CP2) [96–98] is the successor of the CrystalPredictor code developed by Karamertzanis & Pantelides [93, 94]. CP2 targets the Global Search stage and can be used to investigate molecular crystals with integer Z'. To that end, an unbiased candidate generation scheme based on quasi-random, low-discrepancy Sobol' sequences [103] generates starting crystal structures, which are then subject to lattice energy minimization. The HAIEFF model used during lattice energy minimization employs LAM-enabled partial molecular flexibility (for flexible torsions) and an electrostatic description based on atomic-charges. If $\boldsymbol{\theta}^{\text{ind}} = \varnothing$, CP2 can also handle rigid-molecule searches. An overview of the CP2 algorithm is shown in Figure 4.1, which comprises a sequence of LAM Generation followed by a Production Run (the former is skipped for rigid molecules). The salient details of each stage are discussed below.

4.2.1.1 LAM Generation

During the Global Search, a large number of lattice energy minimizations are expected to be performed. Moreover, in the interest of an exhaustive Global Search, the ability to freely sample the entire space of $\boldsymbol{\theta}^{\text{ind}}$ is desirable. As such, all LAMs required to describe $\boldsymbol{\theta}^{\text{ind}}$ are generated (and stored) in advance of the production run and no QM calculations are done concurrent with the search. The LAM generation procedure can be broken into several steps.

First, it is necessary to define $\boldsymbol{\theta}^{\text{ind}}$, and specify sensible bounds for each independent dof. Selection of $\boldsymbol{\theta}^{\text{ind}}$ can supported by 1-dimensional torsion scans to study the effect of the torsion on ΔU_{intra}. Generally, torsions that exhibit shallow energy profiles can be considered as 'flexible'. To manage costs, these conformational studies tend to have limited scope and the choice of $\boldsymbol{\theta}^{\text{ind}}$ (and their flexibility ranges) is always supplemented by chemical intuition. Once the $\boldsymbol{\theta}^{\text{ind}}$ search space is defined, an initial set of LAMs can be generated. Without any prior information about

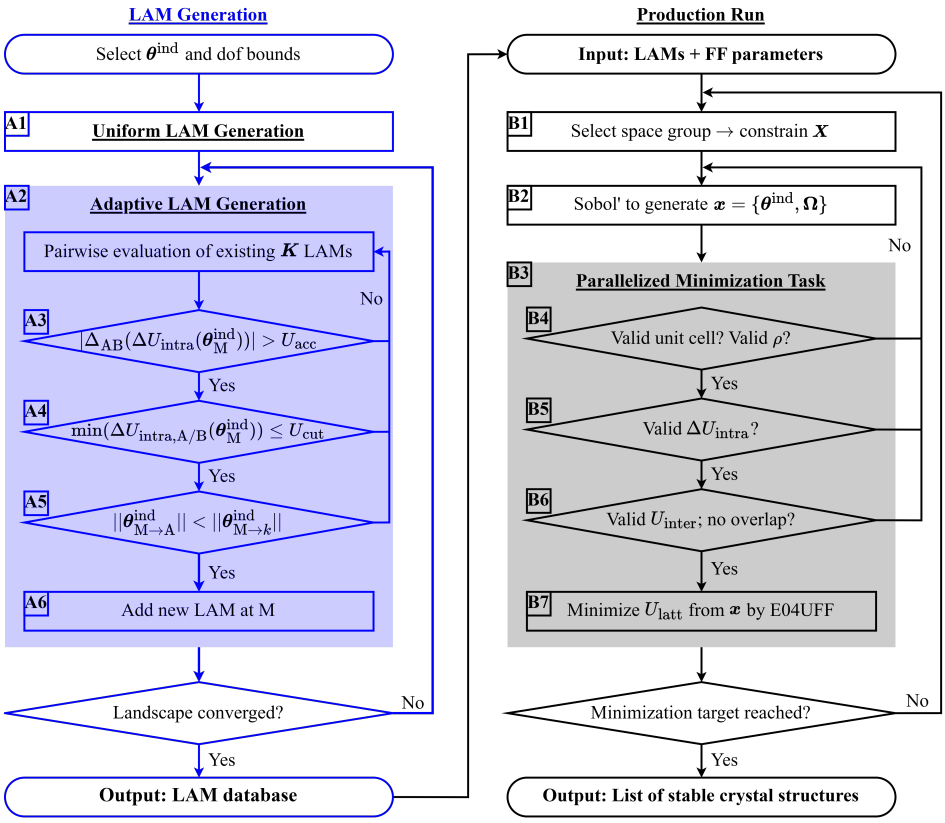


Figure 4.1: Schematic of the CP2 algorithm, involving LAM Generation and the Production Run. Key stages are enumerated and detailed in the main text.

the conformational landscape, a uniform grid is used to sample $\boldsymbol{\theta}^{\text{ind}}$ (Figure 4.1, A1). For molecules with few $\boldsymbol{\theta}^{\text{ind}}$ ($\dim(\boldsymbol{\theta}^{\text{ind}}) \leq 3$), the uniform grid can be populated densely enough for a good description of the conformational landscape, without incurring an impractical number of QM calculations. However, as $\dim(\boldsymbol{\theta}^{\text{ind}})$ grows, it becomes increasingly infeasible to achieve an adequate level of accuracy by uniform sampling alone. To accommodate higher $\dim(\boldsymbol{\theta}^{\text{ind}})$, Sugden *et al.* [97] proposed an adaptive LAM placement scheme.

In systems with high $\dim(\boldsymbol{\theta}^{\text{ind}})$, a sparse uniform grid is first initialized. From here, the goal of the adaptive LAMs is to most-efficiently sample where the existing LAMs give poor predictions. The adaptive algorithm considers the midpoint (M) between two existing LAMs (A and B). The midpoint is chosen as it represents the maximum violation of local-validity for A and B, and should be where both LAMs are most in error without favoring either. If ΔU_{intra} predicted by LAMs A and B ($\Delta U_{\text{intra,A/B}}$) are in poor agreement at $\boldsymbol{\theta}^{\text{ind}}_{\text{M}}$,

$$\left|\Delta U_{\text{intra,A}}(\boldsymbol{\theta}^{\text{ind}}_{\text{M}}) - \Delta U_{\text{intra,B}}(\boldsymbol{\theta}^{\text{ind}}_{\text{M}})\right| > U_{\text{acc}}, \tag{4.2.1}$$

where U_{acc} is the target accuracy, this indicates the need to add a LAM at M. However, two additional caveats apply. First, for the new LAM at M to be accepted, it must fulfill

$$\min(\Delta U_{\text{intra,A}}(\boldsymbol{\theta}^{\text{ind}}_{\text{M}}), \Delta U_{\text{intra,B}}(\boldsymbol{\theta}^{\text{ind}}_{\text{M}})) \leq U_{\text{cut}}, \tag{4.2.2}$$

where U_{cut} is a high-energy cutoff. This reflects the fact that only *observable* molecular conformations (i.e., regions of $\boldsymbol{\theta}^{\text{ind}}$ where ΔU_{intra} is not excessive) are of interest. As such, at least one LAM (A or B) must suggest $\boldsymbol{\theta}^{\text{ind}}_{\text{M}}$ to be energetically-relevant. The second caveat is that

$$||\boldsymbol{\theta}^{\text{ind}}_{\text{M}} - \boldsymbol{\theta}^{\text{ind}}_{\text{A}}|| < ||\boldsymbol{\theta}^{\text{ind}}_{\text{M}} - \boldsymbol{\theta}^{\text{ind}}_{k}||, \qquad \forall k \in \boldsymbol{K}, \quad k \neq \text{A}, \text{B}, \tag{4.2.3}$$

where $\boldsymbol{K}$ is the set of all LAMs that have been evaluated by QM. This distance criterion reflects that poor agreement between A and B is only pertinent if they are the closest LAMs to M. If there exists a LAM (e.g., C) that is closer to M, then C would be the preferred estimator for M, over A and B. Thus over an existing set of $\boldsymbol{K}$ LAMs, these three criteria are applied to all unique LAM pairings (Figure 4.1, A2–A6), and new midpoint LAMs are placed where necessary. The adaptive algorithm is re-ran between rounds of LAM generation (QM calculation), until the landscape converges to the target U_{acc} and no additional LAMs are required. The final set of LAMs are stored in a database that is read-in at the start of the production run. In CP2, it is typical to apply LAM smoothing (Section 3.2.2.3) over all LAMs in the database. Extremely distant LAMs are rendered non-contributing by the weighting of the smoothing algorithm.

4.2.1.2 Production Run

The production run is when CP2 generates candidates and subjects them to lattice energy minimization. The LAM database and FF parameters are provided as inputs. During production, a space group is first selected. The sampling frequency of each space group can either be uniform, or proportional to their observed frequencies [74]. The chosen space group decides if any constraints apply to $\boldsymbol{X}$ (Figure 4.1, B1), from which the decision variable vector is defined as $\boldsymbol{x} = \{\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{\Omega}\}$, where $\boldsymbol{\Omega}$ comprise only unconstrained $\boldsymbol{X}$. A Sobol' sequence is used to generate a trial vector of $\boldsymbol{x}$ (Figure 4.1, B2), and the resulting candidate crystal is subject to several 'sanity' checks. This ensures that (1) the unit cell is not extremely deformed and has a reasonable density (ρ), (2) the initial ΔU_{intra} is not too high, (3) there are no overlapping atoms, and (4) the initial U_{inter} is not excessive (Figure 4.1, B4–B6). This screening prevents minimization of candidates that are likely to fail, take too long to minimize, or are unlikely to minimize to competitive structures. Trial $\boldsymbol{x}$ are generated in the target space group until a valid structure for minimization is found. This is passed to the minimization protocol (Figure 4.1, B7), which employs the NAG E04UFF subroutine to minimize U_{latt} according to,

$$\min_{\boldsymbol{\Omega}, \boldsymbol{\theta}^{\text{ind}}} \Delta U_{\text{intra}}(\boldsymbol{\theta}^{\text{ind}}) + U_{\text{inter}}(\boldsymbol{\Omega}, \boldsymbol{\theta}^{\text{ind}}, \boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}})). \tag{4.2.4}$$

Here, the explicit dependence of U_{inter} on $\boldsymbol{\theta}^{\text{ind}}$ represents the 'direct' effect $\boldsymbol{\theta}^{\text{ind}}$ has on atomic positions in the crystal (including effects from $\boldsymbol{\theta}^{\text{dep}}$). The secondary dependence of U_{inter} on $\boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}})$ represents the 'indirect' effect $\boldsymbol{\theta}^{\text{ind}}$ has on the atomic-charges. This production cycle is repeated until a target number of minimization have been attempted. CP2 has been parallelized using Message Passing Interface (MPI) such that the independent minimization tasks can be distributed over multiple processes (Figure 4.1, B3).

4.2.1.3 Version 2.5.2.1

Throughout this work, I use CrystalPredictorv2.5.2.1, which has underwent several changes.

Code Optimization for Large LAM Databases

Executing LAM smoothing (Section 3.2.2.3) with a sizable LAM database (~$\mathcal{O}(10^4)$) requires iterating through extremely large and deeply nested FOR loops (e.g., iterating over each LAM, for each atom-in-molecule, for Z' molecules). Past versions of CP2 were extremely inefficient at these evaluations due to: (1) suboptimal memory allocation in arrays, and (2) explicit implementation of these nested loops. In CPv2.5.2.1, I address these inefficiencies by:

1. Refactoring the large arrays that hold LAM information, to conform with the column-major memory ordering in Fortran90. Nested FOR loops using these arrays have also been reordered to access memory locations contiguously, resulting in faster computations.
2. Substituting explicitly written FOR loops with intrinsic subroutines from the Basic Linear Algebra Subprograms (BLAS) Fortran libraries. These subroutines are highly optimized to execute various vector and matrix operations, like those involved in LAM smoothing. This substitution can lead to significant code speed-up.

Reformulation of CP2 Flexible Gradients

The E04UFF subroutine used in lattice energy minimization can be sensitive to errors in the user-provided gradients, which prevent tight convergence to a minimum. In such cases, CP2 flags the structure as an invalid lattice energy minimum. In some CSP studies, it was observed that as little as 1% of attempted minimizations were returning valid minima. It was determined that this stemmed from errors in intermediate quantities used for the calculation of U_{inter} gradients with respect to $\boldsymbol{\theta}^{\text{ind}}$. In CPv2.5.2.1, I have re-derived and re-implemented the analytical expressions for these quantities (Appendix D).

Revised Implementation of Shift Algorithm

In systems with large molecules and/or $Z' > 1$, there can be an extremely high propensity to generate unsound $\boldsymbol{x}$ that have invalid intermolecular separation/overlapping atoms. Repeated rejection of these trial $\boldsymbol{x}$ (Figure 4.1, B6) lead to long periods without any $\boldsymbol{x}$ being minimized and a significant reduction of the Global Search production rate (i.e., minimizations completed over time). To address this, Zhang [71] proposed the 'Shift Algorithm'. Given a vector $\boldsymbol{x}$ that exhibits 'molecular clashing' (i.e., unsound intermolecular separation), the Shift Algorithm attempts to separate the clashing molecules by expanding the unit cell, while preserving molecular fractional COM positions. Crucially, the Shift Algorithm employs *anisotropic* scaling of each cell length to most-efficiently separate the molecules and minimize cell expansion. This is critical as excessive call enlargement can lead to longer minimization times that negate the speed-up from 'salvaging' $\boldsymbol{x}$. In Zhang's implementation [71], space group constraints on a, b, and c were neglected, thus only triclinic and monoclinic lattice systems could undergo the protocol. In CPv2.5.2.1, I have re-implemented the Shift Algorithm to account for all space group constraints (Appendix E).

Performance of CPv2.5.2.1

CPv2.5.2.1 has been systematically tested on several molecular systems and compared to the last stable version of CrystalPredictor, CPv2.4.4. Firstly, the modified gradients were validated using E04UFF's gradient-check utility. When requested, second-order forward finite-difference gradients are computed for each element of the decision variable vector according to

$$\frac{\partial F}{\partial x}(x_0) = \frac{4F(x_0 + \delta x) - 3F(x_0) - F(x_0 + 2(\delta x))}{2(\delta x)}, \tag{4.2.5}$$

where F is the objective function value (lattice energy), x is the decision variable being tested, x_0 is its value at the test point, and δx is the finite-difference perturbation. Each finite-difference gradient is compared to its analytical gradient counterpart, and from the relative error the subroutine reports if the analytical gradient is adequately accurate. This test was performed on the experimental structures for three systems: Molecule XXXI from the 7th Blind Test [49] (CSD Refcode: $\text{ZEHGAY}_{\text{maj}}$, $Z' = 1$), l-alanine (CSD Refcode: LALNIN23, $Z' = 1$), and cyproconazole (CSD Refcode: RUTTEH, $Z' = 2$). The results for gradient elements involving $\boldsymbol{\theta}^{\text{ind}}$ are shown in Table 4.1. In all cases, more satisfactory gradients are achieved using CPv2.5.2.1 than CPv2.4.4. For LALNIN23, gradients with respect to two θ^{ind} continue to be flagged as incorrect using CPv2.5.2.1. Quantitative comparison to the finite-difference gradients reveal that the CPv2.5.2.1 gradients are less erroneous, relative to the CPv2.4.4 gradients. Detailed investigation of these gradient elements also indicates that their error stems primarily from $\frac{\partial \Delta U_{\text{intra}}}{\partial \theta^{\text{ind}}}$, rather than $\frac{\partial U_{\text{inter}}}{\partial \theta^{\text{ind}}}$, which is unrelated to the reformulated gradients. Overall, evidence suggests that the $\boldsymbol{\theta}^{\text{ind}}$ gradients in CPv2.5.2.1 are more accurate than in previous CP2 versions.

To test the other CPv2.5.2.1 changes, Global Searches were performed for two relatively large $Z' = 2$ systems: RUTTEH and cannabinol (CSD Refcode: CANNOL). The RUTTEH search involves two distinct molecules (both with 38 atoms) in the asymmetric unit cell, with 9,280 and 10,651 LAMs respectively, while the CANNOL search has two identical molecules (with 49 atoms) in the asymmetric unit cell, with 21,807 LAMs. For each search, a target of 400 minimizations was requested. Three versions of CP2 were tested: v2.4.4, v2.5.2 (which does not use the refactored LAM arrays and FOR loops), and v2.5.2.1. Searches with the latter two versions were attempted with and without the Shift Algorithm. To mitigate performance variation from job allocation to different compute nodes, each search was repeated 30 times and statistics are averaged over all attempts (Table 4.2).

Table 4.1: E04UFF gradient test results for CPv2.4.4 and CPv2.5.2.1. The number of satisfactory $\frac{\partial F}{\partial \theta^{\text{ind}}}$ elements, determined by the E04UFF subroutine, are reported.

Molecule	$\dim(\boldsymbol{\theta}^{\text{ind}})$	CPv2.4.4	CPv2.5.2.1
$\text{ZEHGAY}_{\text{maj}}$	3	3	3
LALNIN23	3	0	1
RUTTEH (Molecule 1)	6	0	6
RUTTEH (Molecule 2)	6	0	6

Table 4.2: Average Global Searches statistics for RUTTEH and CANNOL: number of failed $\boldsymbol{x}$ to reach a valid candidate, time spent generating the candidate, time spent minimizing the candidate, overall production rate, and proportion of minimizations returning stable minima.

	CPv2.4.4	CPv2.5.2 (Shift Off)	CPv2.5.2.1 (Shift Off)	CPv2.5.2 (Shift On)	CPv2.5.2.1 (Shift On)
RUTTEH					
Attempts to candidate	204.32	201.85	211.14	107.51	107.51
Time to candidate (s)	146.65	154.45	186.42	82.58	90.38
Time to minimise (s)	711.86	756.78	346.37	746.32	366.53
Production rate (min s^{-1} $proc^{-1}$)	0.0012	0.0012	0.0021	0.0013	0.0026
Stable proportion (%)	0.00	12.27	12.78	11.99	13.27
CANNOL					
Attempts to candidate	58.98	57.94	67.12	17.51	17.49
Time to candidate (s)	90.87	93.33	134.51	28.67	36.60
Time to minimise (s)	1618.83	1646.65	570.1	1595.31	581.19
Production rate (min s^{-1} $proc^{-1}$)	0.0006	0.0006	0.0017	0.0007	0.0019
Stable proportion (%)	53.68	51.87	52.50	47.88	50.17

Without the Shift Algorithm, all three CP2 versions expend roughly the same effort to generate a valid candidate $\boldsymbol{x}$, although CPv2.5.2.1 is marginally slower. I believe this should be an artefact of the relatively small-scale test, since there are no meaningful differences in the code affecting this stage. In the CP2 versions with the modified gradients, the RUTTEH search yields a significant improvement in the proportion of stable minima, while in the CANNOL search there is only some negligible variation relative to CPv2.4.4. For CANNOL, I expect that over a larger number of minimizations, these variations should diminish. Most notably, CPv2.5.2.1 exhibits a significant reduction in minimization time versus the other two versions, which can only be explained by the code restructuring. The average minimization time reduces by about half for RUTTEH and two-thirds for CANNOL, while their respective production rates increase almost proportionately. When the Shift Algorithm is active, candidate generation rates accelerate significantly in both CPv2.5.2 and CPv2.5.2.1. Critically, the cell expansions have not manifested in any appreciable increase to the minimization times, such that production rates are overall faster with the Shift Algorithm. There is no reason to believe that the Shift Algorithm should have any bearing on the proportion of stable minima. While some variation in this statistic is observed here, this should diminish over more minimizations.

In summary, the changes made in CPv2.5.2.1 appear to bolster both the speed and reliability of the CP2 code, and I continue to use it throughout this work.

4.2.2 Crystal Structure Optimizer – Rigid Molecules (CSO–RM)

The Crystal Structure Optimizer codes were developed by Bowskill [2] to support a range of CSP activities. Crystal Structure Optimizer – Rigid Molecules (CSO–RM) underpins this by facilitating *rigid* minimization of molecular crystals with integer Z', using either point-charge

or multipole HAIEFF models. For a fixed molecular conformation, CSO–RM minimizes U_{inter} with respect to the (unconstrained) rigid-body variables, $\boldsymbol{\Omega}$. Given an input crystal structure, CSO–RM first requests *single-point* QM calculations for molecules in the asymmetric unit cell. This is used to derive the atomic-charges or multipoles, $\boldsymbol{Q}$, associated with the input molecular conformation(s). From here, CSO–RM minimizes U_{inter} with respect to $\boldsymbol{\Omega}$ using the QM-derived $\boldsymbol{Q}$ and a user-specified set of empirical FF parameters. In CSO–RM, this minimization occurs in two steps. First, *quasi-Newton* optimization is done using the E04UFF subroutine. Following this, one can request a more rigorous *Newton* optimization to refine the quasi-Newton solution. The Newton refinement has been implemented in-house and is explained in detail by Bowksill [2]. Analogous to the popular DMACRYS code [250], CSO–RM can be used to conduct rigid-refinement of CP2-generated Global Search structures. However, unlike DMACRYS, CSO–RM outputs several additional numerical quantities (e.g., FF parameter gradients) and achieves tighter convergence in the rigid minimization through the Newton refinement. These features make CSO–RM suitable for interfacing with our other CSP programs, as explained below.

In this work, I primarily use CSO–RMv1.4.1. This version is identical to Bowskill's original release [2], but has added capabilities to use other functional forms in the HAIEFF FF. These changes will be relevant in Chapter 6, when FF parametrization is performed.

4.2.3 Crystal Structure Optimizer – Flexible Molecules (CSO–FM)

Crystal Structure Optimizer – Flexible Molecules (CSO–FM) is the flexible-minimization counterpart to CSO–RM. CSO–FM's role is the refinement of CP2-generated Global Search structures, with multipole electrostatics and LAM-enabled molecular flexibility. CSO–FM can be seen as the successor to the CrystalOptimizer code [95], but CSO–FM calls on CSO–RM rather than DMACRYS. While CP2 and CSO–FM both address flexible lattice energy minimization (albeit with different electrostatic models), in practice the two programs behave quite differently. One difference is that in CP2 $\boldsymbol{\theta}^{\text{ind}} \in \{\text{torsions}\}$, whereas in CSO–FM *all* cdofs can be used in $\boldsymbol{\theta}^{\text{ind}}$. If $\boldsymbol{\theta}^{\text{ind}} = \boldsymbol{\theta}$ (i.e., $\boldsymbol{\theta}^{\text{dep}} = \varnothing$), CSO–FM can treat flexibility atomistically. Other important aspects of the CSO–FM algorithm are explained below, with reference to Figure 4.2.

4.2.3.1 CSO–FM Bilevel Problem Formulation

In CSO–FM, the problem shown in Equation 4.2.4 is reformulated as a bilevel optimization,

$$\begin{aligned} &\min_{\boldsymbol{\theta}^{\text{ind}}} \Delta U_{\text{intra}}(\boldsymbol{\theta}^{\text{ind}}) + U_{\text{inter}}(\boldsymbol{\Omega}^*(\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}})), \boldsymbol{\theta}^{\text{ind}}, \boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}})) \\ &s.t. \\ &\boldsymbol{\Omega}^* = \arg\min_{\boldsymbol{\Omega}} U_{\text{inter}}(\boldsymbol{\Omega}; \boldsymbol{\theta}^{\text{ind}}, \boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}})). \end{aligned} \tag{4.2.6}$$

In the outer problem, U_{latt} is minimized with respect to only the independent cdofs, $\boldsymbol{\theta}^{\text{ind}}$, subject to U_{inter} being a rigid-body minimum ($\boldsymbol{\Omega}^*$) at $\boldsymbol{\theta}^{\text{ind}}$. This constraint is enforced by the inner problem, where $\boldsymbol{\theta}^{\text{ind}}$ (and the associated LAM-predicted $\boldsymbol{Q}$) are fixed parameters in the

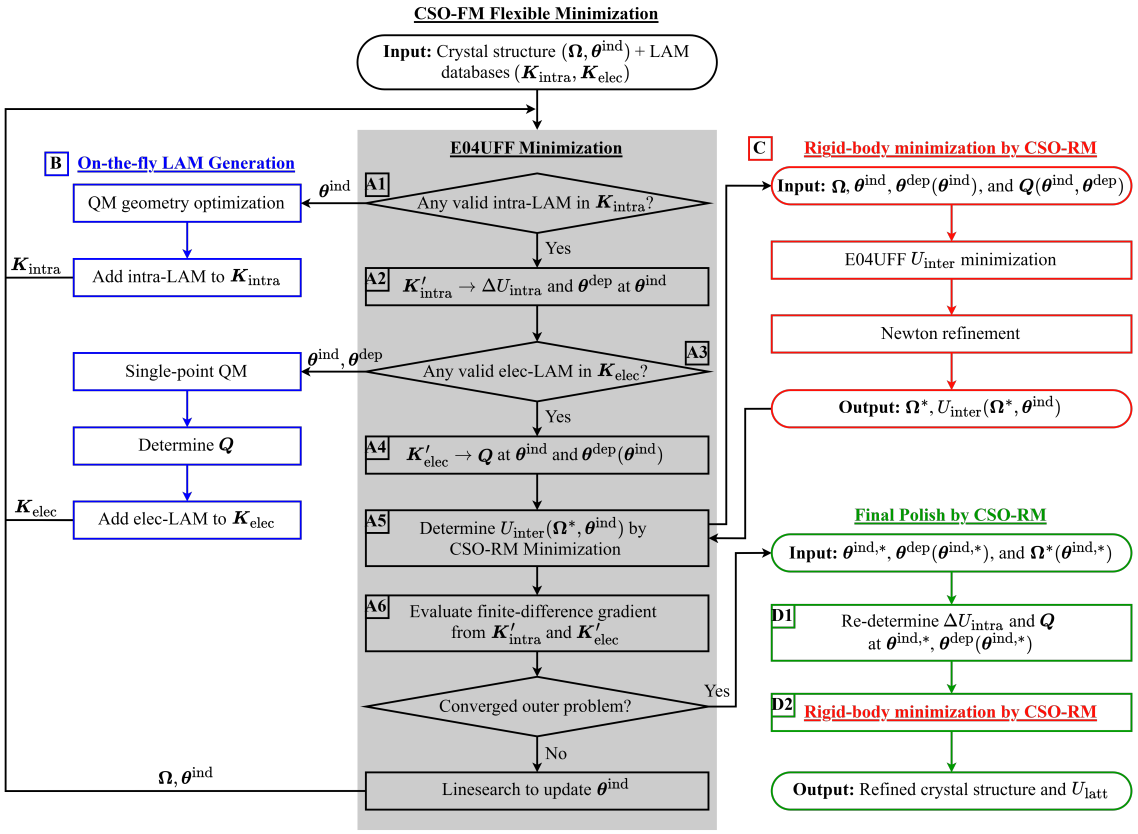


Figure 4.2: Schematic of the CSO–FM algorithm that employs on-the-fly LAM generation and Final Polish by CSO–RM. Key stages are enumerated and detailed in the main text.

rigid minimization of U_{inter}. For clarity, I have omitted the dependencies of U_{inter} and $\boldsymbol{Q}$ on $\boldsymbol{\theta}^{\text{dep}}$. CSO–FM uses the E04UFF subroutine to manage the outer problem. Where U_{inter} and other gradient quantities are needed, CSO–RM subroutines are called to solve the inner problem.

4.2.3.2 CSO–FM LAM Implementation

CSO–FM also relies on the LAM framework to describe molecular flexibility. However, unlike CP2, CSO–FM does not pre-compute LAMs over the *entire* $\boldsymbol{\theta}^{\text{ind}}$ space. This would be inefficient since there is no interest to explore all regions of $\boldsymbol{\theta}^{\text{ind}}$ during Refinement. Instead, it is more likely that the refinement only traverses a local region near the input conformation(s). This assumes the Global Search model is reasonably accurate such that the input crystal structure (and molecular conformation) is near a minimum on the Refinement energy landscape. As such, CSO–FM uses 'on-the-fly' LAM generation.

At each step of the outer minimization, CSO–FM first queries if any LAMs relevant to the current $\boldsymbol{\theta}^{\text{ind}}$ exist in the LAM database (Figure 4.2, A1 and A3). For $\boldsymbol{K}$ LAMs in the database, the relevance of LAM k is determined by its relative distance (d_k) to the current conformation,

$$d_k = \sqrt{\sum_n \left[\frac{\theta_n - \theta_{k,n}}{\Delta_{\text{tol}}}\right]^2}, \qquad \forall\ k \in \boldsymbol{K}. \tag{4.2.7}$$

n is the index of the cdof being considered, θ_n is the dof value at the current conformation, $\theta_{k,n}$ is the corresponding dof value at LAM k, and Δ_{tol} is a user-specified tolerance for dof deviations. Suggested values for Δ_{tol} are 5.0° for torsion and bond angles, and 0.1 Å for bond lengths [95]. If $d_k \leq 1.0$, LAM k is deemed relevant to the current point and added to a valid LAM set, $\boldsymbol{K'}$. All LAMs in $\boldsymbol{K'}$ are used with the smoothing algorithm to determine $\boldsymbol{\theta}^{\text{dep}}$ and ΔU_{intra} (Figure 4.2, A2), or $\boldsymbol{Q}$ (Figure 4.2, A4) at $\boldsymbol{\theta}^{\text{ind}}$. When constructing $\boldsymbol{K}'_{\text{intra}}$ to describe $\boldsymbol{\theta}^{\text{dep}}$ and ΔU_{intra}, n only considers the set of $\boldsymbol{\theta}^{\text{ind}}$, while $\boldsymbol{K}'_{\text{elec}}$ used to describe $\boldsymbol{Q}$ evaluate n over $\boldsymbol{\theta}^{\text{ind}}$ and $\boldsymbol{\theta}^{\text{dep}}$. This reflects that the rotation of multipoles about their local axis systems is dependent on both $\boldsymbol{\theta}^{\text{ind}}$ *and* $\boldsymbol{\theta}^{\text{dep}}$. Due to their distinct treatment, separate intramolecular ($\boldsymbol{\theta}^{\text{dep}}$ and ΔU_{intra}) and electrostatic ($\boldsymbol{Q}$) LAM databases are curated by CSO–FM (Figure 4.2, B). Moreover, a corollary of this treatment is that determination of $\boldsymbol{\theta}^{\text{dep}}(\boldsymbol{\theta}^{\text{ind}})$ is a prerequisite to assess Equation 4.2.7 for the electrostatic LAMs. If $\boldsymbol{K}'_{\text{intra}}, \boldsymbol{K}'_{\text{elec}} \neq \varnothing$, $\boldsymbol{\theta}^{\text{dep}}(\boldsymbol{\theta}^{\text{ind}})$ and $\boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{\theta}^{\text{dep}}(\boldsymbol{\theta}^{\text{ind}}))$ can be provided to CSO–RM for evaluating $U_{\text{inter}}(\boldsymbol{\Omega}^*(\boldsymbol{\theta}^{\text{ind}}); \boldsymbol{\theta}^{\text{ind}})$ (Figure 4.2, A5 → C).

If at the end of the LAM query, $\boldsymbol{K'} = \varnothing$ (i.e., there are no relevant LAMs in the database), CSO–FM exits the minimization protocol and requests any necessary QM/$\boldsymbol{Q}$ calculation (Figure 4.2, B). For an intramolecular LAM, a QM geometry optimization is required, while an electrostatic LAM only needs a single-point evaluation of the electronic structure at $(\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{\theta}^{\text{dep}}(\boldsymbol{\theta}^{\text{ind}}))$. Once complete, the new LAMs are appended to the database, and CSO–FM re-enters the minimization protocol from the last $\boldsymbol{\theta}^{\text{ind}}$ (and $\boldsymbol{\Omega}$).

4.2.3.3 CSO–FM Gradients with respect to Independent Dofs

Solving the outer problem in Equation 4.2.6 requires values of $\frac{\partial \Delta U_{\text{intra}}}{\partial \theta^{\text{ind}}}$ and $\frac{\partial U_{\text{inter}}}{\partial \theta^{\text{ind}}}$. Analytical expressions for the former can be derived from the LAM equations [95]. Conversely, obtaining closed-form expressions for the latter is complicated by the need to conduct the rigid-body minimization (Figure 4.2, C). As such, it is more convenient to obtain $\frac{\partial U_{\text{inter}}}{\partial \theta^{\text{ind}}}$ by (central) finite-differences. Considering the case of a single θ^{ind},

$$\begin{aligned}\frac{\partial U_{\text{inter}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))}{\partial \theta^{\text{ind}}} \approx & \frac{U_{\text{inter}}(\boldsymbol{\Omega}^*(\theta^{\text{ind},+}, \boldsymbol{Q}(\theta^{\text{ind},+})), \theta^{\text{ind},+}, \boldsymbol{Q}(\theta^{\text{ind},+}))}{2(\delta\theta^{\text{ind}})} \\ & - \frac{U_{\text{inter}}(\boldsymbol{\Omega}^*(\theta^{\text{ind},-}, \boldsymbol{Q}(\theta^{\text{ind},-})), \theta^{\text{ind},-}, \boldsymbol{Q}(\theta^{\text{ind},-}))}{2(\delta\theta^{\text{ind}})},\end{aligned} \tag{4.2.8}$$

where $\delta\theta^{\text{ind}}$ is a small perturbation of θ^{ind}, and $\theta^{\text{ind},\pm} = \theta^{\text{ind}} \pm \delta\theta^{\text{ind}}$. Evaluating Equation 4.2.8 is relatively costly since rigid minimizations are performed in both the forward and backward perturbative steps. Instead, from chain-rule, it can be shown that

$$\begin{aligned}\frac{\partial U_{\text{inter}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))}{\partial \theta^{\text{ind}}} = & \left.\frac{\partial U_{\text{inter}}}{\partial \theta^{\text{ind}}}\right|_{\boldsymbol{\Omega}^*, \boldsymbol{Q}} + \left.\frac{\partial U_{\text{inter}}}{\partial \boldsymbol{Q}}\right|_{\boldsymbol{\Omega}^*, \theta^{\text{ind}}} \left[\frac{\partial \boldsymbol{Q}}{\partial \theta^{\text{ind}}}\right] \\ & + \cancel{\left.\frac{\partial U_{\text{inter}}}{\partial \boldsymbol{\Omega}^*}\right|_{\theta^{\text{ind}}, \boldsymbol{Q}}} \left[\frac{\partial \boldsymbol{\Omega}^*}{\partial \theta^{\text{ind}}} + \frac{\partial \boldsymbol{\Omega}^*}{\partial \boldsymbol{Q}} \frac{\partial \boldsymbol{Q}}{\partial \theta^{\text{ind}}}\right],\end{aligned} \tag{4.2.9}$$

wherein the second term is eliminated by the inner problem's optimality criterion, $\frac{\partial U_{\text{inter}}}{\partial \boldsymbol{\Omega}^*} = 0$. The legitimacy of this simplification is bolstered by the tight convergence to optimality attainable with CSO–RM's Newton refinement. From Equation 4.2.9, the required $\frac{\partial U_{\text{inter}}}{\partial \theta^{\text{ind}}}$ term is no longer dependent on $\boldsymbol{\Omega}^*$, and Equation 4.2.8 can be simplified as

$$\begin{aligned}\frac{\partial U_{\text{inter}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))}{\partial \theta^{\text{ind}}} \approx & \frac{U_{\text{inter}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind},+}, \boldsymbol{Q}(\theta^{\text{ind},+}))}{2(\delta\theta^{\text{ind}})} \\ & - \frac{U_{\text{inter}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind},-}, \boldsymbol{Q}(\theta^{\text{ind},-}))}{2(\delta\theta^{\text{ind}})}.\end{aligned} \tag{4.2.10}$$

With this, only *single-point* evaluations of U_{inter} at the unperturbed rigid-body minimum, $\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))$, need to be performed, accounting for the dependence of U_{inter} on θ^{ind} and $\boldsymbol{Q}$. These single-point U_{inter} evaluations can be done using the CSO–RM subroutines.

In practice, analytical ΔU_{intra} gradients have not been implemented in CSO–FM, and finite-differences are conducted over the entire lattice energy,

$$\begin{aligned}\frac{\partial U_{\text{latt}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))}{\partial \theta^{\text{ind}}} \approx & \frac{U_{\text{latt}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind},+}, \boldsymbol{Q}(\theta^{\text{ind},+}))}{2(\delta\theta^{\text{ind}})} \\ & - \frac{U_{\text{latt}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind},-}, \boldsymbol{Q}(\theta^{\text{ind},-}))}{2(\delta\theta^{\text{ind}})}.\end{aligned} \tag{4.2.11}$$

When assessing Equation 4.2.11, the $\boldsymbol{K}'_{\text{intra}}$ and $\boldsymbol{K}'_{\text{elec}}$ LAMs from the preceding function evaluation are reused to update ΔU_{intra}, $\boldsymbol{theta}^{\text{dep}}$ and $\boldsymbol{Q}$ in response to $\delta\theta^{\text{ind}}$ (Figure 4.2, A6). Since $\delta\theta^{\text{ind}}$ are small, minor violation of the LAM validity range from the finite-difference perturbations can be tolerated. Conversely, updating $\boldsymbol{K}'_{\text{intra}}$ and $\boldsymbol{K}'_{\text{elec}}$ at each perturbation is impractical (especially if new LAMs are demanded), and may introduce undesirable numerical behavior.

4.2.3.4 Final Polish

The E04UFF subroutine iterates its minimization cycle, updating $\boldsymbol{\theta}^{\text{ind}}$ via linesearches until the outer problem converges. At that point, CSO–FM enters a 'Final Polish' phase, which aims to reduce inaccuracies that arise from using the LAM model. During the Final Polish, the conformation of the molecule is fixed to the E04UFF-determined optimum, i.e., $(\boldsymbol{\theta}^{\text{ind},*}, \boldsymbol{\theta}^{\text{dep}}(\boldsymbol{\theta}^{\text{ind},*}))$. At this conformation, ΔU_{intra} and the electrostatic features (multipoles) are re-determined (Figure 4.2, D1). The updated $\boldsymbol{Q}$ and $(\boldsymbol{\theta}^{\text{ind},*}, \boldsymbol{\theta}^{\text{dep}}(\boldsymbol{\theta}^{\text{ind},*}))$ are provided to CSO–RM for a rigid-body minimization, initialized from the last set of optimum rigid-body variables, i.e., $\boldsymbol{\Omega}^*(\boldsymbol{\theta}^{\text{ind},*})$, (Figure 4.2, D2). From this, a 'polished' U_{latt} and crystal structure is returned. If the LAM validity ranges have been well-specified such that the LAM model is an adequate local surrogate model for molecular flexibility, we typically do not expect large deviations at the Final Polish stage.

In this work, I use CSO–FMv1.3.1 and CSO–FMv1.3.2. Both versions are identical to the original CSO–FM release, except for some minor changes that enable CSO–FM to interface with CSO–RMv1.4.1. CSO–FMv1.3.2 fixes errors encountered in CSO–FMv1.3.1 related to the treatment of rigid molecules with PCM.

4.2.4 CrystalDynamics

While CSP_0 is the most widely practiced form of CSP today, the Imperial MSE CSP Suite offers some capabilities to perform CSP_thd-type evaluations, using the CrystalDynamics program [67, 232]. As the present work focuses on CSP_0 HAIEFF models, I touch only briefly on this program. Using CrystalDynamics, vibrational contributions to crystal free energy are determined based on a HAIEFF model and lattice dynamics theory. In its current implementation [67], CrystalDynamics employs a harmonic approximation to model the various crystalline vibrational contributions, including: zero-point vibrations, vibrations from thermal energy, and vibrational entropy (F_{vib} in Equation 2.1.3). Under the harmonic approximation, the input structure into CrystalDynamics must be a static lattice energy minimum. As such, a target crystal structure must first be minimized with the desired HAIEFF model in CSO–RM/FM, before being evaluated in CrystalDynamics with an identical energy model. At the static U_{latt} minimum, the dynamical matrix is constructed from second-derivatives quantities of ΔU_{intra}, U_{FF}, and U_{elec}, relative to atomic positions. From the dynamical matrix, phonon frequencies, dispersion curves, and phonon density of states may be evaluated. Finally, the vibrational contribution at a given temperature, $F_{\text{vib}}(T)$, is determined from the phonon frequencies by numerical integration. When $F_{\text{vib}}(T)$ is combined with the static U_{latt}, the crystal's Helmholtz free energy is obtained, while further addition of the PV term gives the Gibbs free energy.

Currently, CrystalDynamics [67] is capable of conducting single-point free energy evaluations for rigid/flexible molecules from CSO–RM/FM. HAIEFF models based on Buckingham potential FFs and multipole electrostatics have been implemented into CrystalDynamics. Interested readers are directed to the work of Konstantinopoulos [67] for details about CrystalDynamics' theory, implementation, and performance.

4.3 Parameter Estimation for HAIEFF Force-fields

On top of the routine CSP tasks handled by the Imperial MSE CSP Suite, a focal point of this work will be the parametrization of HAIEFF FFs. To that end, our group has developed the CrystalEstimator program [1, 2] for large-scale parameter estimation of HAIEFF FFs. In this section, I present the mathematical formulation of the parameter estimation problem and derive the requisite gradient expressions for its solution. Lastly, I outline the algorithmic implementation of the CrystalEstimator program [1, 2].

4.3.1 Parameter Estimation Problem Formulation

The parameter estimation formulation has been discussed in detail by Bowskill and co-workers [1, 2]. Here, I highlight only the salient details.

The goal of the parameter estimation is to find an optimal set of FF parameters that minimizes the error in predicted crystal geometries and energies, relative to the training data. Following the framework of Refs [169, 213, 251], we opt to embed explicit energy minimizations in

the parameter estimation, such that the quality of the predicted energies/geometries are assessed only after energy minimization of each crystal structure. While crystal geometries are allowed to vary in this formulation, it is desirable to limit these changes to $\boldsymbol{\Omega}$ and exclude conformational changes from the model. As shown before, modeling molecular flexibility is non-trivial since $\boldsymbol{\theta}$ affects both ΔU_{intra} and $\boldsymbol{Q}$. Constructing detailed (LAM) models for these QM-derived quantities is ultimately impractical at the scale required for reliable parameter estimation. Moreover, permitting conformational changes in the parameter estimation would inadvertently cause the empirical FF to compensate for errors in the intramolecular model [2]. These errors tend to be more system-specific, thus this would likely harm FF transferability.

Altogether, to avoid these issues, all molecular geometries are fixed to their crystalline conformations in the reference structures ($\boldsymbol{\theta}^{\text{ref}}$), and only the *intermolecular* energy is optimized over the rigid-body variables. As will be explained in Chapter 5, I use periodic DFT-D generated training data for the parameter estimation, for which it is trivial to decouple U_{inter} from U_{latt}. The parameter estimation energy minimization is thus given by

$$\min_{\boldsymbol{\Omega}} U_{\text{inter}}(\boldsymbol{\Omega}; \boldsymbol{p}, \boldsymbol{\theta}^{\text{ref}}), \tag{LM}$$

where U_{inter} contains the electrostatic and FF contributions. The quantities of relevance to the parameter estimation are: (1) the geometric variables at the optimum ($\boldsymbol{\Omega}^*$),

$$\boldsymbol{\Omega}^*(\boldsymbol{p}, \boldsymbol{\Omega}^\dagger, \boldsymbol{\theta}^{\text{ref}}) = \arg\min_{\boldsymbol{\Omega}} U_{\text{inter}}(\boldsymbol{\Omega}; \boldsymbol{p}, \boldsymbol{\theta}^{\text{ref}}), \tag{4.3.1}$$

and (2) the corresponding optimized intermolecular energy (U^*_{inter}),

$$U^*_{\text{inter}}(\boldsymbol{p}, \boldsymbol{\Omega}^\dagger, \boldsymbol{\theta}^{\text{ref}}) = U_{\text{inter}}(\boldsymbol{\Omega}^*, \boldsymbol{p}, \boldsymbol{\theta}^{\text{ref}}). \tag{4.3.2}$$

Since the crystal energy surface is highly non-convex and has many local minima, the solutions to Problem LM are dependent on the chosen starting point ($\boldsymbol{\Omega}^\dagger$). Since the optimal FF parameters aim to reproduce the training data, it is most sensible to set $\boldsymbol{\Omega}^\dagger = \boldsymbol{\Omega}^{\text{ref}}$, where $\boldsymbol{\Omega}^{\text{ref}}$ are the reference values of the rigid-body variables. Henceforth, the 'ref' and '*' superscripts denote values obtained from the reference dataset and from solving Problem LM, respectively.

To assess the quality of the geometry predictions, Gatsiou and co-workers [213, 251] proposed the following. For a structure, s, geometry residuals ($\boldsymbol{G}_s$) are defined for the set of lattice parameters, $\boldsymbol{X}_s$, and (fractional) atomic positions ($\boldsymbol{Y}_s$). The former are explicitly contained in $\boldsymbol{\Omega}_s$, while the latter can be calculated from the expressions in Section 3.1. Geometry residuals for $\boldsymbol{X}_s$ are given by the relative deviation of the lattice parameter from its reference value,

$$G_{s,i}(\boldsymbol{\Omega}^*_s; \boldsymbol{\Omega}^{\text{ref}}_s) = \frac{1}{\sqrt{N_{X,s}}} \left(\frac{X^*_{s,i}(\boldsymbol{\Omega}^*_s) - X^{\text{ref}}_{s,i}(\boldsymbol{\Omega}^{\text{ref}}_s)}{X^{\text{ref}}_{s,i}(\boldsymbol{\Omega}^{\text{ref}}_s)} \right), \qquad i = 1, \cdots, N_{X,s}, \tag{4.3.3}$$

where $N_{X,s}$ is the number of unconstrained lattice parameters in $\boldsymbol{\Omega}_s$. To account for different $N_{X,s}$ between structures (depending on their lattice system, Table 3.1), the relative deviation

is normalized by $\sqrt{N_{X,s}}$. Geometry residuals for $\boldsymbol{Y}_s$ are given by the deviations in (relative) fractional atomic positions in the asymmetric unit cell,

$$G_{s,N_{X,s}+i}(\boldsymbol{\Omega}_s^*;\boldsymbol{\Omega}_s^{\text{ref}}) = \frac{1}{\sqrt{3(N_{r,s}-1)}}(Y_{s,i}^*(\boldsymbol{\Omega}_s^*) - Y_{s,i}^{\text{ref}}(\boldsymbol{\Omega}_s^{\text{ref}})), \qquad i = 1,\cdots,3(N_{r,s}-1), \tag{4.3.4}$$

where $N_{r,s}$ is the total number of atoms in the asymmetric unit cell, and the fractional atomic positions are defined relative to the first atom in the asymmetric unit cell,

$$\boldsymbol{Y}_s = \begin{pmatrix} \hat{x}_{2,s} - \hat{x}_{1,s} \\ \hat{y}_{2,s} - \hat{y}_{1,s} \\ \hat{z}_{2,s} - \hat{z}_{1,s} \\ \vdots \\ \hat{x}_{N_{r,s},s} - \hat{x}_{1,s} \\ \hat{y}_{N_{r,s},s} - \hat{y}_{1,s} \\ \hat{z}_{N_{r,s},s} - \hat{z}_{1,s} \end{pmatrix}. \tag{4.3.5}$$

Note that $\boldsymbol{Y}_s$ has $3(N_{r,s}-1)$ elements since the first atom is excluded. To account for distinct $N_{r,s}$ between structures, the residuals in Equation 4.3.4 are normalized by $\sqrt{3(N_{r,s}-1)}$. Finally, the quality of energy prediction is quantified by an energy residual (E_s),

$$E_s(U_{\text{inter},s}^*;U_{\text{inter},s}^{\text{ref}}) = \frac{U_{\text{inter},s}^* - U_{\text{inter},s}^{\text{ref}}}{U_{\text{inter},s}^{\text{ref}}}, \tag{4.3.6}$$

which reflects the relative error in the predicted intermolecular energy.

From here, the merit function contribution from structure s (MF_s) can be defined as the weighted sum-of-squares of all the residuals,

$$MF_s(\boldsymbol{\Omega}_s^*, U_{\text{inter},s}^*) = \frac{1}{2}\left[\omega_s^g \boldsymbol{G}_s(\boldsymbol{\Omega}_s^*)^T \boldsymbol{G}_s(\boldsymbol{\Omega}_s^*) + \omega_s^e (E_s(U_{\text{inter},s}^*))^2\right], \tag{4.3.7}$$

where ω_s^g and ω_s^e denote the geometry and energy residual weights for structure s. The $\frac{1}{2}$ prefactor does not affect the minimization, but its inclusion simplifies the analytical gradient expressions and provides a link to maximum likelihood estimation. Over NS training structures, the *total* merit function (TMF) is simply the sum of MF_s for all structures in the training data. Finally, the parameter estimation problem is formulated as a bilevel optimization,

$$\begin{aligned}
\bar{\boldsymbol{p}}^* = \quad & \arg\min_{\bar{\boldsymbol{p}}\subseteq\boldsymbol{p}} \quad TMF && = \sum_{s=1}^{NS} MF_s(\boldsymbol{\Omega}_s^*, U_{\text{inter},s}^*) \\
& s.t. \qquad \boldsymbol{\Omega}_s^*(\boldsymbol{p},\boldsymbol{\Omega}_s^\dagger,\boldsymbol{\theta}_s^{\text{ref}}) && = \arg\min_{\boldsymbol{\Omega}_s} U_{\text{inter},s}(\boldsymbol{\theta}_s^{\text{ref}},\boldsymbol{\Omega}_s;\boldsymbol{p}) \\
& \qquad\quad U_{\text{inter},s}^*(\boldsymbol{p},\boldsymbol{\Omega}_s^\dagger,\boldsymbol{\theta}_s^{\text{ref}}) && = U_{\text{inter},s}(\boldsymbol{\theta}_s^{\text{ref}},\boldsymbol{\Omega}_s^*(\boldsymbol{p},\boldsymbol{\Omega}_s^\dagger,\boldsymbol{\theta}_s^{\text{ref}});\boldsymbol{p}) \\
& \qquad\quad \boldsymbol{\Omega}_s^\dagger && = \boldsymbol{\Omega}_s^{\text{ref}} \\
& \qquad\quad \bar{\boldsymbol{p}} && \in [\bar{\boldsymbol{p}}^L, \bar{\boldsymbol{p}}^U]
\end{aligned} \tag{PE}$$

where $\bar{\boldsymbol{p}}$ is the subset of FF parameters to be optimized ($\bar{\boldsymbol{p}} \subseteq \boldsymbol{p}$). $\bar{\boldsymbol{p}}^L$ and $\bar{\boldsymbol{p}}^U$ denote the vectors of lower and upper bounds, respectively, for the fitted parameters. In most cases, $\bar{\boldsymbol{p}}$ can simply

be constrained to be non-negative ($\bar{\boldsymbol{p}}^L = 0$, $\bar{\boldsymbol{p}}^U = 1 \times 10^8$). The bilevel nature of this problem stems from the need to solve Problem LM to obtain the values of $\boldsymbol{\Omega}_s^*$ and $U_{\text{inter},s}^*$.

4.3.2 Derivatives for Solving the Parameter Estimation

The bilevel formulation of Problem PE introduces significant challenges, not least of which is determining accurate gradients for the outer problem. For large-scale parametrization (i.e., simultaneously fitting many parameters with large datasets), finite-difference gradients should be avoided since they require solving Problem LM an impractical number of times: over NS structures, for each $\bar{\boldsymbol{p}}$, in each finite-difference perturbation, and at every minimization step. Non-convexity inherent to Problem LM can also exacerbate numerical noise associated with such finite-difference computations, leading to less-reliable gradients. As such, analytical gradients to Problem PE are desired. Differentiation of MF_s with respect to a fitted parameter, $\bar{p}_k$, gives

$$\frac{\partial MF_s(\boldsymbol{\Omega}_s^*, U_{\text{inter},s}^*)}{\partial \bar{p}_k} = \omega_s^g \left(\frac{\partial \boldsymbol{G}_s(\boldsymbol{\Omega}_s^*)}{\partial \bar{p}_k} \right)^T \boldsymbol{G}_s(\boldsymbol{\Omega}_s^*) + \omega_s^e \left(\frac{\partial E_s(U_{\text{inter},s}^*)}{\partial \bar{p}_k} \right) (E_s(U_{\text{inter},s}^*)), \tag{4.3.8}$$

where k is the index of an estimated parameter in $\bar{\boldsymbol{p}}$. Since $\boldsymbol{\Omega}_s^*$ and $U_{\text{inter},s}^*$ are the solutions to Problem LM, they do not have closed-form analytical expressions. Thus, deriving analytical expressions for $\frac{\partial \boldsymbol{G}_s}{\partial \bar{p}_k}$ and $\frac{\partial E_s}{\partial \bar{p}_k}$ requires leveraging the optimality criterion of the inner problem.

For the energy residual, differentiation of Equation 4.3.6 with respect to $\bar{p}_k$ gives

$$\frac{\partial E_s}{\partial \bar{p}_k} = \frac{1}{U_{\text{inter},s}^{\text{ref}}} \frac{\partial U_{\text{inter},s}^*}{\partial \bar{p}_k}. \tag{4.3.9}$$

From Equations 4.3.1 and 4.3.2, $U_{\text{inter},s}^*$ has an explicit dependency on $\bar{p}_k$, but also an implicit dependence from $\boldsymbol{\Omega}_s^*$ (which is a function of $\bar{p}_k$). As such,

$$\frac{\partial U_{\text{inter},s}^*(\bar{p}_k, \boldsymbol{\Omega}_s^*(\bar{p}_k))}{\partial \bar{p}_k} = \left. \frac{\partial U_{\text{inter},s}^*}{\partial \bar{p}_k} \right|_{\boldsymbol{\Omega}_s^*} + (\nabla_{\bar{p}_k} \boldsymbol{\Omega}_s^*)^T (\nabla_{\boldsymbol{\Omega}_s} U_{\text{inter},s}). \tag{4.3.10}$$

For clarity, I have shown only the dependencies on $\bar{p}_k$. This equation can be simplified in two ways. First, by definition of Problem LM, $\nabla_{\boldsymbol{\Omega}_s} U_{\text{inter},s} = \boldsymbol{0}$. Second, from Equation 3.2.3, only U_{FF} depends on $\boldsymbol{p}$ (and by extension $\bar{p}_k$). This reduces Equation 4.3.10 to

$$\frac{\partial U_{\text{inter},s}^*(\bar{p}_k, \boldsymbol{\Omega}_s^*(\bar{p}_k))}{\partial \bar{p}_k} = \left. \frac{\partial U_{\text{FF},s}^*}{\partial \bar{p}_k} \right|_{\boldsymbol{\Omega}_s^*}, \tag{4.3.11}$$

wherein $\frac{\partial U_{\text{FF},s}^*}{\partial \bar{p}_k}$ is the value of $\frac{\partial U_{\text{FF},s}}{\partial \bar{p}_k}$ determined at the solution of Problem LM, $\boldsymbol{\Omega}_s^*$. The derivative expressions for $\frac{\partial U_{\text{FF},s}}{\partial \bar{p}_k}$, will be derived later in this section.

Deriving expressions for the derivatives of the geometry residuals is slightly more involved. I begin by differentiating Equations 4.3.3 and 4.3.4 with respect to $\bar{p}_k$ to obtain

$$\frac{\partial G_{s,i}}{\partial \bar{p}_k} = \frac{1}{X_{s,i}^{\text{ref}} \sqrt{N_{X,s}}} \frac{\partial X_{s,i}^*}{\partial \bar{p}_k}, \qquad i = 1, \cdots, N_{X,s}, \tag{4.3.12}$$

and

$$\frac{\partial G_{s,i}}{\partial \bar{p}_k} = \frac{1}{\sqrt{3(N_{r,s}-1)}} \frac{\partial Y^*_{s,i-N_{X,s}}}{\partial \bar{p}_k}, \qquad i = N_{X,s}+1, \cdots, N_{X,s}+3(N_{r,s}-1), \tag{4.3.13}$$

respectively. Since $\boldsymbol{X}^*_s \in \boldsymbol{\Omega}^*_s$, it follows that $\frac{\partial X^*_{s,i}}{\partial \bar{p}_k} \in \nabla_{\bar{p}_k} \boldsymbol{\Omega}^*_s$. Unfortunately, $Y_{s,i}$ are derived values from the elements of $\boldsymbol{\Omega}^*_s$. Applying Equation 3.1.17 with the relationship in Equation 3.1.10,

$$\hat{\boldsymbol{r}}_{Ii,s} = \boldsymbol{R}_s^{-1} \boldsymbol{A}^T_{\mathrm{rot},I,s} \boldsymbol{r}^{\mathrm{bf}}_{Ii,s} + \hat{\boldsymbol{r}}_{0,I,s}, \qquad i = 1, \cdots, \eta_{I,s} \quad I = 1, \cdots, Z'_s, \tag{4.3.14}$$

where $\hat{\boldsymbol{r}}_{Ii,s}$ are the fractional coordinates of atom i on molecule I in the asymmetric unit cell used to determine the elements of $\boldsymbol{Y}_s$. The dependence of $\boldsymbol{Y}_s$ on $\boldsymbol{\Omega}^*$ is contained within $\boldsymbol{R}_s^{-1}$ and $\boldsymbol{A}^T_{\mathrm{rot},I,s}$, according to Equations 3.1.8 and 3.1.15 respectively. As such, the elements of $\frac{\partial Y^*_{s,i}}{\partial \bar{p}_k}$ can be expressed as

$$\begin{gathered}\begin{pmatrix} \frac{\partial \hat{x}_{Ii,s}}{\partial \bar{p}_k} & \frac{\partial \hat{y}_{Ii,s}}{\partial \bar{p}_k} & \frac{\partial \hat{z}_{Ii,s}}{\partial \bar{p}_k} \end{pmatrix}^T = \left[\nabla_{\bar{p}_k} \boldsymbol{R}_s^{-1} \boldsymbol{A}^T_{\mathrm{rot},I,s} + \boldsymbol{R}_s^{-1} \nabla_{\bar{p}_k} \boldsymbol{A}^T_{\mathrm{rot},I,s} \right] \boldsymbol{r}^{\mathrm{bf}}_{Ii,s} + \nabla_{\bar{p}_k} \hat{\boldsymbol{r}}_{0,I,s}, \\ i = 1, \cdots, \eta_{I,s}, \quad I = 1, \cdots, Z'_s, \end{gathered} \tag{4.3.15}$$

where

$$\nabla_{\bar{p}_k} \boldsymbol{R}_s^{-1} = \sum_{\Omega = a_s, b_s, c_s, \alpha_s, \beta_s, \gamma_s} \nabla_\Omega \boldsymbol{R}_s^{-1} \frac{\partial \Omega}{\partial \bar{p}_k}, \tag{4.3.16}$$

$$\nabla_{\bar{p}_k} \boldsymbol{A}^T_{\mathrm{rot},I,s} = \sum_{\Omega = \phi_{I,s}, \theta_{I,s}, \psi_{I,s}} \nabla_\Omega \boldsymbol{A}^T_{\mathrm{rot},I,s} \frac{\partial \Omega}{\partial \bar{p}_k}. \tag{4.3.17}$$

Expressions for the $\nabla_\Omega \boldsymbol{R}^{-1}$ have been derived in Table C.2, while $\nabla_\Omega \boldsymbol{A}^T_{\mathrm{rot},I}$ can be obtained by differentiation of Equation 3.1.15 with respect to the Euler angles. Like $\boldsymbol{X}_s$, $\hat{\boldsymbol{r}}_{0,I,s}$ is an explicit element of $\boldsymbol{\Omega}_s$, thus $\nabla_{\bar{p}_k} \hat{\boldsymbol{r}}_{0,I,s} \in \nabla_{\bar{p}_k} \boldsymbol{\Omega}^*_s$.

Common to Equations 4.3.12–4.3.17 is a need for $\nabla_{\bar{p}_k} \boldsymbol{\Omega}^*_s$. To obtain this, I first differentiate the optimality criterion of Problem LM with respect to $\bar{p}_k$,

$$\begin{aligned} &\nabla_{\bar{p}_k} \left[\nabla_{\boldsymbol{\Omega}_s} U_{\mathrm{inter},s}(\bar{p}_k, \boldsymbol{\Omega}^*_s(\bar{p}_k)) \right] = \boldsymbol{0}, \\ &(\nabla^2_{\boldsymbol{\Omega}_s} U_{\mathrm{inter},s})(\nabla_{\bar{p}_k} \boldsymbol{\Omega}^*_s) + \nabla_{\bar{p}_k} \nabla_{\boldsymbol{\Omega}_s} U_{\mathrm{inter},s} = \boldsymbol{0}. \end{aligned} \tag{4.3.18}$$

Recognizing that only U_{FF} has dependence on $\bar{p}_k$, this can be rearranged to give

$$(\nabla^2_{\boldsymbol{\Omega}_s} U_{\mathrm{inter},s})(\nabla_{\bar{p}_k} \boldsymbol{\Omega}^*_s) = -\nabla_{\bar{p}_k} \nabla_{\boldsymbol{\Omega}_s} U_{\mathrm{FF},s}. \tag{4.3.19}$$

Equation 4.3.19 defines a system of linear equations that can be solved for the required values of $\nabla_{\bar{p}_k} \boldsymbol{\Omega}^*_s$. $\nabla^2_{\boldsymbol{\Omega}_s} U_{\mathrm{inter},s}$ is the Hessian matrix of $U_{\mathrm{inter},s}$ with respect to the rigid-body variables. Values for this can be obtained using the expressions in Sections 3.2.1.3 and 3.2.1.4. The remaining variables which have not been derived are $\frac{\partial U_{\mathrm{FF},s}}{\partial \bar{p}_k}$ and $\nabla_{\bar{p}_k} \nabla_{\boldsymbol{\Omega}_s} U_{\mathrm{FF},s}$; analytical expressions for these are formulated below.

Analytical expressions for $\frac{\partial U_{\mathrm{FF}}}{\partial \bar{p}_k}$ are straightforward to obtain. Differentiating Equation 3.2.38 with respect to $\bar{p}_k$ gives

$$\frac{\partial U_{\mathrm{FF}}}{\partial \bar{p}_k} = \frac{1}{2Z} \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} \sum_{\boldsymbol{l}=-\infty}^{\infty} \underbrace{\sum_{J=1}^{Z}}_{J\neq I \text{ if } \boldsymbol{l}=\mathbf{0}} \sum_{j=1}^{\eta_J} \left. \frac{\partial U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\mathrm{FF},P}}{\partial \bar{p}_k} \right|_{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}, \qquad P = \text{BHP or MIP.} \tag{4.3.20}$$

Depending on the functional form of the FF, different expressions must be determined for the pairwise derivative contribution. Derivative quantities for the Buckingham potential can be derived from Equation 3.2.35 as:

$$\left. \frac{\partial U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\mathrm{FF,BHP}}}{\partial A^{t,t'}} \right|_{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} = \exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right), \qquad t = t(Ii), \quad t' = t'(Jj), \tag{4.3.21}$$

$$\left. \frac{\partial U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\mathrm{FF,BHP}}}{\partial B^{t,t'}} \right|_{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} = \frac{A^{t,t'} r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{(B^{t,t'})^2} \exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right), \qquad t = t(Ii), \quad t' = t'(Jj), \tag{4.3.22}$$

$$\left. \frac{\partial U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\mathrm{FF,BHP}}}{\partial C^{t,t'}} \right|_{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} = -\frac{1}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^6}, \qquad t = t(Ii), \quad t' = t'(Jj). \tag{4.3.23}$$

Similar expressions for the Mie potential can be obtained from Equations 3.2.36 and 3.2.37, albeit with slightly more complexity due to the prefactor term:

$$\left. \frac{\partial U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\mathrm{FF,MIP}}}{\partial \varepsilon^{t,t'}} \right|_{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} = C_{\mathrm{Mie}}^{t,t'} \cdot \left[\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{r}}^{t,t'}} - \left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{a}}^{t,t'}} \right], \qquad t = t(Ii), \quad t' = t'(Jj), \tag{4.3.24}$$

$$\left. \frac{\partial U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\mathrm{FF,MIP}}}{\partial \sigma^{t,t'}} \right|_{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} = C_{\mathrm{Mie}}^{t,t'} \cdot \varepsilon^{t,t'} \cdot \left[\frac{\lambda_{\mathrm{r}}^{t,t'} (\sigma^{t,t'})^{\lambda_{\mathrm{r}}^{t,t'}-1}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^{\lambda_{\mathrm{r}}^{t,t'}}} - \frac{\lambda_{\mathrm{r}}^{t,t'} (\sigma^{t,t'})^{\lambda_{\mathrm{a}}^{t,t'}-1}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^{\lambda_{\mathrm{a}}^{t,t'}}} \right], \qquad t = t(Ii), \quad t' = t'(Jj), \tag{4.3.25}$$

$$\left. \frac{\partial U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\mathrm{FF,MIP}}}{\partial \lambda_{\mathrm{r}}^{t,t'}} \right|_{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} = \varepsilon^{t,t'} \left\{ \frac{\partial C_{\mathrm{Mie}}^{t,t'}}{\partial \lambda_{\mathrm{r}}^{t,t'}} \cdot \left[\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{r}}^{t,t'}} - \left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{a}}^{t,t'}} \right] \right.$$
$$\left. + C_{\mathrm{Mie}}^{t,t'} \cdot \left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{r}}^{t,t'}} \ln\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right) \right\}, \qquad t = t(Ii), \quad t' = t'(Jj), \tag{4.3.26}$$

$$\left. \frac{\partial U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\mathrm{FF,MIP}}}{\partial \lambda_{\mathrm{a}}^{t,t'}} \right|_{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} = \varepsilon^{t,t'} \left\{ \frac{\partial C_{\mathrm{Mie}}^{t,t'}}{\partial \lambda_{\mathrm{a}}^{t,t'}} \cdot \left[\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{r}}^{t,t'}} - \left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{a}}^{t,t'}} \right] \right.$$
$$\left. - C_{\mathrm{Mie}}^{t,t'} \cdot \left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{a}}^{t,t'}} \ln\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right) \right\}, \qquad t = t(Ii), \quad t' = t'(Jj), \tag{4.3.27}$$

wherein

$$\frac{\partial C_{\text{Mie}}^{t,t'}}{\partial \lambda_{\text{r}}^{t,t'}} = C_{\text{Mie}}^{t,t'} \cdot \left\{ \frac{1}{\lambda_{\text{r}}^{t,t'}} - \frac{1}{\lambda_{\text{r}}^{t,t'} - \lambda_{\text{a}}^{t,t'}} + \frac{\lambda_{\text{a}}^{t,t'} \ln\left(\frac{\lambda_{\text{a}}^{t,t'}}{\lambda_{\text{r}}^{t,t'}}\right)}{(\lambda_{\text{r}}^{t,t'} - \lambda_{\text{a}}^{t,t'})^2} + \frac{\lambda_{\text{a}}^{t,t'}}{\lambda_{\text{r}}^{t,t'}(\lambda_{\text{r}}^{t,t'} - \lambda_{\text{a}}^{t,t'})} \right\}, \qquad t = t(Ii), \quad t' = t'(Jj), \tag{4.3.28}$$

$$\frac{\partial C_{\text{Mie}}^{t,t'}}{\partial \lambda_{\text{a}}^{t,t'}} = C_{\text{Mie}}^{t,t'} \cdot \left\{ \frac{\ln\left(\frac{\lambda_{\text{r}}^{t,t'}}{\lambda_{\text{a}}^{t,t'}}\right)}{\lambda_{\text{r}}^{t,t'} - \lambda_{\text{a}}^{t,t'}} + \frac{\lambda_{\text{a}}^{t,t'} \ln\left(\frac{\lambda_{\text{r}}^{t,t'}}{\lambda_{\text{a}}^{t,t'}}\right)}{(\lambda_{\text{r}}^{t,t'} - \lambda_{\text{a}}^{t,t'})^2} \right\}, \qquad t = t(Ii), \quad t' = t'(Jj). \tag{4.3.29}$$

$\lambda_{\text{a}}^{t,t'}$ have been fixed to 6 throughout this work. Nevertheless, its derivative expressions are included here for completion.

To obtain the cross-derivatives of $\boldsymbol{\Omega}$ and $\bar{p}_k$, a change from the 'natural' variables (i.e., Euclidean distances) to the rigid-body variables, $\boldsymbol{\Omega}$, must be accommodated. First, I define

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}} \nabla_{\bar{p}_k} U_{\text{FF}} = \frac{1}{2Z} \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} \sum_{\boldsymbol{l}=-\boldsymbol{\infty}}^{\boldsymbol{\infty}} \underbrace{\sum_{J=1}^{Z}}_{J \neq I \text{ if } \boldsymbol{l}=\mathbf{0}} \sum_{j=1}^{\eta_J} \nabla_{\boldsymbol{r}_{\mathbf{0}Ii}} \nabla_{\bar{p}_k} U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\text{FF},P}, \qquad P = \text{BHP or MIP}, \tag{4.3.30}$$

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Jj}} \nabla_{\bar{p}_k} U_{\text{FF}} = \frac{1}{2Z} \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} \sum_{\boldsymbol{l}=-\boldsymbol{\infty}}^{\boldsymbol{\infty}} \underbrace{\sum_{J=1}^{Z}}_{J \neq I \text{ if } \boldsymbol{l}=\mathbf{0}} \sum_{j=1}^{\eta_J} \nabla_{\boldsymbol{r}_{\mathbf{0}Jj}} \nabla_{\bar{p}_k} U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\text{FF},P}, \qquad P = \text{BHP or MIP}, \tag{4.3.31}$$

and

$$\nabla_X \nabla_{\bar{p}_k} U_{\text{FF}} = \frac{1}{2Z} \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} \sum_{\boldsymbol{l}=-\boldsymbol{\infty}}^{\boldsymbol{\infty}} \underbrace{\sum_{J=1}^{Z}}_{J \neq I \text{ if } \boldsymbol{l}=\mathbf{0}} \sum_{j=1}^{\eta_J} \nabla_X \nabla_{\bar{p}_k} U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\text{FF},P}, \qquad X \in \{a, b, c, \alpha, \beta, \gamma\}, \quad P = \text{BHP or MIP}. \tag{4.3.32}$$

It can be shown that

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Ii}} \nabla_{\bar{p}_k} U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\text{FF},P} = -\nabla_{\boldsymbol{r}_{\mathbf{0}Jj}} \nabla_{\bar{p}_k} U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\text{FF},P} = -\nabla_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \nabla_{\bar{p}_k} U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\text{FF},P}, \qquad P = \text{BHP or MIP}, \tag{4.3.33}$$

$$\nabla_X \nabla_{\bar{p}_k} U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\text{FF},P} = (\nabla_X \boldsymbol{R} \cdot \boldsymbol{l})^T \nabla_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \nabla_{\bar{p}_k} U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\text{FF},P}, \qquad X \in \{a, b, c, \alpha, \beta, \gamma\}, \quad P = \text{BHP or MIP}, \tag{4.3.34}$$

and

$$\nabla_{\boldsymbol{r}_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \nabla_{\bar{p}_k} U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\text{FF},P} = \frac{(\boldsymbol{r}_{\mathbf{0}Jj} + \boldsymbol{R} \cdot \boldsymbol{l} - \boldsymbol{r}_{\mathbf{0}Ii})}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \frac{\partial^2 U_{\mathbf{0}Ii,\boldsymbol{l}Jj}^{\text{FF},P}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj} \partial \bar{p}_k}, \qquad P = \text{BHP or MIP}. \tag{4.3.35}$$

The cross-derivative terms, $\frac{\partial^2 U^{\mathrm{FF},P}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\partial \bar{p}_k}$, can be determined by differentiating (with respect to $r_{\mathbf{0}Ii,\boldsymbol{l}Jj}$) the previously derived first-derivatives of the BHP and MIP. For the BHP, these are

$$\frac{\partial^2 U^{\mathrm{FF,BHP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\partial A^{t,t'}} = -\frac{1}{B^{t,t'}}\exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right), \qquad t = t(Ii), \quad t' = t'(Jj), \tag{4.3.36}$$

$$\frac{\partial^2 U^{\mathrm{FF,BHP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\partial B^{t,t'}} = \frac{A^{t,t'}}{(B^{t,t'})^2}\left(1-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right)\exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right), \qquad t = t(Ii), \quad t' = t'(Jj), \tag{4.3.37}$$

$$\frac{\partial^2 U^{\mathrm{FF,BHP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\partial C^{t,t'}} = \frac{6}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^7}, \qquad t = t(Ii), \quad t' = t'(Jj), \tag{4.3.38}$$

and for the MIP

$$\frac{\partial^2 U^{\mathrm{FF,MIP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\partial \varepsilon^{t,t'}} = C^{t,t'}_{\mathrm{Mie}}\cdot\left[\frac{\lambda_{\mathrm{a}}^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{a}}^{t,t'}} - \frac{\lambda_{\mathrm{r}}^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{r}}^{t,t'}}\right], \qquad t = t(Ii), \quad t' = t'(Jj), \tag{4.3.39}$$

$$\frac{\partial^2 U^{\mathrm{FF,MIP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\partial \sigma^{t,t'}} = C^{t,t'}_{\mathrm{Mie}}\cdot\varepsilon^{t,t'}\cdot\left[\frac{(\lambda_{\mathrm{a}}^{t,t'})^2}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\frac{(\sigma^{t,t'})^{\lambda_{\mathrm{a}}^{t,t'}-1}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^{\lambda_{\mathrm{a}}^{t,t'}}} - \frac{(\lambda_{\mathrm{r}}^{t,t'})^2}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\frac{(\sigma^{t,t'})^{\lambda_{\mathrm{r}}^{t,t'}-1}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^{\lambda_{\mathrm{r}}^{t,t'}}}\right], \qquad t = t(Ii), \quad t' = t'(Jj), \tag{4.3.40}$$

$$\frac{\partial^2 U^{\mathrm{FF,MIP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\partial \lambda_{\mathrm{r}}^{t,t'}} = \varepsilon^{t,t'}\left\{\frac{\partial C^{t,t'}_{\mathrm{Mie}}}{\partial \lambda_{\mathrm{r}}^{t,t'}}\cdot\left[\frac{\lambda_{\mathrm{a}}^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{a}}^{t,t'}} - \frac{\lambda_{\mathrm{r}}^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{r}}^{t,t'}}\right]\right.$$
$$\left. - C^{t,t'}_{\mathrm{Mie}}\cdot\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{r}}^{t,t'}}\left[\frac{1+\lambda_{\mathrm{r}}^{t,t'}\ln\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right]\right\}, \qquad t = t(Ii), \quad t' = t'(Jj), \tag{4.3.41}$$

$$\frac{\partial^2 U^{\mathrm{FF,MIP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\partial \lambda_{\mathrm{a}}^{t,t'}} = \varepsilon^{t,t'}\left\{\frac{\partial C^{t,t'}_{\mathrm{Mie}}}{\partial \lambda_{\mathrm{a}}^{t,t'}}\cdot\left[\frac{\lambda_{\mathrm{a}}^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{a}}^{t,t'}} - \frac{\lambda_{\mathrm{r}}^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{r}}^{t,t'}}\right]\right.$$
$$\left. + C^{t,t'}_{\mathrm{Mie}}\cdot\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)^{\lambda_{\mathrm{a}}^{t,t'}}\left[\frac{1+\lambda_{\mathrm{a}}^{t,t'}\ln\left(\frac{\sigma^{t,t'}}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right)}{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\right]\right\}, \qquad t = t(Ii), \quad t' = t'(Jj). \tag{4.3.42}$$

Finally, following the definitions given in Equation 3.2.84 for $\boldsymbol{\Omega}$, $\boldsymbol{\mu}$, and $\boldsymbol{\nu}$, the cross-derivatives with respect to $\boldsymbol{\Omega}$ can be obtained as

$$\frac{\partial^2 U_{\mathrm{FF}}}{\partial \mu_i \partial \bar{p}_k} = \nabla_{\mu_i} \boldsymbol{r}^T \nabla_{\boldsymbol{r}} \nabla_{\bar{p}_k} U_{\mathrm{FF}}, \qquad \mu_i \in \boldsymbol{\mu}, \tag{4.3.43}$$

and

$$\frac{\partial^2 U_{\mathrm{FF}}}{\partial \nu_i \partial \bar{p}_k} = \nabla_{\nu_i} \boldsymbol{r}^T \nabla_{\boldsymbol{r}} \nabla_{\bar{p}_k} U_{\mathrm{FF}}, + \nabla_{\nu_i} \boldsymbol{X}^T \nabla_{\boldsymbol{X}} \nabla_{\bar{p}_k} U_{\mathrm{FF}}, \qquad \nu_i \in \boldsymbol{\nu}. \tag{4.3.44}$$

Expressions for $\nabla^T_{\mu_i} \boldsymbol{r}$, $\nabla^T_{\nu_i} \boldsymbol{r}$, and $\nabla_{\nu_i} \boldsymbol{X}^T$ were derived in Section 3.2.1.3. Values obtained from the two preceding equations directly form the elements of $\nabla_{\bar{p}_k} \nabla_{\boldsymbol{\Omega}_s} U_{\mathrm{FF},s}$.

4.3.3 The CrystalEstimator Algorithm

Having described the mathematical formulation of the parameter estimation problem, I now describe its implementation in the CrystalEstimator program [1, 2].

The CrystalEstimator (CE) algorithm is illustrated in Figure 4.3. Most simply, CE will attempt to minimize the TMF from an initial vector of parameter estimates, $\bar{\boldsymbol{p}}^{[0]}$, using the E04USA subroutine. At the kth major iteration of this optimization, function and gradient evaluations of the TMF are conducted by corresponding calls to CSO–RM (Figure 4.3, D1 and D2). $\bar{\boldsymbol{p}}$ is altered based on linesearches performed by the E04USA subroutine, and this cycle repeats until optimality is reached. Unfortunately, except for small-scale parametrization, this basic implementation usually cannot contend with the complexity of Problem PE. To achieve the desired scale of parameter estimation, several adaptations have been added to CE.

Firstly, the bilevel structure of Problem PE results in discontinuous behavior on the TMF hypersurface [2, 251], likely due to the non-convexity of Problem LM. As $\bar{\boldsymbol{p}}$ are varied in the outer problem, the inner U_{inter} energy surface is altered and (for a given $\boldsymbol{\Omega}^{\dagger}_s$) Problem LM may converge to distinct minima. Drastic changes in $\boldsymbol{\Omega}^{*}_s$ and $U_{\mathrm{inter,s}}(\boldsymbol{\Omega}^{*}_s)$ as a result of these topological changes in the inner problem's solution surface will propagate as discontinuities in the outer problem. To suppress these discontinuities, CE has facility to employ 'dynamic starting points'. When this is used, $\boldsymbol{\Omega}^{\dagger,[k]}_s = \boldsymbol{\Omega}^{\mathrm{ref}}_s$ is applied only in the first TMF evaluation by E04USA (i.e., when $k = 0$). In all subsequent steps, $\boldsymbol{\Omega}^{\dagger,[k]}_s$ is dynamically updated to the optimal geometry obtained in the preceding TMF evaluation ($\boldsymbol{\Omega}^{*,[k-1]}_s$) (Figure 4.3, C1). This biases each TMF evaluation to find the same minima in the inner problem, reducing the propensity of confronting TMF discontinuities. Admittedly, this is only a 'band-aid' solution, and discontinuities can still be encountered under other circumstances [1, 2]. Nevertheless, using dynamic starting points significantly improves the consistency of Problem LM's solution.

Recognizing that in Problem PE, $\boldsymbol{\Omega}^{\dagger}_s = \boldsymbol{\Omega}^{\mathrm{ref}}_s$, it is necessary to accommodate the drift in $\boldsymbol{\Omega}^{\dagger}_s$ stemming from dynamic starting points. To ensure congruence with Problem PE, following the termination of E04USA, the TMF is re-evaluated using the original reference geometries as starting points (Figure 4.3, B2). In short, if E04USA terminates after k major iterations, a $(k{+}1)^{\mathrm{th}}$ TMF evaluation is conducted from $\boldsymbol{\Omega}^{\mathrm{ref}}_s$. If E04USA had converged to a local minimum, and its final TMF value is consistent with the re-evaluated value, this indicates that the drift in $\boldsymbol{\Omega}^{\dagger}_s$ was negligible and Problem PE has been solved (Figure 4.3, B3).

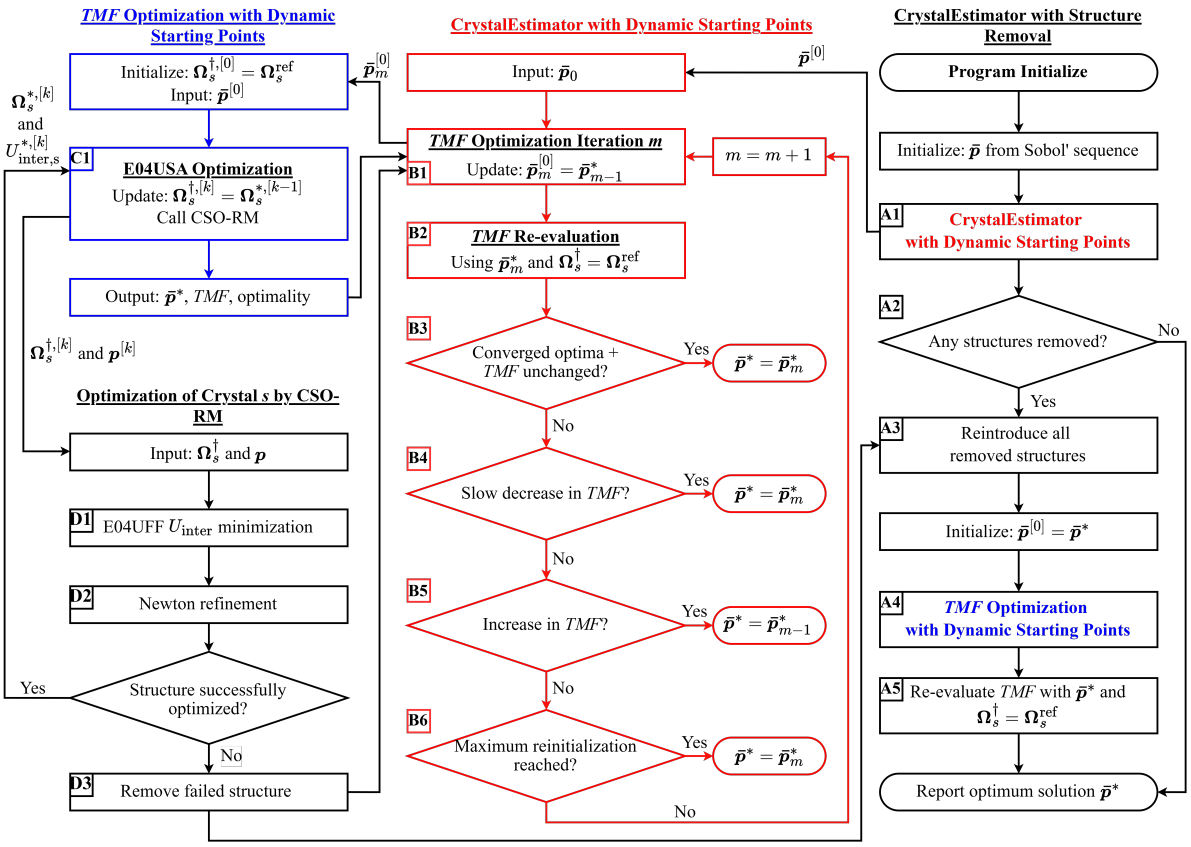


Figure 4.3: Schematic of the CrystalEstimator algorithm with dynamic starting points, algorithm reinitialization, and temporary structure removal. Key stages are enumerated and detailed in the main text.

More often, in large parameter estimation problems, either the E04USA algorithm is unsuccessful in converging to a local minimum and/or the TMF changes significantly upon re-evaluation. In this case, TMF optimization is reattempted, wherein the mth TMF optimization is initialized from the parameter estimates obtained in the $(m-1)^{\text{th}}$ optimization ($\bar{\boldsymbol{p}}_m^{[0]} = \bar{\boldsymbol{p}}_{m-1}^*$) (Figure 4.3, B1). Before the $(m+1)^{\text{th}}$ attempt, some additional termination criteria are considered (Figure 4.3, B4–B6). First, if the re-evaluated TMF value from the mth and $(m-1)^{\text{th}}$ optimization attempt decreases only marginally, or if the TMF value from the mth attempt increases from the $(m-1)^{\text{th}}$ attempt, CE will terminate. Both of these cases indicate that CE is likely in close proximity to the optimum $\bar{\boldsymbol{p}}$, and to reduce computational costs it is most prudent to just accept the best solution found so far. Whilst termination by these criteria do not resolve the non-convergence of E04USA, it is important to reiterate the disconnect between Problem PE and the E04USA minimization, due to dynamic starting points. The E04USA minimizations do not directly solve Problem PE and merely drive $\bar{\boldsymbol{p}}$ towards the true optimum. As such, convergence of the re-evaluated TMF value is a better proxy for Problem PE. Lastly, if a user-specified reinitialization limit is reached, CE will also terminate. Unless one of these four termination criteria are met, the TMF optimization is reattempted. In each attempt, the dynamic starting points are reinitialized so that the first TMF evaluation by E04USA uses $\boldsymbol{\Omega}_s^{\dagger,[0]} = \boldsymbol{\Omega}_s^{\text{ref}}$.

During the TMF optimization, training structures may inadvertently fail CSO–RM minimization with the intermediate parameter values, such that their associated residuals cannot be evaluated. In such cases, CE will (temporarily) remove the problematic structures from the training data, and restart the last TMF optimization attempt with a reduced training set (Figure 4.3, D3 → B1). This enables the parameter estimation to progress, and ideally 'escape' from regions of $\bar{\boldsymbol{p}}$ where those structures could not be minimized. Upon convergence (B3–B6) with the reduced training set, all previously removed structures are reintroduced into the training data. Thereafter, a final TMF optimization (and re-evaluation) is attempted (Figure 4.3, A3–A5). No further allowance for reinitialization of the TMF optimization is given at this point, and the final set of parameter estimates is reported.

The CE algorithm is relatively efficient due to the implementation of analytical gradient expressions. However, the need to solve Problem LM for every training structure can become a serious bottleneck when large training sets are used. For practicable parameter estimation, significant speed-up is achieved in two ways. Firstly, as only rigid minimizations are being performed, $\boldsymbol{Q}$ (point-charges or multipoles) for all training structures can be generated and stored, in advance of the parameter estimation. These are provided as inputs to CE such that U_{elec} can be evaluated without any QM calculations during the parameter estimation. Secondly, MPI parallelization has also been implemented into CE. Using MPI, the minimization of training structures can be distributed over multiple processes and independently executed [2].

Problem PE is extremely non-convex [213, 251], hence to find good parameter estimates it is inadequate to execute the protocol shown in Figure 4.3 only once. Instead, CE is designed to be used within a multistart workflow, wherein multiple instances of the parameter estimation are initialized from different parameter starting vectors, $\bar{\boldsymbol{p}}^{[0]}$. A Sobol' sequence [103] is used to generate $\bar{\boldsymbol{p}}^{[0]}$ by sampling the parameter space centered around a reference set of $\bar{\boldsymbol{p}}$. The choice of reference $\bar{\boldsymbol{p}}$ is discussed in more detail later. Since Sobol' sequences are deterministic, $\bar{\boldsymbol{p}}^{[0]}$ can be independently generated, and individual parameter estimations may be attempted in parallel. Once all parameter estimations have terminated, their individual solutions are clustered into distinct minima on the basis of their TMF and $\bar{\boldsymbol{p}}^{*}$ values. The $\bar{\boldsymbol{p}}^{*}$ which gives the best TMF value is accepted as the optimal parameter set.

The algorithmic performance of the CE program has been studied elsewhere [1, 2]. In this work, I will primarily use CEv1.4.2.2. This version is identical to the implementation described in Refs [1, 2], except for minor changes that enable interfacing with CSO–RMv1.4.1.

4.4 Summary

In this chapter, I have described the various programs and libraries that will be used throughout the rest of this work. Central to this is the Imperial MSE CSP Suite, which are the set of programs developed by our group for various CSP activities. The Imperial MSE CSP Suite employs the HAIEFF energy model and mathematical formalism detailed in Chapter 3 to conduct lattice energy minimizations across different stages of the hierarchical CSP workflow. In service

of this, these codes interface with a range of programs/libraries, internal and external to the suite. Finally, I have discussed the parameter estimation formulation that will be used in the subsequent chapters. This has been implemented into our auxiliary program, CrystalEstimator, that is used for large-scale parameter estimation of HAIEFF FFs.

Having described the theory and programs underpinning the present work, I move on to address the issues identified in Chapter 2. In the next two chapters, I first explore the problems with HAIEFF FF parametrization discussed in Section 2.5.1.

Chapter 5

Force-Field Parametrization: Enriching the Reference Data

As summarized in Section 2.6, there are three key deficiencies related to HAIEFF FF parametrization (and deployment) that I hope to address in the current work:

1. HAIEFF FF parametrization is viewed as an inaccessible task. Without reliable tools to conduct parameter estimation and difficulty assembling adequate data for FF fitting, there is an aversion towards routine FF-fitting.
2. Perhaps a consequence of this aversion towards routine FF-fitting, very little evolution in the HAIEFF FF model definition (e.g., atom-typing, FF functional form) has been attempted. This is in contrast to other FF-based methods, where there are no shortage of distinct FFs. Likely, there are unrealized improvements in FF-performance that could be achieved with some exploration in this area.
3. Aversion to FF-fitting also contributes to the suboptimal deployment of FFs. Rather than fit a suitable FF, *ad hoc* utilization of the available FFs is common.

In this chapter, I address the curation of suitable reference data, while the other two issues are tackled later. The results in this chapter adapt this author's contributions from the published work in Ref [1].

5.1 Background

To an extent, accessibility to FF parametrization has been addressed by previous work in our group. The CrystalEstimator algorithm [1, 2] was developed to enable routine large-scale FF parametrization. Using CrystalEstimator, solving parameter estimation problems involving the simultaneous fitting of > 60 FF parameters, ~450 crystal structures, and ~28,500 data points were demonstrated to be possible [1, 2]. This represents a significant increase in problem size relative to previous attempts at HAIEFF FF parametrization [169, 213].

To use CrystalEstimator for the fitting of *transferable* HAIEFF FFs, an adequately large and diverse dataset must be available. To that end, Bowskill [2] curated a dataset of 445 organic crystal structures (denoted 'DB–445' here) spanning the common organic elements (H, C, N, O, F, S, Cl), and encompassing crystal structures with and without intermolecular hydrogen

bonds. These crystal structures were evaluated by periodic DFT-D, such that a complete set of geometry and energy data was determined for all 445 structures. As discussed before, the decision to curate synthetic data is justified in several ways. The scale of parameter estimation to be performed requires significantly more (geometric and energetic) data than can be realistically obtained from experimental sources alone, especially if one also factors in the (often) poor quality of experimental data. Thermal effects inherent to experimental data are also undesirable when fitting FFs intended for use in CSP_0. Finally, using DFT-D as the 'ground truth' is closer aligned with the multistage CSP workflow, wherein periodic DFT-D is typically employed for the final ranking of polymorphs. While the DB–445 dataset [2] forms a respectable foundation of reference data, there remains room for development. To fit transferable FFs, diversity in the reference data is paramount. However, Bowskill's approach towards curating his dataset was mostly based on subjective judgment and selection of crystals containing molecules from a range of 'typical compound classes' (e.g., amines, acids etc.) [2]. To my knowledge, no quantitative measure of diversity was employed in the curation of the DB–445. Also, structures with intermolecular hydrogen bonds are underrepresented in the DB–445 dataset (only ~30%), despite hydrogen bond design-principles being important to synthetic chemistry (e.g., in synthon-based cocrystal design [63, 252]). Here, I aim to more-formally quantify the diversity of the reference data and use this to guide the addition of new data that fills 'gaps' in the DB–445 [2].

5.2 Quantifying Chemical Diversity

5.2.1 Atomic Environments

Since HAIEFF FFs are based on the pairwise interaction between atom types, a natural choice to visualize chemical diversity is the pairwise matrix of atomic environments. To fit widely transferable FFs, this analysis should be done over *all* atomic environments of interest. A 'high-level' matrix using only the intended atom types would obfuscate if any atomic environments are underrepresented by their encompassing atom type. As such, I define a set of atomic environments as follows (Table 5.1). The atomic environment label begins with the element of the target atom. The three subscripts that follow correspond to its coordination number, valency, and the bonded-neighbor element respectively. '*' represent wildcards and indicate that a feature is disregarded in the atomic environment definition. Atomic environments defined in this manner cover a wide range of chemical environments/groups. Critically, identifying crystal structures containing each atomic environment is relatively easy to query using the associated SMARTS strings in conjunction with the CCDC's Python API.

Performing this diversity quantification on the DB–445 [2], gaps in this dataset become apparent (Figure 5.1). As discussed before, there is an underrepresentation of hydrogen bonding crystal structures, as indicated by the large proportions of yellow and orange entries for $H_{*,*,N}$ and $H_{*,*,O}$. This is particularly true of hydrogen bonding systems containing 'less common' organic elements like F, S, and Cl. Sulphur-containing crystals are sparse across the entire

Table 5.1: Description of the atomic environments (AEs) used for diversity analysis.

AE	SMARTS	Description	Chemical Environment
$C_{2,*,*}$	[#6X2]	Doubly-coordinated C	sp carbons
$C_{3,*,*}$	[#6X3]	Triply-coordinated C	sp^2 carbons
$C_{4,*,*}$	[#6X4]	Quadruply-coordinated C	sp^3 carbons
$O_{1,*,*}$	[#8X1]	Singly-coordinated O	Carbonyls/Esters/Carboxylic acids
$O_{2,*,*}$	[#8X2]	Doubly-coordinated O	Hydroxyls/Ethers/Esters/Carboxylic acids/Oxazoles
$N_{1,3,*}$	[#7X1v3]	Singly-coordinated N, valency 3	Nitriles
$N_{2,3,*}$	[#7X2v3]	Doubly-coordinated N, valency 3	Aromatic nitrogens
$N_{3,3,*}$	[#7X3v3]	Triply-coordinated N, valency 3	Amines/Amides
$N_{3,5,*}$	[#7X3v5]	Triply-coordinated N, valency 5	Nitros
$F_{*,*,*}$	[#9X1]	Singly-coordinated F	Regular fluorines
$Cl_{*,*,*}$	[#17X1]	Singly-coordinated Cl	Regular chlorines
$S_{*,*,O}$	[#16]=[#8]	S double-bonded to O	Sulfoxides/Sulfones/Sulfonamides
$S_{1,*,*}$	[#16X1]	Singly-coordinated S	Thiocarbonyls/Thioamides
$S_{2,*,*}$	[#16X2]	Doubly-coordinated S	Thiols/Sulfides/Disulfides/Thioesters/ Thioazoles
$H_{*,*,N}$	[H][#7]	H bonded to N	Amines/Amides
$H_{*,*,O}$	[H][#8]	H bonded to O	Hydroxyls
$H_{*,*,S}$	[H][#16]	H bonded to S	Thiols
$H_{*,*,C}$	[H][!#7;!#8;!#16]	H not bonded to N, O, or S	All other Hs are carbon-bonded

matrix, while thiol-containing structures (indicated by the presence of $H_{*,*,S}$) are completely absent. Finally, even among the 'more common' organic elements (e.g., C, N, and O), there is clear bias towards certain atomic environments and chemical moieties. In short, the atomic environment (AE) matrix corroborates my concerns with Bowskill's dataset: chemical diversity is unevenly sampled and the DB–445 should be supplemented with more data.

5.2.2 Intermolecular Hydrogen Bonding

In Bowskill's work [2], structures with intermolecular hydrogen bonds were identified solely from the presence of nitrogen- or oxygen-bonded hydrogens. While it is expected that these systems form intermolecular hydrogen bonds, this is not guaranteed, particularly in crystals where there might be a significant energy penalty to break an intramolecular hydrogen bond in service of forming an intermolecular hydrogen bond. In lieu of this, I identify the presence of intermolecular hydrogen bonds using the 'hbonds' crystal property, accessible from the CCDC's Python API. 'hbonds' generates a list of unique crystallographic intermolecular hydrogen bonds from two geometric criteria: X–H···Y distance and angle. These geometric criteria are commonly used to define hydrogen bonds and are generally considered a practical way to evaluate hydrogen bond formation [253]. In my analysis, I have kept the default thresholds for X–H···Y distance, but expanded the lower bound of the X–H···Y angle tolerance from 120° to 100°. Among the DB–445 structures [2], 132 contain $H_{*,*,N}$ and/or $H_{*,*,O}$, yet using 'hbonds' only 126 are flagged with intermolecular hydrogen bonds. This reiterates that the DB–445 dataset has underrepresentation

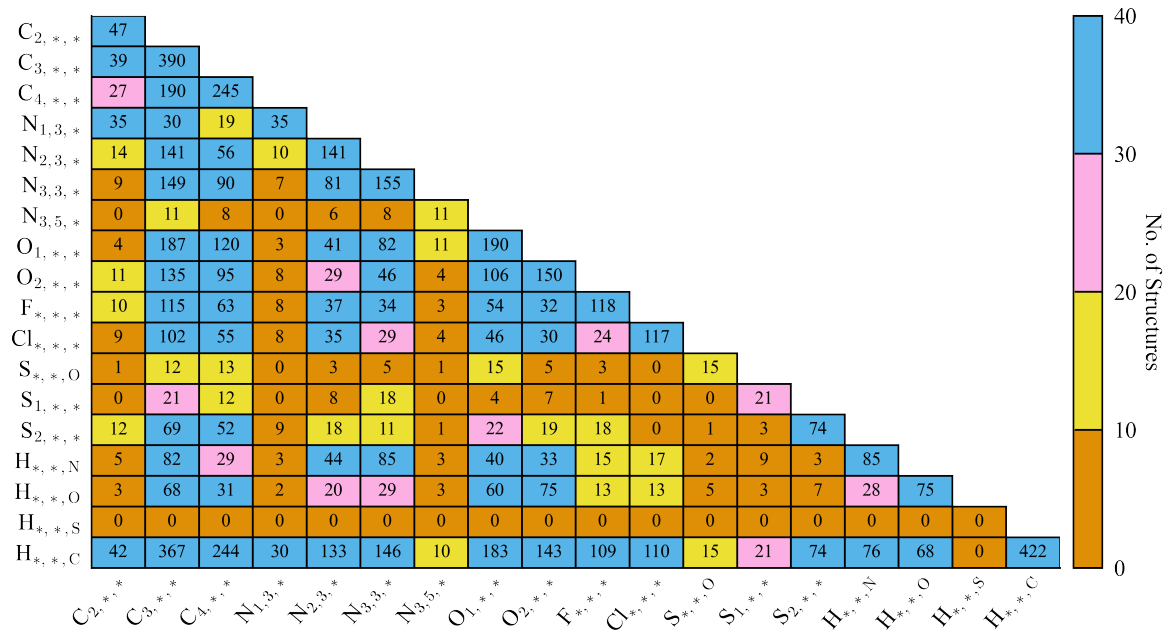


Figure 5.1: Tally of structures containing each pairwise atomic-interaction within the DB–445 dataset [2]. Each matrix entry corresponds to the total number of structures which contain the pair of atoms on the row and column. Cells are colored orange (<10), yellow ($10 \leq$ and ≤ 19), pink ($20 \leq$ and ≤ 29), and blue (≥ 30) according to their structure counts.

of systems forming intermolecular hydrogen bonds.

5.3 Generating Atomic Environment Datasets

5.3.1 DFT-D Methodology

For the collective dataset (i.e., DB–445 + structures appended here) to be useful in parameter fitting, I must apply an identical DFT-D methodology as Bowskill [1, 2] to generate new reference data. Failing to do so would lead to spurious artefacts from inconsistencies in the quantum chemistry method. Bowskill's DFT-D methodology [1, 2] is described below and applied to all candidate crystal structures.

All periodic DFT-D calculations are performed using the Vienna *ab initio* software package (VASPv5.4.4) [245–249]. Due to their good performance in various benchmarking studies [132, 143], the TPSS [130] meta-GGA functional and D3 [85] dispersion-correction are chosen alongside PAW pseudo-potentials. A large plane-wave basis set energy cutoff of 1000 eV is used with a tight Γ-centred **k**-point mesh of at most $0.025 \times 2\pi$ Å^{-1} reciprocal-distance between each **k**-point. Each candidate crystal structure is subject to geometry relaxation, varying its unit cell parameters and atomic positions until all forces are less than 0.01 eV Å^{-1}. Following convergence of the geometry optimization, DFT-D computations are restarted, sometimes multiple times, until the number of geometry-optimization iterations in a single run is relatively small (<20). These repeat attempts help to minimize the effects of Pulay stresses, that arise when the cell volume

changes, and ensure the geometry and energy are tightly converged.

To obtain the desired reference energy values used in Equation 4.3.6, it is also necessary to evaluate the monomer energies for all molecules in the asymmetric unit cell. From the relaxed crystal structure, each distinct molecule, m, in the asymmetric unit cell is extracted and placed in a fixed cubic cell. This cell is sized so as to attain a minimum inter-molecule contact distance of at least 16 Å between monomers in adjacent cells. At this separation, intermolecular interactions should contribute negligibly to the monomer energy, while computational costs are kept reasonable with only a moderate cell volume. A single-point energy evaluation is performed to obtain the monomer energy at the crystalline conformation ($U^{\text{crys}}_{\text{mono},m}$), while a geometry optimization (atomic positions only) is conducted to obtain the corresponding gas-phase energy of the monomer ($U^{\text{gp}}_{\text{mono},m}$). Reference energies for a crystal structure can then be obtained as

$$U^{\text{ref}}_{\text{inter}} = \frac{U_{\text{vasp}}}{Z} - \frac{\sum_{m=1}^{Z'} U^{\text{crys}}_{\text{mono},m}}{Z'}, \tag{5.3.1}$$

$$\Delta U^{\text{ref}}_{\text{intra}} = \frac{\sum_{m=1}^{Z'} U^{\text{crys}}_{\text{mono},m}}{Z'} - \frac{\sum_{m=1}^{Z'} U^{\text{gp}}_{\text{mono},m}}{Z'}, \tag{5.3.2}$$

$$U^{\text{ref}}_{\text{latt}} = \Delta U^{\text{ref}}_{\text{intra}} + U^{\text{ref}}_{\text{inter}}, \tag{5.3.3}$$

where U_{vasp} is the optimal energy of the crystal reported by VASP, while $U^{\text{ref}}_{\text{inter}}$, $\Delta U^{\text{ref}}_{\text{intra}}$, and $U^{\text{ref}}_{\text{latt}}$ are the reference intermolecular, intramolecular and lattice energy of the crystal, as defined in the HAIEFF framework. From the output geometries, the symmetry relations for each of the optimized structures are recalculated using PLATON [254]. In some cases, the symmetry of the crystal cannot be reduced to that of the input structure, in which case it remains as $Z' > 1$.

5.3.2 Curation of Crystal Structures

Since synthetic reference data is being generated, hypothetical structures could be used as inputs into the DFT-D calculations. This is often done when fitting TMFFs [99, 100], as relying solely on experimental structures is too limiting for the parametrization of molecule-specific FFs. Here, I follow Bowskill [2] and select only experimentally reported crystal structures from the CSD [255]. From Figure 5.1, I focus data generation efforts towards structures containing AE pairs that are sparse in the DB–445 AE-matrix. To facilitate this, I employ the CCDC Python API [241] to screen the CSD for crystal structures that fulfill several criteria.

Firstly, given that each crystal structure will be subject to periodic DFT-D, it is desirable to limit the chosen crystal structures to a moderate size. As such, crystals with $Z' \leq 1$ and $Z \leq 4$ are preferentially selected. The molecular size in each candidate crystal is also constrained based on their atom counts. Secondly, to limit the flexibility of the candidate molecules, the number of sp_3 carbons present in each molecular structure is restricted. Thirdly, depending on the target pairwise AE interaction, the molecular structures are required to contain (or exclude) specific atomic environments (identified by their SMARTS strings in Table 5.1). Candidates are disqualified if they exhibit disorder and/or are already present in the DB–445; polymorphic forms

Table 5.2: List of atom environment reference datasets obtained from DFT-D computations.

Reference Set	Target Atomic Environments	Number of Structures
AE1	Any carbon but preferentially $C_{2,*,*}$	14
AE2	$S_{*,*,O}$/ $S_{1,*,*}$/Thiols	21
AE3	$S_{*,*,O}$	12
AE4	Any nitrogen	17
AE5	Any oxygen	10
AE6	Nitro groups ($N_{3,5,*}$ + $O_{1,*,*}$)	8
AE7	$S_{1,*,*}$ + $H_{*,*,N}$	12
AE8	$S_{1,*,*}$ + $H_{*,*,O}$	10
AE9	$S_{2,*,*}$ + $H_{*,*,N}$	13
AE10	$S_{2,*,*}$ + $H_{*,*,O}$	16
AE11	$S_{*,*,O}$ + Any nitrogen + Preferentially with $H_{*,*,N}$	23
AE12	$S_{*,*,O}$ + Any oxygen + Preferentially with $H_{*,*,O}$	10
AE13	$S_{*,*,O}$ + Any halogen	25
AE14	$H_{*,*,O}$ + Any nitrogen	61
AE15	$S_{2,*,*}$ + $Cl_{*,*,*}$	14
AE16	$S_{2,*,*}$ + $S_{*,*,O}$	12
AE17	$Cl_{*,*,*}$ + $H_{*,*,O}$ or $H_{*,*,N}$	15
AE18	$F_{*,*,*}$ + $H_{*,*,O}$ or $H_{*,*,N}$	17
Total		310

that have not been previously included are allowed. If a sufficient number of candidates cannot be identified for a given search, some of these constraints are relaxed. The treatment of each structure by DFT-D permits some inaccuracy in the experimental geometry, thus no restrictions are placed on the experimental method used to resolve the crystal structure. Moreover, for each crystal, I require only *heavy*-atom positions to be resolved, while hydrogens can be missing. Any absent hydrogens are placed in standard positions using Mercury's in-built molecule editor. Some structures that require this treatment include 2-mercapto-pyridine-4-carboxylic acid (CSD Refcode: MPYDCY) and cis-2-butene-episulfone (CSD Refcode: BUTSUL10). This is permissible since hydrogen positions determined experimentally are generally considered to be not very accurate, and in any case the hydrogen positions would be 'corrected' by the DFT-D treatment of the crystal structure. When the DFT-D optimized crystal geometries are compared to their experimental counterparts, the CCDC's COMPACK tool [242] also ignores hydrogen positions. Candidates flagged by the Python API, under these criteria, are subject to a final pass of manual screening to weed out highly irregular structures or structures that continue to exhibit large degrees of flexibility. Despite restrictions on the number of sp_3 carbons, excessive flexibility can still emerge if all of the permitted sp_3 carbons are bonded in succession. The finalized list of structures are then used as inputs in the DFT-D workflow.

The AE datasets that are generated in this way are summarized in Table 5.2. The initial atomic environment groups (AE1–AE6) are loosely defined and usually cover multiple or all AEs for a given element. This 'shotgun' approach is employed early in the data curation to rapidly fill large gaps in the AE-matrix. Thereafter, I focus searches towards more-specific

AE-interaction pairs. As discussed before, the atomic environments most sparsely populated in the DB–445 dataset are those involving sulphurs and halogens. This is intuitive since these elements are generally less common in the CSD. Additionally, since the DB–445 dataset has a greater proportion of non-hydrogen bonding crystals, the AE datasets prioritize hydrogen bonding structures, which must contain either oxygen- or nitrogen-bonded hydrogens. These trends are evident from the target AEs used in AE7–AE18. Crucially, the sulphur–chlorine interaction, which was missing from the DB–445, has been introduced here. Finally, the lack of corresponding '$S_{1,*,*}$' searches for AE15 and AE16 is not an oversight. Searches were attempted for '$S_{1,*,*}$ + $F_{*,*,*}$', '$S_{1,*,*}$ + $Cl_{*,*,*}$', and '$S_{1,*,*}$ + $S_{*,*,O}$', but these did not yield a sufficient number of viable candidates, thus these groups were ultimately discarded from the dataset.

The combined dataset of 755 crystal structures (DB–445 + 310 newly generated structures) has been termed the CE–755 dataset, and is published elsewhere [256]. Ref [256] includes all raw data related to the crystal energies, the geometric deviations between the DFT-D optimized structures and experimental structures, and .res files for the DFT-D optimized crystal structures.

5.4 The CE–755 Dataset

The CE–755 dataset comprises 755 unique crystals, that each have a DFT-D resolved crystal structure and reference energy, $U_{\text{inter}}^{\text{ref}}$. This amounts to a total of 46,315 residuals, since each crystal geometry can contribute multiple residuals (lattice parameters + atomic positions, per Equations 4.3.3–4.3.6). While several large databases for inorganic crystal structures are being curated using similar solid-state DFT calculations (e.g., the Open Quantum Materials Database [257, 258]), the same effort has not been matched in organic systems. In the area of organic crystals, DFT-optimized datasets have been developed for benchmarking DFT methods [132, 140], but these are relatively small (< 60 structures). Most likely, the largest collection of published DFT-D optimized organic crystal structures is from the works of van de Streek & Neumann [259, 260], wherein a total of 440 crystal structures were resolved by DFT-D. However, utilization of this dataset is challenging for two reasons. First, the dataset was constructed in two separate works that use different DFT functionals: the PW-91 functional [259] and the PBE functional [260]. As noted before, this complicates the ability to use the dataset collectively. Second, the primary aim of those works was to assess the efficacy of DFT-D in resolving crystal geometries, thus the relaxed crystal geometries were reported without any associated energy values. In their investigation of polymorphism, Cruz-Cabeza *et al.* [21] also performed large-scale optimization of organic crystals using DFT-D. A total of 446 organic crystals were evaluated, but the results of these computations were never formally published as a database.

In summary, to my knowledge, the CE–755 dataset currently constitutes the largest repository of self-consistent, high-quality, DFT-D data for organic crystal structures, comprising both crystal geometries and energies. Before using the CE–755 dataset for HAIEFF FF parametrization, it is desirable to analyze the dataset's chemical profile, and (where possible) validate the accuracy of the DFT-D computations.

5.4.1 Chemical Profile of the Reference Dataset

For the FFs parametrized with the CE–755 to exhibit the desired transferability, it is imperative that the dataset covers a wide breadth of systems. To that end, I first consider some general characteristics of the CE–755 dataset in Figure 5.2.

By design, the CE–755 dataset comprises only H, C, N, O, F, S, and Cl atoms. Since the dataset contains only organic crystals, carbon is ubiquitous in all of the structures, while hydrogen has slightly less than 100% presence. The frequency of halogens and sulphurs remain disproportionately small compared to the other elements, but this is expected since these elements tend to be less common in organic crystals. The space group distribution of the dataset is also of interest. Although the CCDC reports the space group distribution of the CSD [74], those statistics pertain to both organic and inorganic crystal structures. Instead, using the CCDC Python API, I survey the CSD to generate appropriate statistics for just the organic entries. Of the 1,304,168 entries recorded in the CSD at the time of assessment, only 585,654 entries (~44.9%) were flagged as organic crystals with valid space groups (including repeat entries of the same polymorphs). Comparing the CE–755 space group distribution to the distribution from the CSD, the $P2_1/c$ space group is correctly identified as the most frequently observed space group. However, beyond this, the CE–755 space group distribution is at odds with the CSD distribution. Notably, the $C2/c$ space group is the fifth most observed space group in the CSD, yet it represents only a minor portion of the CE–755. Unfortunately, in the curation of the CE–755, structures originally in the $C2/c$ space group often had fractional Z', which our CSP programs would be unable to process. As such, the symmetry of these crystals had to be reduced to achieve integer Z' numbers, inadvertently skewing the space group distribution. Lastly, I consider the 'complexity' of molecules present in the CE–755 dataset. This is inferred from the size of molecules in the dataset (number of atoms in the molecule) and the corresponding number of distinct atomic environments they contain. By nature of the structure curation methodology, the CE–755 favors small molecules (< 20 atoms), with most systems containing between 5–7 distinct atomic environments. The full dataset contains a total of 22 polymorphic pairs and three polymorphic triplets. It should be noted that the CE–755 contains no multicomponent crystals (i.e., cocrystals, hydrates, solvates, and salts are excluded), thus all structures with $Z' > 1$ have multiple units of the same molecule in their asymmetric unit cell.

The availability of pairwise data between the various atomic environments is reevaluated in Figure 5.3. The pairwise matrix of the CE–755 database clearly improves on the corresponding matrix for the DB–445, indicating that the targeted structure curation was successful. In the CE–755, almost all pairwise interactions are represented by at least 10 structures. It should be reiterated that each structure contributes multiple residuals to Problem PE from its crystal energy and geometry (atomic positions + lattice parameters). For the interactions that fail to meet this threshold, I attempt to justify their relative data paucity and its implications. Matrix entries for $N_{3,5,*}$ and $H_{*,*,S}$ exhibit a systematic scarcity of data. Examination of these atomic environments suggests that they participate in chemical moieties that do not hold broad industrial and

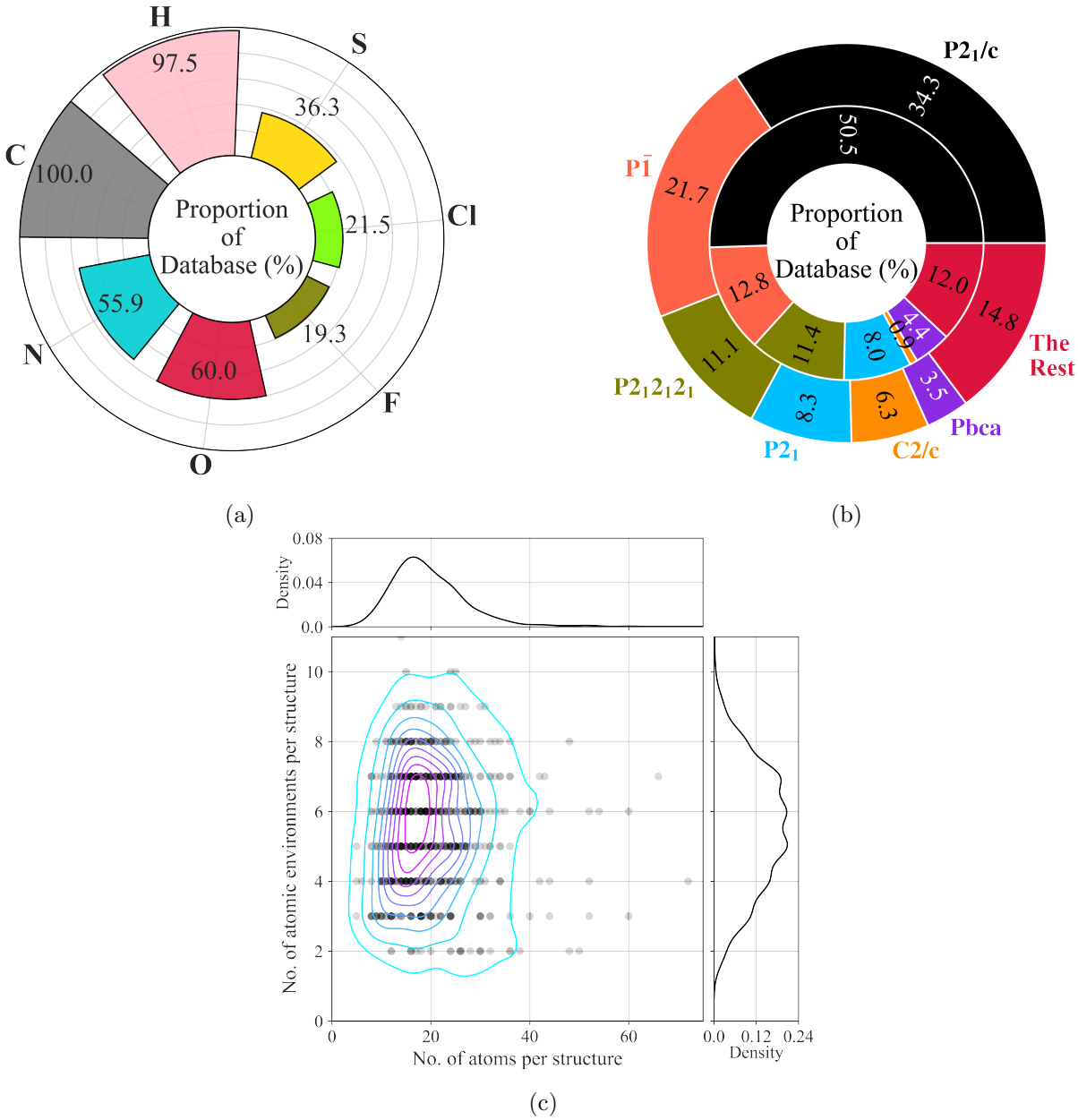


Figure 5.2: (a) The proportion of crystals in the CE–755 containing each element. (b) The proportion of crystals in the CE–755 (inner ring) in the six most frequently observed space groups for organic crystals in the CSD (outer ring). (c) The distribution of molecules based on their composite number of atoms and atomic environments (as defined in Table 5.1). The opacity of the black ● reflect the frequency of compounds with a given number of atoms and atomic environments; this is also indicated by the contour lines and density curves.

scientific interest. As such, it is rational that their presence in the CSD, and by extension the CE–755, is relatively small. For instance, $N_{3,5,*}$ is characteristic of nitro-groups, which are commonly found in compounds used for explosive materials [9, 261]. Such molecules tend to contain nitrogen and oxygen from only their nitro-groups, while rarely containing halogens and sulphur. This is reflected in the spread of data for the $N_{3,5,*}$ pairwise interactions. Likewise, $H_{*,*,S}$ is characteristic of thiol-groups. Thiol-containing compounds have a relatively niche role in phar-

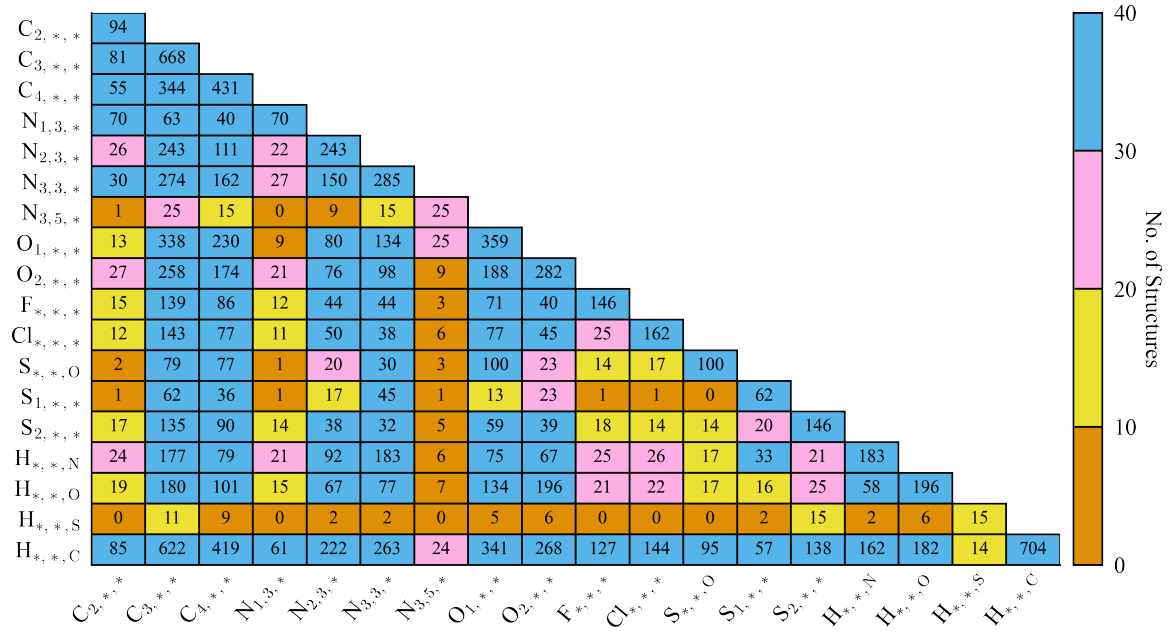


Figure 5.3: Tally of structures containing each pairwise atomic-interaction within the CE–775 dataset. Each matrix entry corresponds to the total number of structures which contain the pair of atoms on the row and column. Cells are colored orange (< 10), yellow (10 ≤ and ≤ 19), pink (20 ≤ and ≤ 29), and blue (≥ 30) according to the structure counts.

maceuticals as chelating agents [262], apart from which there are few solid state applications for organic thiol-compounds. Of the industrially-relevant (solid-state) thiol-compounds, most exhibit significant flexibility (branched, aliphatic compounds with many consecutive sp^3 carbons). As such, when screening the CSD for candidates containing $H_{*,*,S}$, most structures were rejected as highly flexible molecules that violate the target structure curation criteria. Beyond these systematic gaps in the dataset, there are some isolated groups of missing data. As noted before, searches involving $S_{1,*,*}$ with halogens or $S_{*,*,O}$ did not generate a sufficient number of suitable candidates, and thus these were omitted from the dataset. Similarly, searches involving $S_{1,*,*}$ and $S_{*,*,O}$ paired with other uncommon chemical groups, like cyano-groups ($C_{2,*,*}$ + $N_{1,3,*}$) and alkyne-groups ($C_{2,*,*}$ + $C_{2,*,*}$), also yielded an inadequate number of suitable candidates even with relaxed search constraints.

In short, the failure to populate certain parts of the CE–755 AE-matrix can be attributed to two factors: (1) a scarcity of CSD entries containing the target AEs, reflecting a lack of interest in those regions, and (2) the limits of flexibility imposed upon the dataset. The latter implicitly biases against certain atomic environments which are inherently flexible (e.g., $S_{*,*,O}$), but this is unavoidable. Considering my goal is to parametrize industrially-applicable CSP FFs, a lack of data in 'industrially uninteresting' areas is tolerable. If/when these regions of chemical-space become of interest (or if one simply wishes to more evenly populate the AE-matrix), the decision to use synthetic DFT-D data means there is latitude to generate hypothetical structures that can fulfill the AE targets. This has not been done here to maintain consistency with the structure curation philosophy set out by Bowskill [2].

Table 5.3: Total occurrences of each hydrogen bonding motif in the CE–755. A single motif can occur multiple times in a given structure.

	X=N	X=O	X=S	Total
Y=N	144	84	2	230
Y=O	128	165	5	298
Y=S	49	14	15	78
Y=Cl	10	4	-	14
Total	331	267	22	620

Finally, I analyze how well the CE–755 encompasses various hydrogen bonding motifs. In his parameterization attempts with the DB–445, Bowskill [2] identified hydrogen bonding systems as a major challenge for FF fitting. As such, it is desirable that the CE–755 probes a variety of hydrogen bonding motifs, to enable a more robust description of these interactions. Using the CCDC Python API, I screen the CE–755 for crystals containing intermolecular hydrogen bonds. This flagged 314 structures, ~41.6% of the CE–755, which is almost double the number of hydrogen bonding crystals in the DB–445. A more detailed breakdown of the hydrogen bonding motifs present among these 314 structures is given in Table 5.3, according to their hydrogen donor (X) and accepter (Y) elements. Graph set analysis [253, 263] characterizes hydrogen bonding more formally by also considering the repeating hydrogen bond pattern. For simplicity, this has not been done here. The occurrence of the different X–H···Y in the CE–755 largely adheres with chemical intuition. N–H and O–H hydrogen bonded to N and/or O comprise the majority of hydrogen bonds in the CE–755, which may be attributable to two factors. The CSD holds more nitrogen- and oxygen-containing organic compounds than sulphur- and chlorine-containing ones, hence the CE–755 hydrogen bond profile may simply reflect this underlying distribution. Hydrogen bonds involving the period 2 elements also tend to be more stabilizing than their period 3 counterparts, since the former should polarize neighboring hydrogens more. As such, the formation of hydrogen bonds involving nitrogen and oxygen may simply be thermodynamically more favorable than those involving sulphur and chlorine. It is also for this reason that many chemical-design heuristics are based on hydrogen bonds involving oxygen and nitrogen [63, 252]. In any case, it is encouraging to observe that non-canonical hydrogen bonds involving the period 3 elements (i.e., sulphur and chlorine) still have some representation in the CE–755.

Ultimately, the chemical profile of the CE–755, in terms of atomic environments and intermolecular hydrogen bonding, should be adequately broad for the purpose of fitting transferable HAIEFF FFs. While the dataset is sparse in certain data, these tend to be in areas of chemical-space that are of less practical interest.

5.4.2 Validation of the Reference Data

Beyond analyzing the chemical profile of the CE–755, it is also pertinent to validate that the dataset is sufficiently *accurate* to act as reliable reference data. Bowskill [2] validated the

accuracy of the DB–445 against experimentally derived data [142, 213] and concluded that the dataset was of satisfactory quality for parameter fitting. Below, I perform a similar validation over the entire CE–755 dataset.

To begin, I verify that the DFT-D optimized geometries are in adequate agreement with the experimental crystal structures used as inputs for the DFT calculations. COMPACK [242] is used to assess geometric deviations between the experimental and DFT-D optimized structures, with all COMPACK settings kept at their default values (i.e., hydrogen positions are ignored). On average, the geometries of the non-hydrogen bonding and the hydrogen bonding crystals are well-predicted by periodic DFT-D. Their average $RMSD_{15}$ scores are 0.107 Å and 0.113 Å respectively, which in practical terms is a fairly low error. Recognizing that the experimental structures are measured under non-static conditions, while the DFT-D results correspond to static structures, these geometric deviations are within a reasonable threshold.

A similar validation is performed on the calculated lattice energies, against experimentally reported sublimation enthalpies. The experimental data used for this was curated by Gatsiou *et al.* [213] and Robinson *et al.* [158]. In the latter work [158], some crystals are quoted with different CSD Refcodes relative to the structures in the CE–755. In such cases, an $RMSD_{50}$ comparison is done between the two CSD Refcodes to ensure that the data reported in Ref [158] relates to the same structure from the CE–755. This ambiguity was not encountered when using data from Ref [213]. For a fair comparison with the static, 0 K lattice energies generated by DFT-D, the Dulong-Petit law is employed as an approximate correction for finite-temperature effects in the experimental sublimation energies [142, 251]. Between the two experimental databases, the same crystal structure can have several distinct sublimation enthalpies reported. It is beyond the scope of this work to determine the reliability of each experimental source, hence I consider the entire range of reported data. The results of this comparison are summarized in Tables 5.4 and 5.5, while the individual experimental data points are listed in Appendix F.

The TPSS-D3 predicted static lattice energies tend to overbind the crystals, relative to their experimentally-derived lattice energies. This overbinding is on average 3.90–5.24 kJ/mol across all the structures, but is marginally greater for the hydrogen bonding structures (4.70–6.94 kJ/mol), versus the non-hydrogen bonding structures (3.47–4.30 kJ/mol). These trends are aligned with the Bowskill's [2] validation of the DB–445 set. The largest discrepancies in lattice energy are observed for the crystals of melamine (CSD Refcode: MELAMI06) and quinolin-8-ol (CSD Refcode: HXQUIN14), both of which are hydrogen bonding.

In the case of MELAMI06, there are two available experimental data points, which suggest that DFT-D overbinds the crystal by 20.46 kJ/mol [264] and 25.98 kJ/mol [265]. The two measurements were conducted at 432 K [264] and 298.15 K [265] respectively. The reliability of the Dulong-Petit correction when applied to data at elevated temperatures could be a source of error in the experimentally-determined lattice energy. However, in my view, this argument is unconvincing since the two experimental sources have distinct measurement temperatures, yet both predict lattice energies in the same ballpark. Moreover, the reported measurement error of Ref [265] (±4 kJ/mol) would further place the data in good agreement. In short, it is likely

Table 5.4: Comparison of experimentally-derived lattice energies ($U_{\text{latt}}^{\text{exp}} \approx -\Delta H_{\text{sub}}^{\text{exp}} - 2RT$) with the DFT-D computed lattice energies ($U_{\text{latt}}^{\text{DFT}}$) for non-hydrogen bonding structures. The number of measurements associated with each structure is given (#), and if multiple sources give differing experimental values, both $U_{\text{latt}}^{\text{exp}}$ and the deviations in U_{latt} ($\Delta U_{\text{latt}} = U_{\text{latt}}^{\text{exp}} - U_{\text{latt}}^{\text{DFT}}$) are reported as ranges. Positive values of ΔU_{latt} indicate overbinding.

REFCODE	$U_{\text{latt}}^{\text{DFT}}$ (kJ/mol)	#	$U_{\text{latt}}^{\text{exp}}$ (kJ/mol)	ΔU_{latt} (kJ/mol)	Source(s)
ADAMAN08	−69.28	1	−63.96	5.32	[264]
ANTCEN16	−113.00	1	−102.86	10.14	[269]
BENZEN06	−55.02	1	−49.36	5.66	[270]
BIPYRL04	−96.88	3	[−86.86, −79.96]	[10.03, 16.93]	[271–273]
BNZQUI03	−76.24	4	[−71.66, −67.72]	[4.58, 8.52]	[274–276]
BZPHAN01	−124.38	1	−110.96	13.42	[277]
CLMETH03	−33.15	1	−34.11	−0.96	[278]
CRYSEN	−138.77	1	−128.36	10.41	[279]
CUBANE	−56.48	1	−60.16	−3.68	[280]
CUKCIU03	−54.38	1	−40.12	14.26	[281]
CYHEXO	−86.79	1	−79.96	6.83	[282]
DCLANT10	−123.74	1	−119.65	4.09	[283]
DCLBQN	−82.73	1	−74.86	7.88	[276]
DITHAN02	−84.60	3	[−73.86, −67.96]	[10.75, 16.65]	[284, 285]
DMSULO04	−83.79	1	−81.96	1.83	[286]
DNBENZ11	−85.32	1	−91.96	−6.63	[287]
DTENYL02	−96.43	1	−90.16	6.27	[288]
FABPON11	−138.61	1	−128.26	10.35	[289]
HCLBNZ13	−100.14	1	−95.46	4.69	[283]
HEXANE01	−51.67	1	−53.76	−2.09	[290]
HXMTAM10	−85.17	2	[−84.56, −79.86]	[0.61, 5.31]	[291, 292]
NAPHTA31	−83.65	14	[−78.16, −68.96]	[5.49, 14.69]	[279, 285, 293–304]
OCTANE10	−67.34	1	−71.69	−4.35	[290]
PENTAN01	−43.33	1	−44.38	−1.05	[290]
PERLEN06	−148.57	1	−136.50	12.07	[305]
PHENAN08	−106.09	1	−95.86	10.23	[301]
PHENAZ04	−106.40	1	−100.19	6.22	[306]
PNCLBZ	−93.14	1	−91.98	1.16	[158]
PRMDIN01	−65.09	1	−63.26	1.83	[158]
PTOXEC	−76.73	2	[−92.86, −92.82]	[−16.12, −16.09]	[158, 307]
PYRAZI01	−65.26	1	−61.16	4.10	[308]
PYRENE03	−113.74	1	−105.26	8.49	[279]
SUCANH12	−78.54	2	−85.66	−7.12	[158, 309]
TCYETY01	−85.42	1	−86.16	−0.73	[310]
TETDAM03	−70.45	1	−67.16	3.29	[311]
TOXOCN	−72.01	2	[−84.56, −84.45]	[−12.55, −12.44]	[158, 312]
TRIPHE13	−133.93	1	−125.06	8.88	[279]
TRITAN03	−109.71	1	−98.16	11.56	[313]
TRIZIN01	−59.88	4	[−61.85, −59.14]	[−1.97, 0.74]	[285, 314]
TROXAN	−56.98	1	−61.56	−4.58	[315]

Table 5.5: Comparison of experimentally-derived lattice energies ($U_{\text{latt}}^{\text{exp}} \approx -\Delta H_{\text{sub}}^{\text{exp}} - 2RT$) with the DFT-D computed lattice energies ($U_{\text{latt}}^{\text{DFT}}$) for hydrogen bonding structures. The number of measurements associated with each structure is given (#), and if multiple sources give differing experimental values, both $U_{\text{latt}}^{\text{exp}}$ and the deviations in U_{latt} ($\Delta U_{\text{latt}} = U_{\text{latt}}^{\text{exp}} - U_{\text{latt}}^{\text{DFT}}$) are reported as ranges. Positive values of ΔU_{latt} indicate overbinding.

REFCODE	$U_{\text{latt}}^{\text{DFT}}$ (kJ/mol)	#	$U_{\text{latt}}^{\text{exp}}$ (kJ/mol)	ΔU_{latt} (kJ/mol)	Source(s)
ACETAC07	−73.15	1	−70.68	2.46	[316]
AMBNAC12	−125.91	2	−121.06	4.85	[317, 318]
AMPYRE	−104.43	3	[−93.06, −92.06]	[11.37, 12.37]	[319, 320]
AMPYRM11	−96.08	1	−81.46	14.62	[319]
BAGFIY01	−102.83	1	−90.46	12.37	[321]
BENZAC12	−102.67	1	−94.96	7.71	[322]
BESKAL04	−127.39	1	−135.36	−7.96	[323]
BZDMAZ02	−111.69	5	[−111.96, −99.26]	[−0.27, 12.43]	[285, 324–326]
CLBZAC01	−116.31	2	[−111.26, −105.86]	[5.05, 10.45]	[327, 328]
FORMAC02	−67.58	1	−65.06	2.51	[264]
FULPIM	−96.27	1	−93.36	2.91	[329]
HXQUIN14	−91.43	5	[−113.96, −92.86]	[−22.52, −1.42]	[266–268]
IMAZOL06	−91.12	1	−86.36	4.76	[324]
MALIAC12	−117.02	1	−110.36	6.66	[330]
MELAMI06	−150.94	2	[−130.48, −124.96]	[20.46, 25.98]	[264, 265]
NAPOAC01	−127.98	1	−118.56	9.42	[331]
PFBZAC	−92.12	1	−96.46	−4.34	[332]
SALMID04	−115.45	2	[−106.86, −104.26]	[8.59, 11.19]	[333, 334]
SUCACB03	−132.44	1	−128.06	4.38	[335]
TRAZOL03	−91.28	2	[−89.05, −87.96]	[2.22, 3.32]	[158, 336]
XAGWIM	−112.43	1	−105.76	6.67	[337]
ZZZKPE03	−103.02	1	−91.6	11.42	[338]

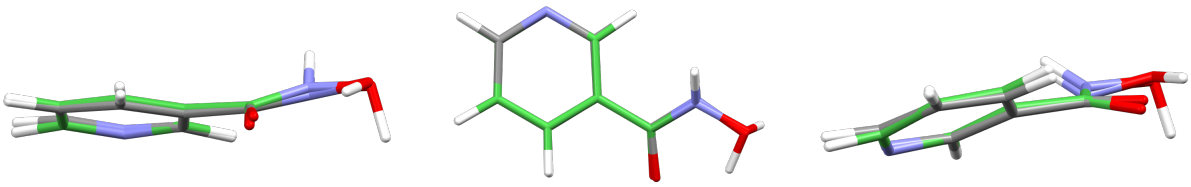
that the DFT-D determined static lattice energy for MELAMI06 may be too negative. In the case of HXQUIN14, there are five experimental measurements for this structure [266–268], all measured at 298.15 K. These experimental reports can be grouped into two lots: a pair [267, 268] with sublimation enthalpies within 0.2 kJ/mol of each other, and three measurements [266] with sublimation enthalpies within 1.6 kJ/mol of each other. The average sublimation energies of these two groups differ by ~20.1 kJ/mol, which explains the large energy range for HXQUIN14. Overall, all of the data concur that the DFT-D derived lattice energy underbinds HXQUIN14, but the extent of this error is unclear due to the conflicting experimental reports.

Across all of the experimental data, the unbiased standard deviation of the energy errors is rather large at 8.11 kJ/mol. This indicates that there are significant, non-systematic differences between the experimental and DFT-D results. This standard deviation for the CE–755 is more than twice that reported by Bowskill [2] when comparing DB–445 results against the experimentally derived static lattice energies from the X23 test set [142]. In my view, this large variability in energy errors (relative to experiment) should not be interpreted as an indictment of the accuracy of the CE–755 data. Rather, it stresses the challenges in determining static lattice energies

from experimental sublimation enthalpies. Where multiple reports of experimental sublimation enthalpy are found in Tables 5.4 and 5.5, there is often significant inconsistency among their reported values, which muddles the effective use of these measurements. This complication is further compounded by the inaccuracies associated with applying a thermal correction to obtain static lattice energies. As noted by Bowskill [2], both Gatsiou *et al.* [213] and Reilly & Tkatchenko [142] derived static lattice energies from experimental sublimation enthalpies, yet the values they determined for the same structure could differ by as much as ~10 kJ/mol. Altogether, there are many challenges associated with obtaining accurate static lattice energies from experimental sublimation enthalpies, justifying the decision to use synthetic data instead.

Lastly, I analyze the intramolecular energies predicted by DFT-D. Formally, the reference intermolecular energies are determined by removing the crystalline conformation's monomer energies from the cell energy (Equation 5.3.1). However, this can also be interpreted as the removal of the intramolecular deformation cost, $\Delta U_{\text{intra}}^{\text{ref}}$, from the DFT-D lattice energy (Equation 5.3.3). By design, the CE–755 should predominantly contain rigid molecules, hence $\Delta U_{\text{intra}}^{\text{ref}}$ contributions are not expected to be large. This is the case for the non-hydrogen bonding structures, which have an average intramolecular contribution of 1.77 kJ/mol, with $\Delta U_{\text{intra}}^{\text{ref}}$ spanning 0.08–9.97 kJ/mol. In most of these crystals, the equilibrium molecular conformation is only marginally strained by the packing forces of the crystalline environment, leading to relatively minor intramolecular contributions to the lattice energy. In the rare cases where this contribution is larger, this is usually associated with molecules that exhibit moderate flexibility. Whilst this contradicts the selection principles of the structure curation, it is sometimes unavoidable given that some atomic environments are inherently flexible.

In contrast, hydrogen bonding structures in the CE–755 exhibit a much higher average intramolecular contribution of 10.79 kJ/mol, covering a range of 0.50–38.90 kJ/mol. In many cases, large intramolecular energy penalties emerge despite seemingly small conformational changes between the crystalline and *in vacuo* conformations. Evidently, these small molecular adjustments are highly destabilizing towards the isolated molecule, but are tenable if they enable the formation of an intermolecular hydrogen bond that compensates for the intramolecular cost. Bowskill



(a) (b) (c)

Figure 5.4: RMSD_1 overlay of the FUZZAB molecule in its equilibrium conformation (grey) and crystalline-phase conformation (green), viewed from (a) the side of the aromatic ring's plane, (b) orthogonal to the aromatic ring's 'face', and (c) along the axis of the C–N bond. In the equilibrium conformation, the O–H is almost planar with the carbonyl oxygen while in the crystalline-phase, the O–H orientates orthogonally to the carbonyl oxygen.

[2] previously discussed the case of acetic acid, wherein a $\Delta U_{\text{intra}}^{\text{ref}}$ of 12 kJ/mol was observed from the mild extension of the acid group's O–H bond. A more extreme case is when the conformational adjustment comes at the expense of disrupting an existing *intramolecular* hydrogen bond. This is the situation for nicotinohydroxamic acid (CSD Refcode: FUZZAB), which has the largest predicted intramolecular energy contribution (38.90 kJ/mol). In Figure 5.4, I have performed an RMSD_1 comparison of the equilibrium conformation of the FUZZAB molecule with its crystalline conformation. When only heavy elements are assessed, the RMSD_1 score is relatively small: 0.098 Å. This negligible discrepancy in heavy atom positions is evident from the overlaid molecules in Figure 5.4. However, closer inspection of these overlaid molecules reveal a significant re-orientation of the acid group hydrogen. In the equilibrium conformation, the acid group (O–H) hydrogen orientates along the same plane as the nearby carbonyl group, likely facilitating the formation of an intramolecular hydrogen bond. On the other hand, when placed into the crystalline environment, the acid group turns away, breaking this planarity and re-orientating the O–H bond almost orthogonally to the carbonyl group. In this crystalline conformation, the O–H hydrogen is now able to form intermolecular hydrogen bonds with a neighboring pyridine group, but at a significant intramolecular cost.

Overall, it appears that even small adjustments of hydrogens on hydrogen bonding moieties (e.g., acids, alcohols, amines) can contribute an appreciable intramolecular cost to the lattice energy. Potentially, there is some degree of DFT delocalization error [83, 84] contributing to an exaggeration of the intramolecular penalty, but in any case the DFT-D results are treated as the 'ground truth'. By extension, the hydrogen bonding structures will likely have a significant, counterbalancing intermolecular cohesive force. Given that hydrogen bonding interactions are predominantly electrostatic in nature [71, 183, 189], this foreshadows difficulties in using this data to parametrize HAIEFF FFs for hydrogen bonding systems. Since the FF functional form has canonically been chosen to describe RD-type interactions, the model may struggle if the electrostatic term does not accurately capture hydrogen bonding and significant proportions of the hydrogen bonding interactions are instead embedded into the FF.

Ultimately, I have shown that the CE–755 dataset has satisfactory chemical diversity for the parametrization of transferable HAIEFF FFs. In terms of reference data accuracy, both crystal geometries and energies are generally well-predicted by the DFT-D computations. In any case, there are sufficient problems associated with employing experimental data that I consider it nonviable for the current work. Admittedly, there are some concerning signs related to the reference intermolecular energies that could pose a challenge during the FF parametrization. However, this is perhaps another indicator that greater investigation into HAIEFF FFs is required.

5.5 Train/Test Partition

With the CrystalEstimator program [1, 2] and the CE–755 dataset [1, 256], large-scale parametrization of HAIEFF FFs can be rigorously investigated. A focal point of this work will be assessing the transferability of these FFs, thus some salient points are discussed below.

One approach to evaluate model transferability is to employ k-fold cross-validation (k-CV), as proposed by Bowskill [2]. In k-CV, the reference data is randomly partitioned into k equal-sized subsets. $k-1$ subsets are chosen as training data, while the holdout set is used to validate the performance of the parameter estimates. This cross-validation is repeated for k combinations of training and validation sets (i.e., each subset is held-out once), and model transferability is inferred by comparing the average training and validation results over the k-folds. Although k-CV is a popular methodology, several factors make it unsuitable for the current work. Firstly, k-CV makes the implicit assumption that every data point is drawn from the same underlying distribution (i.e., all structures draw from the same parameter space), and any random partitioning of the data can provide an unbiased test of model transferability. However, in a chemical dataset, this assumption breaks down since atomic environments and FF atom types are non-uniformly distributed across structures (e.g., some structures may not use FF parameters related to certain elements, while others do). As such, k-CV is liable to generate imbalanced folds that unfairly reflect transferability. Secondly, k-CV requires training FFs for k independent training sets. For large diverse datasets like the CE–755, good parameter estimates require exploration of the solution space through multistart optimization. However, this becomes resource prohibitive if performed for all k folds.

In lieu of this, I opt to partition the reference data into fixed training and test sets (TRN and TST respectively). An 80:20 split between the training (604 structures) and test set (151 structures) was chosen, so as to balance the need for training data, while having sufficient rigor in the transferability test. An important aspect of selecting test set structures is avoiding the aforementioned imbalanced construction. Recognizing that the CE–755 was built from groups of chemically related systems (AE1–18 in Table 5.2, and RS1–34 in the DB–445 [2]), I leverage this to implement a simple form of stratified sampling. For a group k, containing N_k structures, I randomly sample N_k^{TST} crystals to be placed in the test set, where $N_k^{\mathrm{TST}} = 0.20N_k$. By exclusion, any structure not in the test set is instead part of the training set. The TRN/TST structures obtained this way are tabulated in Appendix G.

To verify that the TRN/TST partitioning is sound, I analyze the split achieved across the atomic environments (Figure 5.5), following the AE definitions in Table 5.1. Rather than tabulating the raw number of structures with each pairwise interaction, each matrix entry in Figure 5.5 corresponds to the *proportion* of CE–755 structures placed in the test set that contain the corresponding pairwise AE interaction. This is given by

$$\text{Matrix entry in row } n \text{ and column } m \ (p_{n,m}) = \frac{N_{n,m}^{\mathrm{TST}}}{N_{n,m}^{\mathrm{CE755}}} \times 100, \tag{5.5.1}$$

where $N_{n,m}^{\mathrm{TST}}$ is the number of structures in the test set containing the atomic environments n and m, and $N_{n,m}^{\mathrm{CE755}}$ is the corresponding number of structures in the entire CE–755. For the majority of entries in Figure 5.5, the proportion of pairwise interactions in the test set is within ±10% of the desired 20% split. Cases where a 50:50 split is obtained can be attributed to pairwise interactions for which the CE–755 only has two structures. In such a situation, an even

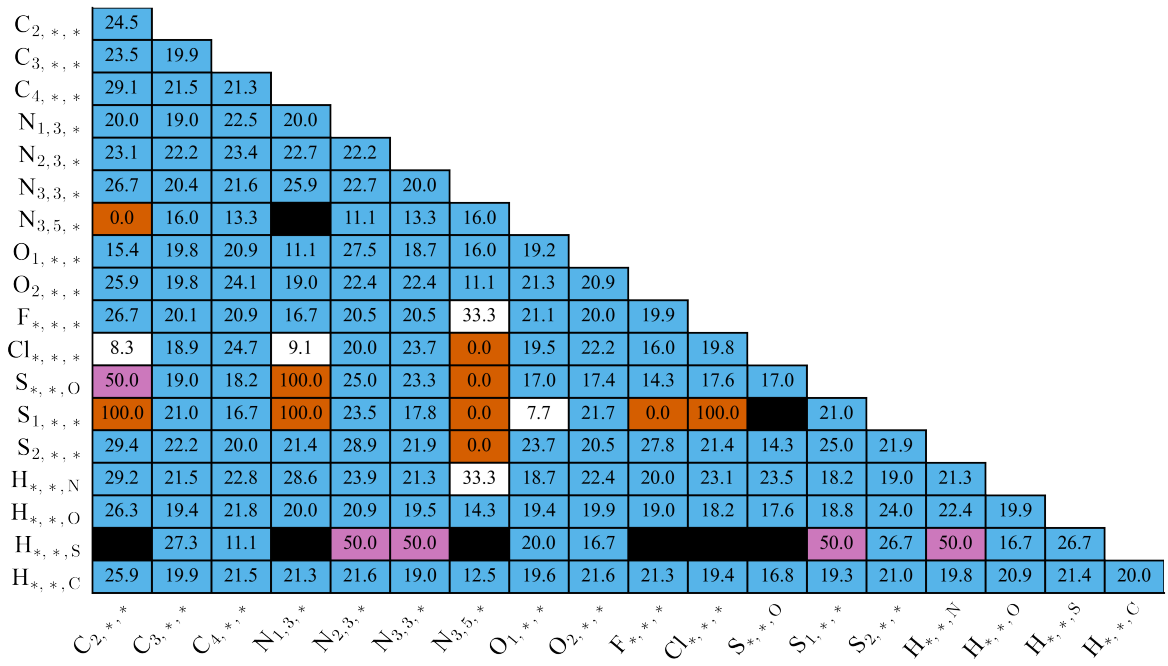
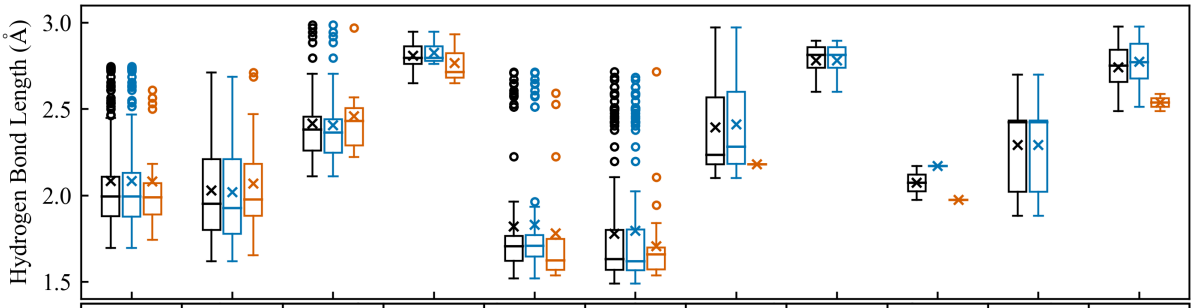


Figure 5.5: Atomic environment matrix showing the proportion of test set structures that contain cross-interactions between atomic environments n and m, as a percentage ($p_{n,m}$) of the number of structures in the CE–755 containing atomic environments n and m (Equation 5.5.1). Colors correspond to: $p_{n,m} = 0$ or $p_{n,m} = 100$ (orange); $p_{n,m} = 50$ (pink); $p_{n,m} \in [10, 30]$ (blue); all other $p_{n,m}$ (white), and black if there are no structures in the CE–755 with that interaction.

split is the best possible TRN/TST division that could be achieved. Cases where 0 or 100% of structures are in the test set can mostly be attributed to the CE–755 dataset containing only one structure with that pairwise interaction. For these interactions, a better TRN/TST division could not be achieved without adding more structures to the dataset. '$N_{3,5,*}$/$Cl_{*,*,*}$' ($N_{n,m}^{\text{CE755}} = 6$), '$N_{3,5,*}$/$S_{*,*,O}$' ($N_{n,m}^{\text{CE755}} = 3$), and '$N_{3,5,*}$/$S_{2,*,*}$' ($N_{n,m}^{\text{CE755}} = 5$) are the exceptions to this. Despite there being > 1 structure in the CE–755 dataset containing these pairwise interactions, none of these structures are assigned to the test set. Several other pairwise interactions also fail to attain the desired data split, in spite of the CE–755 containing multiple structures with these interactions (white entries in the matrix). There is no clear justification for this, other than the unfortunate result of chance during random structure selection.

Considering that I have employed only a rudimentary form of stratified sampling, it is commendable that the resulting TRN/TST split is relatively close to the target, bearing in mind the inherent limitations where data is scarce. Given that crystal structures can have extremely skewed profiles of pairwise atomic environments (e.g., one structure could contain many different pairwise interactions, while another structure has significantly fewer pairwise interactions), obtaining a better TRN/TST split would require significantly more effort than is warranted.

Lastly, I analyze the intermolecular hydrogen bonding motifs in the training and test sets, using the CCDC Python API. Among the TRN and TST structures, 248 out of 604 (41.4%) and 66 out of 151 (43.7%) are flagged, respectively, as containing intermolecular hydrogen bonds.



	N-H N	N-H O	N-H S	N-H Cl	O-H N	O-H O	O-H S	O-H Cl	S-H N	S-H O	S-H S
N_{Total}	144	128	49	10	84	165	14	4	2	5	15
N_{TRN}	106	105	41	7	68	133	13	4	1	5	13
N_{TST}	38	23	8	3	16	32	1	0	1	0	2

Figure 5.6: Hydrogen bond length distribution for different motifs in the training set (blue), test set (orange), and entire CE–755 dataset (black). Each boxes' bounds represent the 1st and 3rd quartiles, with whiskers extending to ±1.5x the interquartile range and outliers given by ○. Medians and means are indicated by ━ and × respectively. The counts of each motif in the entire CE–755 (N_{Total}), the training set (N_{TRN}), and the test set (N_{TST}) are tabulated below.

Even without any explicit constraints during the TRN/TST partitioning, the test set contains ~20% of all hydrogen bonding structures from the CE–755 (62.8 structures out of 314). I evaluate the distribution of hydrogen bond lengths for each hydrogen bonding motif (defined based on donor/acceptor identities) in the CE–755, TRN, and TST datasets (Figure 5.6). Largely, the occurrences of each motif in the training and test sets are close to the target 80:20 division. Moreover, the distribution of hydrogen bond lengths in the test set lie reasonably well within the corresponding distribution of the training set, reflecting that the former is representative of the latter. Some exceptions to this occur for O–H···S, S–H···N, and S–H···S. Like the atomic environment interactions, poor division of the hydrogen bonding motifs is usually the result of the CE–755 having very few occurrences of a motif. This is exacerbated here since the presented statistics reflect the occurrences of a given hydrogen bonding motif, rather than the number of structures containing the motif. Since a single structure can contain several distinct hydrogen bond motifs and/or multiple occurrences of the same motif, it is inherently more challenging to obtain good TRN/TST splits since moving a single structure between sets can result in large swings in the motif counts.

Ultimately, I would assert that the assembled training and test sets (Appendix G) are adequate for the goals of this work. Across the different metrics used to characterize the training and test sets' chemical profiles, there is generally good adherence to the target TRN:TST division. This trend only breaks down in the limit of data scarcity in the CE–755, which prevents effective partitioning of the available structures. Overall, I believe the current test set does give a reasonably balanced representation of the training data, such that meaningful appraisals of FF transferability will be possible following FF parametrization. Expending more effort to improve the TRN/TST division is unlikely to add much more value to the robustness of the transferability test.

5.6 Summary

In this chapter, I sought to improve the availability of diverse reference data for transferable FF parametrization. Expanding on the foundation laid by the DB–445 dataset [2], I have developed and presented the CE–755 dataset, which can be accessed from Ref [256]. By using quantitative metrics to assess the chemical diversity in the dataset and using this to guide structure curation, the CE–755 dataset achieves a more comprehensive coverage of the industrially-relevant chemical space, compared to the DB–445. To my knowledge, the CE–755 currently represents the largest repository of high quality, DFT-D resolved crystal geometries and energies for organic molecular crystals. Validation of the theoretical data in the CE–755 indicates that it should be of sufficient quality for use in parametrizing reliable HAIEFF FFs. However, the validation results also suggest that FFs formulated solely for describing RD-type interactions may prove to be suboptimal, if hydrogen bonding is inadequately described by the electrostatic interactions. Results from Bowskill's previous parametrization attempts also allude to this [2], underscoring the need for greater exploration of HAIEFF FF parametrization. To that end, I have partitioned the CE–755 dataset into an independent training and test set. Crucially, the test set was demonstrated to be a reasonable representation of the training set, in terms of their chemical profiles. As such, the test set should provide meaningful appraisals of transferability for any newly parametrized FFs.

Despite the developments made in this chapter, I do not believe that the CE–755 dataset is close to reaching a point of 'convergence' from the perspective of encompassing chemical diversity, and in principle one can never have 'too much' data. However, further expansion of a general dataset like the CE–755 may have diminishing benefit towards bolstering HAIEFF accuracy if the associated empirical-component of the model has insufficient modeling degrees of freedom to accommodate the wider breadth of chemical diversity (Bowskill's struggles are an example of this). For this reason, further expansion of the dataset should be considered only in parallel with development of the empirical FF component (e.g., more-flexible FF formulations, more atom types). This is studied in the subsequent chapter wherein I will employ the CrystalEstimator program, in tandem with the partitioned CE–755 training and test sets, to investigate the parametrization and deployment of HAIEFF FFs. As will be shown in Chapter 6, model accuracy is still limited by the HAIEFF model formulation and as such expanding the CE–755 may not be a priority at this juncture. Throughout the rest of this thesis, the CE–755 dataset will also form a basis of reference structures for other work.

Chapter 6

Force-Field Parametrization: Fitting Transferable Force-Fields

In Chapter 5, I curated the CE–755 reference dataset for the parametrization of transferable HAIEFF FFs. Bowskill [2] had previously re-parametrized the FIT FF with the DB–445 dataset using the BHP and atom types defined according to Refs [4, 166]. Unfortunately, to obtain reasonable parameter values, Bowskill had to exclude energy residuals for crystal structures containing hydrogen bonds, leading to large errors in their predicted intermolecular energies and raising doubts about the usefulness of his FFs. Bowskill's anecdotal experience highlights a wider issue with the current suite of transferable HAIEFF FFs: their lack of evolution, resulting in suboptimal FF performance. In this chapter, I investigate different modifications to the form of the HAIEFF FF which may enable the development of a model that accurately represents crystal energies and geometries. The transferability of the most promising FFs will be verified using the CE–755 test set, and their ability to resolve polymorphic energy differences will be analyzed using the polymorphic families from the CE–755. Finally, I will consider how model performance changes when the FFs are used with an *ab initio* model that is distinct from the training model. This will provide insight on the implications of *ad hoc* FF deployment.

6.1 Parameter Estimation Methodology

Details of the HAIEFF model (Chapter 3) and parameter estimation (Section 4.3) have been described earlier. Here, I provide only key methodological details relevant to this chapter.

6.1.1 Quantum Mechanical and Electrostatic Models

To derive the atomic-charges/multipoles used in the electrostatic calculations for each structure, ground state charge densities at $\boldsymbol{\theta}^{\text{ref}}$ are evaluated using G09. An 'Ultrafine' integration grid is used during the SCF calculations, without imposing molecular symmetry. The PBE0 hybrid functional [123–125] ('PBE1PBE' in G09) is used with the 6-311G(d,p) basis set [339–341]. Employing a hybrid functional helps reduce DFT delocalization error through the inclusion of HF exact exchange in the DFT energies [83, 84]. The chosen basis set includes polarization functions [341], which enable a better description of bonding, particularly when the charge

distribution is significantly distorted from the atomic distribution. This can be important for hypervalent elements like sulphur [111]. Diffuse functions have been excluded since they are usually unnecessary unless the electron density might disperse far from the nucleus (e.g., anionic systems) [111]. From experience, this level of theory (denoted 'PBE0/6-311G(d,p)') provides reliable results over a wide range of systems, at a reasonable computational cost.

The large $\Delta U_{\text{intra}}^{\text{ref}}$ identified in Section 5.4.2 suggests that induction may play a significant role in intermolecular stabilization of hydrogen bonding (HB) crystal structures. As such, I also consider the implicit embedding of induction by applying the PCM model [190] during SCF calculations. Here, I have chosen to use a continuum dielectric constant of $\varepsilon_{\text{PCM}} = 11$. While $\varepsilon_{\text{PCM}} \approx 3$ is suggested to be more representative of organic crystals [48, 191, 218, 219], there are several practical considerations justifying a higher ε_{PCM}. Firstly, anecdotal experience from various CSP practitioners suggest that using larger ε_{PCM} usually leads to better CSP outcomes. Cooper *et al.* [188] and Price *et al.* [48] have used $\varepsilon_{\text{PCM}} = 11$ in charged systems like salts and zwitterions, while Day and colleagues have commented that $\varepsilon_{\text{PCM}} > 3.0$ may be appropriate for salt and hydrate systems [48]. Results from Bowskill [2] corroborate that using *any* value of ε_{PCM} is preferable to not using PCM at all. It has been suggested that higher values of ε_{PCM} could be compensating for the failure of PCM to capture highly directional, close-range polarizing effects like hydrogen bonding [188]. Secondly, in my experience, the flexible CSP methods developed in our group fail when studying zwitterionic systems, like amino acids, unless $\varepsilon_{\text{PCM}} \geq 11$. At this dielectric value, the amino acid's zwitterionic structure is preserved during LAM calculations, even without constraints on the α-amino group's N–H bond length [188, 191]. Ultimately, I aim to fit FFs complementary to the QM model, hence small errors associated with an exaggerated ε_{PCM} can likely be absorbed and corrected by the empirical FF parameters.

Once the SCF charge density has converged, either atomic-charges are fit using the HLYGAt [173] method in G09, or GDMA (v2.2) [178–181] is used to derive distributed multipoles, up to rank 4 (i.e., hexadecapoles). It should be noted that all of our CSP codes use 'RADIUS H 0.65' as a GDMA input, which controls the relative weight associated with hydrogens in GDMA's quadrature distribution scheme. In GDMAv2, RADIUS 0.65 Å was the default for *all* atoms, but this has been reduced to 0.325 Å for hydrogens from GDMAv2.2 onward. This reflects that hydrogen should occupy less 'volume' on the quadrature-grid and attract less electron density during the multipole distribution [183]. Our decision to utilize a hydrogen radius of 0.65 Å is an artefact of using older GDMA versions, but as noted by Stone [182]: "any reasonable choice of RADIUS will give multipoles that describe the electrostatic potential accurately". Results from Welch [183] corroborate this: at high enough multipole rank (≥2), values of the hydrogen radius between 0.30–0.65 Å all give similarly good representation of the molecular ESP. It is beyond the scope of the current work to discuss the merits or flaws in this choice, but it is an important caveat for readers who may wish to use my FFs.

I have chosen an electrostatic real-space cutoff distance of 10.0 Å, while the maximum Ewald truncation error is set to 1×10^{-6} kJ/mol. These settings have previously been shown to offer a good balance of speed and accuracy in the real-space and reciprocal-space summations [2].

6.1.2 Force-field Formulation

A goal of this chapter is to study different forms of the HAIEFF FF. Previously, I introduced the Buckingham potential (BHP, Equation 3.2.35) and the Mie potential (MIP, Equation 3.2.36). Due to the good physical foundation for its repulsive and attractive components, the BHP represents the 'status quo' for HAIEFF FFs today [2, 4, 169]. However, Bowskill [2] found that FFs based on the native BHP were ineffective at describing HB structures in the DB–445. He suggested that the unphysical short-range attractive divergence of the BHP made it ill-suited for describing intermolecular forces beyond RD-type interactions, particularly HB interactions which are much more attractive. While normally 'insulated' by the repulsive wall, the unphysical region can inadvertently become 'exposed' as the BHP parameters ($A^{t,t'}$, $B^{t,t'}$, $C^{t,t'}$) are varied during a parameter estimation, which may lead to numerical problems. Unlike the BHP, the MIP is sometimes thought to have too 'hard' of a repulsive wall, since its repulsive term does not reflect the exponential decay expected of electron densities. However, it is also because of this that the MIP does not exhibit the same unphysical behavior at short-range, potentially making it more suitable for absorbing errors from the intermolecular model into the FF term. MIPs have not been extensively explored in the context of HAIEFF models for CSP. In the Sixth Blind Test [48], several groups using FF models employed fixed 9–6 or 8–6 Mie potentials, but all groups using HAIEFF models resorted to the BHP.

In between the BHP and MIP, there is an intermediate option that seeks to address the erroneous short-range behavior of the BHP, while retaining its desirable traits. This involves augmenting the native BHP with a damping function (f_6), giving a *damped* BHP (dBHP),

$$U^{\mathrm{FF,dBHP}}_{\mathbf{0}Ii,\mathbf{l}Jj}(r_{\mathbf{0}Ii,\mathbf{l}Jj}) = A^{t,t'} \exp\left(-\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right) - f_6 \frac{C^{t,t'}}{r^6_{\mathbf{0}Ii,\mathbf{l}Jj}}, \qquad t = t(Ii), \quad t' = t'(Jj), \tag{6.1.1}$$

where $A^{t,t'}$, $B^{t,t'}$, and $C^{t,t'}$ are the dBHP model parameters. At long-range, the r^{-n} description of dispersion forces is theoretically sound, but as $r \to 0$ the model wrongly predicts infinite attraction, failing to account for short-range effects that arise from electron overlap (e.g., exchange repulsion and charge penetration). Damping functions are (semi)empirical approaches to 'fix' this erroneous behavior by 'turning off' the attractions as $r \to 0$. Stone has described several damping functions [221], but notable examples in CSP include Chai & Head-Gordon damping [85] and Becke–Johnson damping used in the D3-correction [86], as well as Tang–Toennies damping in the DIFF energy model [92, 165].

It has been suggested that, in practical terms, the choice of damping function is relatively inconsequential to model performance, since minor physical deficiencies in the damping function can be accommodated by fitted model parameters [85, 86]. Detailed selection of the damping function is thus unnecessary, so long as a 'reasonable' function is employed. Since the Tang–Toennies [342] damping function has been successfully demonstrated in other intermolecular CSP models [92, 165], I have opt to use it in the current work. The C_6 Tang–Toennies [342] damping function is given by

$$f_6(\beta^{t,t'}, r_{\mathbf{0}Ii,\boldsymbol{l}Jj}) = 1 - \exp\Big(-\beta^{t,t'} \cdot r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\Big) \sum_{k=0}^{6} \left[\frac{(\beta^{t,t'} \cdot r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^k}{k!} \right], \qquad t = t(Ii), \quad t' = t'(Jj), \tag{6.1.2}$$

where $\beta^{t,t'}$ is an atom type, pairwise specific damping factor. At large interatomic distances ($r_{\mathbf{0}Ii,\boldsymbol{l}Jj} \to \infty$), $f_6 \to 1$ and the damping function effectively has no influence on long-range dispersion. As $r_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ decreases, f_6 reduces smoothly to zero at $r_{\mathbf{0}Ii,\boldsymbol{l}Jj} = 0$, mitigating the unphysical short-range attraction of the BHP. An appealing trait of the Tang–Toennies damping function is that it has only one model parameter, $\beta^{t,t'}$, making it easier to handle. To avoid additional model complexity, it is preferable not to fit this parameter. While theoretical approaches for determining molecule-specific β^{mol} have been proposed [71, 189], they are poorly justified: since damping reflects the effects of charge overlap, different atom pairs should exhibit distinct damping behavior. For this reason, I opt to approximate $\beta^{t,t'} = (B^{t,t'})^{-1}$, as suggested by Stone [221]. The approximate relationship between these parameters can be reasoned from the fact that the $B^{t,t'}$ parameter tunes the exponential-decay of repulsions between atoms, which also results from charge overlap. In this manner, atom type/atom type specific dispersion damping is implemented without any increase in the number of FF model parameters.

To employ this augmented form of the BHP in CSP calculations, derivatives like those obtained for the BHP and MIP have been summarized in Appendix H. For the U_{FF} summation, I employ a real-space cutoff distance of 15.0 Å, as suggested by Bowskill [2]. For all intermolecular interactions, a spline fraction of 10% is also applied.

6.1.3 Programmatic Settings

Bowskill [2] has studied optimal programmatic settings for use with CrystalEstimator (Section 5.2 of Ref [2]). Here, I largely adhere to his recommendations, as summarized in Table 6.1, with a few differences briefly discussed below. Bowskill [2] suggested a maximum of five CSO–RM Newton refinement steps (Figure 4.3, D2). This limit is used for parameter fitting with

Table 6.1: Summary of CrystalEstimator (CE) program settings utilized.

CE Lattice Energy Minimization (Inner Loop) Settings	Value
NAG E04UFF Optimality Tolerance	1×10^{-12}
Newton Refinement Optimality Tolerance	1×10^{-20}
Newton Refinement Hessian Update Method	BFGS
Newton Refinement Iteration Limit	5 or 30
CE Parameter Estimation Minimization (Outer Loop) Settings	**Value**
NAG E04USA Optimality Tolerance	1×10^{-8}
Maximum Structure Removal	20
Dynamic Starting Points	Active
Maximum Permitted Reinitializations	12 or 20

point-charge electrostatics, but it is increased to 30 for estimations with multipole electrostatics. In the latter case, the E04UFF subroutine sometimes fails to converge tightly with its quasi-Newton algorithm (Figure 4.3, D1), necessitating more Newton refinement steps to reach the desired convergence of U^*_{inter}. Tightly converging U^*_{inter} is a critical aspect of suppressing discontinuities in the TMF solution surface, thus 30 was chosen as a sufficiently large value that quashes frequent reports of failed energy minimization and structure removal. Bowskill [2] also recommends a maximum of five CrystalEstimator reinitialization attempts (Figure 4.3, B6→B1). In the current work, possibly due to the larger number of training structures, I have observed that this value is often not enough to reach a satisfactory convergence criteria in the parameter estimation. As such, the reinitialization limit is increased to either 12 or 20, where appropriate. As will be explained in Section 6.1.5, the value of this parameter is mostly inconsequential.

6.1.4 Parameter Bounds and Multistart Settings

CrystalEstimator employs multistart optimization to explore the non-convex solution surface of Problem PE (Section 4.3.3). For a parameter estimation involving $N_{\bar{\boldsymbol{p}}}$ decision variables and N_{Sobol} seeds, a Sobol' sequence [103] is used to initialize the vector $\boldsymbol{\mathcal{S}}_s \in [-0.5, 0.5]^{N_{\bar{\boldsymbol{p}}}}$ for seed s. The qth element of the corresponding starting decision variable vector, $\bar{\boldsymbol{p}}_s^{[0]}$, is determined by

$$\bar{p}_q^{\text{ubd}} = (1 + X_q^{\text{ubd}}) \times \bar{p}_q^{\text{ref}}, \qquad q = 1, \cdots, N_{\bar{\boldsymbol{p}}}, \tag{6.1.3}$$

$$\bar{p}_q^{\text{lbd}} = (1 - X_q^{\text{lbd}}) \times \bar{p}_q^{\text{ref}}, \qquad q = 1, \cdots, N_{\bar{\boldsymbol{p}}}, \tag{6.1.4}$$

$$\bar{p}_{s,q}^{[0]} = \frac{\bar{p}_q^{\text{ubd}} + \bar{p}_q^{\text{lbd}}}{2} + (\bar{p}_q^{\text{ubd}} - \bar{p}_q^{\text{lbd}}) \times \mathcal{S}_{s,q}, \qquad s = 1, \cdots, N_{\text{Sobol}}, \quad q = 1, \cdots, N_{\bar{\boldsymbol{p}}}, \tag{6.1.5}$$

where X_q^{ubd} and X_q^{lbd} are user-specified relative upper and lower search bounds for $\bar{p}_q$, while $\bar{p}_q^{\text{ref}}$ is its reference value. Equations 6.1.3 and 6.1.4 define the sampling bounds for $\bar{p}_q$, denoted $\bar{p}_q^{\text{ubd}}$ and $\bar{p}_q^{\text{lbd}}$; the vector element $\mathcal{S}_{s,q}$ describes a deviation from the centre of this range (i.e., $\frac{\bar{p}_q^{\text{ubd}} + \bar{p}_q^{\text{lbd}}}{2}$). When $s = 1$, $\boldsymbol{\mathcal{S}}_1 = \mathbf{0}$, such that the centre-point is sampled. To better accommodate parameter bounds, I have modified this implementation, as explained below.

(Damped) Buckingham Potential

The model parameters of the BHP and dBHP, $\boldsymbol{p}^{(\text{d})\text{BHP}} = \{A^{t,t'}, B^{t,t'}, C^{t,t'}\}$, do not have any one-to-one relation to the physical features of the pair potential. As such, I cannot impose any meaningful constraints on these model parameters and only require them to be positive,

$$\bar{\boldsymbol{p}}^{(\text{d})\text{BHP}} > \mathbf{0}. \tag{6.1.6}$$

In this case, I use the FIT FF [4] as a convenient set of reference parameters, i.e., $\bar{\boldsymbol{p}}^{\text{ref}} = \bar{\boldsymbol{p}}^{\text{FIT}}$. Following Bowskill's recommendation [2], $X^{\text{ubd/lbd}} = 40\%$ is used for the $A^{t,t'}$ and $C^{t,t'}$ parameters. It is known that the BHP (and by extension dBHP) is extremely sensitive towards the $B^{t,t'}$ parameter [192, 193, 199, 213]. Hence, if $B^{t,t'}$ is to be parametrized, I have found that

setting $X^{\text{ubd/lbd}} = 3\%$ helps generate 'reasonable' starting points from the FIT FF reference. Since no additional constraints are imposed on the FF parameters, the standard Sobol' sequence will not generate invalid starting $\bar{\boldsymbol{p}}_s^{[0]}$, and no further measures are required. For estimations involving the dBHP and BHP, $N_{\text{Sobol}} = 512$ is used.

Mie Potential

Unlike the (d)BHP, there are several parameter constraints which can (and should) be imposed for the MIP. Firstly, it is necessary to constrain

$$\lambda_{\text{r}}^{t,t'} - \lambda_{\text{a}}^{t,t'} \geq 1 \times 10^{-8}, \tag{6.1.7}$$

where 1×10^{-8} acts as a small buffer to prevent $\lambda_{\text{r}}^{t,t'} = \lambda_{\text{a}}^{t,t'}$. If the two exponents equate, the entire MIP becomes zero-valued, leading to discontinuities in the parameter estimation. This constraint also ensures that the repulsive exponent exceeds the attractive exponent, allowing the MIP to have a sensible shape. On top of this, the formulation of the MIP (Equation 3.2.36) imparts clear interpretable relationships between the FF parameters and the features of the pair potential. This can be leveraged to limit the parameter search space to a physically sensible region. Given that $\sigma^{t,t'}$ corresponds to the $r_{\mathbf{0}Ii,\boldsymbol{l}Jj}$-intercept of the pair potential, I constrain

$$\sigma^{t,t'} \geq \sigma_{\text{lbd}}^{t,t'}, \tag{6.1.8}$$

where $\sigma_{\text{lbd}}^{t,t'}$ is a 'reasonable' lower bound for $\sigma^{t,t'}$. The values of $\sigma_{\text{lbd}}^{t,t'}$ and their method of determination are described in Appendix I. Other intuitive bounds that I have defined, include:

$$\sigma^{t,t'} \leq 15, \tag{6.1.9}$$

$$\lambda_{\text{r}}^{t,t'} \leq 60, \tag{6.1.10}$$

$$1 \times 10^{-8} \leq \varepsilon^{t,t'} \leq 1 \times 10^{8}. \tag{6.1.11}$$

In this work, all $\lambda_{\text{a}}^{t,t'}$ are fixed to a value of 6, which is the correct exponent for the leading London dispersion term [221]. As such, no bounds have been defined for $\lambda_{\text{a}}^{t,t'}$.

Unlike the (d)BHP, there are no existing reference HAIEFF FF parameters for the MIP. As such, I have generated a Mie-analogue of the FIT FF for this purpose. This procedure and the reference MIP parameters are detailed in Appendix I; here I cover the salient details. First, $\lambda_{\text{a}}^{\text{FIT},t,t'}$ is fixed to 6.0. Next, $\sigma^{\text{FIT},t,t'}$ is determined from the $r_{\mathbf{0}Ii,\boldsymbol{l}Jj}$-intercept of the corresponding FIT pair potential. Care is taken to disregard the intercept associated with the short-range unphysical divergence of the BHP. Finally, $\lambda_{\text{r}}^{\text{FIT},t,t'}$ and $\varepsilon^{\text{FIT},t,t'}$ are optimized to reproduce the FIT pair potential. This is done for points that start slightly below $\sigma^{\text{FIT},t,t'}$ (so that part of the repulsive wall is sampled) up to an appropriately long-range. With this reference potential, I again specify the search bounds. $X^{\text{ubd/lbd}} = 40\%$ is chosen for $\varepsilon^{t,t'}$ and $\lambda_{\text{r}}^{t,t'}$, as these parameters tend to have more moderate effects on the shape of the pair potential. Like the BHP

$B^{t,t'}$ parameter, it can be intuited that the MIP will be sensitive towards the value of $\sigma^{t,t'}$, since it translates the entire potential along the distance-axis. From testing, I find that this is indeed the case and $X^{\text{ubd/lbd}} = 5\%$ is more appropriate for $\sigma^{t,t'}$.

These additional physical constraints imposed on the MIP parameters (Equations 6.1.7–6.1.11) complicate the generation of $\bar{\boldsymbol{p}}_s^{[0]}$. If a $\bar{p}_q^{\text{ref}}$ is relatively close to a constraint limit, it is likely that the standard approach to determine sampling bounds will infringe into this infeasible region. $\bar{\boldsymbol{p}}_s^{[0]}$ generated from these bounds are liable to violate the physics-informed constraints, leading to failure of the parameter estimation. To mitigate this possibility, CrystalEstimator has been modified to adapt the sampling bounds in accordance with any user-imposed constraints. Specifically, the sampling bounds $\bar{p}_q^{\text{ubd/lbd}}$ are taken as the more restricted value between the constraint bound ($\bar{p}_q^{\text{con}}$) and the user-specified sampling bound:

$$\bar{p}_q^{\text{ubd}} = \min\left((1 + X_q^{\text{ubd}}) \times \bar{p}_q^{\text{ref}}, \bar{p}_q^{\text{con}}\right), \qquad q = 1, \cdots, N_{\bar{\boldsymbol{p}}}, \tag{6.1.12}$$

$$\bar{p}_q^{\text{lbd}} = \max\left((1 - X_q^{\text{lbd}}) \times \bar{p}_q^{\text{ref}}, \bar{p}_q^{\text{con}}\right), \qquad q = 1, \cdots, N_{\bar{\boldsymbol{p}}}. \tag{6.1.13}$$

Under these circumstances, the sampling space may be re-centered, such that the first Sobol' seed ($\boldsymbol{\mathcal{S}}_1 = \mathbf{0}$) no longer corresponds to $\bar{\boldsymbol{p}}^{\text{ref}}$. As it is desirable to attempt the parameter estimation from the reference parameters, I have modified CrystalEstimator to accept a zeroth Sobol' seed, which defaults to $\bar{\boldsymbol{p}}_{s=0}^{[0]} = \bar{\boldsymbol{p}}^{\text{ref}}$. Thus, parameter estimation with the MIP use $N_{\text{Sobol}} = (512 + 1)$.

6.1.5 Parameter Estimation Termination Strategy

FF parameter estimation is initialized from 512 (or 513) $\bar{\boldsymbol{p}}_s^{[0]}$ vectors. From those that terminate without error, CrystalEstimator selects the best parameter estimates, $\bar{\boldsymbol{p}}_{\text{MS}}^*$, which have the lowest final TMF value. The associated seed from which $\bar{\boldsymbol{p}}_{\text{MS}}^*$ emerged is termed the Minimum Sobol (MS). If the MS's optimization terminates by criteria B3–B5 (Figure 4.3), $\bar{\boldsymbol{p}}_{\text{MS}}^*$ should be at, or very near to, a local optimum of Problem PE, and I accept it as a solution. However, a different strategy (Figure 6.1) is employed if the MS's optimization terminate for any other reason.

$\bar{\boldsymbol{p}}_{\text{MS}}^*$ is rejected if the optimization from the MS terminates due to reaching the reinitialization limit (Figure 4.3, B6), or following structure reintroduction (Figure 4.3, A5). In the former, further improvement to the TMF could be achieved if more reinitializations were permitted. In the latter, because the CrystalEstimator algorithm attempts no further reinitialization after structures are reintroduced, it is unclear if additional reinitialization could improve the TMF. In both cases, there is a possibility of further improvement to the TMF if the parameter estimation sequence was allowed to continue. As such, rather than accept $\bar{\boldsymbol{p}}_{\text{MS}}^*$, I opt to restart a secondary round of parameter estimation, substituting $\bar{\boldsymbol{p}}^{\text{ref}}$ with the last set of $\bar{\boldsymbol{p}}_{\text{MS}}^*$. The same sampling bounds and constraints are employed in the restarted parameter estimation and this procedure is repeated until a satisfactory termination criteria is achieved at a (new) MS. To manage computation costs, all secondary parameter estimations are limited to only 128 (or 129 for the MIP) Sobol' seeds. In practice, this does not usually exceed more than three rounds. In this

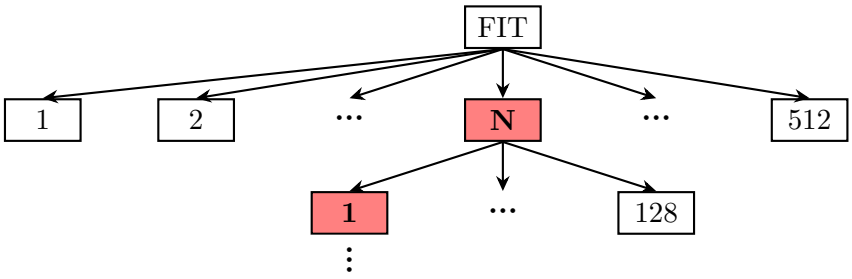


Figure 6.1: The parameterization termination strategy, illustrated for the (d)BHP. Parameter estimation initializes from 512 Sobol' seeds, using $\bar{\boldsymbol{p}}^{\text{ref}} = \bar{\boldsymbol{p}}^{\text{FIT}}$. If the Minimum Sobol (MS) (red) has an unsatisfactory termination status, the parameter estimation is repeated with 128 seeds, using the preceding $\bar{\boldsymbol{p}}^{*}_{\text{MS}}$ as $\bar{\boldsymbol{p}}^{\text{ref}}$. This continues until a satisfactory termination status is achieved at the MS. Estimations using the MIP add a zeroth seed at each stage.

manner, I ensure that for each attempted parametrization, the final accepted solution is likely to have exhausted all feasible progress in the parameter estimation. It is also for this reason that the exact value of CrystalEstimator's reinitialization limit is relatively unimportant.

6.1.6 Fitting Schemes for FF Training

To investigate the efficacy of different force-field models, I perform parameter estimation using several schemes. Common to all of these, are the following:

- Both homoatomic ($t = t'$) and heteroatomic ($t \neq t'$) interaction parameters are optimized. Given the relative abundance of reference data, there is little reason to determine the heteroatomic interaction parameters via combining rules, especially since these rules lack robust physical justification [221]. This is corroborated by other works [2, 169, 213], which show parametrization of the heteroatomic interactions benefits FF performance.
- For all non-hydrogen bonding (NHB) structures, $\omega_s^e = 0.5$ and $\omega_s^g = 1.0$ are employed for the energy and geometry residual weights respectively, as suggested by Bowskill [2].
- For all hydrogen bonding (HB) structures, $\omega_s^g = 1.0$ is also employed. ω_s^e is varied depending on whether it is desired to include the energy residuals for these structures.
- Finally, all heavy elements are represented by only one atom type, based on their proton number. Hydrogens are represented by more atom types, as discussed below.

Seven fitting schemes are considered in this work (Table 6.2). For each, parameter fitting is conducted with and without PCM, using both point-charge and multipole electrostatics, totaling 28 distinct parameter estimations. Below, I describe each of these schemes.

Scheme 1 employs the BHP with omission of energy residuals for the HB structures ($\omega^e_{s,\text{HB}} = 0$). Two hydrogen types are defined: a polarized hydrogen (H_p) is used for nitrogen- and oxygen-bonded hydrogens, and a generic hydrogen (H) is used in all other cases. All but two BHP model parameters are parametrized. It is commonly reported [192, 193, 199, 213] that the $B^{t,t'}$ parameters are extremely difficult to estimate due to the strong correlation between

Table 6.2: Summary of the parameter estimation schemes investigated.

Scheme	FF	$\omega^{e}_{s,\mathrm{HB}}$	H Types	$\bar{\boldsymbol{p}}$	$\boldsymbol{p}_{\mathrm{fixed}}$	$\dim(\bar{\boldsymbol{p}})$
1	BHP	0.00	$\mathrm{H} + \mathrm{H_p}$	$A^{t,t'} + C^{t,t';t,t'\neq \mathrm{H_p}}$	$B^{t,t'} = B^{\mathrm{FIT},t,t'}$; $C^{\mathrm{H_p},t'} = 0$	64
2	BHP	0.50	$\mathrm{H} + \mathrm{H_p}$	$A^{t,t'} + C^{t,t';t,t'\neq \mathrm{H_p}}$	$B^{t,t'} = B^{\mathrm{FIT},t,t'}$; $C^{\mathrm{H_p},t'} = 0$	64
3	BHP	0.50	$\mathrm{H} + \mathrm{H_p}$	$A^{t,t'} + C^{t,t'}$	$B^{t,t'} = B^{\mathrm{FIT},t,t'}$	72
4	dBHP	0.50	$\mathrm{H} + \mathrm{H_p}$	$A^{t,t'} + C^{t,t'}$	$B^{t,t'} = B^{\mathrm{FIT},t,t'}$	72
5	dBHP	0.50	$\mathrm{H} + \mathrm{H_O} + \mathrm{H_N}$	$A^{t,t'} + C^{t,t'}$	$B^{t,t'} = B^{\mathrm{FIT},t,t'}$	90
6	dBHP	0.50	$\mathrm{H} + \mathrm{H_O} + \mathrm{H_N}$	$A^{t,t'} + B^{t,t'} + C^{t,t'}$	-	135
7	MIP	0.50	$\mathrm{H} + \mathrm{H_O} + \mathrm{H_N}$	$\varepsilon^{t,t'} + \sigma^{t,t'} + \lambda_{\mathrm{r}}^{t,t'}$	$\lambda_{\mathrm{a}}^{t,t'} = 6$	135

$A^{t,t'}$ and $B^{t,t'}$, coupled with the pair potential's sensitivity to $B^{t,t'}$. On this basis, the $B^{t,t'}$ parameters are fixed to their corresponding values from the FIT FF ($B^{\mathrm{FIT},t,t'}$). It has also been suggested [2, 169] that the electron density around $\mathrm{H_p}$ atoms can be expected to be low, due to polarization by neighboring nitrogens/oxygens. Consequently, dispersion interactions involving $\mathrm{H_p}$ should be practically non-existent, since they cannot exhibit electron correlation. As such, all dispersion parameters for interactions involving the $\mathrm{H_p}$ atom type ($C^{\mathrm{H_p},t'}$) are fixed to 0. Scheme 1 corresponds exactly to the fitting approach of Bowskill [2].

Schemes 2 and 3 explore the feasibility of incorporating the energy residuals for HB structures into the parameter fitting, while using the BHP. Adhering to the value used for NHB structures, Scheme 2 applies $\omega^{e}_{s,\mathrm{HB}} = 0.50$, but preserves all other Scheme 1 settings. Building on this, Scheme 3 reintroduces $C^{\mathrm{H_p},t'}$ into the parameter estimation. While the reasoning for $C^{\mathrm{H_p},t'} = 0$ is sound, it assumes that the FF term describes only RD-type interactions. This neglects the fact that the (empirical) FF also tacitly absorbs errors from other parts of the intermolecular model, and enforcing $C^{\mathrm{H_p},t'} = 0$ may ultimately hinder model performance. To study the benefits of using dispersion damping, the dBHP is introduced in Scheme 4. Notably, Schemes 3 and 4 have the same number of fitting parameters due to fixing $\beta^{t,t'} = (B^{t,t'})^{-1}$. In the W99 FF [192, 193, 199] (and its revision [169]), it was deemed inadequate to have only one polarized hydrogen type and different types were defined based on their encompassing chemical moiety. In Scheme 5, I probe this need by distinguishing hydrogens bonded to nitrogen ($\mathrm{H_N}$) and hydrogens bonded to oxygen ($\mathrm{H_O}$). To achieve the best model performance, it is desirable to parametrize $B^{t,t'}$. However, for reasons discussed above, this seems infeasible using the BHP. While the dBHP does not decouple $A^{t,t'}$ and $B^{t,t'}$, the implicit introduction of $B^{t,t'}$ into the dispersion term, via the damping function, may restrain the values of $B^{t,t'}$ and alleviate the numerical difficulty in their estimation; this is considered in Scheme 6. Finally, in Scheme 7, I consider an 'overhaul' of the FF by switching from the (d)BHP to the MIP. As Schemes 6 and 7 have the same model complexity (i.e., $\dim(\bar{\boldsymbol{p}})$), I use this to compare the efficacy of the MIP and the dBHP.

6.1.7 Assessment Metrics

To appraise FF accuracy, I employ several assessment metrics. From CrystalEstimator, the energy and geometry residual values for each training structure are reported at the optimized

parameter values. These scores provide a high-level understanding of the model accuracy and the numerical difficulty of the parameter estimation, for a given parametrization.

However, beyond evaluating aggregate performance, these residuals can be difficult to interpret in practical terms and offer limited physical insight. To assess the geometry prediction accuracy, CSP practitioners more often rely on COMPACK [242] RMSD_N scores. Here, I consider RMSD_{15} scores between the calculated crystal structures and their corresponding CE–755 reference geometries. COMPACK is applied with the CCDC's default settings, except the distance/angle tolerances are increased to 40%/40°. If a 15-molecule match is unattainable, these thresholds are (simultaneously) increased in increments of 10 until a valid RMSD_{15} is obtained. In the most extreme case, 70%/70° tolerances are used. While using such loose tolerances can be inaccurate [243], only a few structures (~2–3 out of 755) per parametrization require this treatment, hence I do not expect major inaccuracies in the averaged RMSD_{15} scores. A more alarming mode of COMPACK failure arises due to 'molecular fusing'. When the intermolecular model is extremely imbalanced, unreasonably small intermolecular distances can arise in the optimized crystal geometry. COMPACK interprets this unnatural interatomic separation as the presence of a covalent bond, fusing neighboring molecules together. COMPACK (by default) cannot compare systems which it identifies as different molecules, hence the optimized crystal will match 0 molecules with the reference, and an error is flagged. While there are workarounds to this issue, these involve significant manual treatment. Instead, I report these cases as 'Failed Structures', reflecting the gross error in their crystal geometries.

To evaluate the U_{inter} accuracy, I consider the deviations in U_{inter} (dU_{inter}) given by

$$dU_{\text{inter}} = U_{\text{inter}}^{\text{calc}} - U_{\text{inter}}^{\text{ref}}, \tag{6.1.14}$$

where $U_{\text{inter}}^{\text{calc}}$ is the U_{inter} value calculated with the parameter estimates and $U_{\text{inter}}^{\text{ref}}$ is the reference U_{inter} from the CE–755. $dU_{\text{inter}} > 0$ indicates underbinding of the crystal by the model, while $dU_{\text{inter}} < 0$ indicates overbinding. Average deviations (AD) and average absolute deviations (AAD) are easily determined, while the unbiased standard deviation in dU_{inter} (SD) is given by

$$\text{SD} = \sqrt{\frac{\sum_{s=1}^{NS}(dU_{\text{inter},s} - \text{AD})^2}{NS - 1}}, \tag{6.1.15}$$

where NS is the number of evaluated structures. A large AD can indicate systematic errors in the predictions and physical deficiencies in the model. Yet, CSP is primarily focused on *relative* polymorphic energy differences, hence models with large systematic errors can still be useful. To an extent, the same is also true of the AAD (a large AD will lead to a large AAD). Lowering the SD is most likely to improve CSP outcomes, since it reduces random errors in the model.

6.2 Force-field Training

The optimized parameters for the different fitting schemes (and electrostatic models) are tabulated in Appendix J. Below, I analyze the training set performance of each scheme.

6.2.1 Parameter Estimation Residual Scores

The training (TRN) set residual scores are presented in Figure 6.2. In all of the fitting schemes, the use of multipole electrostatics leads to lower energy and geometry residual scores, versus their point-charge counterparts. By extension, the objective function (OF, used interchangeably with TMF) values of the multipole models are lower across the board. Theoretical [221] and empirical [47, 48, 93, 166, 167] evidence corroborate the inadequacies of the atomic-charge model. Since the electron density around atoms is non-isotropic, it is sensible that embedding this anisotropy into the electrostatic description would improve the residual scores.

Generally, the use of PCM also leads to improved performance for both electrostatic models. This nett improvement seems to be predominantly driven by better accuracy in U_{inter}; the geometry residuals benefit less from PCM, and in some cases actually worsen. For the multipole models, applying PCM lowers energy residual scores by 0.06–0.09 units across all schemes, except for Scheme 1 which decreases by only ~0.03 units. For the point-charge models, Schemes 2–7 exhibit a 0.03–0.10 unit reduction in the energy residual score, except for Scheme 6 where the energy residual score increases when using PCM. The behavior of Scheme 6 is likely attributable to the numerical difficulty in obtaining effective parameter estimates with this scheme, rather than a physical deficiency with the model. Scheme 1 presents with a more modest reduction in the energy residual score. It can be inferred that applying PCM benefits both NHB and HB crystals, but more so the latter: Schemes 2–7, which fit to both NHB and HB energy residuals,

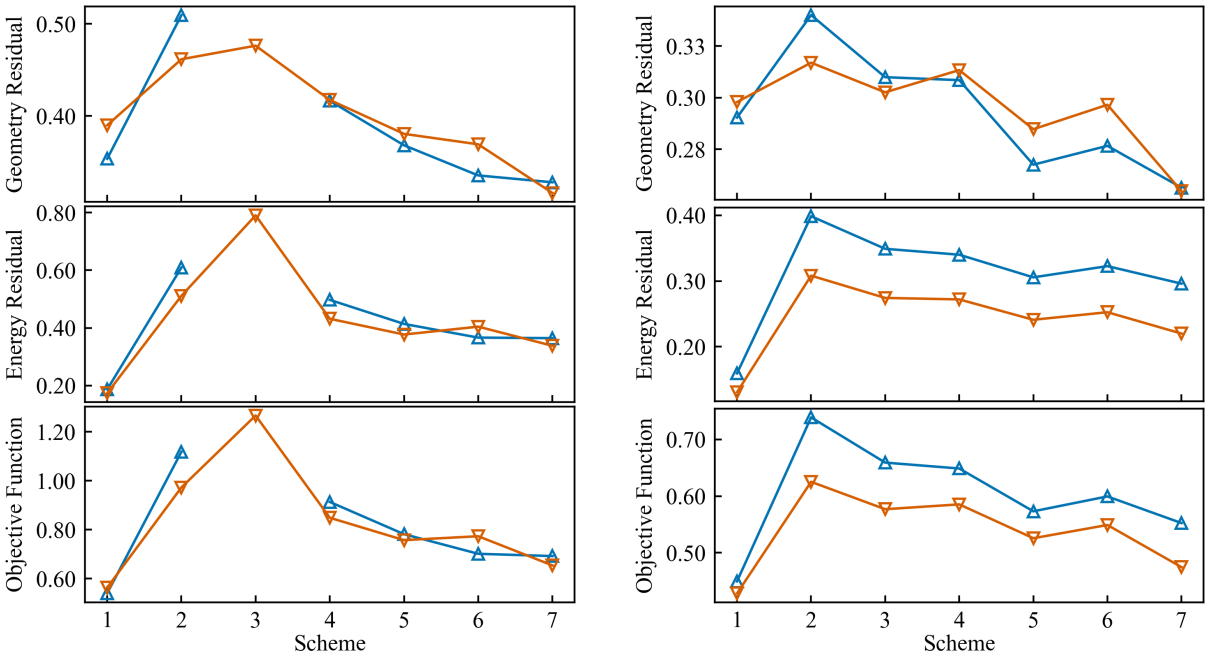


Figure 6.2: Sum of residual scores across the training set for the different fitting schemes, using (a) point-charge electrostatics, and (b) multipole electrostatics. Results derived without and with PCM are given by the blue △ and orange ▽ respectively. The gap in (a) is due to the parameter estimation failing over all attempted Sobol' points and no solution being found.

present with a disproportionately greater reduction in the energy residual scores compared to Scheme 1, which omits the HB energy residuals. From the energy residuals alone, the exact nature of this improvement is unclear, but as evidenced later, it is likely related to the implicit inclusion of induction effects achieved through PCM. The effect of PCM on the geometry residuals is slightly erratic, hence I reserve the discussion about geometry accuracy for later.

At this stage, it is also instructive to examine the numerical behavior of the parameter estimation. Given a physically reasonable model, as more fitted parameters are introduced, a least-squares optimization should always result in a set of model parameters that are better than (or as good as) the same model fitted with fewer parameters. The OF values largely adhere to this principle, with three exceptions. First, the low OF scores of Scheme 1 are simply the result of using fewer residuals than the other schemes during the parameter estimation. Thus, Scheme 1 should not be compared with the other models based on training set residual scores alone.

Second, Schemes 5 ($\dim(\bar{\boldsymbol{p}}) = 90$) and 6 ($\dim(\bar{\boldsymbol{p}}) = 135$) both employ the dBHP with identical atom types, but the latter also includes the estimation of $B^{t,t'}$. Despite solving a more-relaxed optimization problem, Scheme 6 always fails to achieve a smaller OF value than Scheme 5 (except for the "Point-charge + No PCM" case). Like its BHP counterpart, the dBHP still has a poorly conditioned solution surface when using Scheme 6. While the $B^{t,t'}$ parameters are embedded into the dispersion term through the damping function, the strong coupling between $A^{t,t'}$ and $B^{t,t'}$ continues to dominate the parameter estimation and the Scheme 6 FFs often have unphysically large $A^{t,t'}$ values. Overall, this poor numerical behavior hinders the progress of the gradient-based optimization, leading to suboptimal parameter estimates. Like the BHP [2, 166, 169, 192, 213], $B^{t,t'}$ cannot be reliably parametrized even when using the dBHP. Crucially, the MIP used in Scheme 7 has the same number of fitted model parameters as the dBHP in Scheme 6, yet there is no evidence that the MIP suffers from the same numerical challenges. In fact, the OF values obtained with Scheme 7 are consistently the lowest, for a given electrostatic model. This supports that the MIP is physically reasonable for use in HAIEFF models, and has the added benefit of being more numerically amenable to parametrization than the (d)BHP.

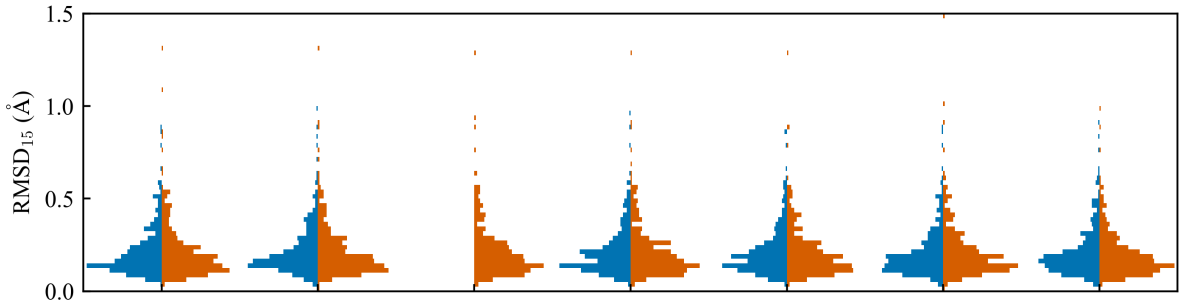
Lastly, Schemes 2 ($\dim(\bar{\boldsymbol{p}}) = 64$) and 3 ($\dim(\bar{\boldsymbol{p}}) = 72$) both employ the BHP with identical atom types, but the latter permits non-zero $C^{\mathrm{H_P},t'}$ values. Using the point-charge model, Scheme 3 completely fails when PCM is not utilized and has a larger OF value than Scheme 2 when PCM is employed. The trend across the different parametrizations indicate that multipoles and PCM (independently) contribute to better fidelity in the electrostatic model. When they are omitted, the empirical FF shoulders a greater burden to absorb these model errors into the FF parameters. However, the BHP was not formulated to describe Coulombic interactions, hence assimilation of these large errors is likely leading to physically unsound parameters that expose the deficiencies of the pair potential (i.e., the short-range divergence of the BHP). This culminates in the complete failure of the model (the failed Scheme 3), or poor numerical behavior during optimization (Scheme 2 outperforming Scheme 3). The 'over-burdening' of the FF term is supported by the fact that when multipole electrostatics are used, the relationship between Schemes 2 and 3 adheres to expectations based on their respective $\dim(\bar{\boldsymbol{p}})$.

6.2.2 CSP Error Metrics

To rationalize the results of the model training, I consider a more nuanced view of energy and geometry errors. In this section, I discuss the $RMSD_{15}$ scores and dU_{inter} for point-charge models (Figures 6.3 and 6.4 respectively) and multipole models (Figures 6.5 and 6.6 respectively).

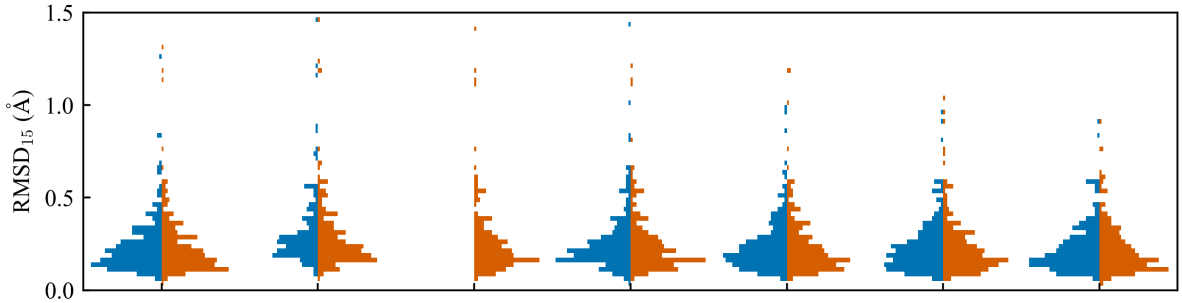
6.2.2.1 Non-hydrogen Bonding (NHB) Crystal Structures

For the NHB systems, crystal geometries are predicted well for all the models that were successfully fitted: average $RMSD_{15}$'s do not exceed 0.220 Å ("Scheme 4 + Point-charges + PCM"), and critically none of the structures present with molecular fusing. Across all of the fitted models, the distribution of $RMSD_{15}$ scores is similar, with a largely unimodal distribution of errors, that is skewed towards lower $RMSD_{15}$ values. The peak of the distributions is usually near/slightly below the average $RMSD_{15}$, and the majority of structures have $RMSD_{15} \leq 0.50$ Å (a common



Scheme	1	2	3	4	5	6	7
Fail	0 | 0	0 | 0	0 | 0	0 | 0	0 | 0	0 | 0	0 | 0
Avg.	0.213 | 0.216	0.219 | 0.217	N/A | 0.217	0.217 | 0.220	0.215 | 0.214	0.205 | 0.215	0.209 | 0.200
Max.	0.896 | 1.319	0.983 | 1.321	N/A | 1.287	0.952 | 1.287	0.872 | 1.294	0.881 | 1.489	0.918 | 0.979

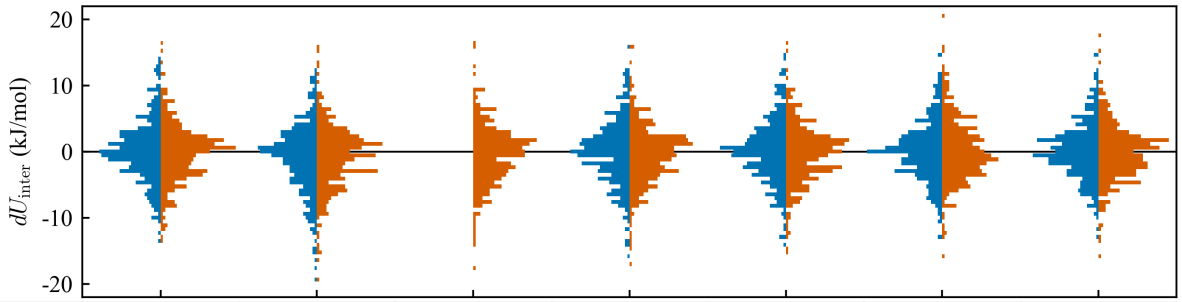
(a)



Scheme	1	2	3	4	5	6	7
Fail	0 | 0	86 | 23	0 | 30	17 | 7	3 | 0	0 | 0	0 | 0
Avg.	0.239 | 0.243	0.316 | 0.289	N/A | 0.259	0.262 | 0.256	0.242 | 0.243	0.235 | 0.238	0.228 | 0.224
Max.	1.264 | 1.310	1.468 | 1.465	N/A | 1.419	1.428 | 1.201	0.985 | 1.198	0.967 | 1.043	0.911 | 0.923

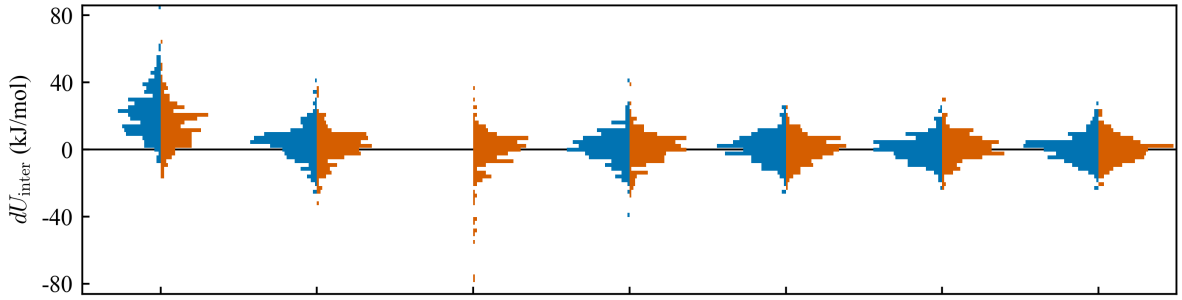
(b)

Figure 6.3: Histograms of $RMSD_{15}$ scores (Å) across Schemes 1–7 for (a) NHB, and (b) HB structures in the TRN set using point-charge electrostatics. The tables contain the number of failed structures, average $RMSD_{15}$, and maximum $RMSD_{15}$. Failed structures are omitted from the average. Results derived without and with PCM are given in blue and orange respectively, and are separated by '|'. Scheme 3 fails without PCM.



Scheme	1	2	3	4	5	6	7
AD	0.22 | -0.22	-0.91 | -0.84	N/A | -0.39	-0.10 | -0.44	0.07 | -0.30	0.17 | 0.28	0.23 | -0.15
AAD	3.54 | 3.42	4.01 | 3.71	N/A | 3.49	3.75 | 3.46	3.58 | 3.45	3.44 | 3.74	3.33 | 3.23
SD	4.69 | 4.52	5.31 | 4.87	N/A | 4.61	4.95 | 4.57	4.74 | 4.54	4.55 | 4.90	4.39 | 4.28

(a)



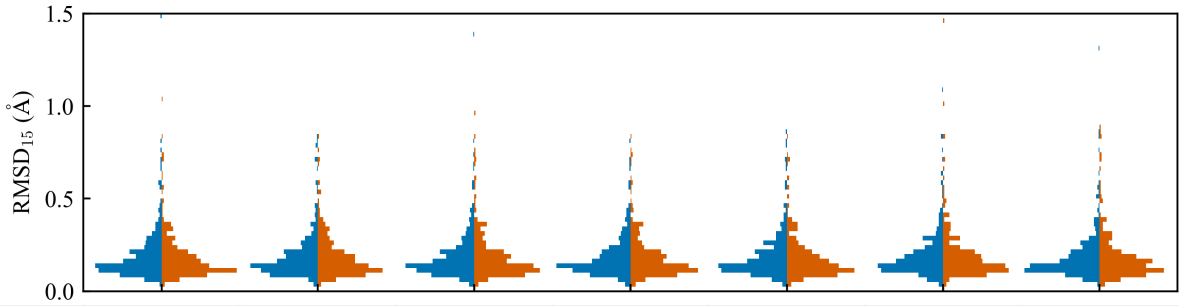
Scheme	1	2	3	4	5	6	7
AD	22.12 | 13.46	4.00 | 1.87	N/A | -1.79	1.37 | 0.70	0.97 | 0.32	0.49 | 1.21	0.70 | 0.19
AAD	22.34 | 14.98	8.47 | 7.69	N/A | 9.24	7.39 | 6.98	6.60 | 6.29	5.94 | 6.34	6.18 | 6.01
SD	13.87 | 12.54	10.15 | 9.93	N/A | 14.32	9.70 | 9.16	8.36 | 7.96	7.68 | 8.10	8.00 | 7.66

(b)

Figure 6.4: Histograms of dU_{inter} (kJ/mol) across Schemes 1–7 for (a) NHB, and (b) HB structures in the TRN set using point-charge electrostatics. The tables contain the AD, AAD, and SD statistics. Results derived without and with PCM are given in blue and orange respectively, and are separated by '|'. Scheme 3 fails without PCM.

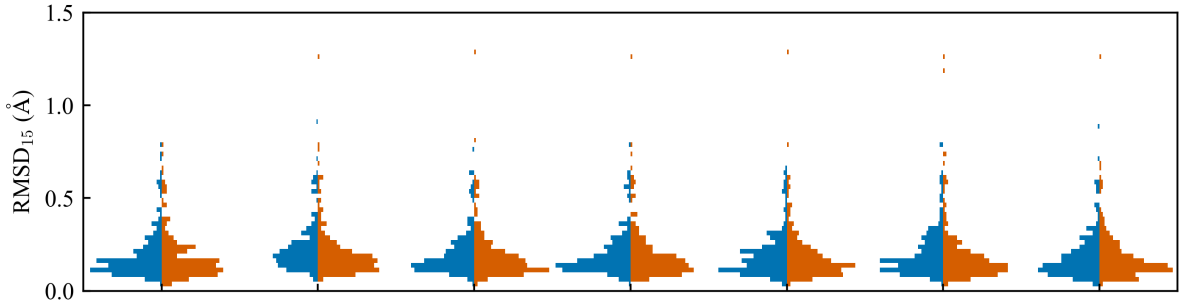
benchmark for identifying structural matches in CSP). Above this threshold, there are a handful of structures with poorly predicted geometries forming the tails of the distribution. Fortunately, these cases are relatively sparse in the NHB systems. The largest contributor to the accuracy of the geometry predictions appears to be the choice of electrostatic model: average RMSD_{15} scores reduce from 0.200–0.220 Å in the point-charge models to 0.183–0.194 Å in the multipole models. It does seem intuitive that there would be a strong correlation between the packing geometry and the anisotropy of the intermolecular forces. However, using multipole electrostatics does not necessarily lead to a reduction in the maximum RMSD_{15} scores, suggesting that the outliers in RMSD_{15} may have other deficiencies limiting their geometry accuracy. For the same electrostatic model, the different fitting schemes have only a marginal effect on the geometrical accuracy. Likewise, the application of PCM has little effect on the geometrical fidelity.

Intermolecular energies for the NHB structures are also predicted well across all of the parametrizations. Generally, the distributions of dU_{inter} are (almost) symmetric, uni-modally distributed, and centered close to 0. Advancing to more-complex fitting schemes, using PCM,



Scheme	1	2	3	4	5	6	7
Fail	0 | 0	0 | 0	0 | 0	0 | 0	0 | 0	0 | 0	0 | 0
Avg.	0.191 | 0.184	0.189 | 0.187	0.187 | 0.187	0.188 | 0.185	0.189 | 0.184	0.194 | 0.193	0.183 | 0.186
Max.	1.498 | 1.028	0.829 | 0.842	1.375 | 0.966	1.514 | 0.839	0.875 | 0.841	1.085 | 1.461	1.325 | 0.884

(a)



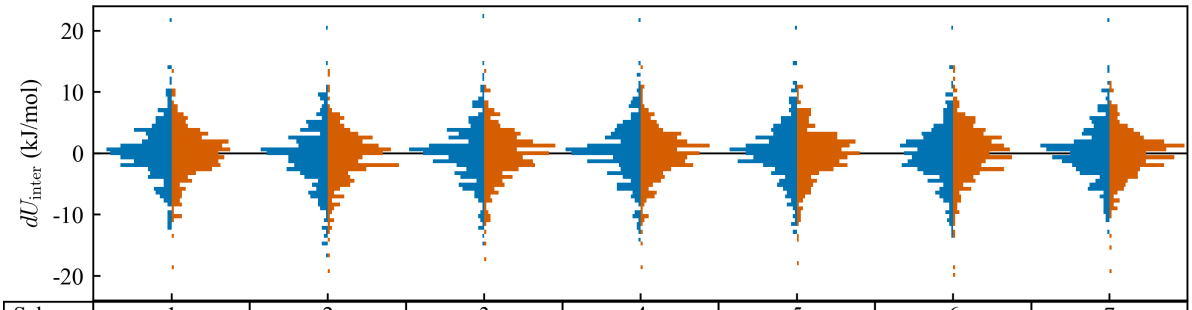
Scheme	1	2	3	4	5	6	7
Fail	0 | 0	39 | 3	24 | 0	2 | 0	0 | 0	0 | 0	0 | 0
Avg.	0.192 | 0.194	0.243 | 0.212	0.208 | 0.200	0.204 | 0.204	0.196 | 0.194	0.204 | 0.191	0.187 | 0.183
Max.	0.786 | 1.681	1.770 | 1.695	1.758 | 1.697	1.514 | 1.704	0.674 | 1.691	0.794 | 1.259	0.899 | 1.269

(b)

Figure 6.5: Histograms of $RMSD_{15}$ scores (Å) across Schemes 1–7 for (a) NHB, and (b) HB structures in the TRN set using multipole electrostatics. The tables contain the number of failed structures, average $RMSD_{15}$, and maximum $RMSD_{15}$. Failed structures are omitted from the average. Results derived without and with PCM are given in blue and orange respectively, and are separated by '|'.

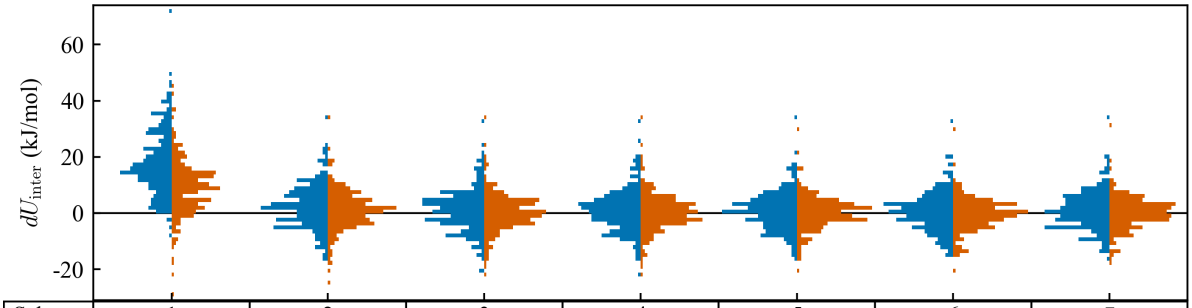
and/or switching to multipole electrostatics all improve the predictions of U_{inter}, but this gain in accuracy is modest at best. For instance, "Scheme 1 + Point-charges + No PCM" and "Scheme 7 + Multipoles + PCM" can be reasoned to be the most distinct models, yet AAD and SD decrease by only 0.61 kJ/mol and 0.78 kJ/mol, respectively, from the former to the latter.

It is instructive to consider the effects of PCM on the NHB structures more closely. Generally, the use of PCM encourages stronger cohesive energies, which manifests as a decrease in ADs. As discussed in Section 2.5.2.1, the application of PCM during SCF calculations polarizes the electron density of the molecule, implicitly embedding induction effects into the resulting charge density [188, 191]. This would lead to greater stabilization contributed by U_{elec}, relative to when PCM is not used, explaining the reduction in ADs. This begs the question of how important the choice of ε_{PCM} is for model performance. Choosing a dielectric constant that exceeds what is 'typical' for organic crystals, one might expect over-polarization of the molecule and a strong bias for overbinding; yet, in the NHB structures this is not observed. Using PCM, AD decreases only



Scheme	1	2	3	4	5	6	7
AD	0.21 | -0.14	-0.52 | -0.68	0.08 | -0.20	0.12 | -0.20	0.15 | -0.17	-0.33 | 0.10	0.12 | -0.19
AAD	3.28 | 3.06	3.63 | 3.25	3.40 | 3.12	3.39 | 3.12	3.33 | 3.06	3.45 | 3.32	3.23 | 2.93
SD	4.49 | 4.07	4.91 | 4.34	4.68 | 4.18	4.65 | 4.19	4.55 | 4.13	4.61 | 4.43	4.37 | 3.91

(a)



Scheme	1	2	3	4	5	6	7
AD	17.92 | 9.10	2.51 | 1.05	1.11 | 0.19	1.03 | 0.19	0.79 | 0.03	0.55 | -0.52	0.77 | 0.04
AAD	18.07 | 10.86	6.43 | 5.27	5.92 | 4.94	5.74 | 4.91	5.19 | 4.40	5.40 | 4.29	5.33 | 4.27
SD	11.51 | 9.97	7.91 | 7.09	7.71 | 6.85	7.53 | 6.77	6.80 | 6.03	7.04 | 5.83	6.94 | 5.81

(b)

Figure 6.6: Histograms of dU_{inter} (kJ/mol) across Schemes 1–7 for (a) NHB, and (b) HB structures in the TRN set using multipole electrostatics. The tables contain the AD, AAD, and SD statistics. Results derived without and with PCM are given in blue and orange respectively, and are separated by '|'.

marginally, and AAD/SD actually improve despite the exaggerated value of ε_{PCM}. To an extent, the empirical FF parameters would compensate for the errors introduced by using $\varepsilon_{\text{PCM}} = 11$. However, given that the HB structures (in Schemes 2–7) exhibit much larger reductions in AD when using PCM, relative to the NHB structures, this suggests that the effects of PCM are also inherently influenced by the molecular chemistry. I rationalize this as follows.

The large continuum dielectric bolsters the polarizability of only the surrounding 'crystalline environment'. However, during the SCRF calculation, each molecule's inherent polarizability acts as a limiting factor in its interaction with the PCM field. Since most NHB structures do not contain highly polarizable atomic environments, the induced response of each molecule to the PCM field is relatively restrained and only marginal enhancement of the cohesive energy is observed. This view aligns with the parameter estimation results of Bowskill [2]: U_{inter} accuracy improves significantly when progressing from a model without PCM to one using $\varepsilon_{\text{PCM}} = 3$, but there is diminishing gain in model accuracy from increasing ε_{PCM} further. In the first 'regime', the continuum has a large inductive effect on the molecule, but in the second 'regime'

the molecule resists further polarization, even as the continuum becomes more polarizable. Altogether, the error associated with an excessive $\varepsilon_{\mathrm{PCM}}$ value is likely smaller than one might expect, such that the empirical FF can accommodate it. Thus, the benefits of using PCM can be reaped even if the chosen $\varepsilon_{\mathrm{PCM}}$ is not commensurate with the molecule's polarizability.

Overall, all of the fitted models achieve similarly acceptable accuracy in describing the geometries and energies of the NHB crystals. The RMSD_{15} scores for most structures are in line with accepted CSP standards, while the energy error statistics are similar to/better than those of the dispersion-corrected DFT models benchmarked by Nyman *et al.* (AAD ~3.8–4.8 kJ/mol and SD ~4.1–5.9 kJ/mol, although the DFT models consider U_{latt} and only U_{inter} is considered here) [66]. The largest improvement in model accuracy emerges when progressing from point-charge to multipole electrostatics, which is manifested predominantly in reduced RMSD_{15} scores, likely from introducing anisotropy. While increasing the FF complexity does benefit accuracy, these gains are minor. Likewise, using PCM benefits the NHB systems only minutely, but it is notable that (within reason) there is some flexibility in choosing the $\varepsilon_{\mathrm{PCM}}$ value.

6.2.2.2 Hydrogen Bonding (HB) Crystal Structures

It is the HB crystals for which the inadequacies of the BHP emerge. Scheme 1 considers the same fitting approach as Bowskill [2], which crucially, omits energetic data for the HB systems. Using Scheme 1, the geometry predictions for the HB crystals are of relatively good quality, only slightly worse than the NHB systems and without any molecular fusing. However, limiting the fitting data in this manner comes at the serious expense of the U_{inter} accuracy. All models derived using Scheme 1 grossly underbind the HB systems, as indicated by their large, *positive* ADs (and large AADs). Compared to Scheme 1 results for the NHB systems, the ADs for the HB systems are 1–2 orders of magnitude larger, while the AADs are larger by a factor of 3–6.

The propensity for the Scheme 1 models to underbind the HB crystals can be reasoned as follows. Electrostatic stabilization from hydrogen bonding interactions should have a significant induction contribution [71, 183, 189]. Yet, the HAIEFF electrostatic model that relies on the isolated-molecule assumption completely neglects polarization, which is induced by the crystalline environment. Thus, the electrostatic model will predict a shortfall in cohesive energy, that (without fitting to the energy residuals) the FF is unable to correct for, resulting in the positive ADs. When an implicit induction model like PCM is applied, the Scheme 1 AD (and AAD) decrease dramatically (~1.6–2x) due to the corresponding enhancement of U_{elec}. To a lesser extent, Scheme 1 also suggests that some of the underbinding in U_{elec} is contributed by the electrostatic model itself: AD and AAD decrease by ~4 kJ/mol when switching from the point-charge model to the multipole model. Alarmingly, when using Scheme 1, the SD in dU_{inter} for the HB systems is untenably large. Whilst the utilization of PCM and/or multipole electrostatics does reduce the SD, even in the best case it exceeds 9.97 kJ/mol. This suggests that the predicted U_{inter} are susceptible to random errors which far exceed expected polymorphic energy differences ($\leq$ 4 kJ/mol in most cases [21, 24]), rendering relative stability rankings com-

pletely unreliable. In fairness, a CSP study is usually focused on a specific molecular system and likely benefits from systematic error cancellation (i.e., the 'effective' SD is probably smaller). Nevertheless, there is a clear need to reduce the energy errors for the HB systems.

Scheme 2 attempts to achieve this by introducing the HB energy residuals into the parameter fitting training data. Doing so allows the empirical FF to 'correct' the U_{elec} deficiencies, leading to a marked reduction in energy errors. However, as alluded to in Section 6.2.1, when the FF absorbs too much error, the resulting FF parameters are extremely unphysical. In this case, many $A^{t,t'}$ involving H_p are inadvertently set to zero, which is likely a consequence of fixing $C^{\text{H}_\text{p},t'} = 0$. By prohibiting non-zero $C^{\text{H}_\text{p},t'}$, interactions involving H_p cannot contribute any attractive stabilization and the only recourse for the parameter fitting to address crystal underbinding is to reduce repulsive interactions (hence $A^{\text{H}_\text{p},t'} \to 0$). The trade-off for exploiting $A^{\text{H}_\text{p},t'}$ to improve energy accuracy manifests in the geometry predictions. The Scheme 2 FFs have a poor balance between repulsions and attractions such that molecular fusing is rampant. The extent to which FF error-assimilation is leveraged (and correspondingly the number of fused structures) is commensurate with the fidelity of the underlying electrostatic model. In the worst case ("Point-charges + No PCM"), 86 HB structures exhibit fusing. As the electrostatic model improves, either by using PCM and/or multipoles, the U_{elec} shortfall lessens and the number of fused HB structures decreases: using multipoles with PCM, only three structures fuse.

Although the "Scheme 2 + Multipoles + PCM" model offers a reasonable balance of energy and geometry accuracy for HB systems, this is suitable for only the Refinement stage of the CSP workflow. For the Global Search stage, a reliable point-charge model is required – which Scheme 2 fails to provide. Recognizing the error-correction role that H_p interactions play in the intermolecular model, Scheme 3 permits $C^{\text{H}_\text{p},t'} \geq 0$, contrary to the recommendation of Refs [2, 169]. For the multipole models, this proves to be beneficial. Relative to Scheme 2, there are fewer fused structures (none with PCM) and all the energy error metrics decrease. Conversely, the point-charge models derived using Scheme 3 exhibit a worse performance than with Scheme 2, despite the availability of more fitting parameters. When PCM is used, the energy errors and the number of fused structures all increase. Several key HB pair potentials (e.g., $\text{O}/\text{H}_\text{p}$, $\text{N}/\text{H}_\text{p}$, and $\text{H}_\text{p}/\text{H}_\text{p}$) are described by extremely small $A^{t,t'}$ and non-zero $C^{t,t'}$. Whilst $C^{\text{H}_\text{p},t'} > 0$ is an intended consequence of Scheme 3, the continued preference for small/zero-valued $A^{t,t'}$ results in pair potentials that are wholly attractive (i.e., the attractive term dominates at *all* interatomic distances). This egregious balance between repulsions and attractions leads to the uptick in molecular fusing and poor accuracy in predicted U_{inter}. When PCM is not used, a greater U_{elec} shortfall will arise and one can expect the severity of these unphysical pair potentials to be magnified. Indeed, I observe crystal structures frequently 'collapse' during their rigid minimization and persistent failure ultimately triggers CrystalEstimator to abort. This occurs over all Sobol' seeds, resulting in no optimal parameter estimates.

Both Schemes 2 and 3 suggest that the BHP cannot provide a reliable FF when the electrostatic model is too coarse (i.e., atomic-charges). In an attempt to compensate for the electrostatic model's inadequacy, the unphysical region of the BHP parameter space is explored during the

parameter estimation, resulting in unsound FFs. Attempting to constrain the BHP parameters away from this unphysical region is cumbersome since $A^{t,t'}$ and $C^{t,t'}$ do not have a clear relationship with features of their pair potential. Instead, I propose that augmenting the underlying form of the FF may be a more effective strategy to improve the model. In Scheme 4, simply by substituting the BHP for the dBHP, the parameter estimation (and model) performance is significantly enhanced, despite no other changes from Scheme 3. A solution for the point-charge model without PCM is found successfully, while molecular fusing is lessened for all models. The theoretical basis of the damping function (i.e., diminishing attractions in the region of electron overlap) is clearly supported by the improved performance achieved by Scheme 4.

In spite of the previous modifications, molecular fusing remains prevalent in the point-charge models, especially when PCM is not used. Motivated by the trends in Figure 5.6, I consider the adequacy of using only one polarized-hydrogen type. Over the entire CE–755, it is observed that hydrogen bonds involving oxygen-bonded H (H_O) are significantly shorter than their nitrogen-bonded counterparts (H_N). For instance, when the hydrogen-acceptor is nitrogen or oxygen, H_O forms hydrogen bonds of ~1.6 Å while H_N forms hydrogen bonds of ~1.9 Å. This aligns with the experimentally reported interatomic distances for HB organic crystals, including carboxylic acids, amides, and alcohols [343]. The inherent difference in hydrogen bond lengths is substantial enough to invalidate lumping these two hydrogens together; thus, in Scheme 5, I distinguish H_O and H_N. The effect of this is marked: almost all occurrences of molecular fusing are eliminated in both the point-charge and multipole models (without PCM, the former still has 3 cases of fusing). Between Schemes 4 and 5, the SDs decrease by ≥1.2 kJ/mol and ~0.7 kJ/mol for the point-charge and multipole models respectively.

As discussed earlier, any attempt to further parametrize the dBHP model parameters (i.e., Scheme 6) is prone to poor numerical behavior and subpar parameter estimates. Little to no advantage in model accuracy is extracted by fitting $B^{t,t'}$ in the BHP/dBHP. With the dBHP, estimating the value of $\beta^{t,t'}$ could be an option to increase model flexibility, without facing the numerical challenge of parametrizing $B^{t,t'}$. However, in my view, the dBHP represents a 'band-aid' solution for the deficiencies of the BHP, and ineffective selection of $\beta^{t,t'}$ would undo any benefits of the current dBHP model. Instead, in Scheme 7, I consider if using the MIP may be a more efficient means to bolster model performance. Since $\lambda_r^{t,t'}$ is required to exceed $\lambda_a^{t,t'}$, the MIP will *always* have a repulsive wall to resist molecular fusing. Moreover, the depth of the attractive well is directly tied to $\varepsilon^{t,t'}$, such that extremely attractive interactions (like hydrogen bonds) can be described without jeopardizing the repulsive region of the pair potential. The results with Scheme 7 corroborate that these characteristics of the MIP do lead to the desired improvement in model predictions, both for geometry and energy. Comparing Schemes 5, 6, and 7, the energy accuracy of the MIP slightly edges out the dBHP models, but crucially the MIP suppresses all occurrence of molecular fusing across all fitted models.

To summarize, the challenge in describing HB systems with HAIEFF models stems from inadequacies in the electrostatic model. To achieve good energy predictions, the FF must be leveraged to shore up U_{elec} shortfalls, but without a suitably formulated FF, there is a strong

trade-off between geometry and energy accuracy. This is apparent for the BHP, which exhibits unphysical behavior when used to describe strongly attractive interactions, like hydrogen bonds. By using a dBHP or MIP, and recognizing the inherent differences between H_O/H_N, I obtain FFs that provide satisfactory accuracy, in both geometry and energy, for HB systems.

Overall, I put forward that Schemes 5 and 7 are best suited for deriving FFs capable of describing NHB and HB crystals across all stages of the CSP workflow. Both schemes are numerically amenable to parameter fitting and provide reasonable accuracy for all electrostatic models. The efficacy in describing HB systems is slightly worse than the NHB systems, but this is expected given the additional flaws in the electrostatic model which affect the former. For all systems, the training results strongly support the use of PCM, even with 'large' ε_{PCM}, since this seems to benefit model performance without any obvious drawbacks. As such, the remaining cases of molecular fusing observed in the "Scheme 5 + Point-charges + No PCM" model can be neglected, since one would be remiss not to use its PCM-inclusive alternative instead.

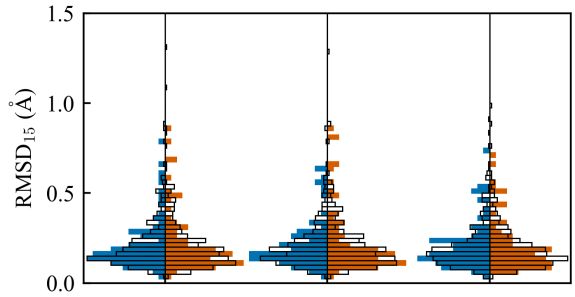
6.3 Force-field Transferability Validation

Since the fitted HAIEFF FFs are intended to be used transferably, verifying that they retain their accuracy on structures beyond the training set is vital. Bowskill [2] has asserted that the assimilation of large errors into the FF can compromise transferability, hence this validation is paramount. To that end, a subset of the CE–755 was withheld from FF training (Section 5.5). Each structure in this test (TST) set is fixed at its crystalline conformation from the CE–755, and the corresponding electrostatic features (atomic-charges/distributed-multipoles) are derived using the methodology in Section 6.1.1. The test structures are subject to rigid-body minimization of U_{inter}, with the complementary FFs parametrized in the preceding section. FF transferability is then judged by whether the distribution of RMSD_{15} scores and dU_{inter} deteriorate significantly between the training and test sets. The change in error (ΔE) between the training and test sets is determined according to

$$\Delta E = E_{\text{TST}} - E_{\text{TRN}}, \qquad E \in \{\text{AD}, \text{AAD}, \text{SD}, \text{Avg. RMSD}_{15}, \text{Max. RMSD}_{15}\}, \tag{6.3.1}$$

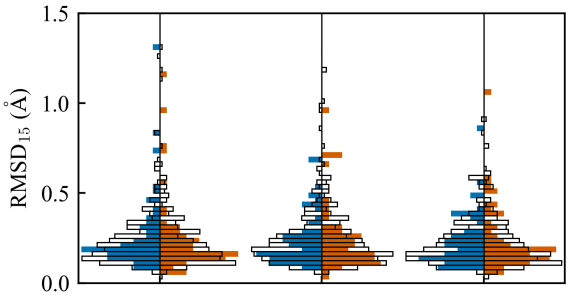
where E_{TST} and E_{TRN} are the error metrics for the test and training set respectively. When extrapolating the model to the test set, $\Delta E > 0$ is expected, but this should be modest for transferable models. This test is performed with the FFs fitted by Schemes 1, 5, and 7.

The RMSD_{15} distributions for the training and test sets are compared in Figures 6.7 and 6.8. In all cases, there are no obvious shifts in the RMSD_{15} distribution. While the average RMSD_{15} usually increases in the test set, indicated by $\Delta\text{Avg.} > 0$, the gain is acceptable in practical terms: the largest ΔAvg. is only 0.023 Å (HB structures using "Scheme 7 + Multipoles + No PCM"). The degree of FF transferability (for geometry predictions) is mostly unaffected by the underlying QM/electrostatic model: changes in the QM/electrostatic model, for a given scheme, do not exhibit any clear trends in ΔAvg. Likewise, none of the fitting schemes seem to preserve



Scheme	1	5	7
ΔAvg.	0.007 | 0.007	0.001 | 0.012	0.004 | 0.008
ΔMax.	-0.100 | -0.447	-0.238 | -0.443	-0.175 | -0.259

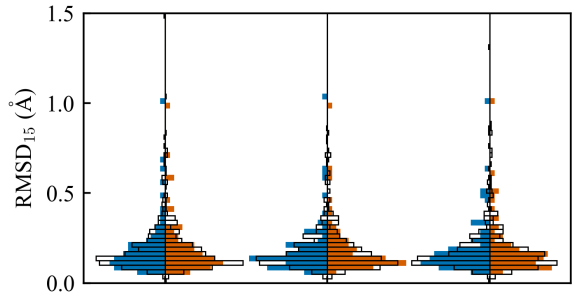
(a)



Scheme	1	5	7
ΔAvg.	0.014 | -0.004	-0.004 | -0.003	0.000 | 0.008
ΔMax.	0.051 | -0.150	-0.297 | -0.237	-0.051 | 0.137

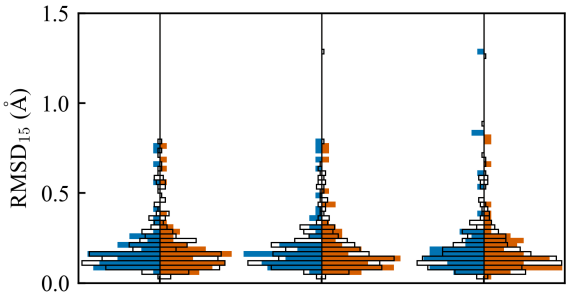
(b)

Figure 6.7: Histograms of $RMSD_{15}$ scores (Å) across Schemes 1, 5, and 7 for (a) NHB, and (b) HB structures using point-charge electrostatics. The tables contain the *change* in average and maximum $RMSD_{15}$. TST set results derived without and with PCM are given in blue and orange respectively, and are separated by '|'. TRN set results are overlaid as unfilled bars.



Scheme	1	5	7
ΔAvg.	0.007 | 0.008	0.003 | 0.009	0.008 | 0.022
ΔMax.	-0.473 | -0.048	0.154 | 0.143	-0.308 | 1.063

(a)

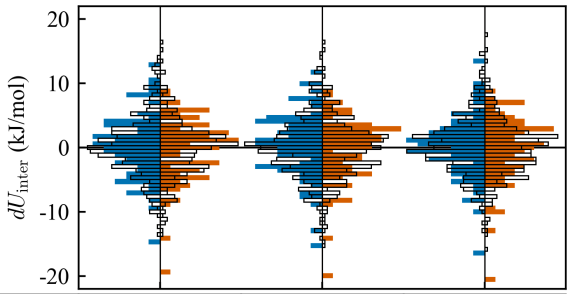


Scheme	1	5	7
ΔAvg.	0.007 | -0.002	0.020 | 0.001	0.023 | 0.010
ΔMax.	-0.012 | -0.929	0.122 | -0.928	0.397 | -0.446

(b)

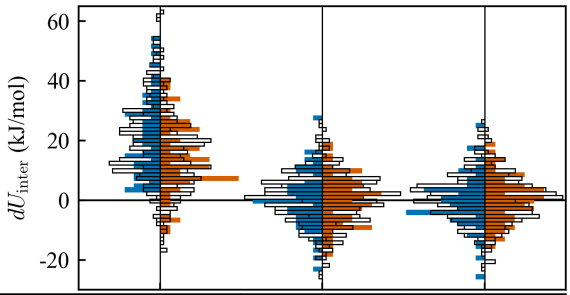
Figure 6.8: Histograms of $RMSD_{15}$ scores (Å) across Schemes 1, 5, and 7 for (a) NHB, and (b) HB structures using multipole electrostatics. The tables contain the *change* in average and maximum $RMSD_{15}$. TST set results derived without and with PCM are given in blue and orange respectively, and are separated by '|'. TRN set results are overlaid as unfilled bars.

their geometry accuracy dramatically better than the others. Ultimately, the results in Figures 6.7 and 6.8 indicate that geometry accuracy is adequately retained from the training to test set. In the worst case, only a small (positive) ΔAvg. is observed and, for all models, the vast majority of test structures still achieve the threshold of $RMSD_{15} \leq 0.50$ Å. It is notable that in moving from the training to test set with a given model, both positive and negative model attributes are retained. Models which led to molecular fusing among HB structures in the training set continue to predict fusing in the test set HB structures (e.g., "Scheme 5 + Point-charges + No PCM" has 3 fused training HB structures, and 1 fused test HB structure). Conversely, models which eliminate fusing in the training set also suppress fusing in the test set (i.e., all other models). This justifies the previous efforts to mitigate molecular fusing in the FF training.



Scheme	1	5	7
ΔAD	-0.002 | 0.028	0.271 | 0.241	0.027 | 0.152
ΔAAD	0.171 | 0.317	0.188 | 0.124	0.165 | 0.374
ΔSD	0.176 | 0.333	0.151 | 0.216	0.298 | 0.556

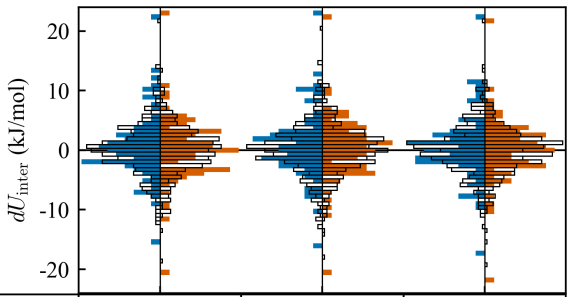
(a)



Scheme	1	5	7
ΔAD	-1.006 | 0.529	-0.879 | 0.009	-0.977 | 0.209
ΔAAD	-1.224 | 0.233	0.267 | 0.088	0.032 | -0.107
ΔSD	-1.206 | -1.072	0.653 | -0.021	0.446 | -0.219

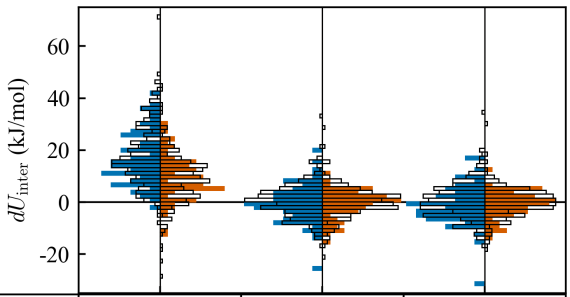
(b)

Figure 6.9: Histograms of dU_{inter} (kJ/mol) across Schemes 1, 5, and 7 for (a) NHB, and (b) HB structures using point-charge electrostatics. The tables contain the *change* in AD, AAD, and SD. TST set results derived without and with PCM are given in blue and orange respectively, and are separated by '|'. TRN set results are overlaid as unfilled bars.



Scheme	1	5	7
ΔAD	0.448 | 0.415	0.514 | 0.648	0.315 | 0.591
ΔAAD	0.466 | 0.598	0.524 | 0.483	0.427 | 0.563
ΔSD	0.891 | 1.227	0.878 | 1.010	0.990 | 1.215

(a)



Scheme	1	5	7
ΔAD	-1.000 | 0.422	-1.283 | -0.654	-1.845 | -0.796
ΔAAD	-1.105 | -0.220	0.252 | -0.461	0.400 | -0.375
ΔSD	-1.313 | -1.127	0.509 | -0.846	0.740 | -0.676

(b)

Figure 6.10: Histograms of dU_{inter} (kJ/mol) across Schemes 1, 5, and 7 for (a) NHB, and (b) HB structures using multipole electrostatics. The tables contain the *change* in AD, AAD, and SD. TST set results derived without and with PCM are given in blue and orange respectively, and are separated by '|'. TRN set results are overlaid as unfilled bars.

The corresponding analysis for dU_{inter} is presented in Figures 6.9 and 6.10. Generally, the energy prediction accuracy is adequately retained between the training and test sets. Across all models tested, none of the energy error metrics (AD, AAD, SD) increase by more than 1.23 kJ/mol (SD for NHB structures in "Scheme 1 + Multipoles + PCM"). There are some indications that the transferability of the energy accuracy is affected by the underlying QM/electrostatic model. When describing the NHB structures, ΔAAD and ΔSD are slightly larger for the models employing multipoles versus point-charges (Figures 6.9a and 6.10a), which could indicate poorer transferability in energy predictions for the multipole models. Among the HB structures, neither AAD nor SD exhibit a similar bias towards either electrostatic model. Conversely, the use of PCM seems to benefit transferability for the HB structures, particularly in Schemes 5 and

Table 6.3: p-values for the Mann–Whitney U test ('MANWHIT'), the 'standard' t-test (HOMO-T), and Welch's t-test (HETERO-T), comparing the distributions of RMSD_{15} and dU_{inter} between the training and test sets. At a 5% significance level, the H_0 is accepted if $p > 0.05$.

Scheme	Type	RMSD_{15} Test	dU_{inter} Test		
		MANWHIT	MANWHIT	HOMO-T	HETERO-T
Point-Charges Without PCM					
1	NHB	0.988	0.935	0.997	0.997
	HB	0.984	0.609	0.595	0.575
5	NHB	0.899	0.423	0.638	0.645
	HB	0.666	0.354	0.456	0.477
7	NHB	0.902	0.650	0.960	0.962
	HB	0.834	0.354	0.385	0.401
Point-Charges With PCM					
1	NHB	0.776	0.572	0.959	0.961
	HB	0.423	0.700	0.757	0.745
5	NHB	0.589	0.255	0.664	0.673
	HB	0.629	0.889	0.994	0.994
7	NHB	0.923	0.298	0.775	0.791
	HB	0.945	0.931	0.843	0.841
Multipoles Without PCM					
1	NHB	0.903	0.824	0.428	0.479
	HB	0.780	0.539	0.522	0.493
5	NHB	0.630	0.635	0.368	0.420
	HB	0.784	0.208	0.181	0.201
7	NHB	0.900	0.788	0.569	0.616
	HB	0.900	0.070	0.062	0.080
Multipoles With PCM					
1	NHB	0.836	0.535	0.427	0.500
	HB	0.839	0.776	0.755	0.738
5	NHB	0.829	0.229	0.216	0.281
	HB	0.636	0.621	0.421	0.381
7	NHB	0.914	0.202	0.241	0.322
	HB	0.924	0.389	0.312	0.279

7 where the HB energy residuals are included in the training. In contrast, the transferability for the NHB structures tends to worsen with PCM. Overall, any biases towards a particular QM/electrostatic model are very subtle and not significant enough to draw concrete generalizations about the relationship between transferability and the QM/electrostatic model.

Unlike Scheme 1, Schemes 5 and 7 both include the HB energy residuals as part of the objective function. It is encouraging to see that for these latter two schemes, the assimilation of energy errors into the FFs does not lead to gross overfitting and the resulting FFs do not exhibit worse transferability (in geometry or energy accuracy) than those fitted by Scheme 1. Ironically, the largest increase in ΔSD comes from a Scheme 1 model. Between Schemes 5 and 7, there is no clear indication that either the dBHP or the MIP has superior FF transferability. It is worth noting that both Schemes 5 and 7 use the same training data, yet the latter involves

the estimation of 45 more empirical parameters. As such, there is an inherently greater risk of Scheme 7 overfitting the model to the training data. The fact that this is not observed is a testament to both the physical suitability of the MIP for describing organic crystals, as well as the adequacy of data diversity in the training and test sets.

From the ΔE values, I conclude that the fitted FFs exhibit satisfactory transferability for the prediction of crystal geometries and cohesive energies. Using Schemes 5 and 7, which were chosen based on their training set performance, average RMSD_{15} values do not increase by more than 0.023 Å between the training and test sets. Similarly, the increases in AAD and SD do not exceed 0.56 kJ/mol and 1.22 kJ/mol respectively.

As a final verification of transferability, I consider two statistical tests: the Mann–Whitney U test (MANWHIT) [344] and the t-test. Applying these to the distribution of RMSD_{15} scores and dU_{inter}, I test the null hypothesis (H_0): that the difference in error distributions between the training and test sets is not statistically significant. The MANWHIT test does not require the sample distributions to be normally distributed, while the t-test does. Depending on whether the samples are homoscedastic or heteroscedastic, either the 'standard' t-test (HOMO-T) or Welch's t-test (HETERO-T) is used. Since the RMSD_{15} scores are clearly non-normally distributed, I perform only the MANWHIT test on this data. In contrast, the dU_{inter} are not obviously skewed, hence I conduct all three tests (MANWHIT, HOMO-T, and HETERO-T). These statistical tests are performed using the 'mannwhitneyu' and 'ttest_ind' functions from the SciPy Python package. Across all models, the associated statistical tests consistently return p-values > 0.050. This indicates that the H_0 is not rejected, and there is no statistically significant difference between the training and test set error distributions. The closest case to rejecting H_0 is dU_{inter} for HB systems using "Scheme 7 + Multipoles + No PCM". Although the statistical tests return p-values close to the 0.050 threshold ($0.062 \leq p \leq 0.080$), Figure 6.10b reflects that ΔAAD and ΔSD are fairly small (< 0.75 kJ/mol), such that the (relatively) small p-value is not of significant concern. These statistical tests corroborate the previous analysis of ΔE, supporting that the model accuracy in the training and test sets is of similar quality.

6.4 Prediction of Relative Polymorphic Energies

The metrics studied in the previous sections are indicative of the *absolute* accuracy of the fitted FFs. However, evaluating accuracy in *relative* polymorphic energies (ΔU_{poly}) is of greater relevance to CSP. To that end, over the set $\boldsymbol{\mathcal{P}}$ of polymorphic pairs in the CE–755, I define $\Delta U_{\text{poly},i}$ for pair i as

$$\Delta U_{\text{poly},i} = U_{\text{inter},h,i} - U_{\text{inter},l,i}, \qquad i \in \boldsymbol{\mathcal{P}}, \tag{6.4.1}$$

where $U_{\text{inter},h,i}$ and $U_{\text{inter},l,i}$ are the intermolecular energies of the less stable (higher energy) and more stable (lower energy) polymorphic forms in pair i respectively. The stability order is based on the $U^{\text{ref}}_{\text{inter}}$ reported in the CE–755, with the reference relative polymorphic energies given by $U^{\text{ref}}_{\text{poly},i} = U^{\text{ref}}_{\text{inter},h,i} - U^{\text{ref}}_{\text{inter},l,i}$. $\boldsymbol{\mathcal{P}}$ contains all unique pairings of polymorphs in the CE-755,

which includes 14 NHB polymorphic pairs and 19 HB polymorphic pairs. Over the set $\boldsymbol{\mathcal{P}}$, three metrics are assessed:

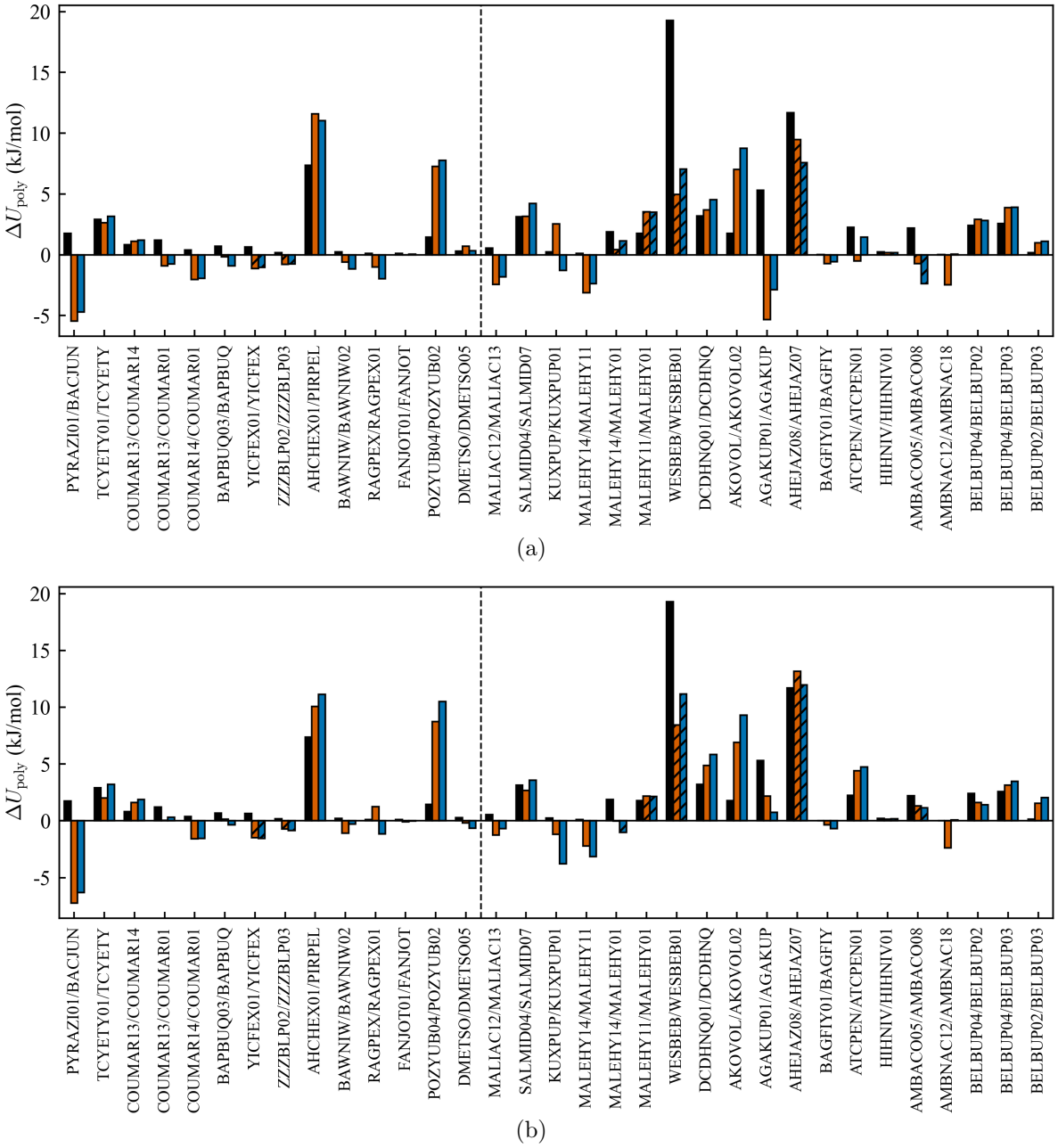
1. Rank Accuracy: Whether the order of relative polymorphic stability is correct. By definition, this is indicated by $\Delta U_{\text{poly},i} > 0$.
2. Deviations in $\Delta U_{\text{poly},i}$ ($d\Delta U_{\text{poly},i}$): This is given by $d\Delta U_{\text{poly},i} = \Delta U^{\text{pred}}_{\text{poly},i} - \Delta U^{\text{ref}}_{\text{poly},i}$. The AD, AAD, and minimum/maximum $d\Delta U_{\text{poly}}$ (Min/Max D) are determined accordingly.
3. Number of pairs with $|d\Delta U_{\text{poly},i}| \leq 5$ kJ/mol: 5 kJ/mol was chosen as an approximate value of 1 SD in dU_{inter}, when applying the fitted FFs to the test set.

$\Delta U_{\text{poly},i}$ values obtained using point-charge (Figure 6.11) and multipole (Figure 6.12) models, paired with their complementary Scheme 5 and 7 FFs, are compared to the reference data in Table 6.4. The discussion of the results apply similarly to the Scheme 5 and 7 models.

When using the point-charge models, only 35–50% of the NHB polymorphic pairs are correctly ranked. While the modeling of HB crystals was suggested to be more challenging, these systems exhibit a higher rank accuracy (63–74%). This is likely attributable to the fact that $\Delta U^{\text{ref}}_{\text{poly},i}$ values are inherently larger in the HB pairs (average $\Delta U^{\text{ref}}_{\text{poly}} = 3.10$ kJ/mol) than NHB pairs (average $\Delta U^{\text{ref}}_{\text{poly}} = 1.30$ kJ/mol). As such, the HB polymorphic pairs have more leeway to be correctly ranked, despite errors in ΔU_{poly}. This is supported by the NHB structures having lower AADs in ΔU_{poly} versus the HB structures, when PCM is not used. When PCM is used, the HB and NHB systems respond in opposite ways. For the NHB pairs, errors in ΔU_{poly} worsen slightly, with AAD increasing and the bounds on the Min/Max $d\Delta U_{\text{poly}}$ widening. Conversely, for the HB pairs AAD decreases and the Min/Max $d\Delta U_{\text{poly}}$ range narrows. Overall, these trends align with the training set results: energy accuracy for the NHB structures is insensitive towards the use of PCM, with only minor changes in ΔU_{poly} accuracy. In contrast, for the HB systems, the improved accuracy of ΔU_{poly} supports that there is benefit to employing PCM.

When using the multipole models, rank accuracy generally improves over the point-charge counterparts: 57–71% of the NHB pairs and 68–79% of the HB pairs are correctly ordered. For the NHB structures, this is reflected in a marked decrease in the AAD and a smaller Min/Max $d\Delta U_{\text{poly}}$ range. For the HB structures, rank accuracy also improves marginally and the range of the Min/Max $d\Delta U_{\text{poly}}$ values tightens (relative to the point-charge models). Oddly, however, the AAD for the HB systems worsens slightly. Overall, paired with the results from the training and test set analysis, I would still conclude that the multipole models offer better accuracy in predicting relative polymorphic energy differences, especially since there is strong theoretical justification for using multipoles. Like the point-charge models, when PCM is used in tandem with the multipole models, accuracy for the NHB systems slightly worsens, while improving for the HB systems (albeit to a lesser extent than in the point-charge case).

On the whole, the trends in ΔU_{poly} accuracy are only tenuously aligned with the results from the training and test set analysis. Whilst the use of multipoles and PCM have some positive impact on the predictions of ΔU_{poly}, this manifests inconsistently over the different error metrics and is fairly marginal. Given the (relatively) small sample size of polymorphic pairs, it may be



(a)

(b)

Figure 6.11: ΔU_{poly} (kJ/mol) when using a point-charge model (a) without PCM, and (b) with PCM. Each polymorphic pair is labeled with their CSD Refcodes ordered as 'More Stable/Less Stable'. $\Delta U_{\text{poly}}^{\text{ref}}$ values are given in black, while $\Delta U_{\text{poly}}^{\text{pred}}$ obtained using the Scheme 5 and 7 models are in orange and blue respectively. Hatched bars indicate that a member of the pair is from the test set. NHB and HB systems are on the left and right of the dashed line respectively.

ill-advised to draw any strong conclusions about the model's ability to predict ΔU_{poly}, since random error is more likely to obfuscate its true accuracy. From the test set, the SD in dU_{inter} is ~5 kJ/mol for the "Multipoles + PCM" models, while all other models have a slightly larger SD. From Table 6.4, the majority of polymorphic pairs have $|d\Delta U_{\text{poly},i}| \leq 5$ kJ/mol. In other words, for most systems, the uncertainty in the model's energy predictions exceeds the residual errors in ΔU_{poly}. It should be noted that the quality of the ΔU_{poly} predictions does not seem strongly

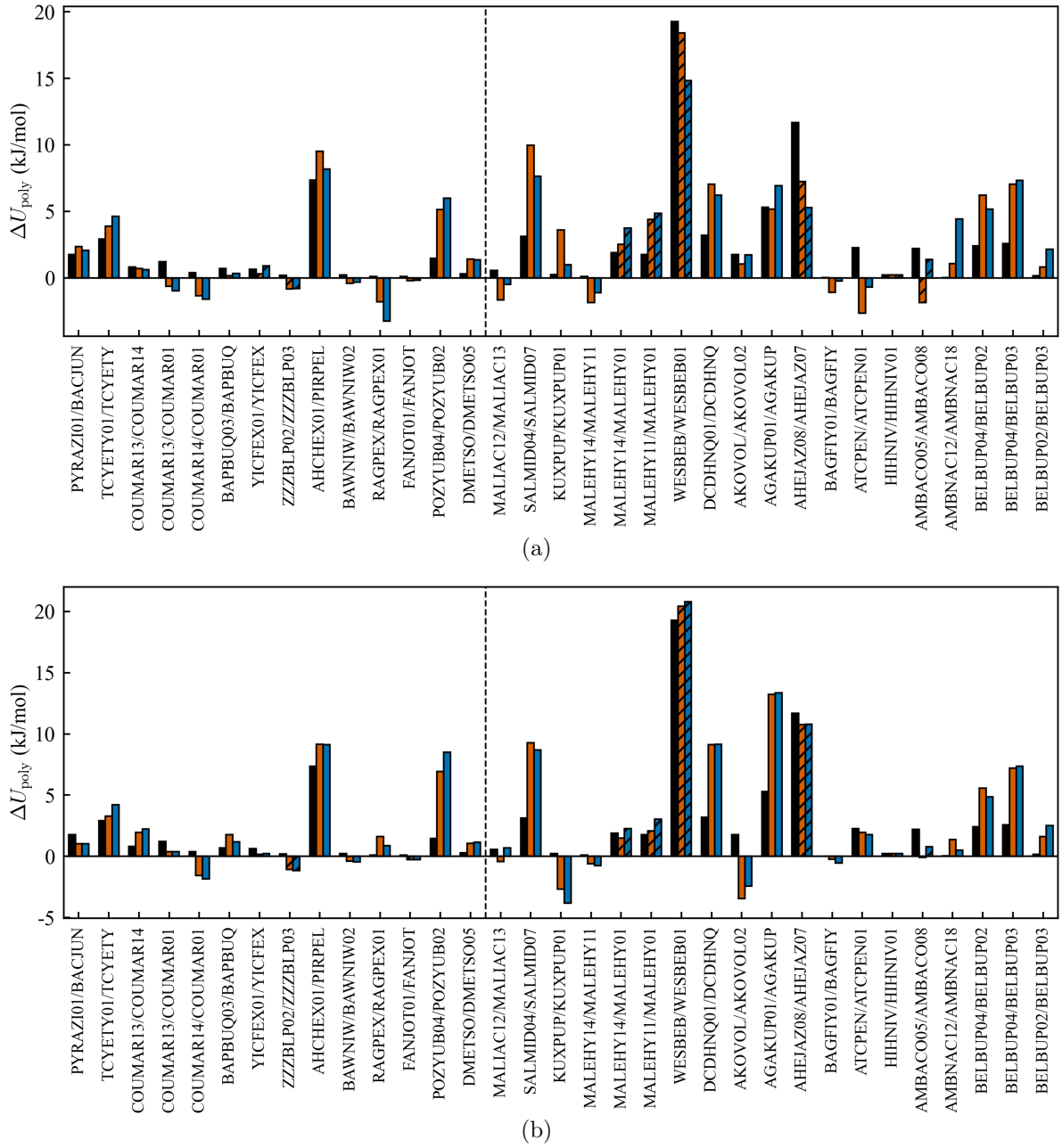


Figure 6.12: ΔU_{poly} (kJ/mol) when using a multipole model (a) without PCM, and (b) with PCM. Each polymorphic pair is labeled with their CSD Refcodes ordered as 'More Stable/Less Stable'. $\Delta U_{\text{poly}}^{\text{ref}}$ values are given in black, while $\Delta U_{\text{poly}}^{\text{pred}}$ obtained using the Scheme 5 and 7 models are in orange and blue respectively. Hatched bars indicate that a member of the pair is from the test set. NHB and HB systems are on the left and right of the dashed line respectively.

influenced by whether both or only one of the polymorphic forms were used in FF training. In fact, visually, the pairs involving 1 training set and 1 test set structure appear better predicted in some cases. While this could also be a result of random error, more likely it further indicates that the fitted FFs are indeed transferable.

Even if systematic trends across the models are masked by the effects of random error, there are six outliers with large errors that are unambiguously affected by the model:

1. **Pyrazine (PYRAZI01/BACJUN):** When using the point-charge models, there is large error in the relative polymorph stability ($d\Delta U_{\mathrm{poly},i} < -6.47$ kJ/mol). This is resolved by switching to the multipole models ($|d\Delta U_{\mathrm{poly},i}| < 0.74$ kJ/mol). Likely, π-interactions associated with the quadrupole moments on aromatic systems [166, 345–347] are poorly described with only atomic-charges. Similar behavior is observed with/without PCM.
2. **Naphthothiadiazole (POZYUB04/POZYUB02):** While relative stability is correctly predicted, $d\Delta U_{\mathrm{poly},i}$ are large for the point-charge models. Without PCM, $d\Delta U_{\mathrm{poly},i} > 5.8$ kJ/mol and PCM exacerbates this ($d\Delta U_{\mathrm{poly},i} > 7.4$ kJ/mol). Like pyrazine, the polyaromatic naphthothiadiazole should be modeled better with multipoles. Indeed, $d\Delta U_{\mathrm{poly},i}$ decrease with multipoles ($d\Delta U_{\mathrm{poly},i} > 3.7$ kJ/mol without PCM, and > 5.5 kJ/mol with PCM), albeit to a lesser extent than for pyrazine. Possibly, the remaining error stems from inadequacies in describing the thiadiazole moiety with the limited range of atom types.
3. **o-Fluoromandelic acid (WESBEB/WEBEB01):** Despite correctly predicting the relative stability order, the point-charge models exhibit poor quantitative accuracy in $\Delta U_{\mathrm{poly},i}$. Without PCM, $d\Delta U_{\mathrm{poly},i} < -12.2$ kJ/mol, while with PCM this improves to $d\Delta U_{\mathrm{poly},i} < -8.1$ kJ/mol. These errors reduce dramatically with the utilization of multipoles: $d\Delta U_{\mathrm{poly},i} < -0.9$ kJ/mol without PCM, and $d\Delta U_{\mathrm{poly},i}$ ~1.3 kJ/mol with PCM. This agrees with the previous assertion that HB interactions are better described with PCM and multipoles. It should be stressed that $\Delta U^{\mathrm{ref}}_{\mathrm{poly}}$ reflects only the relative *intermolecular* energy contribution. In this system, the ability to form/break intramolecular hydrogen bonds leads to highly disparate $\Delta U^{\mathrm{ref}}_{\mathrm{intra}}$ (~20 kJ/mol) between the forms. Their resulting relative U_{latt} (< 1 kJ/mol) is within expected values of polymorphic energy differences [21, 24].
4. **4-Pyridonethione (AKOVOL/AKOVOL02):** Using a point-charge model, the stability order of this pair is correctly predicted, but a large error of $d\Delta U_{\mathrm{poly},i} > 5.1$ kJ/mol is observed, both with and without PCM. Using multipoles, without PCM, the error falls significantly such that $d\Delta U_{\mathrm{poly},i} < 0.8$ kJ/mol. Continuing to apply PCM, however, leads to an overcorrection and the stability ranking is erroneously flipped ($d\Delta U_{\mathrm{poly},i} < -4.2$ kJ/mol).
5. **Benzimidazoquinazoline-6-thione (AGAKUP01/AGAKUP):** The stability ranking is incorrect when point-charges are used without PCM. With PCM applied, the ranking is corrected, but $\Delta U_{\mathrm{poly},i}$ remains in error by ~3.0–4.5 kJ/mol. Using multipoles, the rankings remain correct. Without PCM, the multipole model is acceptably accurate ($d\Delta U_{\mathrm{poly},i} < 1.6$ kJ/mol), but PCM significantly worsens this accuracy, with $d\Delta U_{\mathrm{poly},i}$ ~8.0 kJ/mol.
6. **Octa-3,5-diyn-2,7-diol (AHEJAZ08/AHEJAZ07):** When using either the point-charge or multipole models, relative stability ranking is correctly predicted. However, applying PCM has a significant impact on $d\Delta U_{\mathrm{poly},i}$. Without PCM, $|d\Delta U_{\mathrm{poly},i}| > 2.0$ kJ/mol, while with PCM $|d\Delta U_{\mathrm{poly},i}| < 2.0$ kJ/mol.

I also note unexpected behavior in four HB pairs, wherein the polymorphic relationship is better described with point-charges than multipoles. These include hydroxy-azabenzotriazole (BEL-

Table 6.4: Summary of the key performance metrics in the prediction of ΔU_{poly} for fitting Schemes 5 and 7, over the different models. These include the rank accuracy ('Rank Acc.', i.e., number of correctly ranked pairs), number of pairs with $|d\Delta U_{\text{poly},i}| \leq 5$ kJ/mol ('Within 5 kJ/mol'), the AD in ΔU_{poly}, AAD in ΔU_{poly}, and the minimum/maximum $d\Delta U_{\text{poly}}$ (Min/Max D respectively).

Fitting Scheme	Rank Acc.	Within 5 kJ/mol	AD (kJ/mol)	AAD (kJ/mol)	Min D (kJ/mol)	Max D (kJ/mol)	Rank Acc.	Within 5 kJ/mol	AD (kJ/mol)	AAD (kJ/mol)	Min D (kJ/mol)	Max D (kJ/mol)
	NHB Pairs (No. of pairs: 14)						HB Pairs (No. of pairs: 19)					
Point-Charges Without PCM												
5	5	12	−0.4982	2.0284	−7.2095	5.8038	12	16	−1.6516	2.9605	−14.2970	5.2426
7	6	12	−0.5605	2.0789	−6.4766	6.3002	13	16	−1.2549	2.7083	−12.2069	6.9931
Point-Charges With PCM												
5	7	12	−0.4756	2.1799	−8.9903	7.2868	13	17	−0.7118	2.0626	−10.8487	5.1393
7	5	12	−0.2706	2.2957	−8.0665	9.0485	14	17	−0.5435	2.2885	−8.0910	7.5383
Multipoles Without PCM												
5	8	14	0.0090	1.2077	−1.9009	3.6866	14	18	0.3638	2.5124	−4.9176	6.8459
7	8	14	−0.0859	1.3277	−3.3598	4.5339	15	18	0.6086	2.4176	−6.4060	4.7565
Multipoles With PCM												
5	10	13	0.4147	1.3062	−1.9593	5.4612	13	15	0.9494	2.4238	−5.2278	7.9281
7	10	13	0.4985	1.4442	−2.2296	7.0469	15	16	1.0753	2.3944	−4.1980	8.0506

BUP04/BELBUP02 + BELBUP04/BELBUP03), hydroxybenzamide (SALMID04/SALMID07), and dichloro-dihydroxy-naphthoquinone (DCDHNQ01/DCDHNQ). There is no obvious molecular feature linking these systems, hence it is unclear what is contributing to these results.

In Sections 6.2 and 6.3, I demonstrated that the FFs parametrized by Schemes 5 and 7 have improved accuracy in predicting *absolute* crystal energies, and this performance is transferable to other molecular systems. However, the assessment of *relative* polymorphic energies suggests that these improvements may not be adequate to enable reliable prediction of relative stability. Particularly, random error associated with the transferable FFs is likely still too large to resolve polymorphic energies. It is generally accepted that HAIEFF models based on transferable FFs are suitable for only intermediate stages in the CSP workflow, and not the final CSP ranking [48, 49]. With this in mind, the accurate prediction of relative polymorphic energies is an extremely challenging (and perhaps unrealistic) test of accuracy for the models developed here. Nevertheless, considering the role of these models in the CSP workflow, the reduction in errors achieved with the new FFs are still meaningful. Smaller energy errors enable tighter energy cutoffs in the filtering steps between CSP stages (Figure 2.1), while lower geometry error provide better starting points for optimization in subsequent Refinement stages. This can ultimately contribute to reduced computational load over the course of a CSP study.

6.5 Force-field *Ad Hoc* Deployment

Thus far, I have studied how augmenting the HAIEFF FF can enhance model performance. Critically, the empirical FF plays a large role of correcting errors in the electrostatic component of the HAIEFF model. Since these errors are tied to the details of the underlying *ab initio* model, this raises the question: "*To what extent can parametrized FFs be deployed with QM/electrostatic models other than that used in training?*". *Ad hoc* deployment of HAIEFF FFs is not uncommon, perhaps due to the difficulty associated with FF reparameterization.

To examine how model accuracy responds to changes in the QM/electrostatic model, the FF fitted with "Scheme 7 + Multipole + PCM" is chosen as a reference (i.e., the FF in Table J.27, denoted 'model'). This FF is used with several extrapolated models (i.e., *ab initio* models distinct from the training model, denoted 'extrap.') to conduct rigid minimizations of the test set structures. Finally, the energy and geometry error metrics are compared between the reference and extrapolated models. There is an inherent bias in favor of the reference model for training set structures, hence these are excluded from this analysis. The results of this study are tabulated in Table 6.5 and discussed below.

6.5.1 Sensitivity to Electronic Structure Method

I first consider the effects of altering the electronic structure method (Figure 6.13). The reference model employs multipoles derived from the PBE0 [123–125] charge density. Other DFT functionals that are often used with HAIEFF models include the Perdew–Burke–Ernzerhof (PBE) GGA functional ('PBEPBE' in G09) [123, 124], and the B3LYP hybrid functional [126–128].

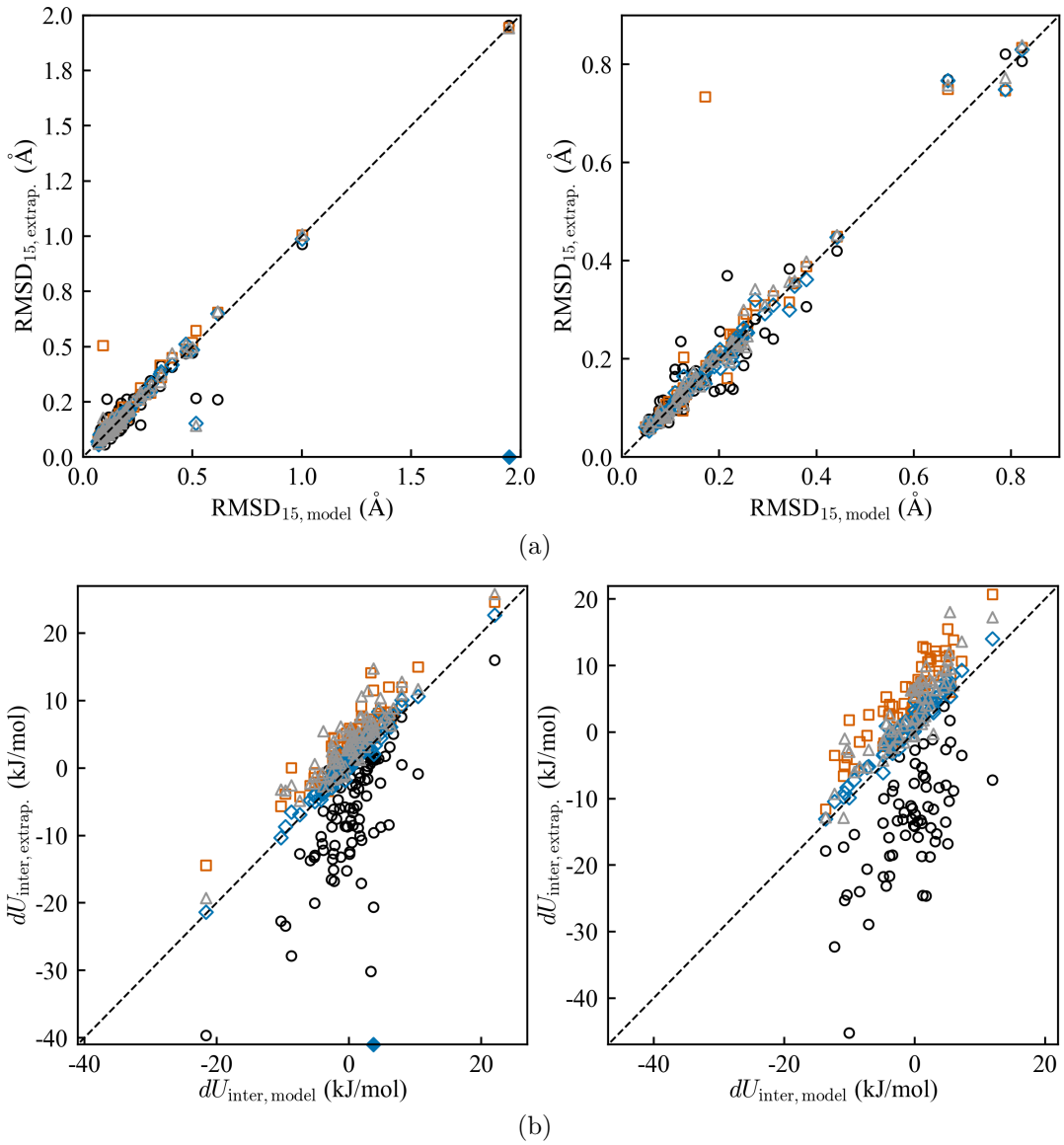


Figure 6.13: Parity plots of (a) $RMSD_{15}$ scores (Å), and (b) dU_{inter} (kJ/mol) between the fitted (PBE0/6-311G(d,p), MIP, rank 4 multipoles, ε_{PCM} = 11) and extrapolated models, for NHB (left) and HB (right) test set structures. The extrapolated models substitute PBE0 with: HF (black ○), PBE (orange □), B3LYP (blue ◇), and MP2 (grey △) during the QM calculations. The blue ◆ in the left plots corresponds to CLMETH03, which failed minimization with B3LYP.

Where DFT delocalization error is expected to be significant, (post-)Hartree–Fock theories like HF [348] and MP2 [349–353] have also been used with HAIEFF models [56, 90, 91, 144].

Irrespective of the electronic structure method, the accuracy of the geometry predictions is largely preserved. Few structures lie far from the 45° parity line, and the greatest increase in $RMSD_{15}$ occurs in 2,3,5,6-tetrafluoropyridin-4-amine (VEPXOG), when using the PBE functional (0.1712 Å → 0.7337 Å). Delocalization error [84] could be affecting the description of the halogenated-aromatic molecule in the VEPXOG structure, which would be most apparent in the PBE method. Overall, the average $RMSD_{15}$ scores do not change significantly in the extrapolated models, and in fact decrease marginally for B3LYP.

Conversely, energy accuracy exhibits strong sensitivity towards changes in the electronic structure method. Relative to U_{inter} from the CE–755, PBE, B3LYP, and MP2 underbind crystals more severely than PBE0, leading to more-positive AD and larger AAD scores. However, for these models, the change in dU_{inter} is fairly uniform, such that their SDs do not differ drastically from the PBE0 reference. The gain in SD relative to the reference model is roughly aligned with how 'dissimilar' the extrapolated model is to PBE0. B3LYP, a hybrid DFT functional, exhibits very little increase, while MP2, a post-HF method, exhibits the most. Unlike these three models, HF presents with large deviations from the PBE0 reference. Relative to U_{inter} from the CE–755, HF is more prone to overbinding crystals leading to a more-negative AD than PBE0. Moreover, there is greater variability in dU_{inter} when using HF, such that its data points are widely spread below the 45° line (Figure 6.13b) and SD increases markedly compared to PBE0.

For practical CSP, there appears to be some leeway to alter the electronic structure method without necessarily suffering a tangible loss in energy accuracy. For PBE, B3LYP, and MP2, predicted U_{inter} deviate uniformly from the PBE0 reference, thus these models should yield similar CSP outcomes (in terms of *relative* energies) to the reference model. However, there is a limit to this type of extrapolation. Pairing the PBE0-fitted FF with too simplistic of an *ab initio* method, like HF, is liable to introduce large, unpredictable errors into the calculated U_{inter}, thereby worsening the reliability of the CSP results. Fortunately, in modern CSP, HF is widely recognized as inadequate and would rarely be selected over the other models tested here.

6.5.2 Sensitivity to Basis Set

Depending on the molecular system of interest, it is typical to tailor the basis set to describe critical molecular features, whilst minimizing computational cost. In view of this, it is desirable to analyze model performance over a range of basis sets (Figure 6.14). As explained in Section 6.1.1, 6-311G(d,p) is a good 'general purpose' basis set, suitable for FF parametrization. In larger molecular systems, 6-311G(d,p) can become cost prohibitive, particularly for flexible CSP wherein *ab initio* geometry optimization is frequently conducted. In such cases, 6-31G(d,p) [341, 354–363] offers a practical alternative by using fewer Gaussian basis functions, but retaining the polarization functions that are useful for describing third-row elements and strained bonding. Going further, the 3-21G [364–369] basis set removes the polarization functions and uses even fewer Gaussian primitives. Conversely, I also consider the 6-311+G(d,p) [339–341, 370] basis set, wherein diffuse functions are added to the 6-311G(d,p) basis set. This would be desirable when describing diffused electron density, such as in anionic systems.

As before, accuracy in predicted crystal geometries is largely indifferent to the choice of basis set. Extrapolation to a highly contracted basis set, 3-21G, leads to the greatest loss in geometry accuracy, with the average RMSD_{15} rising by as much as 0.041 Å in the HB structures. Among the HB structures, large increases in RMSD_{15} (relative to the 6-311G(d,p) reference) are observed for: VEPXOG (0.1712 Å → 0.7267 Å), 3-amino-3-hydroxy-2-nitroprop-2-enamide (NMALAM01, 0.3794 Å → 1.2399 Å), and 2-amino-7-nitrofluorene (AMNIFL, 0.3444 Å → 1.0977

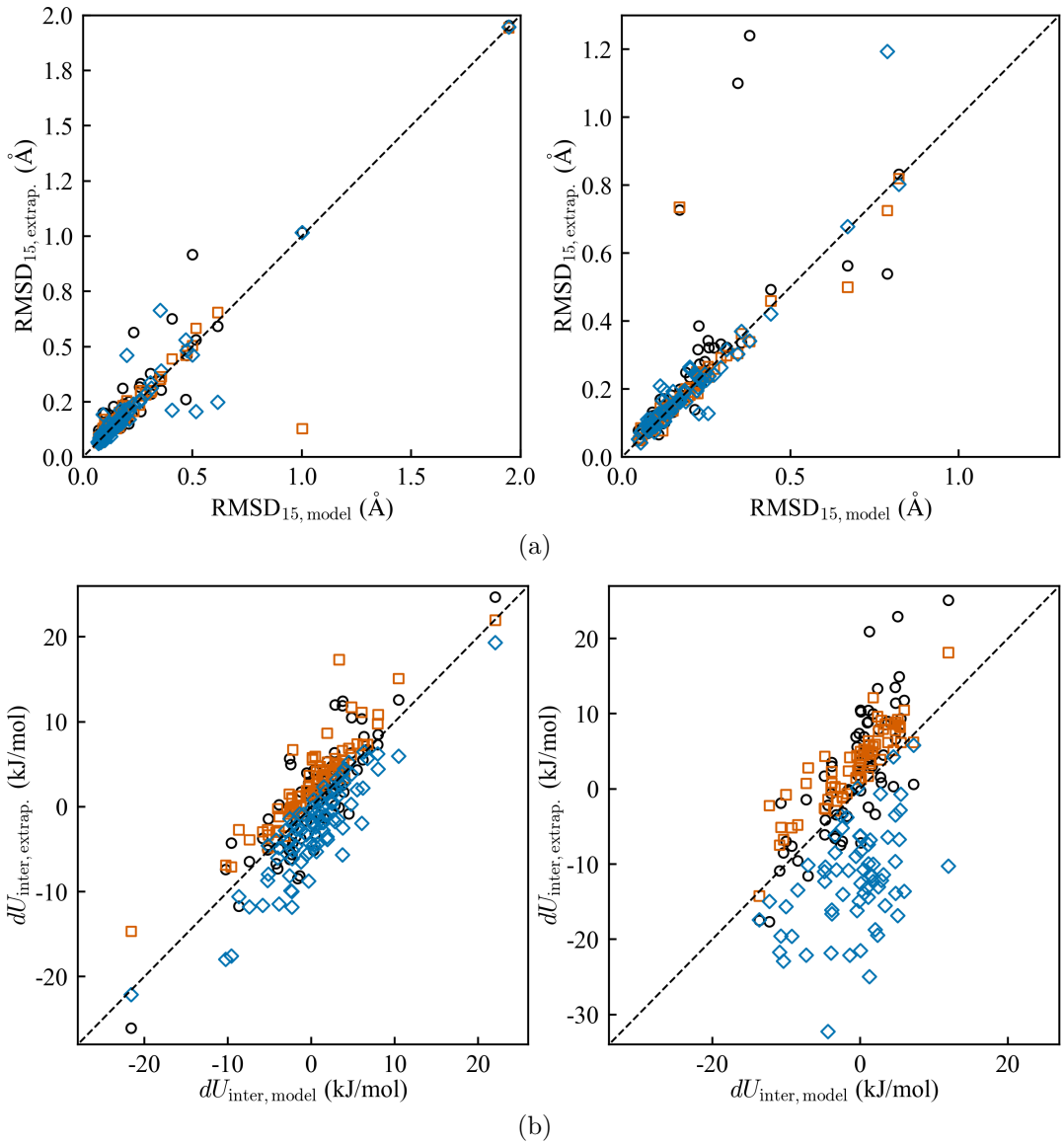


Figure 6.14: Parity plots of (a) $RMSD_{15}$ scores (Å), and (b) dU_{inter} (kJ/mol) between the fitted (PBE0/6-311G(d,p), MIP, rank 4 multipoles, ε_{PCM} = 11) and extrapolated models, for NHB (left) and HB (right) test set structures. The extrapolated models substitute 6-311G(d,p) with: 3-21G (black ○), 6-31G(d,p) (orange □), and 6-311+G(d,p) (blue ◇) during the QM calculations.

Å). The fluoroaromatic VEPXOG molecule likely requires a larger basis set to describe its complex electronic structure, while the nitro-containing NMALAM01 and AMNIFL may be sensitive to the absence of polarization functions. Using the 6-31G(d,p) basis set eliminates the erroneous behavior in these latter two structures, but VEPXOG continues to exhibit poor parity versus 6-311G(d,p) (0.1712 Å → 0.7350 Å). Using 6-311+G(d,p), all three structures regain similar geometric accuracy to 6-311G(d,p), but butan-2-ol's (ZEFPUX) geometry inadvertently worsens (0.7886 Å → 1.1930 Å).

Energy accuracy is more sensitive to these model changes. Generally, reducing the basis set size encourages greater underbinding of the crystals, relative to the U_{inter} from the CE–755. For modest basis set contractions (i.e., 6-31G(d,p)), the discrepancy with 6-311G(d,p) is systematic

and uniform, such that SD changes negligibly. For larger basis set contractions (i.e., 3-21G), additional variability in dU_{inter} emerges and the model's SD increases substantially (+1.3 kJ/mol in NHB structures, and +3.6 kJ/mol in HB structures). In contrast, extending the basis set with diffuse functions (i.e., 6-311+G(d,p)) promotes crystal overbinding, relative to the CE–755 U_{inter}. Like the 3-21G case, greater variance in dU_{inter} arises when using 6-311+G(d,p). For the NHB structures, this is modest and SD increases only marginally (+0.8 kJ/mol). However, for the HB structures, SD rises more substantially by +1.9 kJ/mol.

Overall, model accuracy exhibits greater vulnerability to basis set changes than to changes in the electronic structure method. If only the basis set size is altered (e.g., 6-311G(d,p) → 6-31G(d,p)), model fidelity is mostly preserved. However, changes to the underlying basis set construction (e.g., addition/removal of polarization/diffuse functions) are susceptible to generating spurious errors in both the energy and geometry predictions. This is intuitive, since these augmentations affect the capacity for certain molecular features to be described during the QM calculation. Altogether, a pragmatic parametrization strategy could be to fit independent FFs for each basis set construction, at a moderate basis set size (e.g., 6-311G, 6-311G(d,p), and 6-311+G(d,p)). These FFs should still be serviceable when used with smaller basis sets, as long as their underlying construction is unchanged (e.g., 6-31G, 6-31G(d,p), and 6-31+G(d,p)).

6.5.3 Sensitivity to Electrostatic Method

Over the multistage CSP workflow, point-charge and multipole electrostatic descriptions are typically used in the Global Search and Refinement stages, respectively. Despite this, it is not uncommon to use an identical empirical FF in both stages. For instance, during the Sixth Blind Test [48], several groups employed the FIT FF (which was parametrized using multipole electrostatics [166, 198]) throughout their CSP investigations. Analogously, the multipole-trained reference FF is paired with a point-charge electrostatic model instead (Figure 6.15).

Unlike its response to the electronic structure theory and basis set, geometry predictions are influenced significantly by the choice of electrostatic model. For the NHB crystals, one could be misled by the unchanged average RMSD_{15}. However, from the plot, it is clear that multiple structures are affected by the model change, and it is merely by chance that the structures with improved geometry predictions offset those with worsened predictions. For the HB crystals, the quality of the predicted geometries deteriorates for most structures and the average RMSD_{15} increases by 0.05 Å. U_{inter} predictions also exhibit strong sensitivity to the electrostatic model. When using the point-charge model, the NHB crystals are slightly more prone to underbinding, relative to their CE–755 U_{inter}. While the SD does not increase significantly (+0.3 kJ/mol), the change in dU_{inter} appear quite random from Figure 6.15b. For the HB systems, dU_{inter} unambiguously worsens when using the point-charge model: SD (11.0 kJ/mol) more than doubles from the reference (5.1 kJ/mol), and there is a gross underbinding of the HB crystals.

The responses of the energy and geometry accuracies are largely aligned with trends observed during the FF training. Per Section 6.2, employing multipoles benefits both geometry and energy

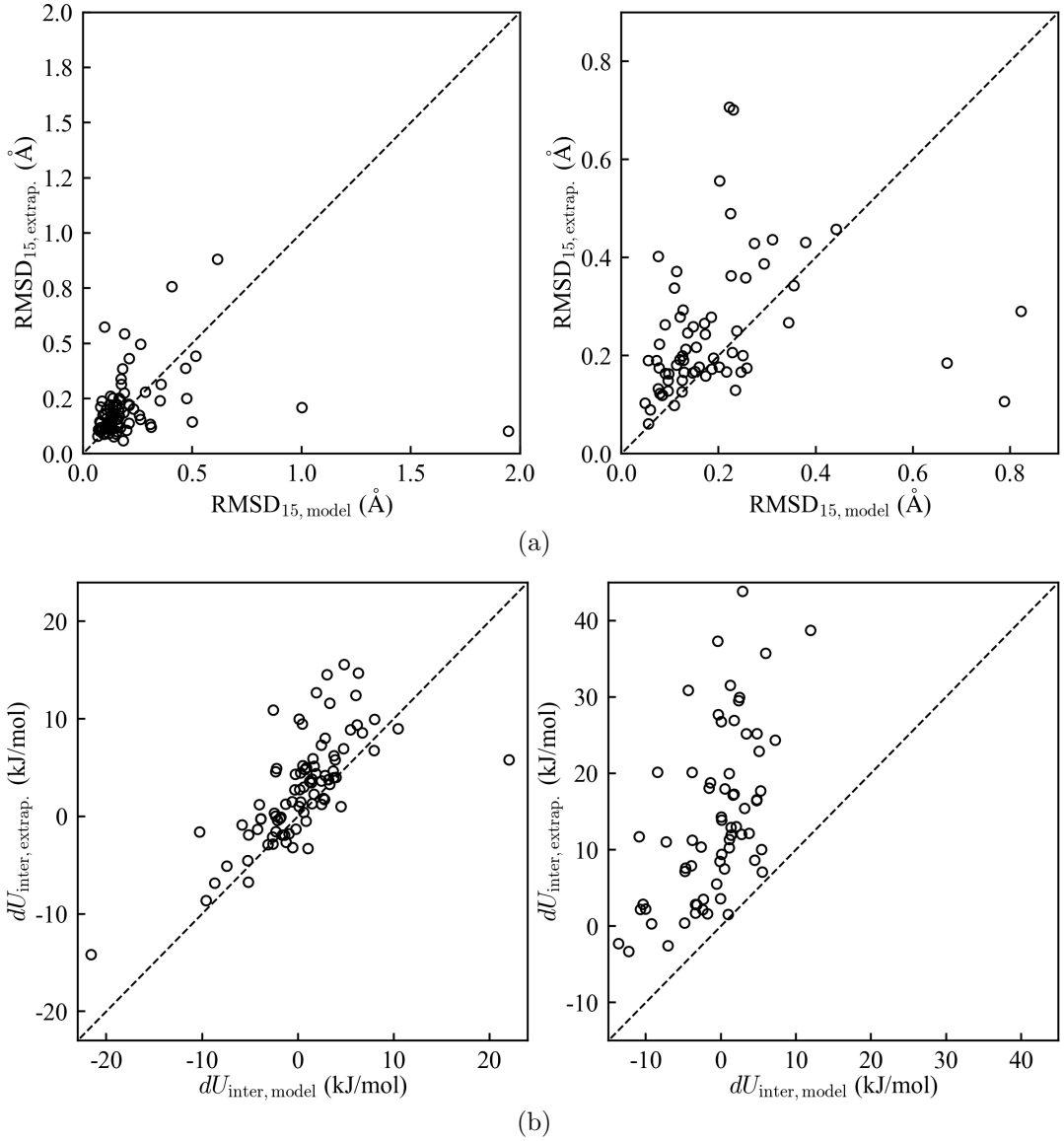


Figure 6.15: Parity plots of (a) $RMSD_{15}$ scores (Å), and (b) dU_{inter} (kJ/mol) between the fitted (PBE0/6-311G(d,p), MIP, rank 4 multipoles, ε_{PCM} = 11) and extrapolated model, for NHB (left) and HB (right) test set structures. The extrapolated model substitutes rank 4 GDMA multipoles with rank 0 G09 HLYGAt atomic-charges (black ○).

predictions, disproportionately so for the HB structures. For the HB structures, switching from point-charges to multipoles significantly enhances stabilization from U_{elec}. Since the reference FF was fitted with multipole electrostatics, it has not been biased by the errors associated with the point-charge model. This explains the worsened model performance and underbinding observed when using this FF with point-charge electrostatics. It can be inferred that the inverse case (i.e., using a point-charge-fitted FF with multipole electrostatics) is liable to double-counting the multipole-contributed cohesive energy, which may not only cause crystal overbinding, but also encourage molecular fusion. In short, the error-correction role of the empirical FF is closely tied to the electrostatic model, thus model accuracy does not extrapolate well, in either direction (i.e., point-charge ↔ multipole), particularly if the system of interest forms hydrogen bonds.

6.5.4 Sensitivity to PCM Settings

The final aspect of the QM model I study is the application of PCM. Purportedly, the environment in organic crystals should exhibit a dielectric constant of ~3 [218, 219], yet it is also reported that CSP outcomes can improve when $\varepsilon_{\mathrm{PCM}}$ exceeds this theoretical value [48, 188]. Because of this, the choice of $\varepsilon_{\mathrm{PCM}}$ can be rather arbitrary in practice. To avoid fitting FFs for each dielectric constant value, it would be useful to demonstrate if a PCM-trained FF exhibits transferability over a range of $\varepsilon_{\mathrm{PCM}}$. In Figure 6.16, I compare models using $\varepsilon_{\mathrm{PCM}}$ = 7, $\varepsilon_{\mathrm{PCM}}$ = 3, and no PCM, against the reference model ($\varepsilon_{\mathrm{PCM}}$ = 11).

In all three extrapolated models, there is almost perfect parity with the reference, in terms of geometry accuracy. For the NHB structures, there is a marginal decrease in average RMSD_{15} as the PCM strength is reduced (and ultimately removed), while for the HB structures the opposite is observed. From the parity plot, differences in geometry errors are negligible in all but a few structures. For cis-2,3-dichloro-1,4-dioxane (CDCDOX), a NHB structure, RMSD_{15} is consistently ~1.0 Å when PCM is used, but this falls sharply without PCM (0.1416 Å). Conversely, for chloromethane (CLMETH03), another NHB structure, geometry accuracy worsens as $\varepsilon_{\mathrm{PCM}}$ falls. At $\varepsilon_{\mathrm{PCM}}$ = 3, the rigid minimization of CLMETH03 fails, and without PCM the minimized geometry cannot generate a valid 15-molecule COMPACK match, even with widened tolerances. For the HB structures, the RMSD_{15} of ZEFPUX decreases from 0.7886 → 0.7819 → 0.7610 → 0.7153 Å as $\varepsilon_{\mathrm{PCM}}$ is lowered from 11 to no PCM, while in 2-methylimidazole (FULPIM), RMSD_{15} increases from 0.6705 → 0.6922 → 0.7421 → 0.8056 Å. It is unclear why these four structures respond distinctly to the PCM changes, but they are ultimately rare outliers.

Like the other extrapolated *ab initio* models, U_{inter} is strongly influenced by the PCM specification. As $\varepsilon_{\mathrm{PCM}}$ is reduced, the ADs of the extrapolated models consistently become more positive, for both NHB and HB structures. From Section 6.2.2, it was suggested that PCM is a proxy for induction and enhances U_{elec} stabilization. As $\varepsilon_{\mathrm{PCM}}$ is reduced, the 'induction' contribution from U_{elec} diminishes accordingly, yet the FF fails to accommodate this since it was conditioned on a higher $\varepsilon_{\mathrm{PCM}}$ value. This translates to the observed increase in underbinding. HB structures, which should have a greater 'induction' contribution, naturally exhibit a larger response compared to the NHB structures. It is notable that, for both NHB and HB crystals, AD increases non-linearly with reduction of $\varepsilon_{\mathrm{PCM}}$. From $\varepsilon_{\mathrm{PCM}}$ = 11 to $\varepsilon_{\mathrm{PCM}}$ = 7, AD increases by only +0.81/+1.52 kJ/mol (NHB/HB respectively). Decreasing further to $\varepsilon_{\mathrm{PCM}}$ = 3, AD rises by an additional +2.41/+4.67 kJ/mol. Finally, 'turning off' PCM causes AD to spike by +4.91/+9.94 kJ/mol. This supports the earlier assertion (Section 6.2.1) that PCM has the greatest influence at low $\varepsilon_{\mathrm{PCM}}$ values, after which there are diminishing effects to further increases in $\varepsilon_{\mathrm{PCM}}$. As AD increases in the extrapolated models, their AADs also rise proportionately. Most importantly is the effect these model changes have on the SD. When using $\varepsilon_{\mathrm{PCM}}$ = 3 or 7, the underbinding occurs uniformly over most structures, such that SD changes negligibly for both the NHB and HB systems (and sometimes even decreases). In contrast, when PCM is omitted from the QM model, dU_{inter} exhibit significantly more variance and SD increases substantially.

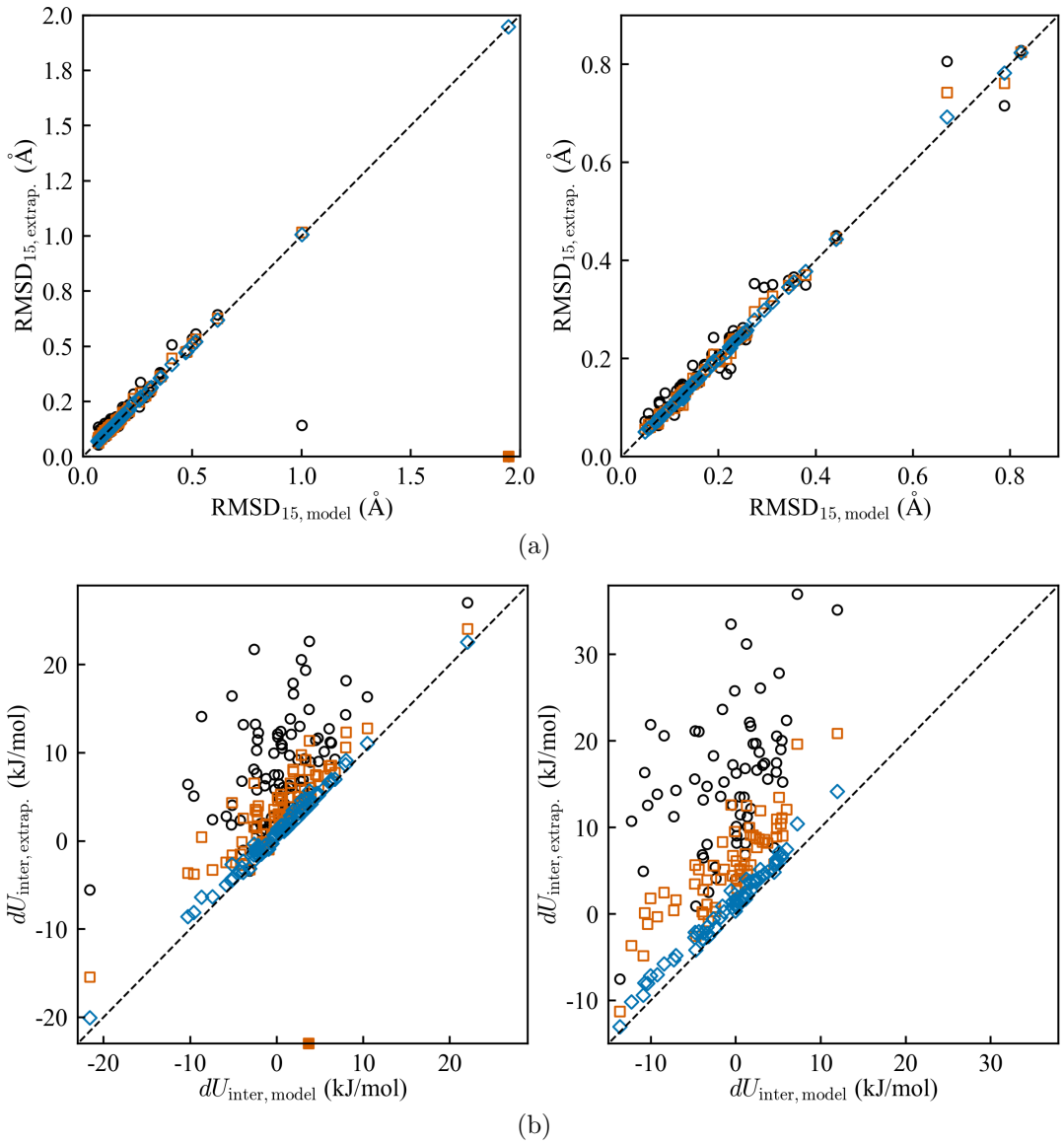


Figure 6.16: Parity plots of (a) $RMSD_{15}$ scores (Å), and (b) dU_{inter} (kJ/mol) between the fitted (PBE0/6-311G(d,p), MIP, rank 4 multipoles, ε_{PCM} = 11) and extrapolated models, for NHB (left) and HB (right) test set structures. The extrapolated models substitute ε_{PCM} = 11 with: no PCM (black ○), and PCM with ε_{PCM} = 3 (orange □) and ε_{PCM} = 7 (blue ◇) during the QM calculations. When optimized without PCM, CLMETH03 cannot achieve a valid $RMSD_{15}$ score (black ●). When ε_{PCM} = 3, the rigid optimization of CLMETH03 failed (orange ■).

Again, this corroborates that model behavior with and without PCM is discontinuous. Once PCM is applied, there is a 'smooth' (but diminishing) effect to adjusting ε_{PCM}.

Ultimately, these results indicate that there is leeway to extrapolate PCM-fitted FFs over different values of ε_{PCM}, within reason. Over the range of ε_{PCM} tested, most structures exhibit no loss of geometry accuracy. Concomitantly, while the predicted dU_{inter} values are sensitive towards ε_{PCM}, their response is shown to be quite uniform. Consequently, predictions of *relative* crystal energies should largely be preserved, irrespective of ε_{PCM}. The transferability of the FF over a range of ε_{PCM} values also suggests that PCM primarily influences the model at low values

Table 6.5: Summary of the key error metrics over the test set structures, for the different models tested. Metrics include the number of failed geometries, average $RMSD_{15}$ ($\overline{RMSD_{15}}$), AD in U_{inter}, AAD in U_{inter}, and SD in dU_{inter}. All errors metrics are relative to the CE–755 reference data. Scores are split as "NHB/HB". The first row contains the statistics for the reference model, which evaluates the test set using the same *ab initio* model applied for training (PBE0/6-311G(d,p), rank 4 multipoles, MIP, PCM with $\varepsilon_{PCM} = 11.0$).

Functional	Basis Set	Electrostatic Model	ε_{PCM}	Failed	$\overline{RMSD_{15}}$ (Å)	AD (kJ/mol)	AAD (kJ/mol)	SD (kJ/mol)
PBE0	6-311G(d,p)	Rank 4 GDMA	11.0	0	0.2078/0.1939	0.3984/−0.7604	3.4883/3.8916	5.1255/5.1312
HF	6-311G(d,p)	Rank 4 GDMA	11.0	0	0.2028/0.1959	−7.1457/−13.4727	8.3731/13.6427	8.6395/8.4096
PBE	6-311G(d,p)	Rank 4 GDMA	11.0	0	0.2205/0.2084	3.6384/5.0867	4.8374/6.4190	5.0538/5.7626
B3LYP	6-311G(d,p)	Rank 4 GDMA	11.0	1*	0.1861/0.1948	1.1708/0.9519	3.7710/4.3732	5.2715/5.3655
MP2	6-311G(d,p)	Rank 4 GDMA	11.0	0	0.2111/0.2000	3.6684/2.8539	4.9037/5.2840	5.3169/6.1287
PBE0	3-21G	Rank 4 GDMA	11.0	0	0.2293/0.2348	0.8695/2.2828	4.6595/7.0728	6.4567/8.6824
PBE0	6-31G(d,p)	Rank 4 GDMA	11.0	0	0.2037/0.1967	2.9212/3.0679	4.4159/4.9099	5.1613/5.2320
PBE0	6-311+G(d,p)	Rank 4 GDMA	11.0	0	0.2044/0.1999	−2.5375/−11.9768	4.6651/12.2847	5.9173/6.9880
PBE0	6-311G(d,p)	Rank 0 HLYGAt	11.0	0	0.2047/0.2439	2.7864/14.0284	4.6494/14.2772	5.3716/11.0283
PBE0	6-311G(d,p)	Rank 4 GDMA	-	1‡	0.1913/0.2039	8.5348/15.3783	8.7958/15.6068	5.9504/8.1470
PBE0	6-311G(d,p)	Rank 4 GDMA	3.0	1§	0.1914/0.1959	3.6199/5.4388	4.6599/6.1918	4.8256/5.3645
PBE0	6-311G(d,p)	Rank 4 GDMA	7.0	0	0.2087/0.1942	1.2093/0.7669	3.5577/4.0550	4.9605/5.0327

* The minimization of CLMETH03 failed and the structure was excluded from all the evaluated statistics for B3LYP.

‡ The minimized CLMETH03 structure could not achieve a valid $RMSD_{15}$ score even at 90%/90° COMPACK tolerances. No fusing was observed. It was excluded from the average $RMSD_{15}$ metric, but included in the energy metrics.

§ The minimization of CLMETH03 failed and the structure was excluded from all the evaluated statistics for $\varepsilon_{PCM} = 3$.

of $\varepsilon_{\mathrm{PCM}}$, with diminishing effect as $\varepsilon_{\mathrm{PCM}}$ increases. FF transferability breaks down when PCM is 'turned off': the no-PCM model behavior is incongruous with the PCM 'regime', thus model reliability will suffer if the PCM-fitted FF is used. An important caveat to these observations is that the extrapolated models employ a FF that was parametrized with a 'more attractive' PCM model (and consequently the reference FF is 'less attractive'). In the inverse case (e.g., taking a FF fitted with $\varepsilon_{\mathrm{PCM}} = 3$ and using it at $\varepsilon_{\mathrm{PCM}} = 11$), the model will 'double-count' some attractive contributions and be more susceptible to molecular fusion. Thus, the 'direction' of the model extrapolation must also be judiciously considered. A pragmatic strategy would be to train FFs at moderate/high $\varepsilon_{\mathrm{PCM}}$ values and, if necessary, use these FFs at lower $\varepsilon_{\mathrm{PCM}}$. If PCM is not intended to be used, a PCM-free FF must be independently parametrized.

6.6 Summary

In this chapter, I have employed the CrystalEstimator algorithm [1, 2] with the CE–755 dataset [1, 256] to parametrize transferable HAIEFF FFs for a range of commonly used CSP electrostatic models. To improve predictions of crystal geometries and intermolecular energies, I have explored a series of alterations to the empirical FF. Augmentations to the form of the FF (BHP → dBHP/MIP) and separation of $\mathrm{H_O}/\mathrm{H_N}$ both proved to be critical changes, particularly for the accurate description of HB crystals. Overall, the best attempts at FF-fitting (Schemes 5 and 7) achieve a significant reduction of intermolecular energy errors in the training set when compared to the most similar attempt at such parametrization, performed by Bowskill [2] (and reflected by Scheme 1). In hydrogen bonding systems, the average absolute dU_{inter} and its SD are reduced by upwards of 4 kJ/mol (depending on the electrostatic model)—a substantial improvement, given that this is of the same magnitude as expected polymorphic energy differences. These gains in energy accuracy have not come at the expense of geometry fidelity, and geometry accuracy is maintained. Furthermore, I have also demonstrated that despite the fitted FFs assimilating significant errors from the electrostatic model, they continue to exhibit transferability.

Despite these improvements, the HAIEFF models are likely still unreliable arbiters of polymorphic energy differences, largely because uncertainty in the model predictions continue to exceed the prediction errors in relative energies. Nevertheless, the improved model accuracy from the newly fitted FFs are still valuable to the earlier stages of the CSP workflow. Given the small margin for error associated with predicting polymorphic energy differences, unnecessary deterioration of model accuracy should be avoided wherever possible. To that end, I have also shown that *ad hoc* FF deployment can be a major source of superfluous model inaccuracy. Although some forms of model extrapolation are less detrimental than others, it remains that the reliability and robustness of the HAIEFF model is maximized when the FF is employed complementarily with its underlying *ab initio* model. Ultimately, congruous use of empirical FFs should be the standard to which all CSP practitioners strive towards.

A recurring theme that has emerged is the inadequacy of the HAIEFF electrostatic model. More specifically, the lack of an accurate induction model is a major source of error for the HB

structures. While PCM has proved to be a convenient first-approximation, it fails to capture the nuances of polarization in the crystal. In the next two chapters, I will consider how a more-detailed induction model can be incorporated into the HAIEFF framework.

Chapter 7

Induction for CSP: Implementing the SCERP Model

Through my study of the HAIEFF force-field, it has emerged that neglecting polarization (used interchangeably with 'induction') can contribute to a significant shortfall of cohesive energy from the electrostatic term. This cascades as a model error that the FF tries to assimilate through its parameters, sometimes with catastrophic results. Herein, I attempt to address this deficiency by integrating an explicit description of induction into the HAIEFF model. This builds upon work done by Zhang [71], who investigated the Self-Consistent Electronic Response to Point-charge (SCERP) induction model [183, 189]. Zhang's work suffers from two shortcomings:

1. While testing his SCERP implementation, several algorithmic flaws were identified, which affect the efficiency and reliability of the SCERP model in our group's CSP programs.
2. His testing of the SCERP-embedded HAIEFF model relied on *ad hoc* deployment of available HAIEFF FFs. As demonstrated, FF-extrapolation is most harmful when it involves changes to the electrostatic model. This casts some doubt on the reliability of his performance assessment.

In this chapter, I will tackle the implementation-related problems. To begin, I introduce the theoretical origin of induction. Thereafter, I detail my implementation of SCERP, comprising: the SCERP protocol (Section 7.2), calculations for rigid systems (Section 7.3), and calculations for flexible systems (Section 7.4). My focus is on bolstering the speed and reliability of the SCERP-embedded calculations. The second issue will be examined in the next chapter.

7.1 Theoretical Origin of Polarization

In isolation (e.g., a gas-phase isolated-molecule QM calculation), the electronic structure of a molecule assumes an unperturbed (or 'static') state. The static electrostatic features (e.g., distributed multipoles) derived from this charge density can be viewed as the 'permanent multipoles' inherent to the molecular structure. Polarization pertains to the perturbation of this static charge density in response to an external electric field, like the PCM field used in Chapter 6. For a molecule in the crystal (referred to as the 'central molecule'), surrounding molecules in the packing also induce an electric field at the central molecule's position. To illustrate the polarizing effect of this crystalline electric field, consider molecules in the crystal described by a

distributed multipole model. Using the classical, two-body Coulomb Law (Equation 3.2.8), the electrostatic potential ($V^{l_1m_1}_{\text{field},\mathbf{0}Ii}$) experienced by the l_1m_1 multipole moment of atom i from the central molecule ($\mathbf{0}I$) is given by

$$V^{l_1m_1}_{\text{field},\mathbf{0}Ii} = \frac{1}{4\pi\varepsilon_0} \sum_{\boldsymbol{l}=-\infty}^{\infty} \underbrace{\sum_{J=1}^{Z}}_{J\neq I \text{ if } \boldsymbol{l}=\mathbf{0}} \sum_{j=1}^{\eta_J} \sum_{l_2} \sum_{m_2} Q^{l_2m_2}_{\boldsymbol{l}Jj} \frac{\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^{l_1+l_2+1}}, \tag{7.1.1}$$

where $Q^{l_2m_2}_{\boldsymbol{l}Jj}$ is a multipole moment on another site in the crystal (not belonging to the central molecule), and $\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}$ is the associated interaction function. The perturbed l_1m_1 multipole moment of atom i from molecule $\mathbf{0}I$ in this electric field ($Q^{\text{total},l_1m_1}_{\mathbf{0}Ii}$) can be expressed as

$$Q^{\text{total},l_1m_1}_{\mathbf{0}Ii} = Q^{\text{static},l_1m_1}_{\mathbf{0}Ii} + \Delta Q^{l_1m_1}_{\mathbf{0}Ii}, \tag{7.1.2}$$

where $Q^{\text{static},l_1m_1}_{\mathbf{0}Ii}$ is the unperturbed static multipole moment, and $\Delta Q^{l_1m_1}_{\mathbf{0}Ii}$ is the induced multipole moment. Determining $Q^{\text{total},l_1m_1}_{\mathbf{0}Ii}$ from $V^{l_1m_1}_{\text{field},\mathbf{0}Ii}$ is non-trivial for two reasons.

Firstly, induction is not a two-body effect and exhibits many-body cooperative behavior [189, 223]. As the central molecule is polarized by the electric field generated by its surrounding molecules, reciprocally the surrounding molecules are polarized by the electric field contributed to by the central molecule. In short, the values of $Q^{l_2m_2}_{\boldsymbol{l}Jj}$ in Equation 7.1.1 depend on the perturbation in $Q^{\text{total},l_1m_1}_{\mathbf{0}Ii}$. This is especially so in a periodic lattice, where all symmetrically-related molecules should have the same electrostatic features (i.e., $Q^{\text{total},l_1m_1}_{\boldsymbol{l}Ii} = Q^{\text{total},l_1m_1}_{\mathbf{0}Ii}$). For this reason, $Q^{\text{total},l_1m_1}_{\mathbf{0}Ii}$ (and $\Delta Q^{l_1m_1}_{\mathbf{0}Ii}$) must come from self-consistent solutions that account for this feedback. Secondly, the expression shown in Equation 7.1.1 is valid only at long-range. In a closely packed crystal, neighboring molecules are expected to experience charge overlap. Short-range effects (like charge penetration and exchange) occur as a result [221], and the Coulomb law description of the electric field breaks down.

In summary, polarization is the response of a molecule to an external electric field and quantifying this effect in a crystal is challenging. To determine $\boldsymbol{Q}^{\text{total}}$, an effective induction model should account for the reciprocal interaction between all molecules in the crystal, as well as the occurrence of quantum effects at short-range.

7.2 The SCERP Protocol

Several approaches have been proposed to determine $Q^{\text{total},l_1m_1}_{\mathbf{0}Ii}$ according to the phenomenon described above. In the WSM method [222–224], distributed polarizabilities are used to evaluate the moments induced in response to $V^{l_1m_1}_{\text{field},\mathbf{0}Ii}$ and short-range effects are approximated by adding a damping function to Equation 7.1.1. This gives Equation 2.5.10, which is solved to self-consistency for all multipole moments. As stated in Section 2.5.2.3, our group has instead opted to adopt the SCERP method [71, 183, 189], wherein a self-consistent solution for the perturbed charge density is determined through a series of QM calculations. During each calculation, a field of point-charges, placed on relevant crystallographic positions, is employed to represent the

periodic images of molecules in the crystal, and only the central molecule is treated with 'real' nuclei. This monomer + field system is evaluated to obtain a perturbed SCF charge density, from which atomic-charges are re-derived and used to update the field-charges. This process repeats until the point-charges converge, with the charge values acting as a proxy to assess the convergence of the perturbed charge density. Once converged, the perturbed charge density can then be used to derive the corresponding induced multipoles.

By substituting all 'non-central' molecules by their associated atomic-charges, the effects of the crystal electric field are captured by the SCERP approach. By solving to self-consistency, the reciprocal nature of polarization is also reflected. Depending on the chosen size of the embedding-environment, the range to which the electric fields' effects are incorporated into the model is modulated. Using the WSM method, Welch [183] had suggested that the electric field contribution from extremely distal sites do not usually induce significant polarization and the field need only be considered up to a modest real-space distance. Using the SCERP method, Zhang [71] also showed that the induction energy contribution converges within a reasonable embedding-range. These studies suggest that the important medium-/long-range effects of polarization are described by SCERP. Since the charge placement and charge values used in SCERP are based on the crystallographic positions and atomic-charges of their equivalent molecules in the crystal, one expects a more faithful representation of the crystalline environment than simply using PCM [190]. In the latter, the electric field is represented by a uniform, polarizable continuum, which interacts with the central molecule only in the vicinity of its molecular cavity. This fails to capture the periodicity and order of the crystalline environment. Using molecular clusters as a reference, Zhang [71] has suggested that the multipoles derived from SCERP give more-accurate molecular ESPs compared to PCM- or static-environment derived multipoles.

As noted by Welch [189], the short-range quantum effects of polarization are only partially reflected in SCERP. Atomic nuclei in close proximity to the central molecule are represented by only their nett charge, without a real electron density. As such, exchange is neglected by the SCERP model. Charge penetration (which emerges from reduced electron shielding during charge overlap) is represented to an extent, but probably without great fidelity. Unlike the WSM method, damping cannot be easily embedded into QM computations to correct for these short-range effects. However, given its highly empirical nature, it is unclear if damping (as enacted in Refs [183, 189]) truly bolsters model fidelity.

The SCERP process has been implemented as a Fortran90 subroutine in CSO–RMv1.5.2 as described below (Figure 7.1).

7.2.1 Algorithmic Implementation

As input, the SCERP process takes a crystal structure, defined by its rigid-body variables, $\boldsymbol{\Omega}$, and the conformations of all molecules in the asymmetric unit cell, $\boldsymbol{\theta}$. An initial guess for the converged point-charges (abbreviated 'pc') is also required. I denote a trial solution for molecule I as $\boldsymbol{Q}_{I,t}^{\text{total,pc}}$, where $t = 0$ at initialization. The choice of $\boldsymbol{Q}_{I,0}^{\text{total,pc}}$ will be discussed

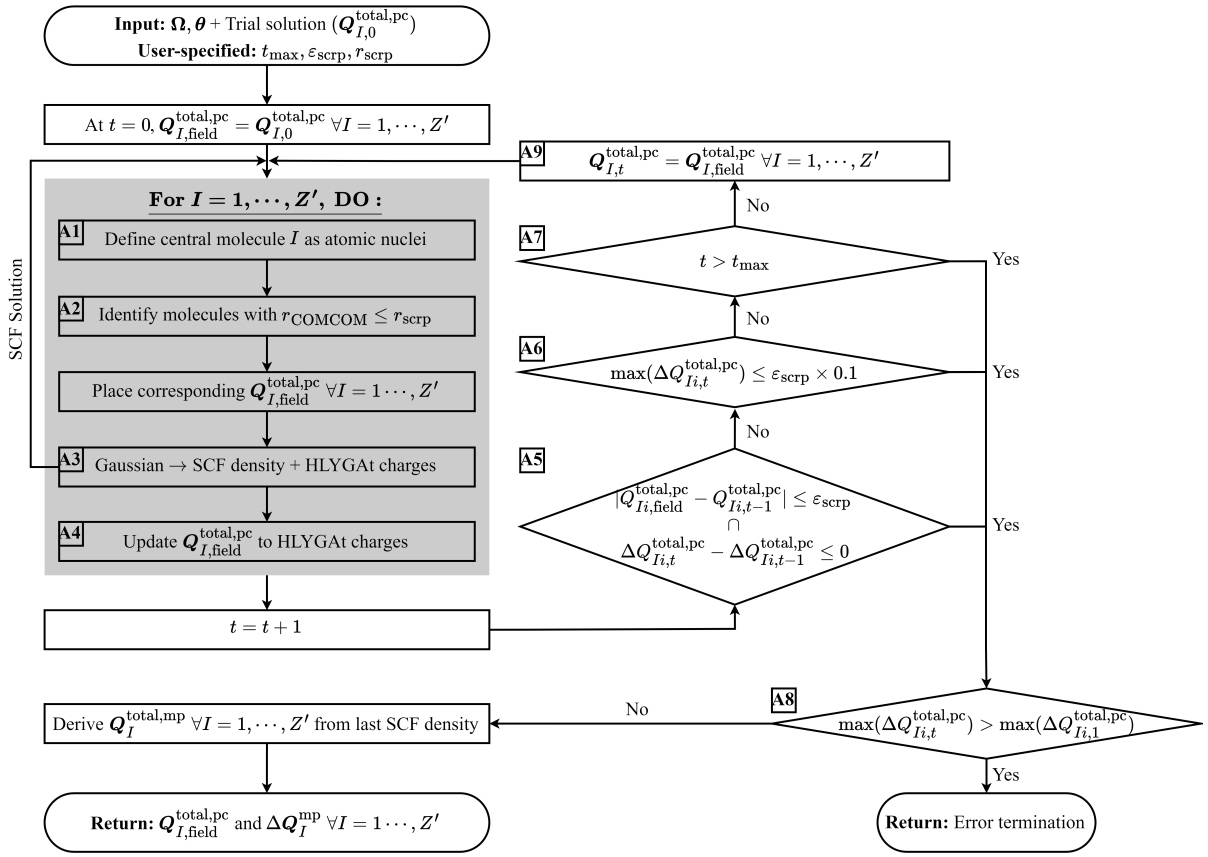


Figure 7.1: Schematic of the SCERP process. Key stages are enumerated and detailed in the main text.

later. At iteration t, the central molecule is first defined with atomic nuclei (Figure 7.1, A1). From $\boldsymbol{\Omega}$ and $\boldsymbol{\theta}$ (and relevant symmetry operations), the crystallographic positions of all other molecules in the infinite lattice are given. The COM-to-COM Euclidean distance (r_{COMCOM}) from these molecules to the central molecule is evaluated and a molecule is deemed relevant only if $r_{\text{COMCOM}} \leq r_{\text{scrp}}$ (Figure 7.1, A2), where r_{scrp} is a user-specified SCERP embedding-radius. All atoms belonging to 'relevant' molecules are replaced by their corresponding SCERP-field atomic-charges ($\boldsymbol{Q}_{I,\text{field}}^{\text{total,pc}}$). This gives a spherical cluster, comprising the central molecule embedded in the point-charge field. Gaussian09 is called to conduct a single-point SCF calculation, after which HLYGAt-fitted atomic-charges are derived from the SCF charge density (Figure 7.1, A3). At this stage, $\boldsymbol{Q}_{I,\text{field}}^{\text{total,pc}}$ is updated to the newly derived HLYGAt charges (Figure 7.1, A4).

For systems with $Z' > 1$, A1–A4 are conducted *sequentially* for each molecule I, i.e., molecule $I = 1$ undergoes QM calculation as the *lone* central molecule, followed by molecule $I = 2$, and so on till $I = Z'$. While QM calculations are not restricted to monomers, independent treatment of molecules is done for several reasons. Firstly, the crystalline environment experienced in an infinite lattice is most faithfully represented when the molecule is embedded *centrally* in a spherical charge field. The isotropy of the sphere reflects the invariance of the lattice in all directions, but this is disrupted by concomitant treatment of molecules, as multiple molecules cannot concurrently be at the spherical centre. Secondly, ESP-fitting methods, like HLY/HLYGAt [173],

usually exclude the regions close to atomic sites from the fitting. This can lead to the 'buried atom' problem, wherein atoms far from the molecular surface lack fitting data and thus have ill-defined charge values. If evaluating multiple molecules together, a similar issue could arise for intermolecular sites that are in close proximity, reducing the accuracy of their charge values and ultimately degrading the quality of the SCERP electric field. Finally, most QM methods exhibit N^x scaling (where N is the number of electrons in the system) [121], thus (in principle) it is more efficient to conduct multiple independent monomer calculations with smaller N.

Termination criteria for the SCERP process are considered at iteration t only after QM calculations for all molecules in the asymmetric unit cell have been completed. A few termination cases are possible. Firstly, graceful convergence (Figure 7.1, A5) occurs when

$$|Q_{Ii,\text{field}}^{\text{total,pc}} - Q_{Ii,t-1}^{\text{total,pc}}| \leq \varepsilon_{\text{scrp}}, \qquad \forall I = 1, \cdots, Z' \quad i = 1, \cdots \eta_I, \tag{7.2.1}$$

and

$$\Delta Q_{Ii,t}^{\text{total,pc}} - \Delta Q_{Ii,t-1}^{\text{total,pc}} \leq 0, \qquad \forall I = 1, \cdots, Z' \quad i = 1, \cdots \eta_I, \tag{7.2.2}$$

are simultaneously fulfilled. Here, $\varepsilon_{\text{scrp}}$ is a user-specified SCERP process convergence threshold and $\Delta Q_{Ii,t}^{\text{total,pc}} = |Q_{Ii,\text{field}}^{\text{total,pc}} - Q_{Ii,t-1}^{\text{total,pc}}|$. The first criterion requires that the change in all charge values between two iterations be below $\varepsilon_{\text{scrp}}$, while the second criterion demands that the change in the charge values be decreasing. A5 represents an extremely robust convergence, but can often be too stringent (particularly due to Equation 7.2.2). As such, a looser case of convergence is allowed if the greatest change in charge values at iteration t does not exceed a tenth of $\varepsilon_{\text{scrp}}$ (Figure 7.1, A6). Beyond this, a user-specified limit on the number of SCERP iterations ($t_{\max}$) can also trigger termination (Figure 7.1, A7). When exiting the SCERP process loop via A5–A7, an additional check (Figure 7.1, A8) is performed to ensure the maximum change in charge values at the final step does not exceed that in the first step. In principle, A8 is relevant to only A7-termination, and if triggered it would suggest that the field-charges have likely *diverged* during the SCERP process. In this case, the SCF density is unusable and error termination is returned. Otherwise, if A8 is passed, the SCERP-perturbed multipoles ($\boldsymbol{Q}_I^{\text{total,mp}}$) can be derived (by GDMA) from the final SCF solution and $\Delta\boldsymbol{Q}_I^{\text{mp}}$ is trivial to obtain from Equation 7.1.2. Although termination by A7 (without triggering A8) indicates that the SCF density is not fully converged, CSP calculations using SCERP will require repeated evaluation of polarization. To avoid impeding such a sequence of calculations, I permit using the unconverged SCF density.

If A5–A7 are not satisfied, the guess for the SCERP-solution is updated (Figure 7.1, A9) and the loop repeats (A9 → A1). To reduce the cost of subsequent QM calculations, 'Guess=Read' is added to the Gaussian input for $t \geq 1$. This requests the SCF calculations to be initialized from the last solution stored in the checkpoint file (A3 → A1). As demonstrated later, this does seem to lead to a meaningful cost reduction.

Table 7.1: SCERP process convergence test specifications.

Test Specification	Value
Level of Theory	PBE0/6-311G(d,p)
$\varepsilon_{\text{scrp}}$	1×10^{-3} a.u.
r_{scrp}	40.0 Å
t_{max}	15
$\boldsymbol{Q}_{I,0}^{\text{total,pc}}$	$\boldsymbol{Q}_{I}^{\text{static,HLYGAt}}$

7.2.2 SCERP Process Convergence

At this stage, it is of interest to test the robustness of the procedure shown in Figure 7.1. To that end, 33 HB crystal structures from the CE–755 dataset are subject to the SCERP process using the program settings shown in Table 7.1. Values for $\varepsilon_{\text{scrp}}$ and r_{scrp} follow the recommendation of Zhang [71], while t_{max} is set to an arbitrarily large value. For these tests, $\boldsymbol{Q}_{I,0}^{\text{total,pc}}$ is initialized at the static values of the HLYGAt-derived atomic-charges (i.e., $\boldsymbol{Q}_{I,0}^{\text{total,pc}} = \boldsymbol{Q}_{I}^{\text{static,HLYGAt}}$).

Of the 33 crystals tested, only 32 successfully converged. The remaining unconverged structure is 1-naphthoic acid (NAPOAC01, $Z' = 1$ with 21 atoms) and the values of its $\boldsymbol{Q}_{I,\text{field}}^{\text{total,pc}}$ are shown in Figure 7.2. Strikingly, not only does the SCERP process fail to converge for this system, but increasingly large oscillations in $\boldsymbol{Q}_{I,\text{field}}^{\text{total,pc}}$ are observed, indicative of the sequence diverging. Several attempts were made to remedy this. Firstly, to provide a better starting guess, $\boldsymbol{Q}_{I,0}^{\text{total,pc}} = \boldsymbol{Q}_{I}^{\text{PCM,HLYGAt}}$ was considered, since this would initiate the SCERP-field in an already-polarized state. Per the guidance of Refs [183, 189], basis sets with diffuse functions were tested, as well as using CHelpG (instead of HLYGAt) to derive atomic-charges. Finally, the value of r_{scrp} was varied to larger and smaller embedding-radii. Unfortunately, none of these efforts resolved the divergent behavior, suggesting that the issue may not be physical, but numerical in nature. As such, I introduce the concept of decelerated SCERP. While $t < t_{\text{slow}}$, where t_{slow} is a user-specified decelerated-iteration limit, rather than a full update of $\boldsymbol{Q}_{I,\text{field}}^{\text{total,pc}}$

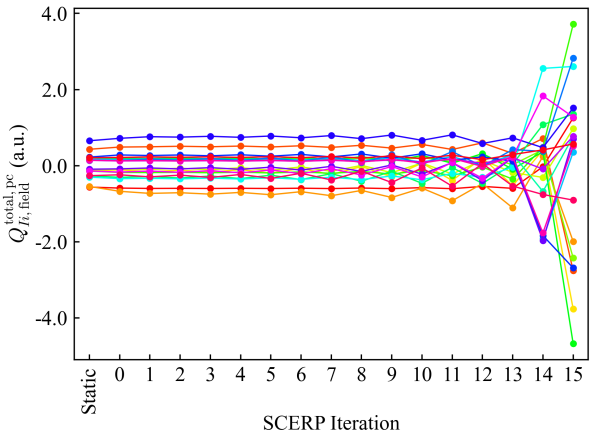


Figure 7.2: $Q_{Ii,\text{field}}^{\text{total,pc}}$ for atoms in NAPOAC01 over the SCERP process (Table 7.1), as a function of iteration number. Each line/color represents a different atom.

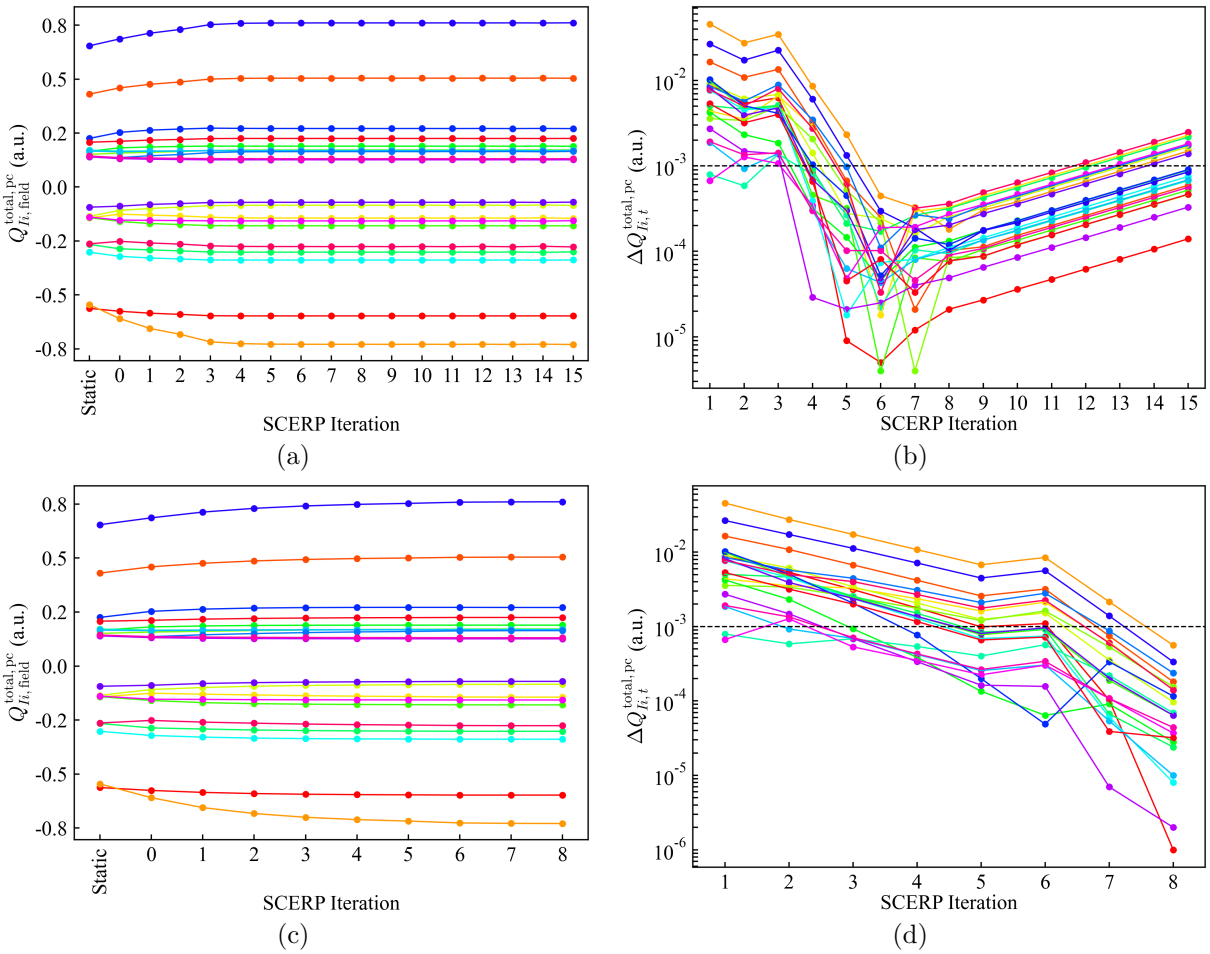


Figure 7.3: $Q_{Ii,\text{field}}^{\text{total,pc}}$ and $\Delta Q_{Ii,t}^{\text{total,pc}}$ for atoms in NAPOAC01 using decelerated SCERP, as a function of iteration number. (a) and (b): $t_{\text{slow}} = 3$; (c) and (d): $t_{\text{slow}} = 6$. Each line/color represents a different atom. Test parameters are given in Table 7.1. The dashed line in (b) and (d) shows $\varepsilon_{\text{scrp}} = 1 \times 10^{-3}$ a.u. $\Delta Q_{Ii,t}^{\text{total,pc}}$ between 'Static' and 'Iteration 0' are omitted.

(i.e., $\boldsymbol{Q}_{I,\text{field}}^{\text{total,pc}} = \boldsymbol{Q}_{I}^{\text{total,HLYGAt}}$ in Figure 7.1, A4), I instead impose a bisection scheme

$$\boldsymbol{Q}_{I,\text{field}}^{\text{total,pc}} = \begin{cases} \frac{1}{2}\left(\boldsymbol{Q}_{I}^{\text{total,HLYGAt}} + \boldsymbol{Q}_{I,t}^{\text{total,pc}}\right), & \text{if } t < t_{\text{slow}}, \\ \boldsymbol{Q}_{I}^{\text{total,HLYGAt}}, & \text{if } t \geq t_{\text{slow}}. \end{cases} \tag{7.2.3}$$

$\boldsymbol{Q}_{I,\text{field}}^{\text{total,pc}}$ is taken as the midpoint of the new HLYGAt-derived atomic-charges ($\boldsymbol{Q}_{I}^{\text{total,HLYGAt}}$) and the last guess of the SCERP-field solution ($\boldsymbol{Q}_{I,t}^{\text{total,pc}}$, note this occurs before t is incremented).

This deceleration scheme is tested on NAPOAC01 using $t_{\text{slow}} = 3$ and $t_{\text{slow}} = 6$ (Figure 7.3), keeping all other specifications at their values in Table 7.1. For $t_{\text{slow}} = 3$, no oscillatory behavior is observed throughout the SCERP process, and the field-charges seem to be converging. However, closer inspection of $\Delta Q_{Ii,t}^{\text{total,pc}}$ reveals otherwise. By iteration 6, while all $\Delta Q_{Ii,t}^{\text{total,pc}}$ fall below 1×10^{-3} a.u. (the specified $\varepsilon_{\text{scrp}}$), not all $\Delta Q_{Ii,t}^{\text{total,pc}}$ are decreasing and $\max(\Delta Q_{Ii,t}^{\text{total,pc}})$ exceeds 1×10^{-4} a.u., preventing termination by either A5 or A6. If this was a stable solution, steps beyond iteration 6 should progress towards convergence, yet $\Delta Q_{Ii,t}^{\text{total,pc}}$ begin to increase again from iteration 7. Given enough steps, divergence will likely occur again and the SCERP process

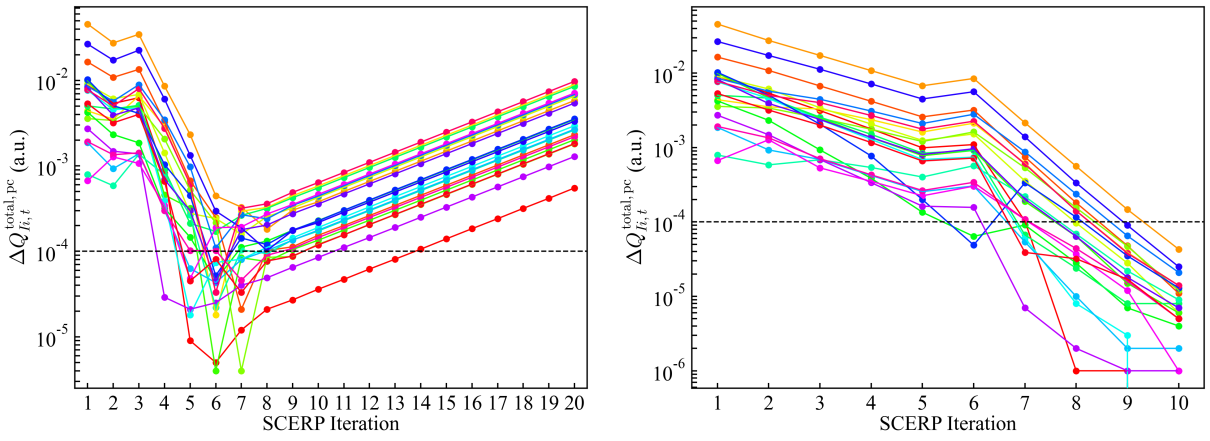


Figure 7.4: $\Delta Q_{Ii,t}^{\text{total,pc}}$ for atoms in NAPOAC01 using decelerated SCERP, as a function of iteration number when (a) $t_{\text{slow}} = 3$, and (b) $t_{\text{slow}} = 6$. Each line/color represents a different atom. Test parameters are given in Table 7.1, but $\varepsilon_{\text{scrp}} = 1 \times 10^{-4}$ a.u. and $t_{\text{max}} = 20$. The dashed line shows $\varepsilon_{\text{scrp}} = 1 \times 10^{-4}$ a.u. $\Delta Q_{Ii,t}^{\text{total,pc}}$ between 'Static' and 'Iteration 0' are omitted.

has not truly converged to a stable solution. In contrast, when $t_{\text{slow}} = 6$, graceful termination by condition A5 is achieved at iteration 8; the condition given by Equation 7.2.2 makes it improbable that this diverges given more iterations (and in fact this system motivated adding the secondary A5 criterion). These tests are repeated with $t_{\text{max}} = 20$ and $\varepsilon_{\text{scrp}} = 1 \times 10^{-4}$ a.u. (to 'force' more iterations for $t_{\text{slow}} = 6$) (Figure 7.4). When $t_{\text{slow}} = 3$, $\Delta Q_{Ii,t}^{\text{total,pc}}$ continue to rise after iteration 15, indicative of divergence. Conversely, when $t_{\text{slow}} = 6$, A5-convergence is achieved, despite the more stringent $\varepsilon_{\text{scrp}}$, reflecting that the solution is indeed stable.

Henceforth, I will use $t_{\text{slow}} = 6$: from testing done over the entire CE–755 dataset, this seems to suppress all cases of divergence. For structures that are natively convergent, the deceleration protocol does not affect their final solutions, but (intuitively) more iterations are spent to reach convergence. In my view, this is a necessary trade-off to ensure robustness of the SCERP process.

7.2.3 SCERP-field Convergence

Having bolstered the reliability of convergence during the SCERP process, I briefly study the value of r_{scrp}. This parameter was investigated by Zhang [71], and I seek to validate his suggested value of $r_{\text{scrp}} = 40$ Å. To that end, two structures, NAPOAC01 and 6H-indolo[2,3-b]quinoxaline (FITZEO01, $Z' = 1$ with 26 atoms) undergo the SCERP process at different values of r_{scrp}, with the test specifications shown in Table 7.2. A set of reference point-charge values is determined using $r_{\text{scrp}} = 50$ Å and $\varepsilon_{\text{scrp}} = 1 \times 10^{-4}$ a.u.; the converged SCERP charges at other values of r_{scrp} are compared to this. In all cases, the SCERP process converges either by A5 or A6, and the final SCERP-field charges for both structures are presented in Figure 7.5. The average absolute percentage deviations (%AADs) and maximum absolute percentage deviations (Max. %ADs) in $\boldsymbol{Q}_{I,\text{field}}^{\text{total,pc}}$, relative to the 50 Å reference, are tabulated in Table 7.3.

For both structures, there are large changes in $Q_{Ii,\text{field}}^{\text{total,pc}}$ between the static state and the

Table 7.2: SCERP-field convergence test specifications.

Test Specification	Value
Level of Theory	PBE0/6-311G(d,p)
$\varepsilon_{\text{scrp}}$	1×10^{-3} or $1 \times 10^{-4\dagger}$ a.u.
r_{scrp}	Between 0.0 (Static) and 50.0† Å
t_{max}	15 or 20†
t_{slow}	6
$\boldsymbol{Q}_{I,0}^{\text{total,pc}}$	$\boldsymbol{Q}_{I}^{\text{static,HLYGAt}}$

† Only used during the calculations conducted at the reference embedding of 50 Å.

smallest SCERP-embedded state. As r_{scrp} continues to increase, there are subtle oscillations in the individual values of $Q_{Ii,\text{field}}^{\text{total,pc}}$, but crucially these oscillations are damped. This indicates that the SCERP-field charges are converging within the test range of r_{scrp}. Oscillatory behavior, with respect to r_{scrp}, was also observed in Zhang's [71] study of U_{ind}. The convergence of $\boldsymbol{Q}_{I,\text{field}}^{\text{total,pc}}$ is further evidenced from Table 7.3: all error metrics decrease as r_{scrp} increases, albeit at a diminishing rate. By r_{scrp} = 30 Å, %AADs are < 1% for both systems, although the Max. %AD remains large for FITZEO01. This large maximum stems from an atomic-charge with relatively small magnitude (~8.6×10^{-3} a.u.) returning a large percentage error, despite its absolute error being small (in fact the error is less than $\varepsilon_{\text{scrp}}$). Ignoring this atom, the Max. %AD at r_{scrp} = 30 Å is only 2.654%. When r_{scrp} = 40 Å, all error metrics are < 1% in both systems.

From this, I assert that Zhang's [71] recommendation of r_{scrp} = 40 Å is adequate to determine the polarizing effects of the SCERP field. There are some indications that convergence is already achieved around r_{scrp} = 30 Å, but to leave some margin for error (particularly to accommodate larger molecules), r_{scrp} = 40 Å is a 'safe' choice. Welch *et al.* [189] suggest that the cost of

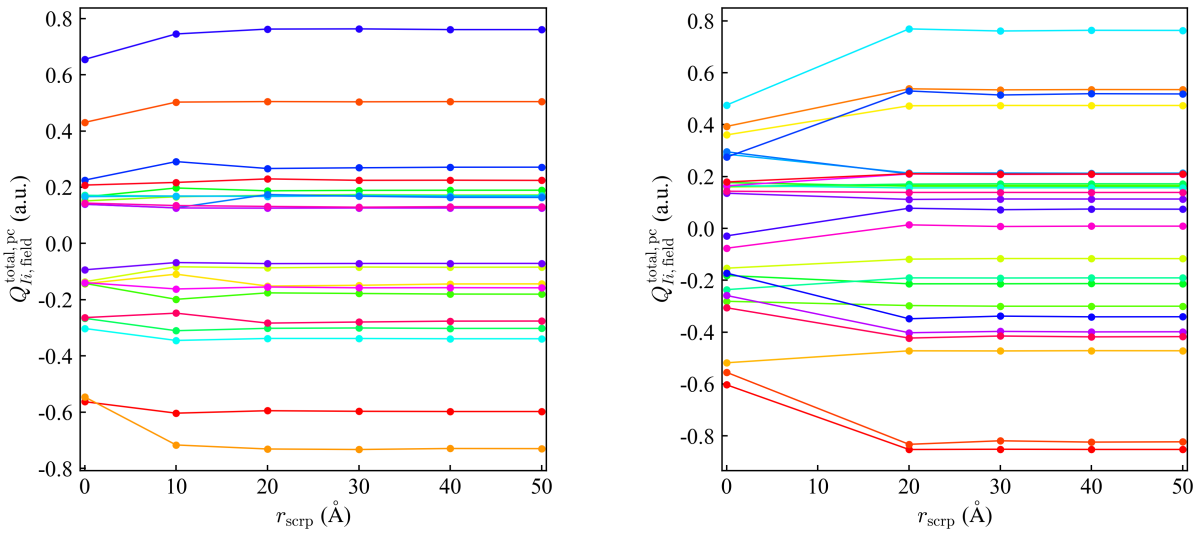


Figure 7.5: Converged $Q_{Ii,\text{field}}^{\text{total,pc}}$ for atoms in (a) NAPOAC01, and (b) FITZEO01 at different r_{scrp} values. Each line/color represents a different atom. r_{scrp} = 0 Å indicates 'Static' conditions. Test parameters are given in Table 7.2.

Table 7.3: Errors for the SCERP-field convergence test, using the 50 Å embedding as a reference.

Structure	Error in $Q^{\text{total,pc}}_{Ii,\text{field}}$	r_{scrp} (Å)				
		Static	10.0	20.0	30.0	40.0
NAPOAC01	%AAD (%)	14.282	5.135	1.565	0.753	0.154
	Max. %AD (%)	61.592	24.045	6.356	3.443	0.555
FITZEO01	%AAD (%)	64.578	-	3.078	0.900	0.116
	Max. %AD (%)	989.517	-	58.394	15.358	0.732

SCERP QM calculations is mostly agnostic to the choice of r_{scrp}. While this may be true of the QM computations themselves, I observe a significant cost to write/read extremely long input/output files to/from Gaussian. Thus, I do not recommend progressing to larger r_{scrp}.

7.3 Rigid-Molecule Calculations with SCERP

To interface the SCERP process with the rest of our CSP programs, I first provide the mathematical formulation of the induction-coupled CSP calculations. Following this, I probe the efficacy of the induction model using single-point, SCERP-embedded energy evaluations. Finally, I describe the algorithmic implementation of SCERP, as it pertains to rigid-body minimizations. All developments have been implemented in CSO–RMv1.5.2 (and its subversions).

7.3.1 Electrostatic Calculations for SCERP

Following the approach of Welch *et al.* [189] and Zhang [71] (and as derived by Stone [221]), the total induction energy (subject to appropriate normalization for the unit cell) is given by

$$U_{\text{ind}} = \frac{1}{2Z}\sum_{I=1}^{Z}\sum_{i=1}^{\eta_I}\sum_{\boldsymbol{l}=-\boldsymbol{\infty}}^{\boldsymbol{\infty}}\underbrace{\sum_{J=1}^{Z}}_{J\neq I \text{ if } \boldsymbol{l}=\boldsymbol{0}}\sum_{j=1}^{\eta_J}\frac{1}{4\pi\varepsilon_0}\sum_{l_1,l_2}\sum_{m_1,m_2}\Delta Q^{l_1m_1}_{\boldsymbol{0}Ii}Q^{l_2m_2}_{\boldsymbol{l}Jj}\frac{\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}}{(r_{\boldsymbol{0}Ii,\boldsymbol{l}Jj})^{l_1+l_2+1}}, \tag{7.3.1}$$

where $\Delta Q^{l_1m_1}_{\boldsymbol{0}Ii}$ is an induced multipole obtained via the SCERP process. For brevity, I have omitted the dependencies of $\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}$. This expression is similar to the functions used for 'regular' electrostatic calculations (Equations 3.2.5 and 3.2.8), hence the existing energy subroutines implemented in CSO–RM/FM should also be applicable to the induction-embedded computations. However, the use of $\Delta Q^{l_1m_1}_{\boldsymbol{0}Ii}Q^{l_2m_2}_{\boldsymbol{l}Jj}$, instead of $Q^{l_1m_1}_{\boldsymbol{0}Ii}Q^{l_2m_2}_{\boldsymbol{l}Jj}$, creates an asymmetry that must be accommodated through judicious manipulation of the subroutine inputs.

To that end, consider a hypothetical case,

$$\begin{aligned}U_{\text{hypo}} =& \frac{1}{2Z}\sum_{I=1}^{Z}\sum_{i=1}^{\eta_I}\sum_{\boldsymbol{l}=-\boldsymbol{\infty}}^{\boldsymbol{\infty}}\underbrace{\sum_{J=1}^{Z}}_{J\neq I \text{ if } \boldsymbol{l}=\boldsymbol{0}}\sum_{j=1}^{\eta_J}\frac{1}{4\pi\varepsilon_0}\\ &\cdot\sum_{l_1,l_2}\sum_{m_1,m_2}\left(Q^{l_1m_1}_{\boldsymbol{0}Ii}+\frac{\Delta Q^{l_1m_1}_{\boldsymbol{0}Ii}}{2}\right)\left(Q^{l_2m_2}_{\boldsymbol{l}Jj}+\frac{\Delta Q^{l_2m_2}_{\boldsymbol{l}Jj}}{2}\right)\frac{\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}}{(r_{\boldsymbol{0}Ii,\boldsymbol{l}Jj})^{l_1+l_2+1}}.\end{aligned} \tag{7.3.2}$$

Expanding this further gives

$$\begin{aligned} U_{\text{hypo}} =& \frac{1}{2Z} \sum_{I=1}^{Z} \sum_{i=1}^{\eta_I} \sum_{\boldsymbol{l}=-\boldsymbol{\infty}}^{\boldsymbol{\infty}} \underbrace{\sum_{J=1}^{Z}}_{J\neq I \text{ if } \boldsymbol{l}=\boldsymbol{0}} \sum_{j=1}^{\eta_J} \frac{1}{4\pi\varepsilon_0} \sum_{l_1,l_2} \sum_{m_1,m_2} \frac{\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}}{(r_{\boldsymbol{0}Ii,\boldsymbol{l}Jj})^{l_1+l_2+1}} \\ & \cdot \left[Q^{l_1m_1}_{\boldsymbol{0}Ii} Q^{l_2m_2}_{\boldsymbol{l}Jj} + \frac{Q^{l_1m_1}_{\boldsymbol{0}Ii} \Delta Q^{l_2m_2}_{\boldsymbol{l}Jj}}{2} + \frac{\Delta Q^{l_1m_1}_{\boldsymbol{0}Ii} Q^{l_2m_2}_{\boldsymbol{l}Jj}}{2} + \frac{\Delta Q^{l_1m_1}_{\boldsymbol{0}Ii} \Delta Q^{l_2m_2}_{\boldsymbol{l}Jj}}{4} \right] \\ =& U_{\text{elec}} + U_{\text{ind}} + U_{\text{error}}. \end{aligned} \tag{7.3.3}$$

From here, the most efficient means to obtain the total electrostatic contribution from the static electrostatic and induction terms ($U_{\text{elec}} + U_{\text{ind}}$) is to deduct U_{error} from U_{hypo}. In practice, this employs two evaluations of the existing CSO–RM/FM electrostatic subroutines:

1. The first uses $\left(Q^{l_1m_1}_{\boldsymbol{0}Ii} + \frac{\Delta Q^{l_1m_1}_{\boldsymbol{0}Ii}}{2}\right)\left(Q^{l_2m_2}_{\boldsymbol{l}Jj} + \frac{\Delta Q^{l_2m_2}_{\boldsymbol{l}Jj}}{2}\right)$ as subroutine inputs to obtain U_{hypo}.
2. The second uses $\left(\frac{\Delta Q^{l_1m_1}_{\boldsymbol{0}Ii}}{2}\right)\left(\frac{\Delta Q^{l_2m_2}_{\boldsymbol{l}Jj}}{2}\right)$ as subroutine inputs to determine U_{error}.

Thereafter, $U_{\text{elec+ind}}$ is trivial to determine. Three evaluations would be needed to obtain U_{elec} and U_{ind} separately [189], but given that evaluating Ewald sums is relatively expensive (especially for multipole models), it is more pragmatic to evaluate $U_{\text{elec+ind}}$ as proposed here.

Since the SCERP process provides $\boldsymbol{\Delta Q}$ for all multipole moments up to the equivalent rank of $\boldsymbol{Q}$, no major modifications are needed in the existing CSO–RM/FM subroutines. Wherever the electrostatic energy (or its gradients) are requested, a second evaluation is appended (with the appropriate input multipole moments for each assessment). A subtraction of the two computed values is then done at the end. In this way, the induction contribution is assessed for multipole interactions up to the same rank as the static electrostatics.

7.3.1.1 Treatment of the Ewald Sum

Here, I address the use of the Ewald sum for evaluating induction. Zhang [71] (citing Welch [183]) stated that the reciprocal-space and self-interaction terms have negligibly small induction contributions. As such, Zhang suggested that $U_{\text{elec+ind}}$ need only be evaluated for the real-space term, while the reciprocal and self-correction terms consider just U_{elec}. In other words:

$$\begin{aligned} U_{\text{elec+ind}} \approx& U_{\text{real}}\left(Q_1 + \frac{\Delta Q_1}{2}, Q_2 + \frac{\Delta Q_2}{2}\right) - U_{\text{real}}\left(\frac{\Delta Q_1}{2}, \frac{\Delta Q_2}{2}\right) \\ & + U_{\text{rec}}\left(Q_1, Q_2\right) + U_{\text{corr}}\left(Q_1, Q_2\right), \end{aligned} \tag{7.3.4}$$

where, for clarity, I have simplified the annotation of the multipole moments. While this is a convenient simplification, in my view this approach is erroneous.

Firstly, Zhang neglects inherent differences between DMACRYS [250] (used by Welch and co-workers [183, 189]) and our own CSO–RM/FM [2] codes. In DMACRYS, Ewald sums are only applied to charge–charge, charge–dipole, and dipole–dipole interactions, thus the omission

of Ewald sum terms is restricted to these contributions. Moreover, Welch's WSM-induction implementation [183, 189] is limited to lower-ranked multipole interactions, due to the difficulty in deriving higher-ranked distributed polarizabilities. Conversely, CSO–RM/FM apply Ewald sums to *all* multipole interactions up to $l_1+l_2 = 4$ [2], and the induction contribution is evaluated to the same rank. It follows that omitting Ewald sum terms in CSO–RM/FM will likely result in a larger error than in DMACRYS. Secondly, as discussed in Section 3.2.1.5, the Ewald sum is *not* a physical partitioning of interaction energy and the contributions from U_{real}, U_{rec}, and U_{corr} are tied to the choice of the Ewald convergence parameter, $\alpha_{l_1+l_2}$. Irrespective of $\alpha_{l_1+l_2}$, U_{real}, U_{rec}, and U_{corr} should sum to the same value [2], but altering $\alpha_{l_1+l_2}$ 'redistributes' the energy contribution among the three terms. As such, it is not definitive that U_{rec} and U_{corr} must be small relative to U_{real} (this may have been true for Welch [183] based on his chosen $\alpha_{l_1+l_2}$).

To illustrate this, I perform single-point energy evaluations of pyrene (PYRENE03 in the CE–755 dataset) using a complete Ewald sum (i.e., no simplification of U_{rec} and U_{corr}) and point-charge electrostatics. Using the specifications in Table 7.4, U_{real}, U_{rec}, and U_{corr} are assessed at different values of α_0. To modulate α_0 in CSO–RM, I alter the Ewald sum truncation error parameter (ε^{max}). As ε^{max} increases, $\alpha_{l_1+l_2}$ will decrease at a fixed real-space cutoff distance (Figure 4.1 in Ref [2]). Evaluations are done with and without induction; the induction contribution is then isolated by deducting the latter's results from the former. From Table 7.5, the sum of the three Ewald terms (denoted U_{sum}) is almost constant for all α_0, which aligns correctly with Figure 4.9 in Ref [2]. Analyzing the induction contribution, it can be seen that the proportions of U_{real}, U_{rec}, and U_{corr} are α_0-dependent and not necessarily real-space dominated. If Zhang's [71] approximation were employed, the error introduced is equal to $U_{\text{rec}} + U_{\text{corr}}$ from the induction contribution. Depending on α_0, this ranges from −0.6842 to −1.8137 kJ/mol, which in any case represents a sizable portion of the total U_{ind} contribution.

In summary, induction-coupled electrostatic calculations are possible with minor changes to the CSO–RM code. $U_{\text{elec+ind}}$ can be obtained via two evaluations of the electrostatic energy, both of which use the the derivations given in Section 3.2.1.1. Critically, I have also shown that for accurate induction-coupled calculations, at all values of ε^{max} and $\alpha_{l_1+l_2}$, a complete Ewald summation should be used. The general implementation described thus far is applicable

Table 7.4: Induction-coupled Electrostatic Ewald summation test specifications.

Test Specification	Value
Level of Theory	PBE0/6-311G(d,p)
Electrostatic model	HLYGAt-fitted atomic-charges
Real-space cutoff distance	10.0 Å
Ewald sum truncation error	1×10^{-20}, 1×10^{-6}, 1×10^{-2} kJ/mol
$\varepsilon_{\text{scrp}}$	1×10^{-3} a.u.
r_{scrp}	40.0 Å
t_{max}	15
t_{slow}	6
$\boldsymbol{Q}_{I,0}^{\text{total,pc}}$	$\boldsymbol{Q}_{I}^{\text{static,HLYGAt}}$

Table 7.5: U_{real}, U_{rec}, and U_{corr} determined from single-point evaluations of PYRENE03 with a point-charge model, using a complete Ewald sum. U_{sum} is the sum of these three contributions.

Electrostatics	α_0	U_{real} (kJ/mol)	U_{rec} (kJ/mol)	U_{corr} (kJ/mol)	U_{sum} (kJ/mol)
Static	0.025080	-4.0736	1.1990	-6.3614	-9.2361
	0.053286	0.1201	7.1902	-16.5314	-9.2211
	0.154262	1.0683	69.9155	-80.2048	-9.2210
Static + Ind.	0.025080	-5.1631	1.0904	-6.9370	-11.0098
	0.053286	-0.1492	7.4002	-18.2491	-10.9981
	0.154262	1.1049	74.1874	-86.2904	-10.9981
Induction	0.025080	-1.0895	-0.1086	-0.5756	-1.7737
	0.053286	-0.2694	0.2101	-1.7177	-1.7770
	0.154262	0.0366	4.2719	-6.0856	-1.7770

to electrostatic models truncated at any multipole rank. However, as one could infer, SCERP is inherently cost prohibitive due its reliance on repeated QM calculation, disqualifying its use in the Global Search stage. In the rest of this work, I will discuss SCERP only in the context of a Refinement model that uses higher-ranked multipoles (rank 4 unless stated otherwise).

7.3.2 Single-point Evaluation of Induction Energy

Before addressing SCERP-coupled rigid-body minimizations, I use the current calculation framework to study the induction model via single-point energy evaluations. Since these calculations do not alter the geometry of the crystal, the predicted electrostatic energies are compared at fixed crystal structures. Crystals from the CE–755 dataset [1, 256] are employed as a test set, comprising 441 NHB and 314 HB crystal structures. For each crystal, single-point energy evaluations are performed with: (1) only static electrostatics (i.e., no PCM or SCERP), (2) PCM-embedded electrostatics ($\varepsilon_{\text{PCM}} = 11$), and (3) SCERP-embedded electrostatics. For each polarization model (PM), the induction contribution is determined as

$$U_{\text{ind,PM}} = U_{\text{elec+ind,PM}} - U_{\text{elec,static}}, \qquad \text{PM} = \text{PCM11}, \text{SCERP}. \tag{7.3.5}$$

For all calculations, GDMAv2.2 is used to generate multipoles up to hexadecapoles (rank 4), from SCF charge densities determined at the PBE0/6-311G(d,p) level of theory in G09. Where relevant, $\varepsilon_{\text{scrp}}$, r_{scrp}, t_{max}, t_{slow}, and $\boldsymbol{Q}_{I,0}^{\text{total,pc}}$ follow Table 7.4. It should be noted that the treatment of induction in Equations 7.1.2 and 7.3.3 (i.e., separating Q^{static} and ΔQ, and using two electrostatic evaluations) is done only with SCERP. While a similar treatment could be applied to the PCM-embedded case, to my knowledge, most CSP codes (including our own) do not do this. As such, I have kept this convention and determine $U_{\text{elec+ind,PCM11}}$ by a single electrostatic evaluation, using $\boldsymbol{Q}^{\text{total}}$ derived from the PCM-embedded QM calculation.

7.3.2.1 SCERP Process Convergence and Compute Times

Building on the tests done in Section 7.2.2, I examine the convergence of the SCERP process over the CE–755. Since all molecules in the asymmetric unit cell are subject to the SCERP process,

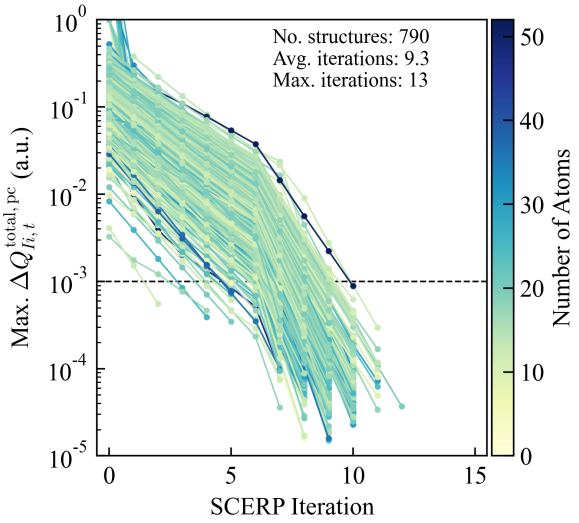


Figure 7.6: The maximum values of $\Delta Q_{Ii,t}^{\mathrm{total,pc}}$ during the single-point SCERP evaluation of structures from the CE–755 dataset. The dashed line indicates $\varepsilon_{\mathrm{scrp}} = 1 \times 10^{-3}$ a.u.

over 755 crystal structures (some of which are $Z' > 1$), there are 790 *molecules* evaluated. From Figure 7.6, the SCERP process converges successfully for all 790 molecules (either by A5 or A6) within the allotted $t_{\mathrm{max}} = 15$, and no cases of divergence are observed. To reach convergence, a maximum of 13 iterations are expended for a single molecule, and on average ~9 iterations are used. Welch *et al.* [189] reported that a self-consistent solution for the WSM method is reached in about 6 iterations, while Zhang's implementation of SCERP converged in about half-a-dozen iterations [71]. Here, the number of iterations used to converge the SCERP process is raised by the deceleration protocol ($t_{\mathrm{slow}} = 6$). Comparing the change in $\max\left(\Delta Q_{Ii,t}^{\mathrm{total,pc}}\right)$ before and after the sixth SCERP iteration, the deceleration protocol clearly contributes to a significant retardation of the SCERP process. As described later, SCERP-embedded energy minimizations require multiple assessments of the SCERP process. As such, I maintain that preserving the robustness of the SCERP process should be prioritized over its speed, otherwise minimizations can be impeded by spurious termination of the SCERP process, if divergence occurs. There is no indication that the convergence rate is related to molecular size.

Due to restrictions in the user license, I am unable to report exact Gaussian compute times for each SCERP iteration; however, their general trends are described here. In general, the expected positive correlation between compute time and molecular size is observed, although the cost of electronic structure methods scales non-linearly with system size [121]. Crucially, for a given molecule, compute times generally decrease over the course of the SCERP process. This is more obvious in larger molecules, where CPU times decrease by as much as half between the first and last SCERP iteration. Conversely, smaller molecules exhibit less reduction in CPU time over the SCERP process. Over the entire dataset, the compute time (per atom) decreases by about 25% between the first and last iteration of each SCERP process. This diminishing cost is likely attributable to the restart of each QM calculation from the preceding SCF solution (Figure 7.1, A3 → A1).

Overall, there is further evidence that the SCERP process is relatively robust. The greater number of iterations needed to reach convergence due to process deceleration is partially offset by the implemented cost saving strategy, particularly for larger molecules.

7.3.2.2 Predicted Induction Energies

To gauge the physical fidelity of the SCERP induction model, I compare predicted U_{ind} against the incumbent induction model used in CSP calculations, PCM (Figure 7.7). The difference in predicted induction energies (ΔU_{ind}) is determined as

$$\Delta U_{\text{ind}} = U_{\text{ind,SCERP}} - U_{\text{ind,PCM11}}, \tag{7.3.6}$$

such that $\Delta U_{\text{ind}} > 0$ indicates the PCM model predicts more U_{ind} stabilization. In Figure 7.7, each structure is colored based on the average absolute ESP (Avg. $|V_{\text{field}}|$) exerted by the (zeroth iteration) SCERP-field charges on the 0.8x van der Waals (vdW) surface of each molecule. The 0.8x surface is arbitrarily chosen as a region where a significant proportion of the molecule's charge density should exist. The ESP at a point p on this surface is given by,

$$V_{\text{field},p} = \frac{1}{4\pi\varepsilon_0} \sum_{s}^{N_s} \frac{1 \times Q_{s,0}^{\text{total,pc}}}{r_{ps}}, \tag{7.3.7}$$

where N_s is the total number of SCERP-field charges, s is the index of a point-charge in the SCERP-field, $Q_{s,0}^{\text{total,pc}}$ is its corresponding charge value at the zeroth SCERP iteration, and r_{ps} is the Euclidean distance between points s and p. $V_{\text{field},p}$ is expressed as the energy resulting from a unit-charge probe on the vdW surface (hence the multiplication by 1). The Avg. $|V_{\text{field}}|$ for a structure is determined as

$$\text{Avg. } |V_{\text{field}}| = \sum_{p}^{N_p} \frac{|V_{\text{field},p}|}{N_p}, \tag{7.3.8}$$

where N_p is the total number of sampling points on the 0.8x vdW surface. If $Z' > 1$, only the first molecule in the asymmetric unit cell is assessed. Avg. $|V_{\text{field}}|$ provides insight into the driving force for polarization exerted by the SCERP-field onto the central molecule. This cannot be inferred from the magnitudes of $Q_{s,0}^{\text{total,pc}}$ alone, since the polarizing force is a cumulative result of the constructive/destructive interplay of all SCERP-field charges.

For both the NHB and HB crystal structures, it is reassuring to observe the expected relationship between U_{ind} and Avg. $|V_{\text{field}}|$. Larger Avg. $|V_{\text{field}}|$ suggest a greater polarizing force on the central molecule(s) from the SCERP-field. Correspondingly, one should expect a greater degree of polarization and more stabilization from U_{ind}. This is generally the case, with the color of the data points darkening towards the lower portion of the plots. It is also worth noting the different scales of Avg. $|V_{\text{field}}|$ for the NHB and HB structures. For the NHB structures, Avg. $|V_{\text{field}}|$ values do not exceed 1.20 eV, while for the HB structures this limit is at ~1.75 eV. This is consistent with chemical intuition: molecules that form hydrogen bonds tend to inherently exhibit greater charge separation (within the molecule itself). This manifests as larger magnitudes in $\boldsymbol{Q}_{I,0}^{\text{total,pc}}$ (which are initialized at $\boldsymbol{Q}_{I}^{\text{static,HLYGAt}}$) and higher Avg. $|V_{\text{field}}|$.

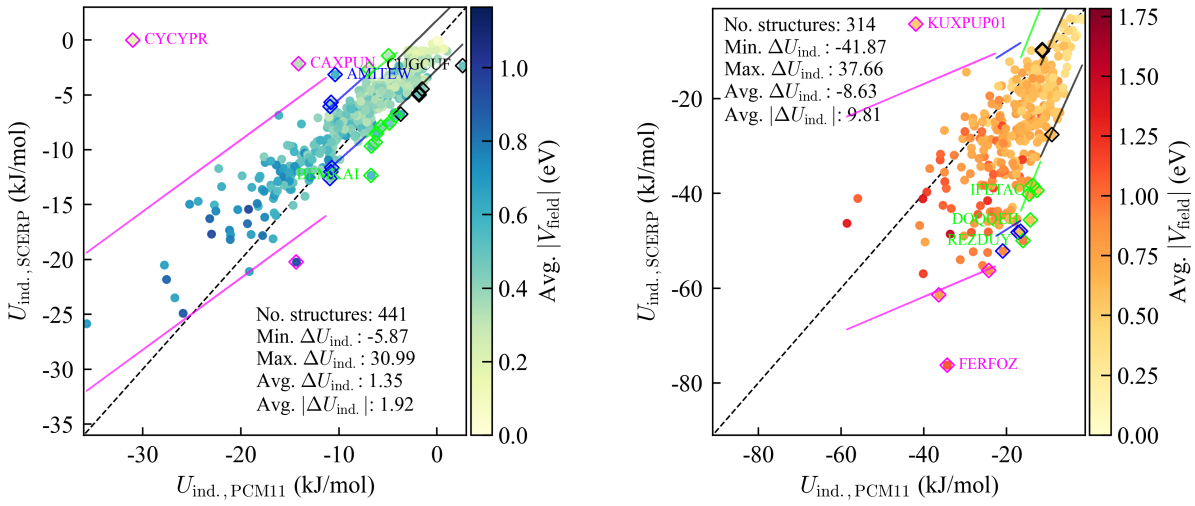


Figure 7.7: Predicted U_{ind} for the PCM- and SCERP-embedded models in (a) NHB, and (b) HB structures. The colors of each point reflect the Avg. $|V_{\text{field}}|$ values. The solid lines represent a $\pm 2\sigma_{\text{std.dev}}$ bound from the best fit line in each quantile. Points outside these bounds are flagged as outliers ($\diamond$), with the top-5 outliers labeled by their CSD Refcodes.

In the NHB crystals, the SCERP-embedded model tends to predict less induction stabilization than the PCM model (Avg. ΔU_{ind} = 1.35 kJ/mol). From intuition, this is a reasonable correction from SCERP. For most organic crystals, $\varepsilon_{\text{PCM}} \approx 3$ should be more representative of the electric field in the crystalline environment [48, 191, 218, 219]. When $\varepsilon_{\text{PCM}} = 11$ is used, the pseudo-'crystalline environment' provided by PCM is too polarizable. I expect this to over-polarize the central molecule and predict excess stabilization from U_{ind}. This is also suggested from Figures 6.4a and 6.6a, wherein the FF training conducted with $\varepsilon_{\text{PCM}} = 11$ leads to negative ADs in the NHB structures (i.e., overbinding). SCERP proves to more accurately represent the crystalline electric field and corrects for this overstabilization. In Chapter 6, I suggested that the over-polarization caused by using $\varepsilon_{\text{PCM}} = 11$ may not be too drastic due to restraint from the molecule's inherent polarizability (such that the FF can still absorb these errors). Figure 7.7a corroborates this, with Avg. $|\Delta U_{\text{ind}}|$ being only 1.92 kJ/mol.

In contrast, for the HB crystals, *greater* stabilization from induction is predicted using the SCERP-embedded model compared to the PCM model (Avg. ΔU_{ind} = −8.63 kJ/mol). On the one hand, this is surprising considering the previous statements regarding the choice of ε_{PCM}: "*Surely,* $\varepsilon_{\text{PCM}} = 11$ *is also excessive for HB systems?*" On the other hand, anecdotal experience has shown that hydrogen-bonding [48] and highly charged [48, 188] systems benefit from using $\varepsilon_{\text{PCM}} > 3$. Cooper *et al.* [188] proposed that this may be because PCM provides only a mean approximation of induction in the crystal, whilst neglecting the details of highly directional polarization effects (e.g., from hydrogen bonds). The current results support this proposition. Through a more-nuanced electronic-embedding approach, the strongly polarizing (local) environment around hydrogen bonds is better represented with SCERP, which leads to enhanced U_{ind} stabilization over the $\varepsilon_{\text{PCM}} = 11$ model. To illustrate this, I study the 4-oxo-1,4-

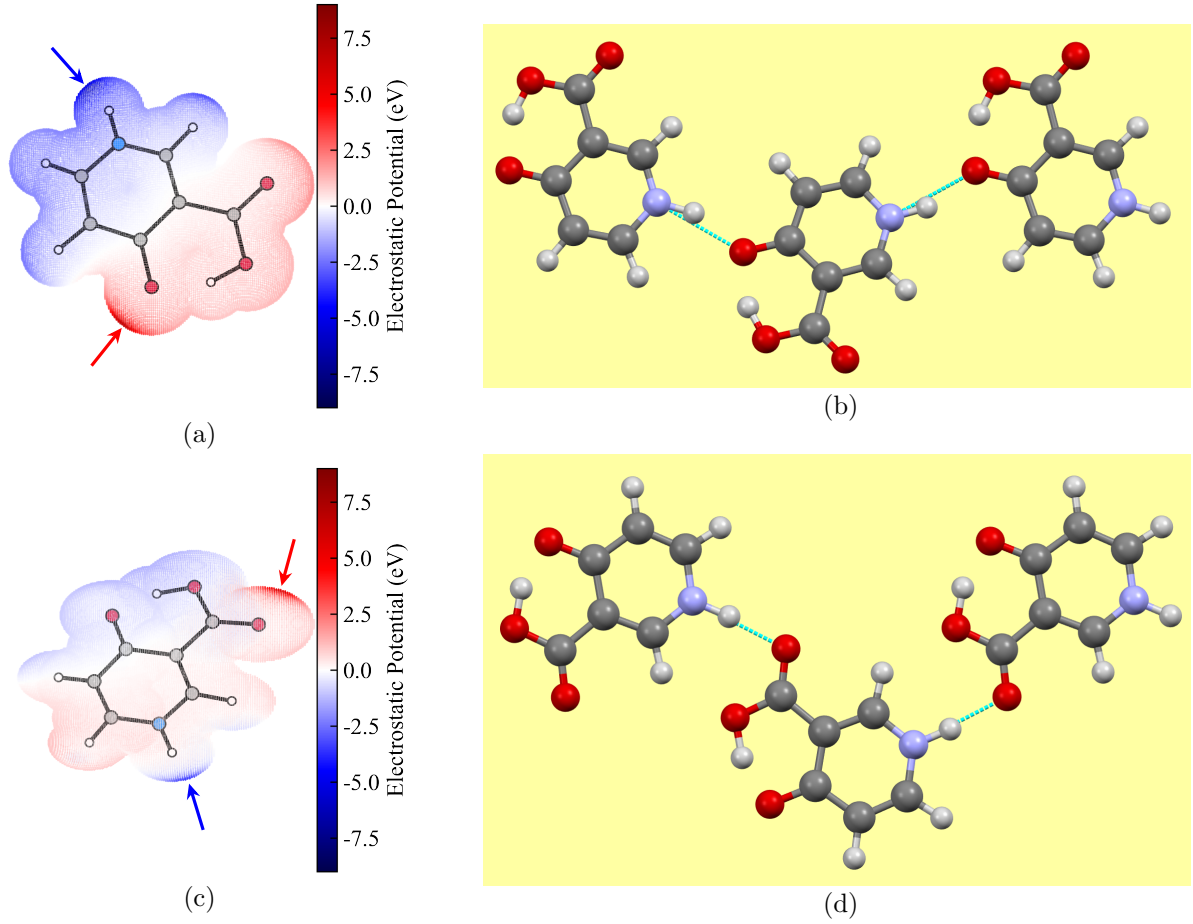


Figure 7.8: (a) and (c): ESP from the SCERP-field on the 0.8x vdW surface of KUXPUP and KUXPUP01 respectively; arrows point to the areas of greatest ESP. (b) and (d): The HB-motif in KUXPUP and KUXPUP01 respectively; hydrogen bonds are given by the cyan dotted lines.

dihydropyridine-3-carboxylic acid system, which has two forms in the CE–755: KUXPUP and KUXPUP01 (Figure 7.8). Despite both polymorphs (effectively) containing the same conformation ($RMSD_1$ = 0.041 Å, including hydrogen positions), the ESPs generated by their respective SCERP-fields are distinct. For each form, the region on the vdW surface experiencing the greatest polarizing force reflects their different hydrogen bonding motifs. Conversely, using PCM, the molecules in both polymorphs would share the same polarization profile, since the PCM-response is packing-agnostic. The validity of the enhanced stabilization predicted by SCERP is corroborated again by the FF training results: the HB crystals tend to be underbound even when using PCM with $\varepsilon_{PCM} = 11$ (positive ADs in Figures 6.4b and 6.6b).

Largely, SCERP appears to rectify deficiencies in the PCM model, but there are cases where the former seems to be erroneous. To identify these cases, each dataset (NHB and HB) is split into four equal quantiles (based on $U_{ind,PCM11}$). This is done to accommodate heteroscedasticity over the full sets. For each group, a best fit line is found by quantile regression, where the line is fit such that X% of the data exists below the line; here, I use $X = 50$. This form of regression is usually more robust to outliers. Finally, I flag outliers as points that are > 2 standard deviations ($\sigma_{std.dev}$) away from each best fit line. The top-5 outliers (greatest multiple of $\sigma_{std.dev}$ away from

their respective best fit lines) are labeled in Figure 7.7. Among the outliers identified, most are (visually) still quite close to the rest of the data. I also disregard FERFOZ as an outlier since it adheres to the larger trend across the entire HB dataset (accounting for heteroscedasticity). Only two systems emerge as convincing outliers: cis-1,2,3-tricyanocyclopropane (CYCYPR) and KUXPUP01. With SCERP, both are predicted to have almost no U_{ind}, in contrast to the PCM model. Two potential sources of this error are considered:

1. The hydrogen 'RADIUS' used in GDMA's quadrature weighting scheme ($r_{\text{H}}^{\text{GDMA}}$). This is discussed in Section 7.3.2.3.
2. The atomic-charge representation of the SCERP-field. A satellite-charge model was explored, but found to be impractical. Interested readers are directed to Appendix K.

Admittedly, neither of these explain the outliers and (at this point) it remains unclear what is causing this inaccuracy. However, given the relative rarity of this spurious behavior, I assert that the SCERP model is adequately robust to warrant further development in our CSP codes. Following the discussion in Section 7.3.2.3, I will complete the implementation of SCERP in CSO–RM and FM. These outliers will be revisited in Chapter 9, after further testing.

7.3.2.3 Evaluation of GDMA RADIUS

In GDMAv2 [182], a grid-based quadrature multipole distribution scheme was introduced [181]. Under this scheme, the weight attributed to an atomic site for a grid point is determined by: (1) the distance between the site and the grid point, and (2) the 'RADIUS' ($r_{\text{X}}^{\text{GDMA}}$) associated with the site's element (X). With a smaller $r_{\text{X}}^{\text{GDMA}}$, the weight falls off faster with distance, leading to less electron density being apportioned to a site. Welch [183] had proposed that hydrogens should use $r_{\text{H}}^{\text{GDMA}}$ = 0.325 Å. Firstly, he intuits that hydrogens occupy less volume in space and thus should hold less electron density. Secondly, when using smaller $r_{\text{H}}^{\text{GDMA}}$, Welch found that molecular ESPs converged faster (i.e., at lower rank) to those obtained using GDMAv1.2-derived multipoles (*cf.* Table 2 in Ref [183]). Finally, Welch noted that U_{elec} predicted from GDMAv2 rank 4 multipoles agree better with GDMAv1.2 results when $r_{\text{H}}^{\text{GDMA}}$ = 0.350 Å instead of 0.650 Å, the former having slightly greater electrostatic stabilization (*cf.* Table 5 in Ref [183]). Contrary to this, when GDMA is called by our group's CSP codes, the programs default to using $r_{\text{H}}^{\text{GDMA}}$ = 0.650 Å (which was the default in GDMAv2.1). Since this has been applied to all results generated up to this point, it is desirable to preserve consistency and continue using $r_{\text{H}}^{\text{GDMA}}$ = 0.650 Å. Here, I discuss the implications of this choice.

The physical intuition Welch proposes is sound, so here I assess only his empirical results. Firstly, rapid (multipole rank) convergence of the molecular ESPs may have been pertinent to Welch since his induction calculations were truncated at lower multipole ranks [183, 189]. However, in this work, all electrostatic interactions involving multipoles (including induction) are truncated at rank 4, hence this is less important. Stone [182] agrees that using $r_{\text{H}}^{\text{GDMA}}$ = 0.325 Å gives multipole moments that align better with chemical intuition, but notes that accurate molecular ESPs are attainable for most values of $r_{\text{H}}^{\text{GDMA}}$. Despite this, Welch still found disparate

U_{elec} predicted with $r_{\text{H}}^{\text{GDMA}}$ = 0.350 Å and $r_{\text{H}}^{\text{GDMA}}$ = 0.650 Å. I posit that this could be related to how DMACRYS conducts its Ewald sums. Using GDMA-derived multipoles with CSO–RM ($r_{\text{H}}^{\text{GDMA}}$ = 0.650 Å), Bowskill showed that the energy contributions from multipole interactions do not converge smoothly with $l_1 + l_2$ (*cf.* Figures 7.4 and 7.5 in Ref [2]). This may be due to the poorer rank convergence associated with $r_{\text{H}}^{\text{GDMA}}$ = 0.650 Å, but in any case his results indicate that significant electrostatic contributions can come from the higher-ranked multipole interactions when $r_{\text{H}}^{\text{GDMA}}$ = 0.650 Å. Welch's U_{elec} shortfall (when using $r_{\text{H}}^{\text{GDMA}}$ = 0.650 Å) could stem from DMACRYS performing Ewald sums only up to dipole–dipole interactions, and failing to account fully for contributions from higher-ranked interactions (particularly the charge–quadrupole interactions, which are also conditionally convergent).

For greater clarity on the influence of $r_{\text{H}}^{\text{GDMA}}$, the previous single-point energy evaluations are repeated using $r_{\text{H}}^{\text{GDMA}}$ = 0.325 Å. In Figure 7.9, I compare the predicted $U_{\text{elec+ind}}$ for the two values of $r_{\text{H}}^{\text{GDMA}}$, across all three cases ('Static', 'PCM11', and 'SCERP'). The difference in their predicted $U_{\text{elec+ind}}$ is determined as

$$\Delta U_{\text{elec+ind}} = U_{\text{elec+ind,PM}}^{0.325} - U_{\text{elec+ind,PM}}^{0.650}, \qquad \text{PM} = \text{Static, PCM11, SCERP}, \tag{7.3.9}$$

such that $\Delta U_{\text{elec+ind}} > 0$ indicates that greater electrostatic stabilization is predicted when $r_{\text{H}}^{\text{GDMA}}$ = 0.650 Å. Like Welch, I observe that the predicted $U_{\text{elec+ind}}$ differ depending on the value of $r_{\text{H}}^{\text{GDMA}}$. However, I find that using $r_{\text{H}}^{\text{GDMA}}$ = 0.650 Å tends to predict greater elec-

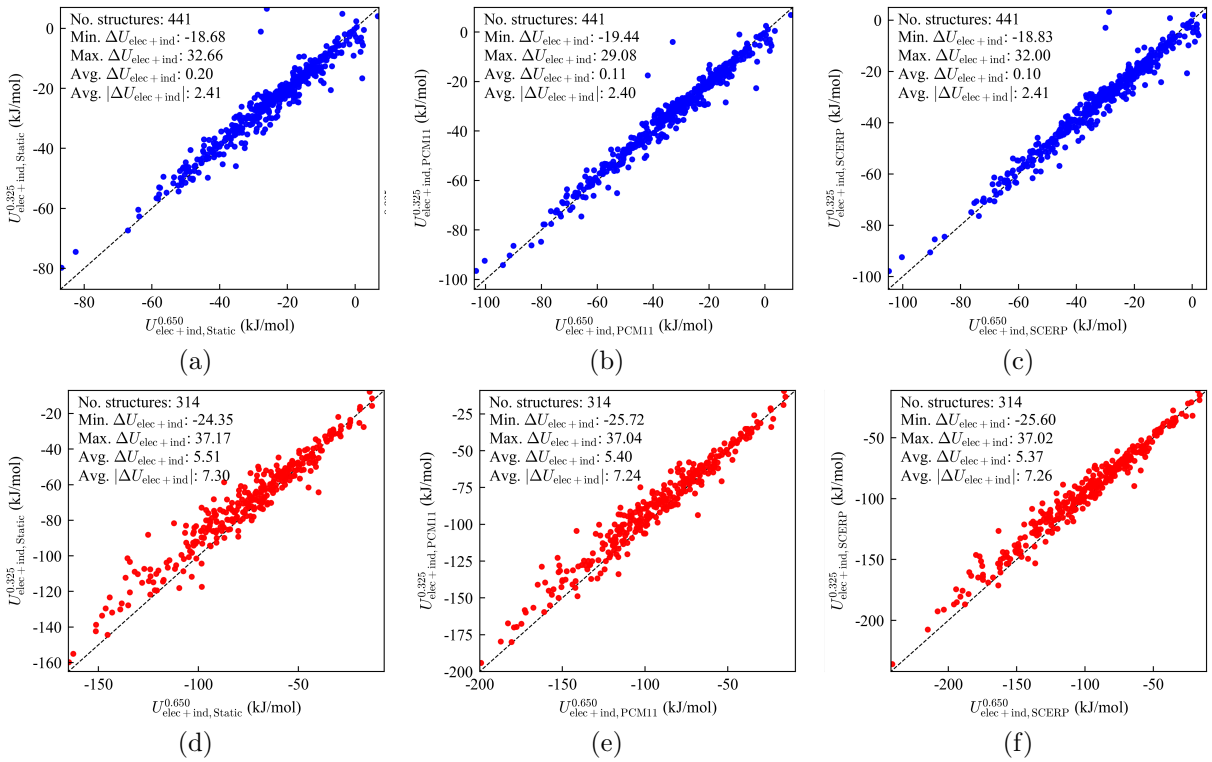


Figure 7.9: Predicted $U_{\text{elec+ind}}$ for $r_{\text{H}}^{\text{GDMA}}$ = 0.650 Å and 0.325 Å. NHB structures: (a) without PCM or SCERP, (b) with PCM (ε_{PCM} = 11), and (c) with SCERP. HB structures: (d) without PCM or SCERP, (e) with PCM (ε_{PCM} = 11), and (f) with SCERP.

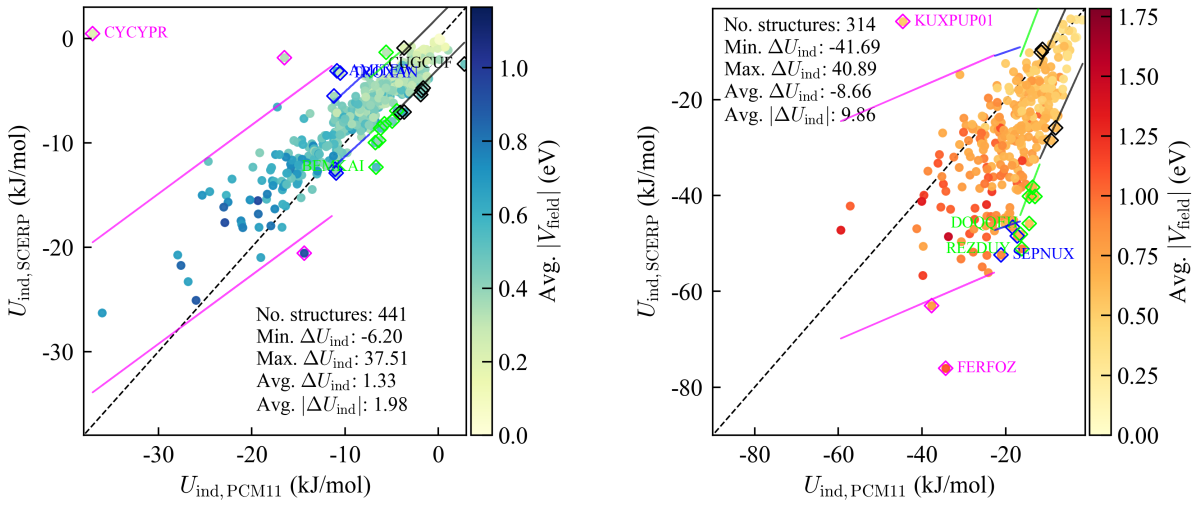


Figure 7.10: Predicted U_{ind} when r_H^{GDMA} = 0.325 Å, for the PCM- and SCERP-embedded models in (a) NHB, and (b) HB structures. The colors of each point reflect the Avg. $|V_{field}|$ values. The solid lines represent a $\pm 2\sigma_{std.dev}$ bound from the best fit line in each quantile. Points outside these bounds are flagged as outliers ($\diamond$), with the top-5 outliers labeled by their CSD Refcodes.

trostatic stabilization than using r_H^{GDMA} = 0.325 Å: average $\Delta U_{elec+ind} > 0$ in both the NHB and HB structures. $\Delta U_{elec+ind}$ also tend to be smaller in the NHB structures, versus the HB structure. The contradictory $\Delta U_{elec+ind}$ behavior predicted here, versus Welch [183], aligns with my proposition regarding the Ewald sum treatment in DMACRYS and CSO–RM. It should also be noted that the comparison done here is conducted on a much larger data set, whereas Welch considered only a single hydantoin case study [183].

From these results, I believe it is justifiable to continue using r_H^{GDMA} = 0.650 Å. Using CSO–RM, greater electrostatic stabilization is usually predicted with r_H^{GDMA} = 0.650 Å, particularly for HB structures (Avg. $\Delta U_{elec+ind}$ ~5 kJ/mol). From the FF fitting, HB crystals exhibit mild underbinding with r_H^{GDMA} = 0.650 Å, hence switching to r_H^{GDMA} = 0.325 Å may exacerbate this error. A key tenet of this work is the complementary utilization of the HAIEFF FF and electrostatic model. Leveraging this, relatively small errors associated with r_H^{GDMA} can be reconciled by a consistent FF parametrization. Finally, I note that altering r_H^{GDMA} has not remedied the outlier behavior observed for CYCYPR and KUXPUP01 (Figure 7.10). Henceforth, I will continue to use r_H^{GDMA} = 0.650 Å. This will facilitate the comparison of results obtained with/without SCERP, and avoid obfuscating the effects of the induction model.

7.3.3 CSO–RM Minimization Protocol

Thus far, I have discussed the steps and formulation leading up to a single-point (SCERP-coupled) energy evaluation. Extending this to energy minimizations is non-trivial due to additional dependencies that SCERP introduces into the intermolecular model. For rigid molecules, the optimization problem solved by CSO–RM (without SCERP) is given by,

$$\min_{\boldsymbol{\Omega}} U_{\text{inter}}(\boldsymbol{\Omega};\boldsymbol{\theta},\boldsymbol{Q}(\boldsymbol{\theta})), \tag{7.3.10}$$

where the rigid-body variables, $\boldsymbol{\Omega}$, are decision variables in the minimization of U_{inter}. Since molecules are treated rigidly, the minimization is conducted at a fixed conformation ($\boldsymbol{\theta}$ is a fixed input parameter, as denoted by its position after the ';'). By the isolated-molecule assumption, the electrostatic features of the molecule are dependent on only its conformation ($\boldsymbol{Q}(\boldsymbol{\theta})$), thus these are also treated as fixed inputs. In short, the point-charges/multipoles employed in a rigid optimization have hitherto been derived once and then fixed throughout the minimization.

Adding induction into the intermolecular model, the optimization problem is altered as

$$\min_{\boldsymbol{\Omega}} U_{\text{inter}}(\boldsymbol{\Omega},\Delta\boldsymbol{Q}(\boldsymbol{\theta},\boldsymbol{\Omega});\boldsymbol{\theta},\boldsymbol{Q}(\boldsymbol{\theta})). \tag{7.3.11}$$

Critically, the induced multipoles also have a dependency on the rigid-body variables: changes to $\boldsymbol{\Omega}$ affect the crystal packing, and by extension the embedding-environment used to generate $\Delta\boldsymbol{Q}$. Attempting to solve this by updating $\Delta\boldsymbol{Q}$ at every trial value of $\boldsymbol{\Omega}$ is impractical for two reasons. Firstly, each update of $\Delta\boldsymbol{Q}$ requires converging the SCERP process. If done at *every* step of the optimization, this translates to an untenable number of QM calculations. Secondly, to complement the update of $\Delta\boldsymbol{Q}$, gradients of $\Delta\boldsymbol{Q}$ with respect to $\boldsymbol{\Omega}$ would be required by the E04UFF subroutine. However, $\Delta\boldsymbol{Q}$ are derived implicitly from the SCERP process and there is no closed-form expression for $\Delta\boldsymbol{Q}(\boldsymbol{\Omega})$. As such, these gradients can be obtained only by finite-differences involving the SCERP process, but this would drastically exacerbate the number of QM computations. In lieu of this, I opt to use an iterative approach to resolve this optimization (Figure 7.11). This procedure is explained in more detail below.

It is important to recognize that CSO–RM has been written as two separate executables. The first (denoted as 'CSORM_Init') conducts QM computations, derives the electrostatic features (e.g., multipole moments) needed in the minimization, and finally stores these features as an intermediate input file. This input is used by the second executable (denoted as 'CSORM_Min') to perform the U_{inter} minimization. This construction reflects that, under the isolated-molecule assumption (and without SCERP), the electrostatic features of rigid molecules can be pre-computed and do not need any further update during the minimization.

My embedding of SCERP into CSO–RM conforms with this existing framework. To represent CSORM_Init, I initialize the iteration count of the CSO–RM protocol at $k = -1$. In this 'priming' step, static multipoles are generated (Figure 7.11, R1), as they would be without SCERP. Additionally, the induced multipoles are also generated by CSORM_Init, via the SCERP process (Figure 7.11, R4 → Figure 7.1). Prior to this, SCERP rigid-body variables are initialized at the input crystal structure (Figure 7.11, R2: $\boldsymbol{\Omega}_{\text{scrp},k=-1} = \boldsymbol{\Omega}$), while the trial solution for the SCERP process is initialized at the static point-charges (Figure 7.11, R3: $\boldsymbol{Q}_{I,0,k=-1}^{\text{total,pc}} = \boldsymbol{Q}_{I}^{\text{static,pc}}$). The HLYGAt-values of $\boldsymbol{Q}_{I}^{\text{static,pc}}$ are derived as part of the preceding static QM calculation (Figure 7.11, R1). Once the induced multipoles, $\Delta\boldsymbol{Q}_{I,k=-1}^{\text{mp}}$ are found, CSORM_Init stores these, together with the static multipoles, in the intermediate input file; CSORM_Init terminates here.

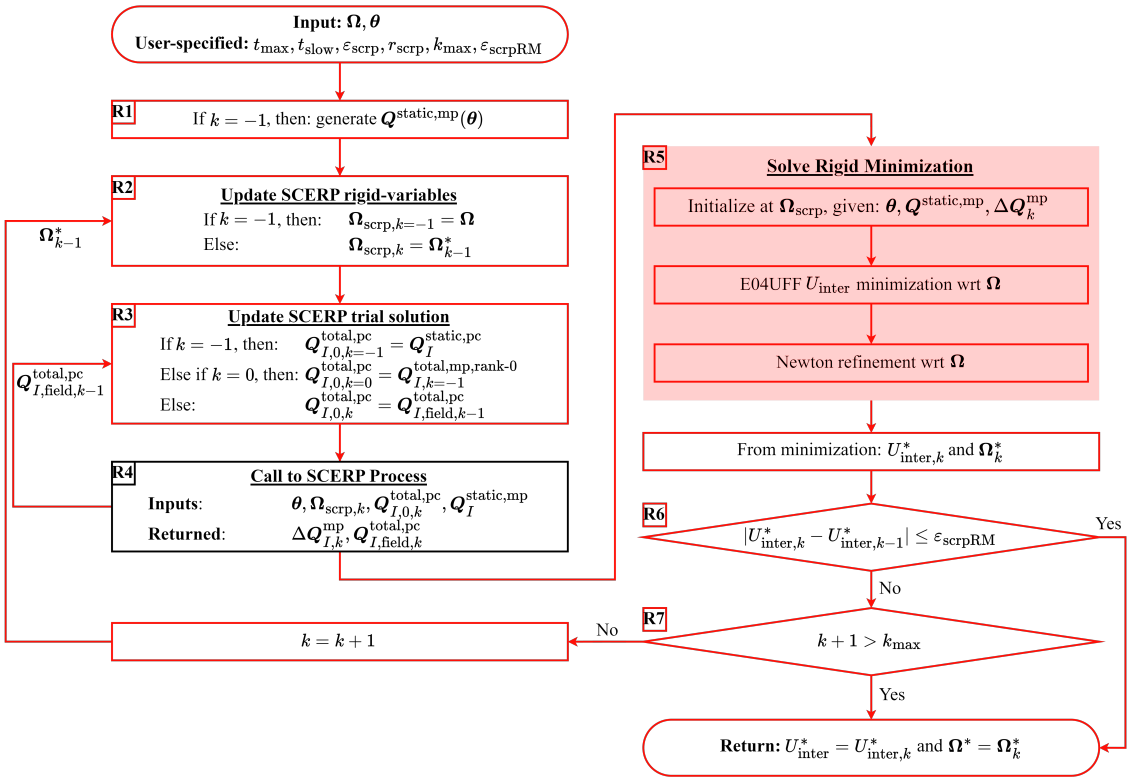


Figure 7.11: Schematic of the CSO–RM protocol, when used with SCERP. Key stages are enumerated and detailed in the main text.

Executing CSORM_Min in the $k = -1$ iteration, steps R1–R4 are skipped and all requisite electrostatic information is read from the CSORM_Init-generated input file. For a given set of induced multipoles ($\Delta\boldsymbol{Q}^{\text{mp}}_k = \Delta\boldsymbol{Q}^{\text{mp}}_k(\boldsymbol{\Omega}_{\text{scrp},k})$), CSORM_Min attempts to solve the rigid minimization problem (Figure 7.11, R5) formulated as

$$\min_{\boldsymbol{\Omega}} U_{\text{inter}}(\boldsymbol{\Omega}; \boldsymbol{\theta}, \boldsymbol{Q}(\boldsymbol{\theta}), \Delta\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{\Omega}_{\text{scrp}})). \tag{7.3.12}$$

In short, the induced multipoles are *fixed* during this minimization. The rigid-minimization proceeds normally (i.e., E04UFF quasi-Newton minimization, followed by Newton refinement), except that the electrostatic calculations are augmented (per Section 7.3.1) to accommodate the additional induction contributions. Solving Equation 7.3.12, $U^*_{\text{inter},k}$ and $\boldsymbol{\Omega}^*_k$ are returned. However, because $\boldsymbol{\Omega}^*_k \neq \boldsymbol{\Omega}_{\text{scrp},k}$, the induced multipoles derived at $\boldsymbol{\Omega}_{\text{scrp},k}$ are no longer valid at $\boldsymbol{\Omega}^*_k$, and as such this is not yet a solution to Equation 7.3.11. To obtain a 'true' solution to Equation 7.3.11, CSORM_Min repeats this protocol (Figure 7.11, R6 → R7 → R2), but initialized from the last rigid-body solution, $\boldsymbol{\Omega}^*_{k-1}$. Note that static multipoles do not need to be re-derived between iterations. As the 'guess' for $\boldsymbol{\Omega}_{\text{scrp},k}$ is refined between steps, it is likely that $\boldsymbol{\Omega}^*_k \to \boldsymbol{\Omega}_{\text{scrp},k}$, in which case $\Delta\boldsymbol{Q}^{\text{mp}}_k(\boldsymbol{\Omega}_{\text{scrp},k})$ will gradually converge; this would be reflected in $U^*_{\text{inter},k}$. On this basis, CSORM_Min terminates the minimization cycle when either: (Figure 7.11, R6) the change in $U^*_{\text{inter},k}$ between iterations is below the specified convergence tolerance, $\varepsilon_{\text{scrpRM}}$, or (Figure 7.11, R7) the maximum number of minimization attempts ($k_{\max}$) is reached.

The final set of $U^{*}_{\text{inter},k}$ and $\boldsymbol{\Omega}^{*}_{k}$ are returned as the solutions to Equation 7.3.11. Zhang [71] demonstrated that for $\varepsilon_{\text{scrpRM}} = 1 \times 10^{-3}$ kJ/mol, this protocol converges in ~5 iterations. From my own testing, I also find that this is usually the case. However, it must be stressed that this iterative strategy does not guarantee convergence, as exemplified in Sections 8.3.4.1 and 9.2.3.

To minimize computational costs where possible, I exploit all available information related to the choice of $\boldsymbol{Q}^{\text{total,pc}}_{I,0,k}$. In Zhang's CSO–RM implementation [71], the SCERP process is always initialized at the static electrostatics (i.e., $\boldsymbol{Q}^{\text{total,pc}}_{I,0,k} = \boldsymbol{Q}^{\text{static,pc}}_{I}$, $\forall k$) — this is suboptimal. In the current implementation, at each SCERP process call (R4), the converged SCERP-field charges ($\boldsymbol{Q}^{\text{total,pc}}_{I,\text{field},k}$) are stored. In succeeding CSO–RM iterations, the SCERP process is then initialized at this last available solution (i.e., $\boldsymbol{Q}^{\text{total,pc}}_{I,0,k} = \boldsymbol{Q}^{\text{total,pc}}_{I,\text{field},k-1}$ at R3), with two exceptions. Firstly, in the $k = -1$ iteration, there is no precedence to initialize $\boldsymbol{Q}^{\text{total,pc}}_{I,0,k=-1}$ at any value except $\boldsymbol{Q}^{\text{static,pc}}_{I}$. Secondly, in the $k = 0$ iteration, due to the separation of CSORM_Init and CSORM_Min, neither $\boldsymbol{Q}^{\text{static,pc}}_{I}$ nor $\boldsymbol{Q}^{\text{total,pc}}_{I,\text{field},k=-1}$ are accessible to CSORM_Min. To avoid performing additional QM calculations, I obtain a viable set of point-charges from

$$\boldsymbol{Q}^{\text{total,pc}}_{I,0,k=0} = \boldsymbol{Q}^{\text{total,mp,rank-0}}_{I,k=-1} = \boldsymbol{Q}^{\text{static,mp,rank-0}}_{I} + \Delta\boldsymbol{Q}^{\text{mp,rank-0}}_{I,k=-1}. \tag{7.3.13}$$

$\boldsymbol{Q}^{\text{total,mp,rank-0}}_{I,k=-1}$ are the *total* rank 0 multipole moments from the electrostatic model derived by CSORM_Init. These are 'reconstituted' from the corresponding static and $k = -1^{\text{th}}$ induced multipoles. With this, each call to the SCERP process is initialized from the best available guess for $\boldsymbol{Q}^{\text{total,pc}}_{I,0,k}$. As will be demonstrated in Chapter 9, this strategy does accelerate the convergence of the SCERP process in later iterations of the CSO–RM minimization cycle.

7.4 Flexible-Molecule Calculations with SCERP

7.4.1 CSO–FM Minimization Protocol

Lastly, I address the extension of SCERP to CSO–FM. This has been implemented in CSO–FMv1.4.0.2, which interfaces with CSO–RMv1.5.2.3. The CSO–FMv1.4.0.2 protocol is mostly unchanged from Figure 4.2. Below, I discuss the salient changes to the algorithm.

1. First is the role of the LAMs in SCERP computations. The existing LAM framework is formulated to build surrogate models based solely on $\boldsymbol{\theta}^{\text{ind}}$. The additional dependencies of $\Delta\boldsymbol{Q}$ on $\boldsymbol{\Omega}$ cannot be captured this way, and it is unclear how a surrogate model of $\Delta\boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{\Omega})$ could be built. As such, in CSO–FMv1.4.0.2, only the static electrostatic features ($\boldsymbol{Q}$) and conformational variables (ΔU_{intra} and $\boldsymbol{\theta}^{\text{dep}}$) are determined from the LAM models (as they would be without SCERP). All induced electrostatic features must be generated via appropriate calls to the SCERP process.
2. To that end, CSO–FMv1.4.0.2 interfaces with the CSO–RM protocol shown in Figure 7.11. When called from CSO–FM (Figure 4.2, C and D2), CSO–RM is initialized from Figure 7.11 R2, with $k = 0$ and $\boldsymbol{\Omega}_{\text{scrp},0}$ equal to the $\boldsymbol{\Omega}$ provided by CSO–FM. The generation

of induced multipoles is handled natively by calls to the SCERP process in the CSO–RM protocol, with static multipoles provided by either the LAMs (Figure 4.2, A4), or an independent static electrostatic evaluation (Figure 4.2, D1).

3. To reduce computational costs, I (again) reuse SCERP process information. At stage C of Figure 4.2, the CSO–RM protocol returns $\boldsymbol{\Omega}^*$, $U_{\text{inter}}(\boldsymbol{\Omega}^*)$, *and* the last set of SCERP-field charges ($\boldsymbol{Q}^{\text{total,pc}}_{I,\text{field},*}$). At the fth CSO–FM iteration, CSO–RM will initialize $\boldsymbol{Q}^{\text{total,pc}}_{I,0,k=0}$ using $\boldsymbol{Q}^{\text{total,pc}}_{I,\text{field},*}$ from the $(f-1)^{\text{th}}$ CSO–RM call, overriding the charge reconstitution normally done at $k = 0$ (Figure 7.11, R3). When $f = 0$, there is no prior $\boldsymbol{Q}^{\text{total,pc}}_{I,\text{field},*}$, and $\boldsymbol{Q}^{\text{total,pc}}_{I,0,k=0}$ is initialized from the static rank 0 multipole moments instead. The stored $\boldsymbol{Q}^{\text{total,pc}}_{I,\text{field},*}$ are also used in the gradient calculations, as discussed below. $\boldsymbol{Q}^{\text{total,pc}}_{I,\text{field},*}$ are not reused in the Final Polish step, and instead $\boldsymbol{Q}^{\text{total,pc}}_{I,0,\text{Polish}}$ are initialized from $\boldsymbol{Q}^{\text{static,HLYGAt}}_{I}$ derived as part of the static QM evaluation of D1 in Figure 4.2.

The most critical change which needs to be addressed is the effect of induction on the treatment of gradients with respect to the $\boldsymbol{\theta}^{\text{ind}}$. This is discussed separately below.

7.4.2 Gradients for CSO–FM with SCERP

Without SCERP, CSO–FM solves the bilevel optimization problem given in Equation 4.2.6. Introducing induction alters the problem as

$$\begin{aligned}
&\min_{\boldsymbol{\theta}^{\text{ind}}} \Delta U_{\text{intra}}(\boldsymbol{\theta}^{\text{ind}}) + U_{\text{inter}}(\boldsymbol{\Omega}^*(\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}})), \boldsymbol{\theta}^{\text{ind}}, \boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}}), \Delta\boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{\Omega}^*(\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}}))))\\
&s.t.\\
&\boldsymbol{\Omega}^* = \arg\min_{\boldsymbol{\Omega}} U_{\text{inter}}(\boldsymbol{\Omega}, \Delta\boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}}, \boldsymbol{\Omega}); \boldsymbol{\theta}^{\text{ind}}, \boldsymbol{Q}(\boldsymbol{\theta}^{\text{ind}})).
\end{aligned} \tag{7.4.1}$$

Note that the inner problem is consistent with the SCERP-embedded CSO–RM minimization problem (Equation 7.3.11). All dependencies on $\boldsymbol{\theta}^{\text{dep}}$ are implicit in the $\boldsymbol{\theta}^{\text{ind}}$ arguments.

Rigorous Gradient Derivation

Applying chain-rule to the intermolecular term in Equation 7.4.1 (assuming one θ^{ind}),

$$\begin{aligned}
&\frac{\partial}{\partial\theta^{\text{ind}}} U_{\text{inter}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}), \Delta\boldsymbol{Q}(\theta^{\text{ind}}, \boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))))\\
&\quad= \frac{\partial U_{\text{inter}}}{\partial\theta^{\text{ind}}}\bigg|_{\boldsymbol{\Omega}^*,\boldsymbol{Q},\Delta\boldsymbol{Q}} + \frac{\partial U_{\text{inter}}}{\partial\boldsymbol{Q}}\bigg|_{\boldsymbol{\Omega}^*,\theta^{\text{ind}},\Delta\boldsymbol{Q}}\left[\frac{\partial\boldsymbol{Q}}{\partial\theta^{\text{ind}}}\right]\\
&\quad\quad+ \frac{\partial U_{\text{inter}}}{\partial\Delta\boldsymbol{Q}}\bigg|_{\boldsymbol{\Omega}^*,\theta^{\text{ind}},\boldsymbol{Q}}\left\{\frac{\partial\Delta\boldsymbol{Q}}{\partial\theta^{\text{ind}}} + \frac{\partial\Delta\boldsymbol{Q}}{\partial\boldsymbol{\Omega}^*}\left[\frac{\partial\boldsymbol{\Omega}^*}{\partial\theta^{\text{ind}}} + \frac{\partial\boldsymbol{\Omega}^*}{\partial\boldsymbol{Q}}\left(\frac{\partial\boldsymbol{Q}}{\partial\theta^{\text{ind}}}\right)\right]\right\}\\
&\quad\quad+ \cancel{\frac{\partial U_{\text{inter}}}{\partial\boldsymbol{\Omega}^*}\bigg|_{\theta^{\text{ind}},\boldsymbol{Q},\Delta\boldsymbol{Q}}}\left[\frac{\partial\boldsymbol{\Omega}^*}{\partial\theta^{\text{ind}}} + \frac{\partial\boldsymbol{\Omega}^*}{\partial\boldsymbol{Q}}\frac{\partial\boldsymbol{Q}}{\partial\theta^{\text{ind}}}\right],
\end{aligned} \tag{7.4.2}$$

where the optimality criterion of the inner problem eliminates the $\frac{\partial U_{\text{inter}}}{\partial\boldsymbol{\Omega}^*}$ term. This is the

induction-coupled analogue to Equation 4.2.9. A critical difference with the induction-free case is that the resulting gradient expression still contains dependencies on $\boldsymbol{\Omega}^*$. Without any further approximation, the finite-difference (FD) computation necessitates evaluating

$$\begin{aligned}
&\frac{\partial}{\partial\theta^{\mathrm{ind}}}U_{\mathrm{latt}}(\boldsymbol{\Omega}^*(\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}})),\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}}),\Delta\boldsymbol{Q}(\theta^{\mathrm{ind}},\boldsymbol{\Omega}^*(\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}})))\\
&\quad\approx\frac{U_{\mathrm{latt}}(\boldsymbol{\Omega}^*(\theta^{\mathrm{ind},+},\boldsymbol{Q}(\theta^{\mathrm{ind},+})),\theta^{\mathrm{ind},+},\boldsymbol{Q}(\theta^{\mathrm{ind},+}),\Delta\boldsymbol{Q}(\theta^{\mathrm{ind},+},\boldsymbol{\Omega}^*(\theta^{\mathrm{ind},+},\boldsymbol{Q}(\theta^{\mathrm{ind},+})))}{2(\delta\theta^{\mathrm{ind}})}\\
&\quad-\frac{U_{\mathrm{latt}}(\boldsymbol{\Omega}^*(\theta^{\mathrm{ind},-},\boldsymbol{Q}(\theta^{\mathrm{ind},-})),\theta^{\mathrm{ind},-},\boldsymbol{Q}(\theta^{\mathrm{ind},-}),\Delta\boldsymbol{Q}(\theta^{\mathrm{ind},-},\boldsymbol{\Omega}^*(\theta^{\mathrm{ind},-},\boldsymbol{Q}(\theta^{\mathrm{ind},-})))}{2(\delta\theta^{\mathrm{ind}})},
\end{aligned}\tag{7.4.3}$$

where rigid-body *minimizations* are done in both FD steps. Without induction, it was already thought of as 'expensive' to conduct minimizations during the FD evaluation. With induction, this is even more acute, since the CSO–RM minimizations demand multiple cycles of SCERP process + optimization. To (slightly) mitigate this cost, CSO–RM calls during the FD evaluation are initialized with $\boldsymbol{Q}^{\mathrm{total,pc}}_{I,\mathrm{field},*}$. At the fth CSO–FM iteration, the function evaluation (Figure 4.2, C) precedes the gradient evaluation (Figure 4.2, A6), thus $\boldsymbol{Q}^{\mathrm{total,pc}}_{I,\mathrm{field},*}$ are always available for the FD computation. $\boldsymbol{Q}^{\mathrm{total,pc}}_{I,\mathrm{field},*}$ are not updated following a FD CSO–RM minimization. From testing, this strategy does slightly reduce the cost of the 'rigorous' FD calculation. However, it is still preferable if CSO–RM minimization be avoided altogether. To that end, some simplifying assumptions are considered.

Approximation 1: Invariant $\boldsymbol{\Delta Q}$

In the simplest case, one can assume that the induced multipoles are completely unchanged following a FD perturbation. All derivatives of $\Delta\boldsymbol{Q}$ are eliminated from Equation 7.4.2, which reduces to an identical gradient expression as the induction-free model (Equation 4.2.9):

$$\begin{aligned}
&\frac{\partial}{\partial\theta^{\mathrm{ind}}}U_{\mathrm{inter}}(\boldsymbol{\Omega}^*(\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}})),\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}}),\Delta\boldsymbol{Q}(\theta^{\mathrm{ind}},\boldsymbol{\Omega}^*(\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}})))\\
&\quad\approx\left.\frac{\partial U_{\mathrm{inter}}}{\partial\theta^{\mathrm{ind}}}\right|_{\boldsymbol{\Omega}^*,\boldsymbol{Q},\Delta\boldsymbol{Q}}+\left.\frac{\partial U_{\mathrm{inter}}}{\partial\boldsymbol{Q}}\right|_{\boldsymbol{\Omega}^*,\theta^{\mathrm{ind}},\Delta\boldsymbol{Q}}\left[\frac{\partial\boldsymbol{Q}}{\partial\theta^{\mathrm{ind}}}\right].
\end{aligned}\tag{7.4.4}$$

By this approximation, the gradients are independent of $\boldsymbol{\Omega}^*$ and the FD requires only single-point energy evaluations, done at $\boldsymbol{\Omega}^*(\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}}))$ and $\Delta\boldsymbol{Q}(\theta^{\mathrm{ind}},\boldsymbol{\Omega}^*(\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}})))$:

$$\begin{aligned}
&\frac{\partial}{\partial\theta^{\mathrm{ind}}}U_{\mathrm{latt}}(\boldsymbol{\Omega}^*(\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}})),\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}}),\Delta\boldsymbol{Q}(\theta^{\mathrm{ind}},\boldsymbol{\Omega}^*(\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}})))\\
&\quad\approx\frac{U_{\mathrm{latt}}(\boldsymbol{\Omega}^*(\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}})),\theta^{\mathrm{ind},+},\boldsymbol{Q}(\theta^{\mathrm{ind},+}),\Delta\boldsymbol{Q}(\theta^{\mathrm{ind}},\boldsymbol{\Omega}^*(\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}})))}{2(\delta\theta^{\mathrm{ind}})}\\
&\quad-\frac{U_{\mathrm{latt}}(\boldsymbol{\Omega}^*(\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}})),\theta^{\mathrm{ind},-},\boldsymbol{Q}(\theta^{\mathrm{ind},-}),\Delta\boldsymbol{Q}(\theta^{\mathrm{ind}},\boldsymbol{\Omega}^*(\theta^{\mathrm{ind}},\boldsymbol{Q}(\theta^{\mathrm{ind}})))}{2(\delta\theta^{\mathrm{ind}})}.
\end{aligned}\tag{7.4.5}$$

I assert that there is no physical justification for this approximation. It is a corollary that if $\boldsymbol{Q}$ vary with respect to the FD perturbation, $\Delta\boldsymbol{Q}$ should also exhibit a response, since both are

results of the molecular charge density. When testing this approximation, I often find gradients are grossly in error, precluding meaningful progress during the optimization. Zhang also found this to be the case [71].

Approximation 2: Invariant $\boldsymbol{Q}^{\text{total}}$

Recognizing that $\boldsymbol{Q}$ and $\Delta\boldsymbol{Q}$ cannot be independently invariant to the FD perturbation, Zhang [71] proposed the assumption that the *total* multipole moments should be invariant following the FD perturbation. From Equation 7.1.2, it follows that

$$\frac{\partial \Delta \boldsymbol{Q}}{\partial \theta^{\text{ind}}} = -\frac{\partial \boldsymbol{Q}}{\partial \theta^{\text{ind}}}, \tag{7.4.6}$$

and applying this to Equation 7.4.2 gives

$$\begin{aligned}
&\frac{\partial}{\partial \theta^{\text{ind}}} U_{\text{inter}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}), \Delta\boldsymbol{Q}(\theta^{\text{ind}}, \boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))) \\
&\quad \approx \frac{\partial U_{\text{inter}}}{\partial \theta^{\text{ind}}}\bigg|_{\boldsymbol{\Omega}^*, \boldsymbol{Q}, \Delta\boldsymbol{Q}} + \frac{\partial U_{\text{inter}}}{\partial \boldsymbol{Q}}\bigg|_{\boldsymbol{\Omega}^*, \theta^{\text{ind}}, \Delta\boldsymbol{Q}} \left[\frac{\partial \boldsymbol{Q}}{\partial \theta^{\text{ind}}}\right] + \frac{\partial U_{\text{inter}}}{\partial \Delta\boldsymbol{Q}}\bigg|_{\boldsymbol{\Omega}^*, \theta^{\text{ind}}, \boldsymbol{Q}} \left[\frac{\partial \Delta\boldsymbol{Q}}{\partial \theta^{\text{ind}}}\right].
\end{aligned} \tag{7.4.7}$$

It is tacitly implied, from Equation 7.4.6, that $\Delta\boldsymbol{Q}$ has no dependency on $\boldsymbol{\Omega}^*$, thus the resulting gradient expression strongly resembles the induction-free case, but with an additional term reflecting the effects of $\Delta\boldsymbol{Q}$. Zhang [71] claims that the latter two terms (pertaining to $\boldsymbol{Q}$ and $\Delta\boldsymbol{Q}$) should cancel out, yet from Equation 7.3.3 (ignoring the U_{error} contribution), it can be inferred that this is not the case, since $\frac{\partial U_{\text{inter}}}{\partial \boldsymbol{Q}} \neq \frac{\partial U_{\text{inter}}}{\partial \Delta\boldsymbol{Q}}$. Nevertheless, this assumption accomplishes the goal of eliminating $\boldsymbol{\Omega}^*$ from the gradient, and the corresponding FD expression use only single-point energy evaluations at $\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))$:

$$\begin{aligned}
&\frac{\partial}{\partial \theta^{\text{ind}}} U_{\text{latt}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}), \Delta\boldsymbol{Q}(\theta^{\text{ind}}, \boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))) \\
&\quad \approx \frac{U_{\text{latt}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind},+}, \boldsymbol{Q}(\theta^{\text{ind},+}), \Delta\boldsymbol{Q}(\boldsymbol{Q}^{\text{total}}(\theta^{\text{ind}}, \boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))), \boldsymbol{Q}(\theta^{\text{ind},+})))}{2(\delta\theta^{\text{ind}})} \\
&\quad - \frac{U_{\text{latt}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind},-}, \boldsymbol{Q}(\theta^{\text{ind},-}), \Delta\boldsymbol{Q}(\boldsymbol{Q}^{\text{total}}(\theta^{\text{ind}}, \boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))), \boldsymbol{Q}(\theta^{\text{ind},-})))}{2(\delta\theta^{\text{ind}})}.
\end{aligned} \tag{7.4.8}$$

The $\Delta\boldsymbol{Q}$ used in these evaluations are determined from the unperturbed total multipoles and the perturbed static multipoles, according to Equation 7.1.2.

Zhang [71] reported that using this gradient approximation gave reasonable minimization results compared to the rigorous FD gradients, but I find this model to be rather dubious. From a physical standpoint, one could make an argument that Equation 7.4.6 reflects a kind of charge balance around each atomic site. However, this becomes increasingly tenuous for the higher-ranked multipole moments, as they do not adhere to analogous conservation rules as the nett charge. Moreover, by this approximation, there is no correlation between $\Delta\boldsymbol{Q}_I$ and $\Delta\boldsymbol{Q}_J$, where I and J are independent molecules in the asymmetric unit cell. This is at odds with the understanding of how polarization manifests in the cell.

Table 7.6: FD modes for SCERP-embedded CSO–FM. For all modes, $\boldsymbol{\theta}^{\text{dep}}$ and $\boldsymbol{Q}$ are updated by their respective LAMs during the FD computation (as they would be without induction).

Mode	Description	FD Eqn.	$\Delta\boldsymbol{Q}$ Update	U_{latt} Evaluation
1	Invariant $\Delta\boldsymbol{Q}$	7.4.5	Fixed at $\Delta\boldsymbol{Q}(\theta^{\text{ind}}, \boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})))$	Single-point at $\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))$
2	Invariant $\boldsymbol{Q}^{\text{total}}$	7.4.8	$\Delta\boldsymbol{Q} = \boldsymbol{Q}^{\text{total}}(\theta^{\text{ind}}, \boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))) - \boldsymbol{Q}(\theta^{\text{ind},\pm})$	Single-point at $\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))$
3	Invariant $\boldsymbol{\Omega}^*$	7.4.10	SCERP process to find $\Delta\boldsymbol{Q}(\theta^{\text{ind},\pm}, \boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})))$	Single-point at $\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))$
4	No simplifications	7.4.3	Full CSO–RM minimization with embedded SCERP processes, in both $\pm\delta\theta^{\text{ind}}$ steps	

Approximation 3: Invariant $\boldsymbol{\Omega}^*$

Building on Approximation 2, the final case I consider assumes that $\boldsymbol{\Omega}^*$ is (relatively) invariant towards small perturbations. For small $\delta\theta^{\text{ind}}$, this implies:

$$\frac{\partial\boldsymbol{\Omega}^*}{\partial\theta^{\text{ind}}} \to 0, \qquad \text{and} \qquad \frac{\partial\boldsymbol{\Omega}^*}{\partial\boldsymbol{Q}} \to 0. \tag{7.4.9}$$

Applying this to Equation 7.4.2, results in the same simplified gradient expression as Approximation 2 (Equation 7.4.7), but with a distinct FD expression:

$$\begin{aligned}
&\frac{\partial}{\partial\theta^{\text{ind}}} U_{\text{latt}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}), \Delta\boldsymbol{Q}(\theta^{\text{ind}}, \boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})))\\
&\quad\approx \frac{U_{\text{latt}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind},+}, \boldsymbol{Q}(\theta^{\text{ind},+}), \Delta\boldsymbol{Q}(\theta^{\text{ind},+}, \boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))))}{2(\delta\theta^{\text{ind}})}\\
&\quad - \frac{U_{\text{latt}}(\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}})), \theta^{\text{ind},-}, \boldsymbol{Q}(\theta^{\text{ind},-}), \Delta\boldsymbol{Q}(\theta^{\text{ind},-}, \boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))))}{2(\delta\theta^{\text{ind}})}.
\end{aligned} \tag{7.4.10}$$

During a FD perturbation, $\Delta\boldsymbol{Q}$ are re-evaluated by a SCERP process conducted at the unperturbed $\boldsymbol{\Omega}^*(\theta^{\text{ind}}, \boldsymbol{Q}(\theta^{\text{ind}}))$ (since $\boldsymbol{\Omega}^*$ is assumed to be invariant). However, to account for the direct and indirect effects of θ^{ind}, the SCERP process is provided with $\theta^{\text{ind},\pm}$, $\boldsymbol{\theta}^{\text{dep}}(\theta^{\text{ind},\pm})$ and $\boldsymbol{Q}(\theta^{\text{ind},\pm})$, that have been updated by the LAMs. Finally, a single-point U_{latt} evaluation is conducted at the unperturbed $\boldsymbol{\Omega}^*$, but with the updated θ^{ind} (and $\boldsymbol{\theta}^{\text{dep}}$), $\boldsymbol{Q}$, and $\Delta\boldsymbol{Q}$.

While this approximation does not completely eliminate the need to do QM calculations during the FD computations, it should be significantly cheaper than the rigorous FD approach. Crucially, only a single-point energy evaluation is required here, hence there is no compounding of rigid-body minimizations and SCERP processes. Moreover, like the rigorous approach, $\boldsymbol{Q}^{\text{total,pc}}_{I,\text{field},*}$ from the same CSO–FM iteration can be reused here to expedite SCERP process convergence. Following the criticisms of Approximations 1 and 2, I believe that this approach should also be the most physically reasonable amongst all of the approximations.

7.4.3 Numerical Limits and Early-Stopping

The proposed FD gradients (Table 7.6) have been implemented in CSO–FMv1.4.0.2, and their efficacy will be examined in greater detail in Chapter 9. Before ending this chapter, I discuss the numerical limits of the flexible minimization, as it pertains to the E04UFF subroutine.

At a set of decision variable values ($\boldsymbol{x}$), the E04UFF subroutine will request function ($F(\boldsymbol{x})$) and gradient ($\boldsymbol{g}(\boldsymbol{x})$) evaluations. From this, the subroutine decides a search direction ($\boldsymbol{p}$), and attempts a trial step, from $\boldsymbol{x}$ to $\boldsymbol{x} + \alpha\boldsymbol{p}$, where α is the step size. If $F(\boldsymbol{x} + \alpha\boldsymbol{p}) > F(\boldsymbol{x})$, the trial step has not reduced the function value, and is rejected. Trial steps are re-attempted until a decreasing step is found, at which point the subroutine may either accept the step, or attempt additional steps regardless. This repeats until convergence is reached. E04UFF considers a few convergence criteria, but I will focus on its first-order optimality criterion. For unconstrained optimizations (like in CSO–FM), this is given by

$$\|\boldsymbol{g}(\boldsymbol{x})\| \leq \sqrt{\tau_{\text{nag}}}\left[1 + \max\left(1 + |F(\boldsymbol{x})|, \|\boldsymbol{g}(\boldsymbol{x})\|\right)\right], \tag{7.4.11}$$

$$\|\boldsymbol{g}(\boldsymbol{x})\| = \sqrt{\sum_v \left(\frac{\partial F}{\partial x_v}\right)^2}, \tag{7.4.12}$$

where τ_{nag} is a user-specified NAG optimality tolerance. Put into words, the norm of $F(\boldsymbol{x})$'s first-derivative vector must be approaching zero. This is judged in a relative sense, by scaling the optimality tolerance to either the function value, or the gradient vector's norm. In CSO–FM, this scaling (almost) always uses $1 + |F(\boldsymbol{x})|$.

For induction-coupled CSP, converging this problem is non-trivial due to the function precision at which $F(\boldsymbol{x})$ (i.e., U_{inter} and U_{latt} in CSO–RM and FM respectively) is determined. This precision is exactly tied to the value of $\varepsilon_{\text{scrpRM}}$. In principle, $\varepsilon_{\text{scrpRM}}$ could be set to extremely small values, but it is often impractical to do so. The energy differences relevant to CSP are usually in the range of 1 to 0.01 kJ/mol. Converging the CSO–RM iterative protocol to significantly higher precision (i.e., smaller $\varepsilon_{\text{scrpRM}}$) will correspond to a non-negligible increase in computational costs, but with little practical value in return. On this basis, Zhang [71] had recommended using $\varepsilon_{\text{scrpRM}} = 0.001$ kJ/mol. Unfortunately, limiting the function precision in this way has other implications on the flexible minimization.

Firstly, the successful convergence of the E04UFF subroutine is predicated on obtaining accurate gradients. A source of inaccuracy is the gradient approximations proposed, which will be studied further in Chapter 9. In the case of the rigorous FD gradients (Table 7.6, Mode 4), where the FD steps require solving an inner CSO–RM minimization, function precision becomes another important source of error. To illustrate this, consider the case where $\dim(\boldsymbol{x}) = 1$ (i.e., only one flexible degree of freedom). Rearranging Equation 7.4.11 gives

$$\Delta F(x) \leq \frac{\Delta x}{180} \times \sqrt{\tau_{\text{nag}}}(2 + |F(\boldsymbol{x})|), \tag{7.4.13}$$

where $\frac{\Delta x}{180}$ is the dimensionless FD step (for torsions and angles) used by CSO–FM, and $\Delta F(x)$ is the corresponding change in U_{latt} from the step. Applying the default CSO–FM parameter values ($\Delta x = 0.001°$, $\tau_{\text{nag}} = 1 \times 10^{-8}$), and assuming $|F(\boldsymbol{x})| \approx 100$ kJ/mol (the typical order of magnitude for U_{latt} is $\sim\mathcal{O}(10^2)$), $\Delta F(x) \leq 5.67 \times 10^{-8}$ kJ/mol is required for the E04UFF subroutine to converge. At the suggested $\varepsilon_{\text{scrpRM}}$, the gradients calculated in the vicinity of the optimum are subject to numerical noise, and the convergence criterion cannot be reliably evaluated. Beyond assessing convergence, gradients are also used to determine $\boldsymbol{p}$. Inaccuracies in the gradients will lead to ineffective search directions, that further inhibit the optimizer's ability to approach the minimum.

Its effects on the gradients aside, the precision of $F(\boldsymbol{x})$ also affects the success of linesearches. This is intuitive since the E04UFF subroutine accepts a step only if $F(\boldsymbol{x})$ decreases. It follows that if a reduction in $F(\boldsymbol{x})$ is smaller than $\varepsilon_{\text{scrpRM}}$, the subroutine may fail to perceive this, and progress in the minimization is erroneously halted. Again, this would be most acute near the optimum, since both the step size and the gradients are likely to be smallest here.

From this analysis, there are clear numerical limits to reliably solving the flexible minimization. Failing to account for these restrictions usually manifests in the E04UFF subroutine expending more trial steps, without making additional progress. In practice, this has been highly disruptive towards generating reliable results and represents a significant waste of resources. To mitigate this, I have found the combination of several strategies to be effective:

1. For flexible minimizations, I tighten the SCERP convergence tolerances to $\varepsilon_{\text{scrp}} = 1 \times 10^{-5}$ a.u. and $\varepsilon_{\text{scrpRM}} = 1 \times 10^{-5}$ kJ/mol. $\varepsilon_{\text{scrp}} = 1 \times 10^{-5}$ a.u. is the limit of SCERP process convergence, since Gaussian reports atomic-charges to only six decimal places (in the standard output). $\varepsilon_{\text{scrpRM}} = 1 \times 10^{-5}$ kJ/mol was found to enable reasonable progress in the flexible optimization, whilst remaining computationally tractable.
2. The E04UFF subroutine has a function precision (ε_{fp}) parameter. When SCERP is used with CSO–FM, this is set to $\varepsilon_{\text{fp}} = 1 \times 10^{-6}$, in accordance to guidelines from the E04UFF user-manual. The chosen value reflects that $F(\boldsymbol{x})$ is typically of $\mathcal{O}(10^2)$ and is accurate to the third decimal place (i.e., precise to six significant figures).
3. To avoid the optimizer 'aimlessly meandering' about when progress cannot be reliably made, τ_{nag} should be selected based on $\varepsilon_{\text{scrpRM}}$. To that end, I generalize Equation 7.4.13 for N flexible torsions, assuming that the gradient elements in $\boldsymbol{g}(\boldsymbol{x})$ are identical:

$$\tau_{\text{nag,min}} \geq \left[\frac{\Delta F \times \sqrt{N}}{\frac{\Delta x}{180}\,(2 + |F|)} \right]^2 . \tag{7.4.14}$$

   The minimum 'sensible' optimality tolerance ($\tau_{\text{nag,min}}$) is determined by setting ΔF as $\varepsilon_{\text{scrpRM}}$ and Δx as the chosen CSO–FM FD step size (in degrees). τ_{nag} is selected with a margin of error from $\tau_{\text{nag,min}}$.
4. In CSO–FMv1.4.0.2, I have implemented a rudimentary early-stopping mechanism. If 8 *consecutive* trial steps are attempted without any decrease in $F(\boldsymbol{x})$, CSO–FM terminates

the E04UFF minimization. This not only eliminates futile linesearch attempts when function precision is limiting, but also cases where $\boldsymbol{g}(\boldsymbol{x})$ are too incorrect to give useful search directions (due to function precision, or otherwise). This was motivated by my observation that the E04UFF subroutine does not necessarily terminate promptly when it cannot find a decreasing step between major iterations. Beyond this point, further search attempts by the subroutine rarely reduce the function value significantly.

7.5 Summary

In this chapter, I have described the embedding of the SCERP induction model into the MSE CSP suite, comprising three key building blocks. First is the SCERP process, wherein induced multipoles are derived through a self-consistent QM procedure. Second is the integration of the SCERP process into CSO–RM, as an iterative rigid-minimization problem. Finally, the CSO–RM protocols used for U_{inter} minimization are interfaced with CSO–FM, for SCERP-coupled flexible minimization. This implementation builds on Zhang's earlier attempt [71], but crucially redresses several methodological shortcomings from his work. The computational efficiency of SCERP-coupled computations has been bolstered through the judicious reuse of information across the different SCERP protocols. The SCERP process should also converge more reliably in the current implementation with the addition of the deceleration protocol. Finally, I have re-visited the application of the Ewald sum for accurate induction calculations.

Applying this revised SCERP implementation over the CE–755 dataset, I have examined the physical validity of the SCERP induction-model. Comparing the induction energies predicted by the SCERP model against a PCM model (with $\varepsilon_{\text{PCM}} = 11$), reasonable corrections to the latter are suggested by the former. As hoped, the polarizing influence of the crystalline environment appears to be reflected with greater nuance using SCERP, particularly in local environments with directional effects like hydrogen bonds. An important detail affecting the prediction of induction energy using SCERP is the choice of $r_{\text{H}}^{\text{GDMA}}$. For consistency with the rest of this work, and contrary to the recommendation of Welch [183], I have opted to use $r_{\text{H}}^{\text{GDMA}} = 0.650$ Å. My results suggest that this parameter does influence predicted U_{ind} values, but not to an extent that convinces me of a strong need to alter its value. A physical deficiency of the SCERP model remains its reliance on an atomic-charge representation. I have attempted to resolve this through a satellite-charge SCERP model (Appendix K), but practical issues limit the viability of this approach. As will be shown later (with more testing), this may not be the most pressing issue with the current SCERP model.

An overarching goal of this work has been to develop an induction model within the HAIEFF paradigm. This chapter has only addressed half of this problem: the reliable calculation of the electrostatic contributions. In the next chapter, I will use the implemented SCERP model to tackle the remaining half: parametrizing a consistent set of HAIEFF FFs compatible with SCERP. Only with a unified HAIEFF model (i.e., complementary electrostatic + FF), can an accurate appraisal of the SCERP methodology be obtained.

Chapter 8

Induction for CSP: Fitting SCERP-Compatible Force-fields

In Chapter 7, I detailed the theory and implementation of the SCERP induction model in our group's CSP codes. However, in light of the risks highlighted in Section 6.5, the SCERP-coupled HAIEFF model is incomplete without a complementary HAIEFF FF. As admitted by Zhang [71], this is a limitation in his appraisal of the SCERP-coupled HAIEFF model. By resorting to an *ad hoc* choice of FF, he likely introduced double-counting of induction effects, thereby undermining the analysis of SCERP model accuracy. In this chapter, I seek to complete the integration of SCERP into the HAIEFF framework by parametrizing SCERP-compatible HAIEFF FFs. In the process, I also aim to gain insight into the physical model as it pertains to the use of electrostatic damping. In Section 8.1, I first describe how the parameter estimation is affected by the inclusion of SCERP. Following this, the concept of electrostatic damping and the questions surrounding its use are introduced (Section 8.2). Through a series of parameter estimations on small- and medium-sized training sets (Sections 8.3), I study the effects of damping on model accuracy. These findings inform my ultimate choice of damping implementation, and in Section 8.3.4, I parametrize a final set of FFs for deployment. As stated in Section 7.3.1.1, SCERP is used only with rank 4 multipole models and all fitted FFs will adhere to this specification.

8.1 Parameter Estimation with SCERP

To conduct parameter estimation of SCERP-compatible HAIEFF FFs, I again utilize CrystalEstimator [1, 2] in tandem with the CE–755 dataset [1, 256]. A key aspect of CrystalEstimator is the embedding of (intermolecular) energy minimizations as part of the parameter estimation objective function (Figure 4.3, D1–D2). However, from Section 7.3.3, solving rigid-body energy minimizations with SCERP (Equation 7.3.11) is non-trivial as it necessitates iteratively updating $\Delta\boldsymbol{Q}$, until the solution to Equation 7.3.12 converges. As explained below, directly inserting this protocol into CrystalEstimator is unlikely to yield good results.

Without SCERP, CrystalEstimator can leverage the rigid-body approximation and isolated-molecule assumption to complete all QM-related computations (i.e., solving SCF calculations, deriving electrostatic features) prior to the parameter estimation. These pre-computed electrostatic features are read-in as inputs and used during CrystalEstimator's parameter estimation

protocol. This is akin to CSORM_Init/CSORM_Min described in Section 7.3.3. The separation of 'QM calculation' and 'energy minimization' is key to CrystalEstimator's ability to handle many training structures, since energy minimizations are relatively inexpensive and amenable to parallelization on single processors. Without SCERP, MPI parallelization is needed for FF parametrization with multipoles, due to the cost of evaluating the electrostatic contribution [1, 2]. With SCERP, one can expect the cost of a *single* energy minimization to roughly double, since the electrostatic contribution requires two equivalent function evaluations. If this is further compounded by the need to do *multiple* energy minimizations *and* QM calculations on-the-fly, large-scale parameter estimation will no longer be realistic. Additionally, a solution to Equation 7.3.11 obtained via the iterative procedure will have limited function precision. This can have significant upstream effects when integrated into a bilevel optimization, like in CSO–FM (Section 7.4.3) and CrystalEstimator. This is particularly true for CrystalEstimator, since the *analytical* gradients of the objective function are formulated under the strict assumption of optimality in the inner rigid optimization (Section 4.3.2). Failing to tightly approach $\nabla_{\boldsymbol{\Omega}} U_{\text{inter}} = \mathbf{0}$ will hamper CrystalEstimator's ability to solve the parameter estimation.

Altogether, there are a host of problems associated with solving the 'complete' SCERP-embedded rigid minimization as part of the parameter estimation, thus pragmatic concessions must be made. To that end, I formulate the rigid-body minimization in CrystalEstimator as

$$\min_{\boldsymbol{\Omega}} U_{\text{inter}}(\boldsymbol{\Omega}; \boldsymbol{p}, \boldsymbol{\theta}^{\text{ref}}, \boldsymbol{Q}(\boldsymbol{\theta}^{\text{ref}}), \Delta\boldsymbol{Q}(\boldsymbol{\theta}^{\text{ref}}, \boldsymbol{\Omega}^{\text{ref}})), \tag{LM$_{\text{SCERP}}$}$$

where $\Delta\boldsymbol{Q}(\boldsymbol{\theta}^{\text{ref}}, \boldsymbol{\Omega}^{\text{ref}})$ are the induced multipoles at the reference conformation and lattice. This is the parameter estimation analogue to Equation 7.3.12, and CrystalEstimator need only solve one iteration of the rigid minimization, with $\Delta\boldsymbol{Q}$ fixed at their reference values (i.e., on-the-fly QM calculations are avoided). The separation of 'QM calculation' and 'energy minimization' is preserved this way, and no changes need to be made to the mathematical form (Section 4.3.2) or algorithmic approach (Section 4.3.3) of the parameter estimation.

Unfortunately, this does create an inherent disconnect between the FF's utilization in the parameter estimation and its deployment in real rigid minimizations. By the proposed approach, the FF is parametrized with $\Delta\boldsymbol{Q}^{\text{mp}}_{I,k=-1}$, and 'extrapolated' over subsequent $\Delta\boldsymbol{Q}^{\text{mp}}_{I,k}$ encountered during the minimization (Figure 7.11). One can expect that this will lead to some deterioration in model accuracy between CrystalEstimator and CSO–RM, and an analysis of FF performance must account for this. In the rest of this chapter, error metrics determined from CrystalEstimator and CSO–RM will be distinguished by 'CE' and 'CSORM' respectively.

8.2 Damping for Electrostatic Interactions

8.2.1 Background

At medium- and long-range interatomic separations, the classical description of intermolecular interactions by r^{-n} multipole expansions is theoretically sound. However, at short-range,

where charge overlap occurs, quantum effects such as charge penetration and exchange repulsion arise, undermining the multipole expansion's validity [221]. To approximate these effects in the classical model, a popular approach has been the use of empirical damping functions, like the Tang–Toennies damping function [342] that was applied to dispersion interactions in Chapter 6. Similarly, several authors [71, 165, 183, 189, 200, 202, 223, 224] have proposed that damping should also be applied to the electrostatic/induction interactions.

While the physical motivation for using a damping function is widely accepted, there is very little consensus on how the damping function should be implemented. For instance, Zhang [71] imposed damping on both the static and induced electrostatic contributions, while other authors discuss damping in the context of only the induction contribution [165, 183, 189, 223, 224]. This emphasis on induction damping likely stems from the iterative protocol used in its calculation [200]. Stabilization from induction encourages closer contacts, which in turn promotes more polarization. Without damping, this positive feedback increases the likelihood of entering a divergent 'spiral'. However, there is also no theoretical justification *not* to damp the static electrostatic contribution. As noted by Misquitta & Stone [202], all of the r^{-n} multipole expansions (static electrostatics, polarization, and dispersions) are subject to the same short-range effects and should be damped accordingly. Opting not to damp the static electrostatic term seems largely motivated by convenience [200, 202]. The multipole interaction ranks that require damping are also unclear. Some authors choose to selectively damp terms for $l_1 + l_2 \geq 2$ [71, 183, 250], which again lacks any rigorous theoretical rationale. Totten *et al.* [371] have suggested that since low-ranked multipole interactions diverge less quickly, damping them is unlikely to have any meaningful influence, except at unphysical interatomic distances.

In summary, there is a good theoretical basis to embed electrostatic damping, particularly where induction is computed iteratively. Yet, due to the empirical nature of the damping function, there is considerable liberty in how the function is implemented. This is particularly true here, since my pursuit of a complementary HAIEFF FF will likely absorb any minor errors in the damping function implementation.

8.2.2 Implementation of Electrostatic Damping

In this work, I adopt the damping approach of Zhang [71], which uses the Tang–Toennies damping function [342]. This function is a popular choice [71, 165, 183, 189, 200, 202, 223, 224], and crucially was also used in the earlier FF parametrization work (Chapter 6, Equation 6.1.2). For convenience, I rewrite a generalized form of the damping function here,

$$f_n(\beta, r_{\mathbf{0}Ii,\mathbf{l}Jj}) = 1 - \exp\left(-\beta \cdot r_{\mathbf{0}Ii,\mathbf{l}Jj}\right) \sum_{k=0}^{n} \left[\frac{(\beta \cdot r_{\mathbf{0}Ii,\mathbf{l}Jj})^k}{k!} \right], \tag{8.2.1}$$

where n is the damping order, and β is the damping parameter. For the electrostatic interactions, n is set to $l_1 + l_2 + 2$, such that rank 0 interactions are damped by f_2, while rank 4 multipole interactions are damped by f_6 [71]. The progression in damping behavior is illustrated in Figure 8.1 and supports Totten *et al.*'s [371] claim that damping is less aggressive (i.e., occurs at shorter

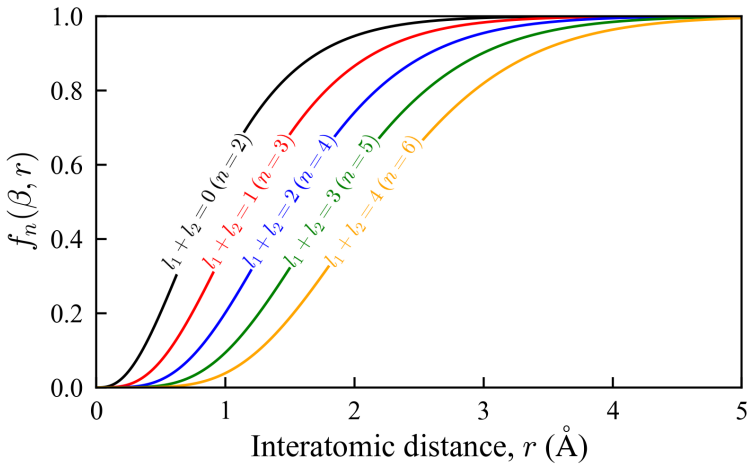


Figure 8.1: The Tang–Toennies damping function, $f_n(\beta = 1.6400\,\text{bohr}^{-1}, r)$, for $n = 2$ to 6.

range) for lower ranked interactions. As such, only multipole interactions with $l_1 + l_2 \geq 2$ are damped [71], i.e., charge–charge and charge–dipole interactions are undamped. Following Refs [71, 183], damping is imposed on only the real-space contribution of the Ewald sum,

$$U^{\text{real,damp}}_{\mathbf{0}Ii,\boldsymbol{l}Jj} = \frac{1}{4\pi\varepsilon_0} \sum_{l_1,l_2} \frac{\Gamma(\frac{1}{2}+l_1+l_2, \alpha_{l_1+l_2}\cdot\pi\cdot(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2)}{\Gamma(\frac{1}{2}+l_1+l_2)} \cdot f_{l_1+l_2+2}(\beta, r_{\mathbf{0}Ii,\boldsymbol{l}Jj}) \cdot \sum_{m_1,m_2} Q^{l_1m_1}_{Ii} Q^{l_2m_2}_{Jj} \frac{\widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^{l_1+l_2+1}}. \tag{8.2.2}$$

In principle, a re-derivation of the Ewald sum is required in order to correctly incorporate the effects of damping into the reciprocal and self-correction terms. However, as noted by Welch [183], this is non-trivial and simply damping the real-space term is likely adequate. From Equation 8.2.2, first- and second-derivatives of $U^{\text{real,damp}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ with respect to $r_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ are trivial to obtain via the chain-rule. The necessary first- and second-derivatives of $U^{\text{real}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ and f_n with respect to $r_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ have been derived in Chapter 3 and Appendix H respectively.

Finally, in the CSO–RM SCERP implementation, $U_{\text{elec+ind}}$ is directly computed through two evaluations of the Ewald sum. It would be cumbersome to decouple U_{elec} and U_{ind} for the purpose of damping one, but not the other. As such, I choose to impose damping on the entire $U_{\text{elec+ind}}$ contribution. As explained before, this is fundamentally sound [202].

8.2.3 The Damping Factor

The preceding decisions related to electrostatic damping are inherently 'hand-wavy', and pursuing a rigorous theoretical justification for them is unlikely to be fruitful. However, the damping factor (β) does have some physical significance, since it modulates the aggressiveness of the damping function. In view of this, I believe more care should be taken in its specification.

Previous authors [71, 165, 183, 189, 202, 250] have proposed the use of a molecule-pair specific damping parameter, $\beta^{\text{AB}}_{\text{mol}}$, determined according to,

$$\beta_{\mathrm{mol}}^{\mathrm{AB}} = \sqrt{2I_{\mathrm{mol}}^{\mathrm{A}}} + \sqrt{2I_{\mathrm{mol}}^{\mathrm{B}}}, \tag{8.2.3}$$

where $I_{\mathrm{mol}}^{\mathrm{X}}$ is the first vertical ionization potential (in Hartrees) of molecule X (X = A or B). While the form of this expression has some physical basis (see derivation by Misquitta *et al.* [224]), from simple chemical intuition, it is clear that the use of $\beta_{\mathrm{mol}}^{\mathrm{AB}}$ is questionable. As described in Section 8.2.1, the physical origin of damping is the charge overlap between atomic sites at short-range. Recognizing that the atomic size of elements are distinct, it can be reasoned that the onset (and aggressiveness) of damping should be dependent on the interacting *atom-pairs*. From this perspective, $\beta_{\mathrm{mol}}^{\mathrm{AB}}$ at best provides a molecular-average representation of charge overlap, but ultimately an atom-pair dependent damping factor, $\beta_{\mathrm{atm}}^{tt'}$, is more physically sound. Authors who used $\beta_{\mathrm{mol}}^{\mathrm{AB}}$ do acknowledge this model deficiency [183, 201, 224, 250], but due to various constraints, the development of $\beta_{\mathrm{atm}}^{tt'}$ has not been studied. To that end, an aim of this chapter is to determine a suitable choice of β, potentially in the form of $\beta_{\mathrm{atm}}^{tt'}$.

Unfortunately there is no straightforward (first-principles based) methodology to derive values for $\beta_{\mathrm{atm}}^{tt'}$. Nevertheless, I put forward that simply using a chemically intuitive set of $\beta_{\mathrm{atm}}^{tt'}$ is preferable to using $\beta_{\mathrm{mol}}^{\mathrm{AB}}$. This is especially so since the HAIEFF FF can absorb small errors in the electrostatic model, such as those arising from imprecise values of $\beta_{\mathrm{atm}}^{tt'}$. To obtain these 'intuitive' values, one avenue is to return to the HAIEFF FF, and correlate $\beta_{\mathrm{atm}}^{tt'}$ to the Buckingham potential's $B^{\mathrm{tt'}}$ parameter. This was done during the earlier FF fitting (Section 6.1.2), with credible results from the damped Buckingham potential. However, tying the values of $\beta_{\mathrm{atm}}^{tt'}$ to the HAIEFF FF is undesirable if one ever wishes to depart from the BHP/dBHP (e.g., using the Mie potential instead). In lieu of this, I propose that intuitive values of $\beta_{\mathrm{atm}}^{tt'}$ (that are agnostic to the HAIEFF FF form) can be obtained as the solution to

$$\%f_6 = 1 - \exp\left(-\beta_{\mathrm{atm}}^{tt'} \cdot r_{\mathrm{Bondi}}^{tt'}\right) \sum_{k=0}^{6} \left[\frac{(\beta_{\mathrm{atm}}^{tt'} \cdot r_{\mathrm{Bondi}}^{tt'})^k}{k!} \right]. \tag{8.2.4}$$

$\%f_6$ is a target damping factor value (in percent) for the damping function evaluated with $n = 6$.

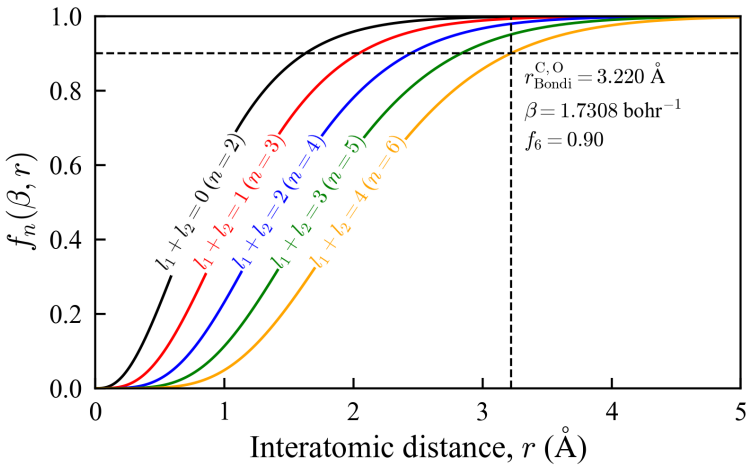


Figure 8.2: The Tang–Toennies damping function, $f_n(\beta = 1.7308\ \mathrm{bohr}^{-1}, r)$, for $n = 2$ to 6.

This target should be achieved at the sum of Bondi radii [3] for atom types t and t' ($r^{tt'}_{\text{Bondi}}$). For instance, between a carbon and and oxygen with Bondi atomic radii 1.70 Å and 1.52 Å respectively, $r^{\text{C,O}}_{\text{Bondi}}$ = 3.22 Å. Targeting $\%f_6$ = 90%, the value of $\beta^{\text{C,O}}_{\text{atm}}$ can be determined as 1.7308 bohr^{-1} (Figure 8.2). The reasoning for this formulation is as follows. The Bondi atomic radii provide approximate values for the size of the different elements/atom types, and charge overlap between two atom types should occur around $r^{tt'}_{\text{Bondi}}$. Given that the damping function is most aggressive for $n = 6$, it is sensible to derive the damping parameter value on the basis of f_6, as this reflects that the earliest onset of damping coincides with $r^{tt'}_{\text{Bondi}}$.

With this, it is trivial to determine $\beta^{tt'}_{\text{atm}}$ using the atom type definitions employed in the HAIEFF FF. The only question that remains is: "*what is an appropriate value of* $\%f_6$*?*" This will be investigated as part of the FF parametrization.

8.3 Force-field Parameter Estimation

8.3.1 General Methodology

To parametrize SCERP-compatible HAIEFF FFs, I adopt a similar parameter estimation methodology to Chapter 6. The salient details of the approach are summarized below.

Firstly, all QM calculations are performed at the PBE0/6-311G(d,p) level of theory, and multipoles up to rank 4 are derived from the SCF charge densities. Unless otherwise stated, PCM is not used in any SCF calculations. Different subsets of the CE–755 are used to generate parameter estimation training sets of varied size. For all parameter estimations, I employ the same parameter bounds and multistart settings described in Section 6.1.4, and adopt the same termination strategy as shown in Figure 6.1. However, depending on the training problem size, a different number of Sobol' points are employed (Table 8.1). Here, I opt to use only fitting Schemes 5 and 7 (Table 6.2), since these provided the most successful FFs in Chapter 6. The dBHP FF (Scheme 5) is used in all parameter fitting attempts, but the MIP FF (Scheme 7)

Table 8.1: Training sets for SCERP-compatible FF fitting. The number of training structures (N_{TRN}), atom types present, number of fitted parameters, and number of Sobol' seeds are shown. Values for Schemes 5 and 7 (only in the final fitting) are given before and after '/', respectively.

Alias	N_{TRN}	Atom types present	dim($\bar{\boldsymbol{p}}$)	N_{Sobol}
NHB47[§]	47	C, H, N	12/-	32/-
HB53[§]	53	C, H, O, N, H_O, H_N	42/-	64/-
TP100[§]	100	C, H, O, N, H_O, H_N	42/-	64/-
TP442[§]	442	C, H, O, N, F, S, Cl, H_O, H_N	88[*]/-	128/-
Final[‡]	604	C, H, O, N, F, S, Cl, H_O, H_N	90/135	256/1[†]

[§] The training set structures are tabulated in Appendix M.1.

[*] The S/Cl interaction is absent from this training set.

[‡] The full CE–755 training/test set is used, as tabulated in Appendix G.

[†] Due to resource constraints, only one Sobol' point is attempted for the Scheme 7 parametrization.

Table 8.2: Summary of CrystalEstimator (CE) program settings used in SCERP-compatible FF parametrization.

CE Lattice Energy Minimization (Inner Loop) Settings	Value
NAG E04UFF Optimality Tolerance	1×10^{-12}
Newton Refinement Optimality Tolerance	1×10^{-20}
Newton Refinement Hessian Update Method	BFGS
Newton Refinement Iteration Limit	12 or 15
CE Parameter Estimation Minimization (Outer Loop) Settings	**Value**
NAG E04USA Optimality Tolerance	1×10^{-8}
Maximum Structure Removal	20
Dynamic Starting Points	Active
Permitted Number of Reinitializations	15
SCERP Process Settings	**Value**
SCERP-field Representation	HLYGAt-charges
SCERP Embedding Radius, r_{scrp}	40.0 Å
SCERP Process Convergence Tolerance, $\varepsilon_{\text{scrp}}$	1×10^{-3} a.u.
SCERP Process Slowed Iterations, t_{slow}	6
Maximum SCERP Process iterations, t_{max}	15

is used with only the full dataset (Section 8.3.4). Finally, the CrystalEstimator optimization settings and the SCERP model settings are tabulated in Table 8.2.

Electrostatic damping is applied only in models that use SCERP-embedded electrostatics. To evaluate $\beta^{\text{AB}}_{\text{mol}}$, $I^{\text{X}}_{\text{mol}}$ are determined via two QM calculations. Firstly, molecule X is subject to a geometry optimization with the molecule's charge set to zero (the CE–755 comprises only neutral molecules), which gives the ground-state conformation and electronic energy for X *in vacuo*. Thereafter, a *single-point* SCF calculation is initialized from the ground-state conformation with X's charge set to +1. The first vertical ionization potential is obtained by deducting the neutral molecule's energy from its cation's energy. From $I^{\text{X}}_{\text{mol}}$, $\beta^{\text{AB}}_{\text{mol}}$ are computed by Equation 8.2.3. To derive $\beta^{tt'}_{\text{atm}}$, Equation 8.2.4 is solved for different pairs of t and t'. For simplicity, I

Table 8.3: Bondi atomic radii [3] used for each atom type, t.

Atom Type (t)	Bondi Radius (Å)
C	1.700
H	1.200
O	1.520
N	1.550
F	1.470
Cl	1.750
S	1.800
H_{O}	1.000
H_{N}	1.000

have defined t and t' according to the atom types used in the HAIEFF FF. Bondi atomic radii for each atom type are tabulated in Table 8.3. For atom types defined solely by their element, this corresponds to the radii reported by Bondi [3]. However, for oxygen- and nitrogen-bonded hydrogens (H_O and H_N respectively), I use a slightly smaller radius of 1.000 Å. This reflects that these polarized atoms are stripped of electron density [169] and thus should have a smaller radius than 'normal' H. $\beta_{\text{atm}}^{tt'}$ have been derived for $\%f_6$ = 70%, 80%, and 90% (Appendix L).

All parameter estimation is conducted using CrystalEstimatorv1.5.2.1, which interfaces with CSO–RMv1.5.2.1. All QM calculations are conducted in Gaussian (G09-d01), and rank 4 multipoles are determined using GDMAv2.2 ($r_{\text{H}}^{\text{GDMA}}$ = 0.650 Å).

8.3.2 Fitting with Small Training Sets

The two smallest training sets are constructed based on the presence of H_O and/or H_N atoms in their member molecules. NHB47 consists of 47 non-hydrogen bonding (NHB) crystals structures, which have no $H_{O/N}$ atoms present. Conversely, HB53 comprises 53 structures containing H_O and/or H_N atoms. By the geometric criteria described in Section 5.2.2, $H_{O/N}$-containing crystals do not necessarily form intermolecular hydrogen bonds; this is the case for two structures in HB53 (CSD Refcodes: HABVIQ and JOSLOU). For simplicity, I refer to the entire HB53 set as consisting of hydrogen bonding (HB) structures. At this stage, my goal is to understand how model performance is influenced by the electrostatic/induction model, and to that end I conduct parametrization for a range of models (Table 8.4). For each energy model (and for each training set), independent parameter estimations are performed to obtain optimal FF parameter estimates complementary to the corresponding model (and training set).

In Figure 8.3, I present the CrystalEstimator residual scores for the different models, at their respective Minimum Sobols. Readers are reminded that the reported residual scores are not normalized by the number of crystal structures in each set. Firstly, it is clear that a complete neglect of induction (i.e., Static) consistently leads to the worst predictions of crystal energies, for both NHB and HB crystals. The improvement in energy accuracy from introducing induction, either through PCM or SCERP, is more drastic for the HB systems than NHB systems: energy residuals decrease by ~0.026×10^{-2} units per structure and ~0.016×10^{-2} units per structure respectively. Generally, the injection of induction also promotes better geometry predictions.

Table 8.4: Scheme 5 FF parameter estimations attempted in the trial FF fitting problems.

Model	Induction Model	Electrostatic Damping
Static	None	None
PCM_{11}	PCM with ε_{PCM} = 11	None
SCRP	SCERP	None
$\text{SCRP}_{\text{mol.}}$	SCERP	$\beta_{\text{mol}}^{\text{AB}}$
$\text{SCRP}_{\text{atm,70}}$	SCERP	$\beta_{\text{atm}}^{\text{tt}'}$, with $\%f_6$ = 70%
$\text{SCRP}_{\text{atm,80}}$	SCERP	$\beta_{\text{atm}}^{\text{tt}'}$, with $\%f_6$ = 80%
$\text{SCRP}_{\text{atm,90}}$	SCERP	$\beta_{\text{atm}}^{\text{tt}'}$, with $\%f_6$ = 90%

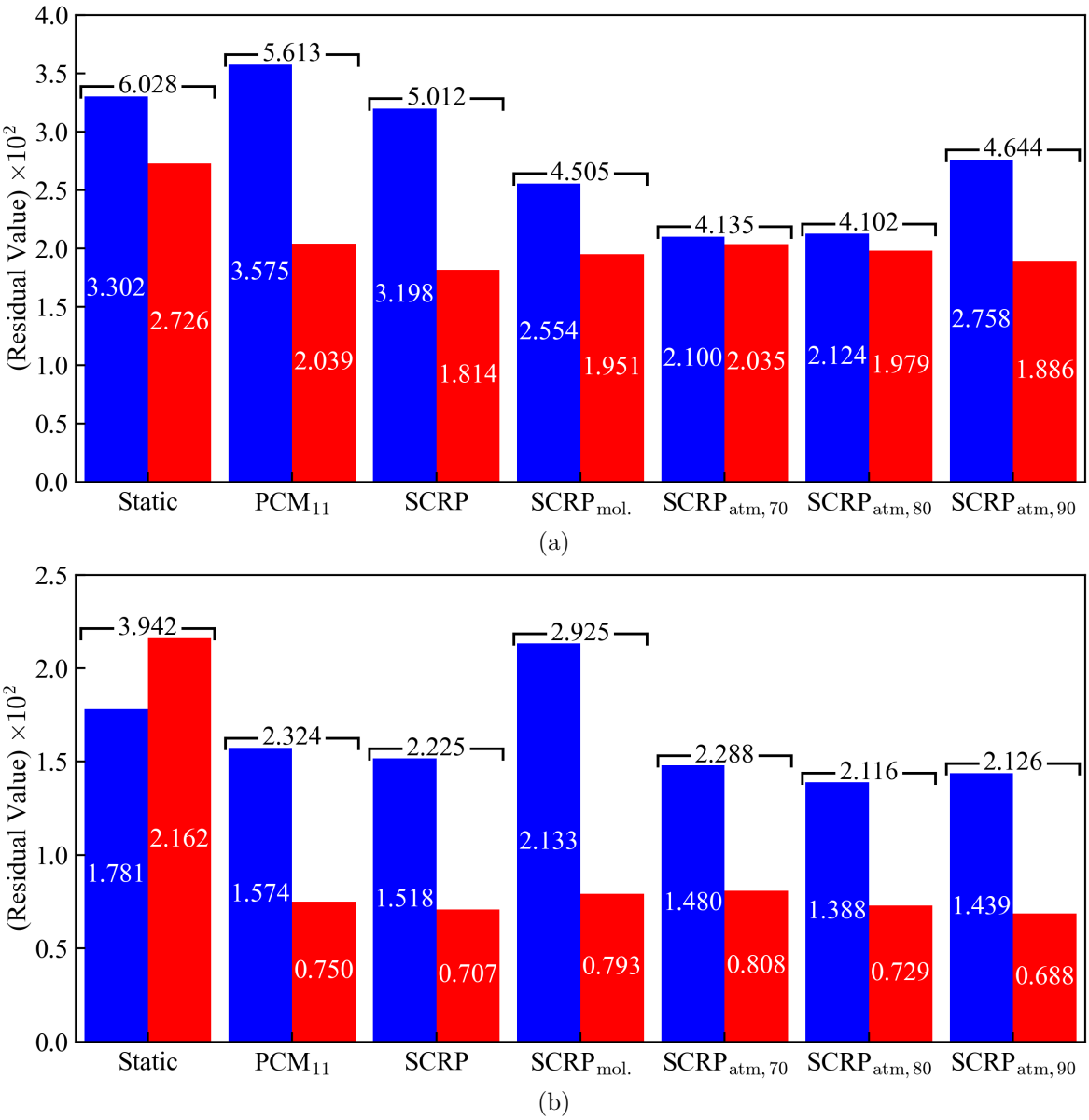


Figure 8.3: Sum of geometry (blue) and energy (red) residual scores for the (a) NHB47, and (b) HB53 training sets. The total objective function value is summed above each pair.

The exceptions to this are: (1) Static → PCM_{11} in NHB47, and (2) Static → $SCRP_{mol.}$ in HB53. In these cases, the deterioration in geometry accuracy is modest relative to the improvement in energy accuracy. As such, model accuracy as a whole (i.e., total objective function value) is *always* bolstered when induction is added to the electrostatic model. This behavior is consistent with Figure 6.2b and supports the notion that induction is important in the modeling of crystal energies [189], particularly for HB structures.

Comparing the induction-embedded models (PCM_{11} versus SCERP variants), there is ambiguity as to whether SCERP's explicit representation of polarization results in superior modeling accuracy. For the NHB47 and HB53 set, the average objective function values for the SCERP variants are 4.480×10^{-2} and 2.336×10^{-2} units respectively. It should be noted that the $SCRP_{mol.}$ model significantly elevates the average in HB53, and excluding it reduces the average to only

2.189×10^{-2} units. From this perspective, the accuracy of the SCERP variants seem to be on par with (or even superior) to that of the PCM_{11} model. However, it is important to recognize that the improved performance of the SCERP models is largely driven by the geometry predictions. This is most apparent from the NHB47 results, wherein the energy residuals fluctuate around a similar value across all of the induction-embedded models ($\sim 1.9 \times 10^{-2}$ units), while the geometry residuals exhibit a marked decrease from the PCM_{11} model to the SCERP models. For the HB53 set, the manifestation of this is more subtle, since the difference in total objective function value (from PCM_{11} to the SCERP models) is less drastic.

Finally, focusing on the SCERP-embedded models, my interests lie in the effects of electrostatic damping. According to Misquitta *et al.* [224], $\beta_{\text{mol}}^{\text{AB}}$ are typically between 1.7–1.9 bohr^{-1} for most organic molecules. From my calculations, I obtain an average $\beta_{\text{mol}}^{\text{AB}}$ ~1.6 bohr^{-1} over the entire CE–755. This lower value could stem from the presence of fluorine, sulphur, and chlorine atoms in the CE–755, whereas the preceding indicative range pertains primarily to carbon-, hydrogen-, nitrogen-, and oxygen-containing molecules [224]. The period three elements should be more diffuse than their period two counterparts and thus experience charge overlap at a longer range. This would manifest in smaller values of $\beta_{\text{mol}}^{\text{AB}}$. In comparison, the average $\beta_{\text{atm}}^{tt'}$ are 1.5, 1.7, and 2.0 bohr^{-1} for $\%f_6$ = 70%, 80%, and 90%, respectively (Table L.1). Since smaller values of β correspond with more aggressive (earlier onset) of damping, I roughly rank the 'strength' of electrostatic damping according to (in descending order):

$$\text{SCERP}_{\text{atm,70}} > \text{SCERP}_{\text{mol.}} > \text{SCERP}_{\text{atm,80}} > \text{SCERP}_{\text{atm,90}} > \text{SCERP}.$$

On this basis, the energy residuals in Figure 8.3 suggest that energy accuracy is inversely related to damping strength, i.e., more aggressive damping leads to poorer energy predictions. Geometry accuracy does not exhibit any obvious correlation with damping strength.

As explained earlier, the model performance reported by CrystalEstimator reflects the accuracy after only the $k = -1$ iteration of a CSO–RM minimization with SCERP. To assess the 'true' performance of the SCERP-embedded models, the fitted FFs are utilized in 'full' rigid minimizations of the training structures by CSO–RM (per Figure 7.11). These minimizations employ congruent FF/electrostatic models and each minimization is initialized at the corresponding reference structure from the CE–755 ($\mathbf{\Omega}_s^{\text{ref}}$). The error metrics post-CE and post-CSORM are compared in Table 8.5.

For the SCERP-embedded models, the current approach to FF parametrization compels a degree of *ad hoc* deployment during the CSO–RM minimization. Consequently, it is unsurprising that all of the error metrics worsen slightly post-CSORM, relative to post-CE. For most structures, geometry accuracy is well-preserved between CE and CSORM. Across the NHB47 and HB53, only two outlier structures (CSD Refcodes: BENZEN06 and OCAPUE) exhibit a significant change in RMSD_{15}, defined as $|\Delta(\text{RMSD}_{15})_{\text{CE}\rightarrow\text{CSORM}}| \geq 0.2$ Å. Conversely, predictions of crystal energy exhibit greater drift following the iterative rigid minimization protocol, especially for the HB crystals. Large changes in energy error, defined as $|\Delta(dU_{\text{inter}})_{\text{CE}\rightarrow\text{CSORM}}| \geq 1.0$ kJ/mol, are observed in many HB53 structures, while for the NHB47 structures the frequency of

Table 8.5: Error metrics for the models trained on NHB47 and HB53. The metrics are reported for the minimized structures post-CE, and (for SCERP-embedded models) post-CSORM. The latter values are given after the '/'. Root-mean-square-deviations (between CE and CSORM) in $RMSD_{15}$ and dU_{inter} scores are also reported for the SCERP-embedded models. The numbers of structures with $|\Delta(RMSD_{15})_{CE\rightarrow CSORM}| \geq 0.2$ Å and $|\Delta(dU_{inter})_{CE\rightarrow CSORM}| \geq 1$ kJ/mol are given in parenthesis. $dU_{inter} = U_{inter}^{calc} - U_{inter}^{ref}$, such that $dU_{inter} > 0$ indicates underbinding.

Model	Avg. $RMSD_{15}$ (Å)	AD (kJ/mol)	AAD (kJ/mol)	SD (kJ/mol)	$RMSD_{RMSD_{15}}$ (Å)	$RMSD_{dU_{inter}}$ (kJ/mol)
NHB47						
Static	0.2059	0.4106	2.9394	4.2059	-	-
PCM_{11}	0.2071	−0.0148	2.8328	3.7990	-	-
SCRP	0.2155/0.2320	−0.3496/−0.1411	2.4073/2.4854	3.2231/3.2583	0.0539 (1)	0.3530 (1)
$SCRP_{mol.}$	0.1904/0.2070	−0.2707/−0.1123	2.5392/2.6076	3.4147/3.4815	0.0662 (1)	0.3341 (1)
$SCRP_{atm,70}$	0.1792/0.2009	−0.0132/0.0255	2.5591/2.6154	3.5518/3.6460	0.0950 (1)	0.3344 (3)
$SCRP_{atm,80}$	0.1783/0.1966	0.0050/0.0170	2.5854/2.6581	3.5426/3.6454	0.0536 (1)	0.3397 (3)
$SCRP_{atm,90}$	0.2007/0.2185	−0.3200/−0.1583	2.4952/2.5765	3.3116/3.3647	0.0664 (1)	0.3231 (1)
HB53						
Static	0.1801	−0.0193	4.5214	5.6190	-	-
PCM_{11}	0.1597	−0.1817	2.6298	3.5643	-	-
SCRP	0.1541/0.1917	−0.0682/0.8467	2.4440/2.9349	3.1141/3.6295	0.2018 (1)	1.8288 (33)
$SCRP_{mol.}$	0.1776/0.2155	−0.0716/−0.3527	2.7819/3.1883	3.4104/3.9122	0.2003 (1)	1.5793 (19)
$SCRP_{atm,70}$	0.1657/0.2030	−0.1236/−0.5630	2.7678/3.2294	3.4634/3.9778	0.2025 (1)	1.6817 (20)
$SCRP_{atm,80}$	0.1579/0.1949	−0.0901/−0.2854	2.6065/2.9622	3.2370/3.6970	0.2035 (1)	1.4809 (17)
$SCRP_{atm,90}$[†]	0.1540/0.1640[*]	−0.5307/−0.5074	2.5581/2.8238	3.1197/3.5592	0.0132[*] (0)	1.5363 (15)

[†] OCAPUE fused during minimization with CSO–RM and has been removed from the computation of all statistics (post-CE and post-CSORM).

[*] OCAPUE exhibits the largest increase in $RMSD_{15}$ in all of the SCERP models ($\Delta(RMSD_{15})_{CE\rightarrow CSORM} \approx +1.5$ Å), hence its omission affects the post-CSORM $RMSD_{15}$ statistics significantly.

this is mostly negligible. This aligns with results from Section 6.5, wherein geometry accuracy is less sensitive to *ad hoc* FF deployment than energy accuracy, and energies for HB systems are most susceptible to change. With that said, the SCERP models exhibit varying degrees of performance drift, with those utilizing electrostatic damping showing a marked reduction. This is reflected in the root-mean-square-deviation (RMSD) of the error metrics, given by

$$\mathrm{RMSD_X} = \sqrt{\frac{\sum_{s=1}^{NS}(\mathrm{X_{CE},}_s - \mathrm{X_{CSORM},}_s)^2}{NS}}, \qquad \mathrm{X} = \mathrm{RMSD}_{15}, dU_{\mathrm{inter}}, \tag{8.3.1}$$

where s is the structure index and NS is total number of structures. Across the board, the damped SCERP models have smaller $\mathrm{RMSD}_{dU_{\mathrm{inter}}}$, than the undamped model. This is commensurate with the damped models having fewer structures with large dU_{inter} drift. The mitigation of performance drift from damping is intuitive: electrostatic damping quenches the driving force for change by reducing the magnitude of the electrostatic/induction interactions.

Accounting for performance drift, all of the SCERP models still outperform the Static case in both training sets, at least in energy predictions. However, among the induction-embedded models, there is closer parity in energy accuracy (*cf.* Table 8.5: AAD and SD). In the NHB47 set, the SCERP models outperform PCM_{11} post-CE and, due to their minimal drift, continue to do so post-CSORM, albeit to a lesser extent. In contrast, in the HB53 set, the large drift in energy accuracy post-CSORM erodes the advantage the SCERP models may have had over the PCM_{11} model (post-CE). The conflicting relationship between damping strength, energy accuracy, and performance drift leads to the $\mathrm{SCRP}_{\mathrm{atm},90}$ model yielding the best energy accuracy among the SCERP variants, although it is (at best) only marginally better than the PCM_{11} model.

From these small training sets, it is clear that induction-embedded models are almost ubiquitously superior to induction-free models. However, despite the evidence from Chapter 7, SCERP does not seem to necessarily spur greater model accuracy over using just PCM. This is especially so given that the parameter estimation is conducted with a restricted SCERP-embedded minimization, resulting in non-negligible performance drift during real FF deployment.

8.3.3 Fitting with Medium Training Sets

The trials done with NHB47 and HB53 have two flaws that limit their appraisal of SCERP. First, each set reflects only a narrow range of chemical diversity, and in fact a third of the target atom types (F, S, and Cl) are absent. Second, a critical trait of the SCERP model is its ability to provide tailored descriptions of the polarizing environment in different chemical systems (and packing arrangements), akin to a PCM model with adaptive $\varepsilon_{\mathrm{PCM}}$. For a fixed $\varepsilon_{\mathrm{PCM}}$, the model error (relative to SCERP) in NHB and HB systems tends to be in opposition (*cf.* Figure 7.7). By segregating NHB and HB systems into distinct training sets, the fitted FF can more easily correct for PCM model errors in 'one direction', narrowing the gap in performance between PCM_{11} and the SCERP variants. In light of this, I progress to moderate-sized training sets, which cover a wider range of chemical diversity and incorporate a mix of NHB and HB structures. The first training set I consider is TP100, which combines NHB47 and HB53 into a single group.

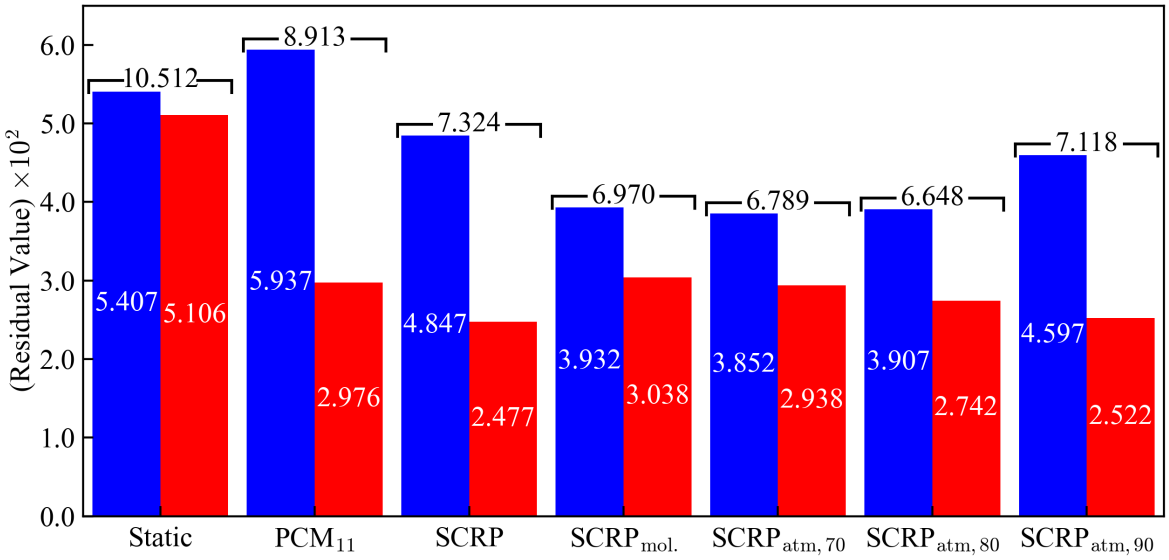


Figure 8.4: Sum of geometry (blue) and energy (red) residual scores for the TP100 training set. The total objective function value is summed above each pair.

This unified set comprises 49 NHB structures and 51 HB structures (by the geometric criteria) since two structures in HB53 do not contain intermolecular hydrogen bonds. The same seven electrostatic models (Table 8.4) are utilized in the fitting of HAIEFF FFs by Scheme 5.

The associated CrystalEstimator training residual values for TP100 are presented in Figure 8.4. Despite having an even split of NHB and HB structures, the TP100 residual scores strongly resemble the NHB47 set, and most of the previous trends resurface here. Firstly, the Static electrostatic model is dominated by (almost) all of the induction-embedded models, in both geometry and energy accuracy. Among the induction-embedded models, the SCERP variants again exhibit superior objective function values, versus PCM_{11}, but this is predominantly geometry-driven. Lastly, between the different SCERP variants energy accuracy is inversely proportional to the aggressiveness of the electrostatic damping. However, geometry accuracy appears to favor the strongly damped models, such that there is a trough in objective function values among $SCRP_{mol.}$, $SCRP_{atm,70}$, and $SCRP_{atm,80}$.

The model performance post-CE and post-CSORM is tabulated in Table 8.6, from which a few things are noted. Firstly, despite the seemingly large variation in geometry accuracy suggested by the residual scores, in practical terms all of the models are similarly competent at predicting crystal packing, with average $RMSD_{15}$ scores < 0.22 Å. Among the SCERP models, geometry predictions again prove to be relatively resistant to drift and the excellent average $RMSD_{15}$ scores are maintained even following the full minimization with CSO–RM. The residual scores may be suggesting larger differences in geometry accuracy between the models since these scores explicitly consider all atomic positions (and lattice parameters), whereas the $RMSD_{15}$ scores ignore hydrogen positions, as is the default in COMPACK. Nevertheless, the $RMSD_{15}$ score is likely a more meaningful (and less abstract) metric for assessing geometry accuracy, since it considers the crystal packing beyond just the central unit cell. Overall, the average $RMSD_{15}$

Table 8.6: Error metrics for the models trained on TP100. The metrics are reported for the minimized structures post-CE, and (for SCERP-embedded models) post-CSORM. The latter values are given after the '/'. Root-mean-square-deviations (between CE and CSORM) in $RMSD_{15}$ and dU_{inter} scores are also reported for the SCERP-embedded models. The numbers of structures with $|\Delta(RMSD_{15})_{CE\rightarrow CSORM}| \geq 0.2$ Å and $|\Delta(dU_{inter})_{CE\rightarrow CSORM}| \geq 1$ kJ/mol are given in parenthesis. $dU_{inter} = U^{calc}_{inter} - U^{ref}_{inter}$, such that $dU_{inter} > 0$ indicates underbinding.

Model	Avg. $RMSD_{15}$ (Å)	AD (kJ/mol)	AAD (kJ/mol)	SD (kJ/mol)	$RMSD_{RMSD_{15}}$ (Å)	$RMSD_{dU_{inter}}$ (kJ/mol)
NHB Structures (49 crystals)						
Static	0.2071	0.1735	2.9762	4.1985	-	-
PCM_{11}	0.2115	−0.1905	2.8618	3.8994	-	-
SCRP	0.1999/0.2141	−0.1318/−0.0209	2.5425/2.7156	3.3843/3.5292	0.0293 (0)	0.6253 (3)
$SCRP_{mol.}$	0.1842/0.2026	−0.3716/−0.4065	2.7760/2.8650	3.5698/3.6824	0.0551 (1)	0.2652 (0)
$SCRP_{atm,70}$	0.1907/0.1995	−0.2085/−0.1711	2.6228/2.6924	3.4825/3.5830	0.0143 (0)	0.2868 (1)
$SCRP_{atm,80}$	0.1917/0.2113	−0.2935/−0.2364	2.6084/2.6855	3.3960/3.4893	0.0811 (1)	0.2823 (1)
$SCRP_{atm,90}$	0.2027/0.2147	−0.2274/−0.1380	2.5193/2.6044	3.2605/3.3400	0.0284 (0)	0.3046 (1)
HB Structures (51 crystals)						
Static	0.1841	0.5880	4.9501	5.7793	-	-
PCM_{11}	0.1743	0.1463	2.9245	3.6976	-	-
SCRP	0.1591/0.1732	0.0373/1.0828	2.4690/2.7804	2.9497/3.3140	0.0302 (0)	1.9365 (30)
$SCRP_{mol.}$	0.1662/0.1790	0.0839/−0.0123	2.9537/3.2573	3.6547/4.1040	0.0267 (0)	1.4981 (19)
$SCRP_{atm,70}$	0.1627/0.2026	−0.0079/−0.2111	2.9147/3.2937	3.6276/4.0838	0.2083 (1)	1.4992 (19)
$SCRP_{atm,80}$	0.1554/0.1940	−0.0105/0.0900	2.6890/2.9503	3.3125/3.6780	0.2053 (1)	1.3930 (15)
$SCRP_{atm,90}$	0.1546/0.1829	−0.0153/0.6179	2.5370/2.6931	3.0952/3.3795	0.0991 (2)	1.4720 (24)

scores observed here, and in Chapter 6, indicate that HAIEFF models are generally reliable at predicting crystal geometries, regardless of the underlying electrostatic model. I believe this aligns with the experience of CSP practitioners, most of whom would regard energy accuracy as a more pressing weakness of HAIEFF methods.

In terms of the energy error metrics, the Static model remains inferior to all induction-embedded models, even the SCERP models that exhibit performance drift. Between the SCERP variants, energy accuracy post-CE continues to favor the models with less aggressive damping, but simultaneously post-CSORM performance drift is mitigated by the use of damping. Through these two conflicting forces, the net effect of damping strength and performance drift on model accuracy ultimately favors the $\text{SCRP}_{\text{atm,90}}$ model. Across the NHB and HB structures, $\text{SCRP}_{\text{atm,90}}$ yields the most competitive combination of AAD and SD (post-CSORM), outperforming all other SCERP variants. As posited, by including both NHB and HB structures in the training set, the benefits of using SCERP, over PCM, emerge more clearly. Post-CE, almost all of the SCERP models outperform PCM_{11}, except for $\text{SCRP}_{\text{mol.}}$ and $\text{SCRP}_{\text{atm,70}}$ in the HB structures. These latter two models likely employ damping too aggressively, to the point that the fidelity of the electrostatic model is undermined. Post-CSORM, these two SCERP models continue to deteriorate in performance, and lose more ground to the PCM_{11} model. In contrast, due to their superior post-CE model accuracy, the other SCERP models edge out the PCM_{11} model (post-CSORM), even with performance drift.

Further comparison between PCM_{11} and the SCERP variants is conducted with TP442, which comprises the entirety of the DB–445 dataset, except for CYCYPR, KUXPUP01, and CAXPUN. From Figure 7.7, these three structures potentially exhibit erroneous behavior with SCERP. The remaining structures include 317 NHB and 125 HB crystals. Crucially, this dataset expands the chemical profile of the training data to include fluorine, sulphur, and chlorine atom types, albeit without any sulphur–chlorine interactions. The preceding trials consistently demonstrate the Static model's inferiority to the induction-embedded models. Additionally, the $\text{SCRP}_{\text{atm,70}}$ model is suggested to damp electrostatic interactions 'too aggressively', such that its energy accuracy is uncompetitive relative to the other $\text{SCRP}_{\text{atm.}}$ models. To expedite testing, I exclude these two models from the fitting with TP442. Although the $\text{SCRP}_{\text{mol.}}$ model has performed similarly to $\text{SCRP}_{\text{atm,70}}$, its distinct construction versus the $\text{SCRP}_{\text{atm.}}$ models warrants further testing with TP442.

The residual scores and performance metrics for the models trained on TP442 are provided in Appendix M.2. As before, there is evidence to indicate that energy accuracy scales inversely with the aggressiveness of damping. Despite this, in the TP100 set, $\text{SCRP}_{\text{mol.}}$ would outperform SCRP and $\text{SCRP}_{\text{atm,90}}$ on the basis of superior geometry accuracy. Using TP442, $\text{SCRP}_{\text{mol.}}$ no longer exhibits this competitive advantage in geometry, and ultimately has the greatest total objective function value among the SCERP models. $\beta_{\text{mol}}^{\text{AB}}$ are suggested to be most valid for simple molecules (containing only H, C, O, and N atoms [250]) without diverse functionalities [224, 250]. With the exploration of greater chemical diversity in TP442, the flaws in using $\beta_{\text{mol}}^{\text{AB}}$ are increasingly exposed. Among the remaining SCERP variants, there is closer parity in model

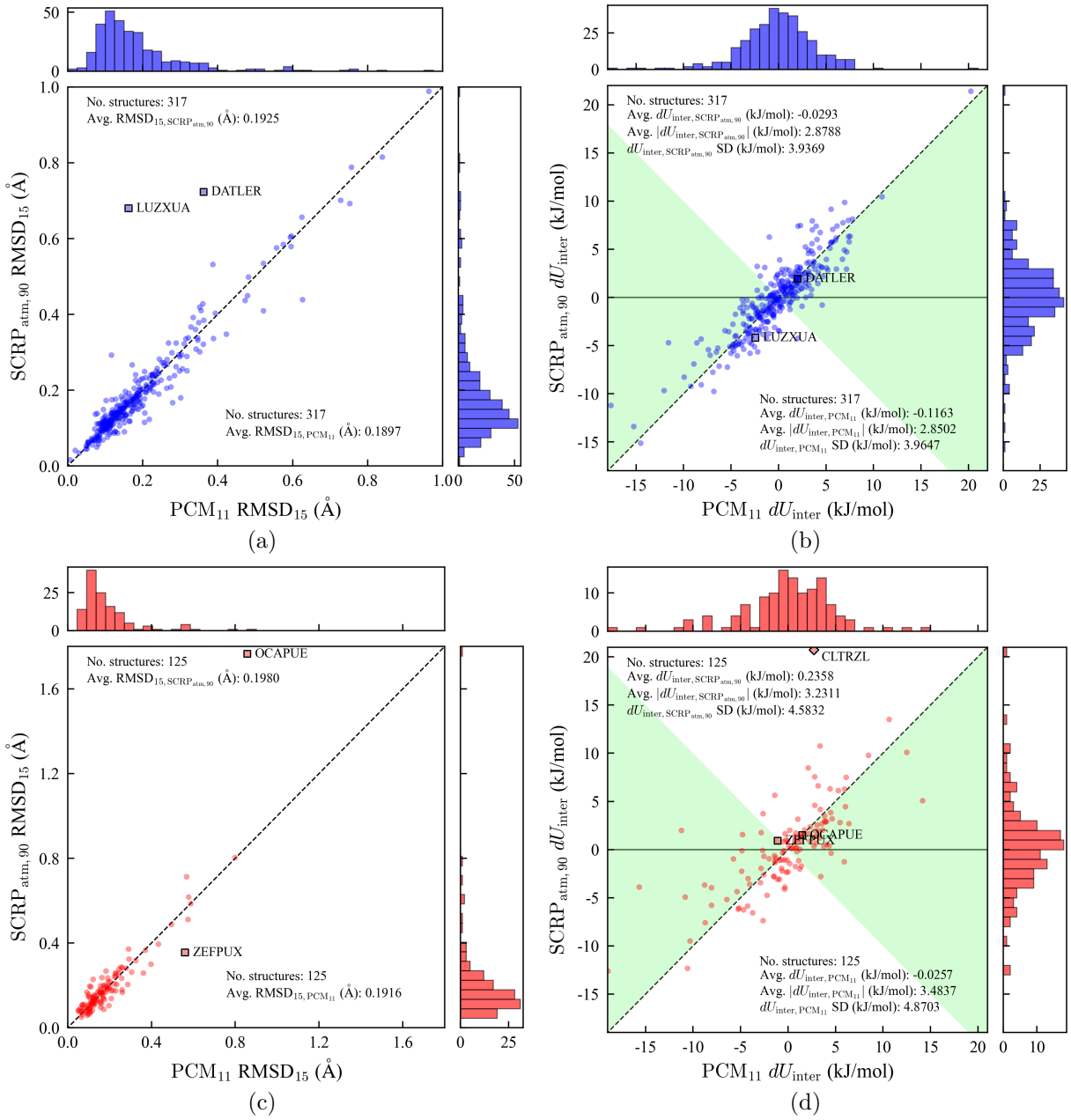


Figure 8.5: TP442 errors for PCM_{11} and post-CSORM $SCRP_{atm,90}$: (a) NHB crystals $RMSD_{15}$, (b) NHB crystals dU_{inter}, (c) HB crystals $RMSD_{15}$, and (d) HB crystals dU_{inter}. Structures with $|\Delta RMSD_{15}| \geq 0.20$ Å and $|\Delta dU_{inter}| \geq 15.0$ kJ/mol are marked with □ and ◇ respectively. Green areas in (b) and (d) denote where the y-axis model has a lower $|dU_{inter}|$. Histograms of each model's errors are shown on the corresponding axes.

performance post-CE, but post-CSORM $SCRP_{atm,90}$ edges out the other models.

Overall, the difference in performance between the SCERP variants span, at most, a few decimal points in kJ/mol (Table M.2), and one could argue that (with use of a complementary FF) the damping implementation will ultimately have minimal impact on practical results. Nevertheless, a critical takeaway is that serviceable $\beta_{atm}^{tt'}$ can be determined purely from intuition. Using these approximate $\beta_{atm}^{tt'}$ achieves model accuracies that are as good as, and often better than, using β_{mol}^{AB}. There is also a good likelihood that $\beta_{atm}^{tt'}$ will extend more success-

fully to complex chemical systems. At this stage, it is worth revisiting the question of whether there is even a need to include damping at all. A more-detailed comparison of the SCRP and $\text{SCRP}_{\text{atm,90}}$ performance (Figure M.2) reveals that the two models behave almost identically, with the exception of a few outliers. When applied in moderation, damping does not seem to promote notable improvements in the model performance, yet it also does not deteriorate model accuracy. Other authors [189, 224] have emphasized that electrostatic damping can play a pivotal role in suppressing divergent behavior in the intermolecular model. While this was not observed here, the results from Chapter 6 do corroborate this claim. As such, I put forward $\text{SCRP}_{\text{atm,90}}$ as a sound implementation of damping with SCERP.

Finally, I compare the performance of PCM_{11} against that of post-CSORM $\text{SCRP}_{\text{atm,90}}$ in the TP442 set (Figure 8.5). While the average RMSD_{15} scores suggest that $\text{SCRP}_{\text{atm,90}}$ has poorer geometry accuracy, closer inspection reveals that for most structures, geometry predictions are of similar quality between the two models. The larger average RMSD_{15} scores for $\text{SCRP}_{\text{atm,90}}$ are driven almost exclusively by a few outliers that exhibit large deterioration in geometry accuracy relative to PCM_{11}. It is unclear why these structures suffer such large degradation in geometry accuracy, but if omitted from the average RMSD_{15} statistic, $\text{SCRP}_{\text{atm,90}}$ performs on par with PCM_{11} in geometry accuracy. Interestingly, for these structures which PCM_{11} and $\text{SCRP}_{\text{atm,90}}$ predict highly disparate RMSD_{15} scores, their corresponding dU_{inter} values differ only marginally. In the NHB structures, $\text{SCRP}_{\text{atm,90}}$ has comparable energy error metrics (AAD and SD) to the PCM_{11} model, while in the HB structures $\text{SCRP}_{\text{atm,90}}$ creeps ahead slightly. Admittedly, it is unclear if sub-kJ/mol reductions in the energy error metrics form strong evidence of SCERP's superiority in a CSP context. CLTRZL is noted to be significantly more underbound (relative to the CE–755 reference U_{inter}) when using $\text{SCRP}_{\text{atm,90}}$ versus PCM_{11}. In the single-point U_{ind} calculations, CLTRZL was not flagged as an outlier, but it does lie quite close to the outlier limit in Figure 7.7b (coordinates: $(-29.8, -14.4)$).

8.3.4 Final Parametrization

Testing thus far has informed how SCERP and electrostatic damping should be applied, and the proposed $\text{SCRP}_{\text{atm,90}}$ model seems to slightly outperform PCM_{11}. For deployment, the final SCERP-compatible FF parametrizations employ the full CE–755 training (TRN) and test (TST) datasets, previously used in Chapter 6 (Appendix G). Here, I parametrize FFs using *both* Schemes 5 and 7 (Table 6.2), and their fitted parameters are tabulated in Appendix M.3.

8.3.4.1 SCERP-Embedded CSO–RM Minimization Performance

Before I discuss the model accuracy with the fitted FFs, I analyze the algorithmic performance of the SCERP-embedded CSO–RM minimization protocol that was illustrated in Figure 7.11. Readers are reminded that the SCERP-embedded CSO–RM minimization attempts to solve Equation 7.3.11 iteratively using intermediate solutions to Equation 7.3.12, until $U^*_{\text{inter},k}$ converges. In Figure 8.6, I present the convergence behavior of the minimization when using the

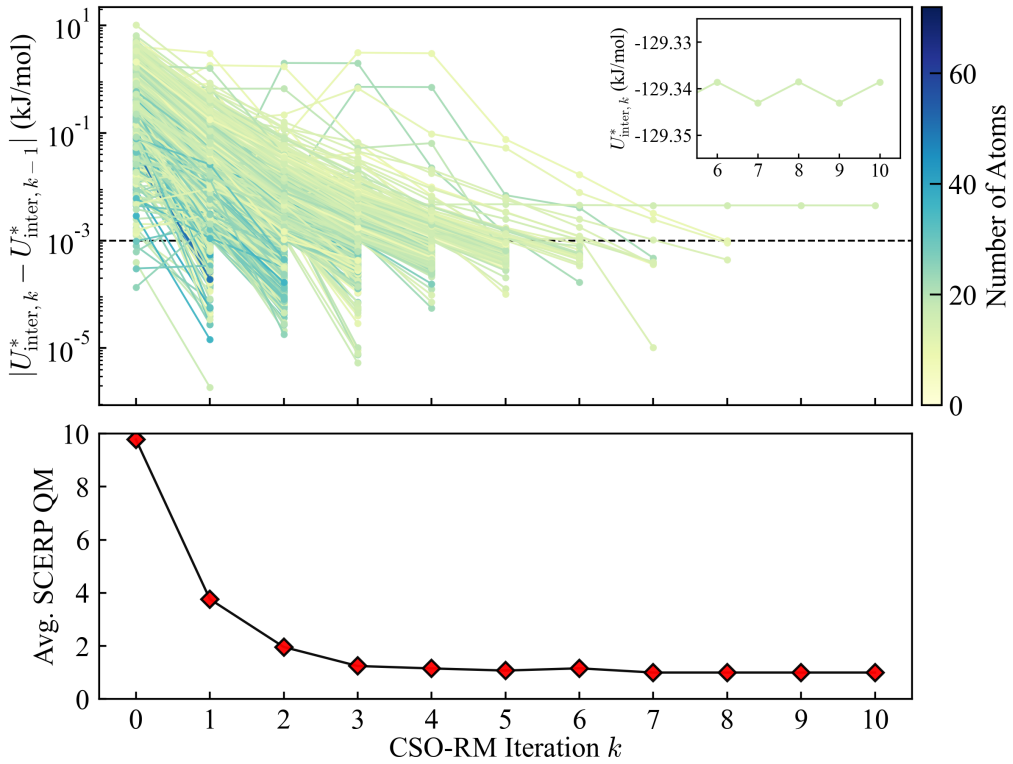


Figure 8.6: Top: convergence of the SCERP-embedded CSO–RM minimization ($\varepsilon_{\text{scrpRM}} = 1 \times 10^{-3}$ kJ/mol (dashed line) and $k_{\max} = 10$) for the $\text{SCRP}_{\text{atm,90}}$ model + Scheme 7 fitted FF over the CE–755 dataset. Inset: $U^*_{\text{inter},k}$ for BEMKAI, which did not converge. Bottom: the average number of QM computations (normalized by Z') to converge the SCERP process at iteration k.

$\text{SCRP}_{\text{atm,90}}$ model paired with its Scheme 7 fitted FF, over the entire CE–755.

Over the 755 crystal structures, 754 successfully converge to a tolerance of 1×10^{-3} kJ/mol within the allotted $k_{\max}$ of 10. On average, 5.04 iterations of CSO–RM minimization are expended to reach U^*_{inter} convergence, including the $k = -1$ iteration. This aligns with the results of Zhang [71], who reported requiring about half-a-dozen iterations to converge CSO–RM with SCERP. There is no indication that the rate of convergence is influenced by the size of the molecules contained within the asymmetric unit cell. The remaining unconverged case corresponds to 6-chloro-4-nitro-1,2,3-benzotrithiole 2-oxide (CSD Refcode: BEMKAI). From the inset plot in Figure 8.6, $U^*_{\text{inter},k}$ oscillates around a value of −129.34 kJ/mol during the last few iterations of CSO–RM for this structure. From testing, these oscillations are sustained, and convergence is unlikely even if $k_{\max}$ was increased. Non-convergence of this nature is an inherent risk of self-consistent procedures (e.g., the CSO–RM minimization), particularly when there is no mechanism for partial updates between iterations. Unlike the decelerated SCERP process, obtaining an intermediate of two crystal structures may not be as straightforward as using an average of two consecutive $\mathbf{\Omega}^*_k$. Ultimately, with the current CSO–RM protocol, oscillatory non-convergence cannot be ruled out. Fortunately, in my experience, the frequency of this is quite rare (e.g., only 1 out of 755 structures) and it is unlikely to affect the ranking of polymorphs as they are of relatively small magnitude.

While it has not been presented, the $\text{SCRP}_{\text{atm,90}}$ model paired with its Scheme 5 fitted FF shares many of these traits. Of 755 structures, 750 converged, expending on average 5.21 iterations of CSO–RM minimization. For two unconverged structures, AKOVOL02 and BELBUP02, additional CSO–RM iterations would likely have led to convergence. Conversely, the structures YIFGAW and JIWWUK exhibit oscillatory non-convergence, similar to BEMKAI. Finally, the OCAPUE structure fused following minimization with the full CSO–RM protocol. As shown in Chapter 6, the Scheme 5 FF does not completely eliminate the possibility of molecular fusing, hence this result is not unreasonable with FF extrapolation across CSO–RM iterations. Overall, I consider the convergence of U^{*}_{inter}, under the current CSO–RM protocol, to be satisfactory.

In Figure 8.6, the cost of converging the SCERP process at each step of the CSO–RM minimization is also shown. Between each iteration k, there is a marked decrease in the number of QM iterations (t from Figure 7.1) used to converge the SCERP process, and in later iterations of CSO–RM, the SCERP process converges in only 1–2 QM calculations. This progressive reduction in SCERP QM cost is likely attributable to reusing $\boldsymbol{Q}^{\text{total,pc}}_{I,\text{field},k-1}$ between CSO–RM iterations (Figure 7.11, R4 → R3). As $\boldsymbol{\Omega}^{*}_{k}$ converges, $\boldsymbol{Q}^{\text{total,pc}}_{I,\text{field},k-1}$ become increasingly good guesses for $\boldsymbol{Q}^{\text{total,pc}}_{I,\text{field},k}$, leading to the non-negligible savings in QM expenditure. It is encouraging to see this simple algorithmic strategy enhance the model's practicality, justifying its extension to the CSO–FM protocol, as described in Section 7.4.1.

8.3.4.2 Training Set Performance

Next, I consider the accuracy of the SCERP-embedded HAIEFF model when paired with its corresponding complementary FF, over the CE–755 training set. I assess the model accuracy only post-CSORM, since this reflects the true performance upon deployment. As the incumbent induction-approximation used in HAIEFF models, the model performance with the PCM_{11}-parametrized FFs from Chapter 6 are used as a basis of comparison here. The Scheme 5 FF + $\text{SCRP}_{\text{atm,90}}$ model is compared against its PCM_{11} counterpart (Figure 8.7), and likewise for the Scheme 7 models (Figure 8.8).

For both fitting schemes, the PCM_{11} and $\text{SCRP}_{\text{atm,90}}$ models largely predict crystal geometries with close parity in RMSD_{15} scores. Over this large training set, there is no indication that accuracy in predicted crystal geometries is significantly enhanced when using $\text{SCRP}_{\text{atm,90}}$, and in fact (like TP442) there are a few structures that deteriorate drastically in RMSD_{15} score when using the SCERP model. Interestingly, in the Scheme 7 FF + $\text{SCRP}_{\text{atm,90}}$ case, there are some structures which also exhibit the opposite behavior, wherein their RMSD_{15} scores *decrease* dramatically with the $\text{SCRP}_{\text{atm,90}}$ model (for clarity I have not marked these with □ in the figure). In any case, these (positive and negative) outliers are very much exceptions to the rule, and as a whole, all of the HAIEFF models (Scheme 5/7 with PCM_{11}/$\text{SCRP}_{\text{atm,90}}$) achieve a similar quality of geometry prediction, with average RMSD_{15} scores around 0.18–0.19 Å. For practical CSP, this precision is acceptable and aligns with the perceived adequacy of HAIEFF models for predicting crystal geometries.

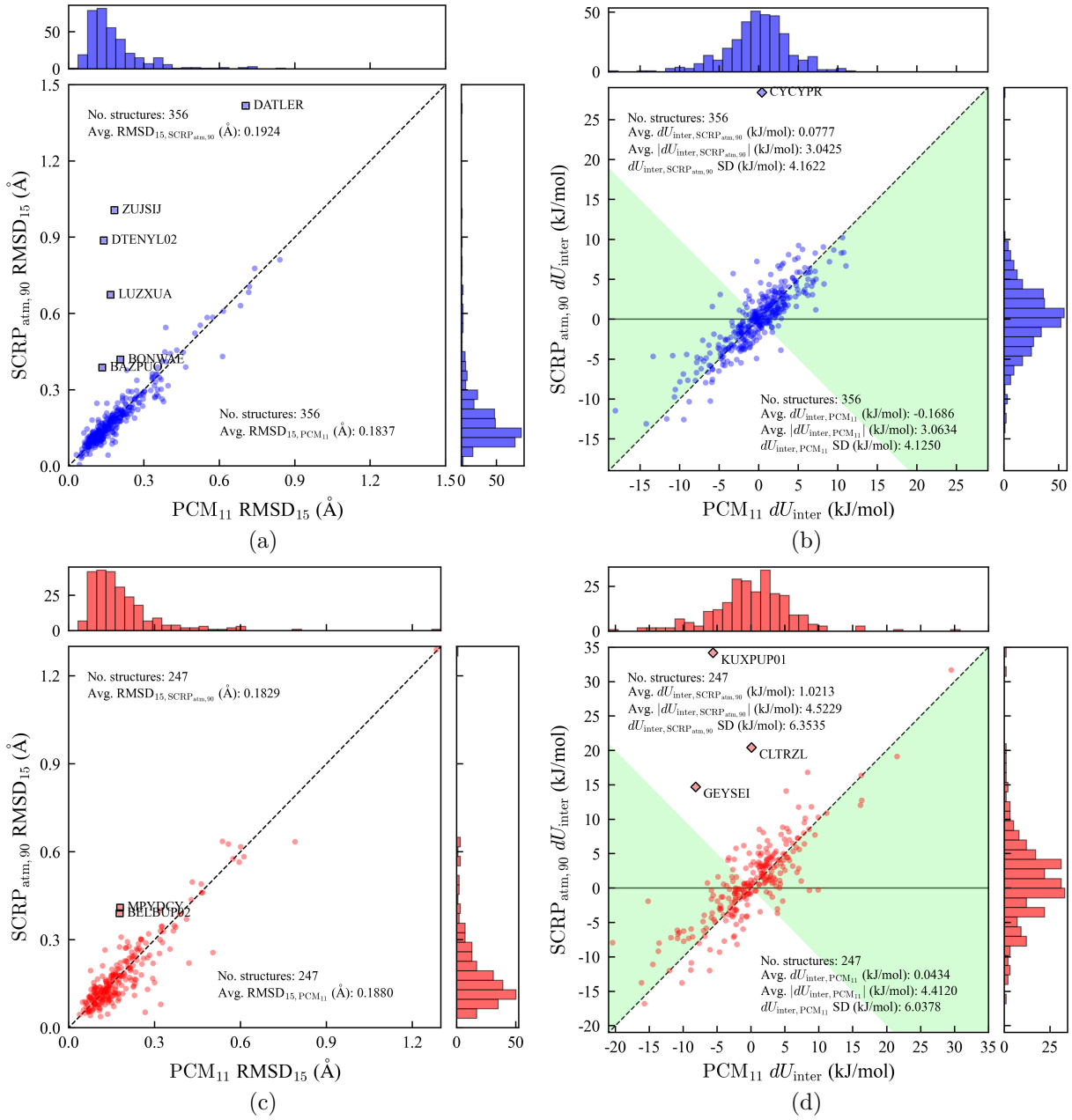


Figure 8.7: CE–755 TRN set errors for PCM_{11} and post-CSORM $SCRP_{atm,90}$, using Scheme 5: (a) NHB crystals $RMSD_{15}$, (b) NHB crystals dU_{inter}, (c) HB crystals $RMSD_{15}$, and (d) HB crystals dU_{inter}. Structures with $\Delta RMSD_{15} \geq 0.20$ Å and $|\Delta dU_{inter}| \geq 15.0$ kJ/mol are marked with □ and ◇ respectively. Green areas in (b) and (d) denote where the y-axis model has a lower $|dU_{inter}|$. Histograms of each model's errors are shown on the corresponding axes. OCAPUE is omitted from the HB statistics, due to its fusion with $SCRP_{atm,90}$.

Shifting focus to energy accuracy, the dU_{inter} predicted by the two induction models are also strongly centred about the 45° parity line, although the HB structures exhibit a larger spread compared to the NHB structures. This greater variation in model performance for the HB structures likely reflects that induction plays a larger role in their stabilization, such that they are more strongly influenced by the details of the polarization model. Overall, crystals that have poorer energy predictions with $SCRP_{atm,90}$ (points in the white regions of the figure) are mostly

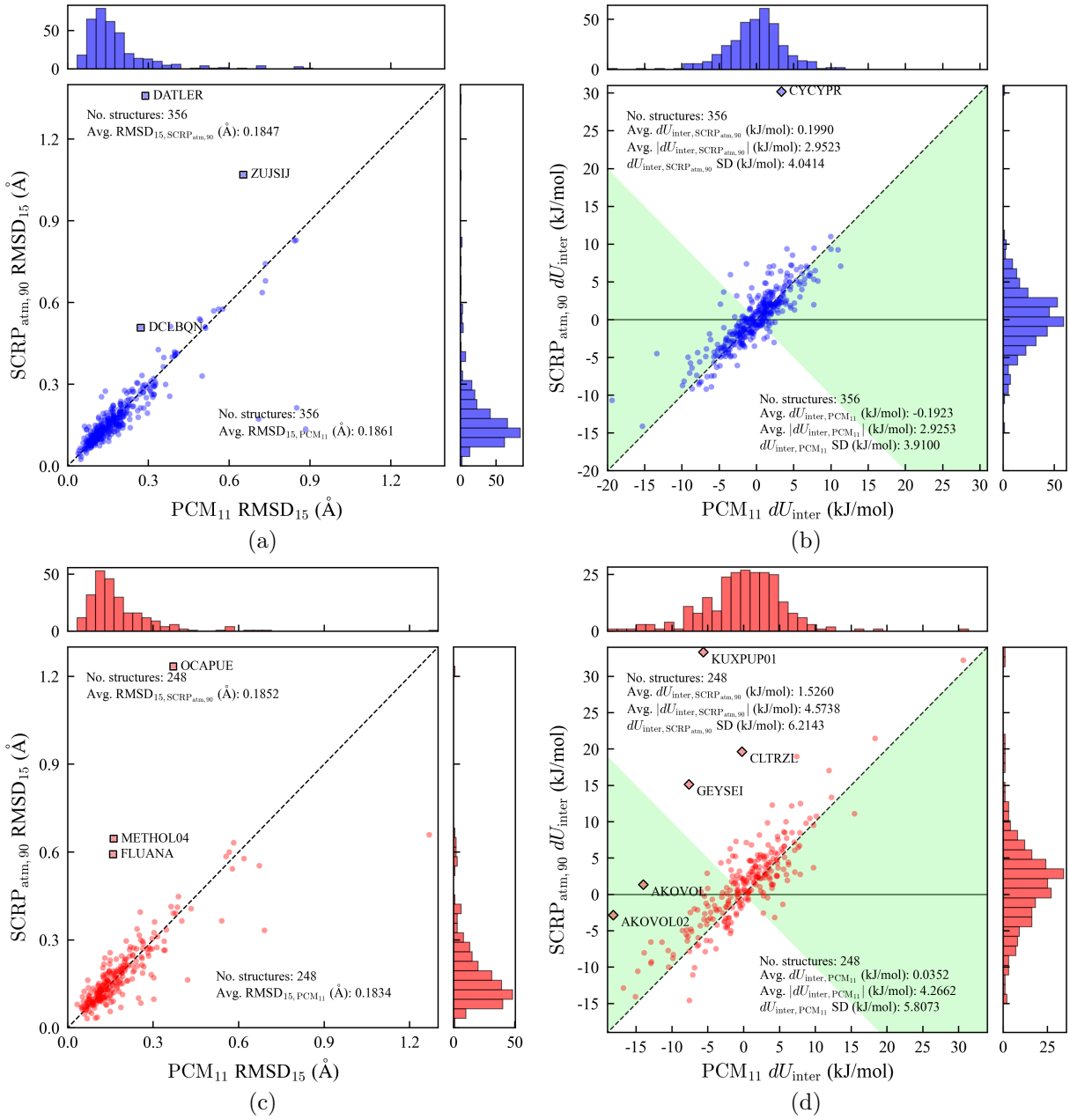


Figure 8.8: CE–755 TRN set errors for PCM_{11} and post-CSORM $SCRP_{atm,90}$, using Scheme 7: (a) NHB crystals $RMSD_{15}$, (b) NHB crystals dU_{inter}, (c) HB crystals $RMSD_{15}$, and (d) HB crystals dU_{inter}. Structures with $\Delta RMSD_{15} \geq 0.20$ Å and $|\Delta dU_{inter}| \geq 15.0$ kJ/mol are marked with □ and ◇ respectively. Green areas in (b) and (d) denote where the y-axis model has a lower $|dU_{inter}|$. Histograms of each model's errors are shown on the corresponding axes.

offset by structures that have enhanced energy predictions (points in the green regions of the figure). As such, there are only small differences in the AAD and SD between the $SCRP_{atm,90}$ and PCM_{11} models, with these metrics slightly favoring the latter.

The slightly superior energy accuracy achieved with the PCM_{11} model is biased by a few structures that exhibit a marked worsening of dU_{inter} when using $SCRP_{atm,90}$: CYCYPR, KUXPUP01, CLTRZL, and GEYSEI. Like in the TP442 set, these dU_{inter} outliers do not correspond to the structures with large gains in $RMSD_{15}$ score. When using $SCRP_{atm,90}$, dU_{inter} for these

four structures are significantly more positive than when using PCM_{11}. This is observed irrespective of the fitting scheme, suggesting that the prediction errors are due to the SCERP model underbinding these crystals considerably. Indeed, during the earlier single-point assessment of U_{ind}, CYCYPR and KUXPUP01 were flagged as erroneous (Figure 7.7). CLTRZL was not originally identified as problematic, but from TP442 was later noted to be near the outlier limit of Figure 7.7b. Only GEYSEI has not been marked as having (potentially) erroneous U_{ind} predictions with SCERP, and on Figure 7.7b it has coordinates (−26.29,−34.27), i.e., SCERP was predicting greater U_{ind} than PCM_{11}. Evidently, the combination of electrostatic damping and structure minimization (as opposed to a single-point evaluation) is afflicting GEYSEI with a similar U_{ind} error when using $SCRP_{atm,90}$. If these large outliers are omitted from the dU_{inter} statistics, the $SCRP_{atm,90}$ model edges out PCM_{11} in accuracy (Table 8.7 and Figure M.3).

Table 8.7: dU_{inter} AAD and SD using PCM_{11} and $SCRP_{atm,90}$ for the CE–755 training set, omitting the outliers: CYCYPR, KUXPUP01, CLTRZL, and GEYSEI.

dU_{inter}	Scheme 5 FF				Scheme 7 FF			
	PCM_{11}		$SCRP_{atm,90}$		PCM_{11}		$SCRP_{atm,90}$	
	NHB	HB	NHB	HB	NHB	HB	NHB	HB
AAD (kJ/mol)	3.0709	4.4096	2.9710	4.2944	2.9241	4.2634	2.8756	4.3519
SD (kJ/mol)	4.1307	6.0411	3.8857	5.8262	3.9109	5.8107	3.7191	5.7254

Finally, I compare the performance of the $SCRP_{atm,90}$ model when paired with its respective Scheme 5 and 7 fitted FFs. Between the two fitting schemes, the Scheme 7 FF does slightly outperform its Scheme 5 counterpart, although the differences are subtle across the error metrics. This is consistent with the Chapter 6 results, wherein the 'more complex' FF (Scheme 7 MIP, $\dim(\bar{\boldsymbol{p}}) = 135$) tends to outperform the 'less complex' alternative (Scheme 5 dBHP, $\dim(\bar{\boldsymbol{p}}) = 90$). Critically, the Scheme 5 model has an occurrence of molecular fusing (OCAPUE), which is suppressed when using the Scheme 7 model. This is also consistent with the observations from Chapter 6, wherein the ability of the dBHP to suppress molecular fusing is somewhat transient, since its repulsive walls are still susceptible to spurious disappearance at certain values of $A^{t,t'}$ and $C^{t,t'}$. Conversely, the MIP is constrained ($\lambda_r^{tt'} > \lambda_a^{tt'}$) to *always* have repulsive walls that reliably counteract molecular fusing.

8.3.4.3 Test Set Performance

To validate the transferability of the parametrized SCERP-compatible FFs, I consider the post-CSORM model accuracy in the CE–755 test set. As before, the modeling performance of $SCRP_{atm,90}$ paired with the Scheme 5 (Figure 8.9) and the Scheme 7 (Figure 8.10) fitted FFs are compared against their respective PCM_{11} counterparts.

Using either fitting scheme with the $SCRP_{atm,90}$ model, geometry accuracy is preserved well such that the average $RMSD_{15}$ scores do not exhibit any noticeable increase between the training and test sets. The test set geometries predicted using the $SCRP_{atm,90}$ model are mostly on par

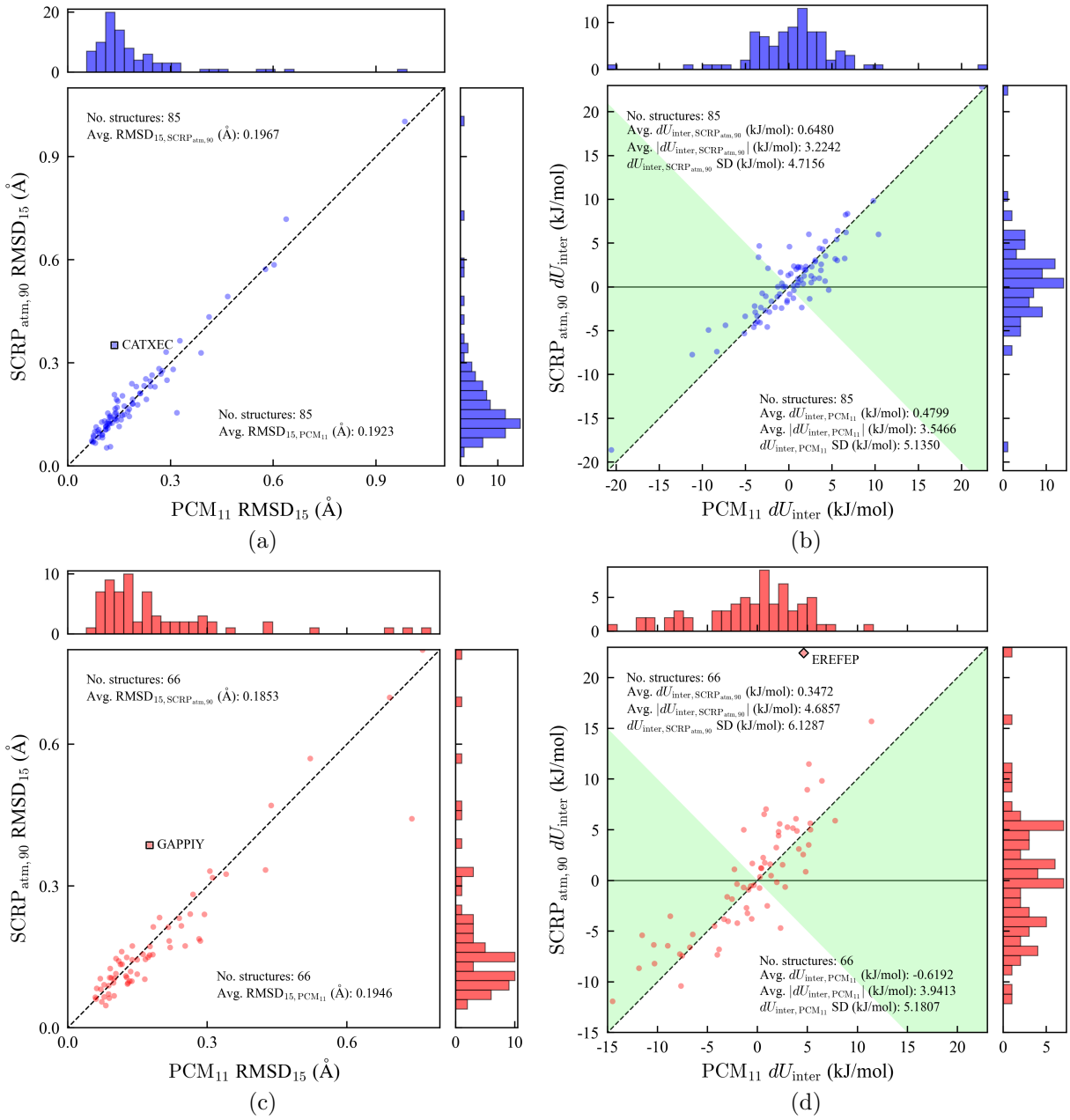


Figure 8.9: CE–755 TST set errors for PCM_{11} and post-CSORM $SCRP_{atm,90}$, using Scheme 5: (a) NHB crystals $RMSD_{15}$, (b) NHB crystals dU_{inter}, (c) HB crystals $RMSD_{15}$, and (d) HB crystals dU_{inter}. Structures with $\Delta RMSD_{15} \geq 0.20$ Å and $|\Delta dU_{inter}| \geq 15.0$ kJ/mol are marked with □ and ◇ respectively. Green areas in (b) and (d) denote where the y-axis model has a lower $|dU_{inter}|$. Histograms of each model's errors are shown on the corresponding axes.

in $RMSD_{15}$ score, relative to the PCM_{11} model. Two structures do exhibit much larger $RMSD_{15}$ scores with the $SCRP_{atm,90}$ model, but similar geometry outliers are present in the training set results as well. This suggests that poor FF transferability is not the cause for these aberrations, but rather they are likely linked to the $SCRP_{atm,90}$ model itself.

In the assessment of the test set dU_{inter}, the $SCRP_{atm,90}$ model does not predict any large outliers (relative to PCM_{11}) among the NHB structures, but one HB structure (EREFEP) is flagged. EREFEP, similar to GEYSEI, did not present with strongly erroneous behavior in

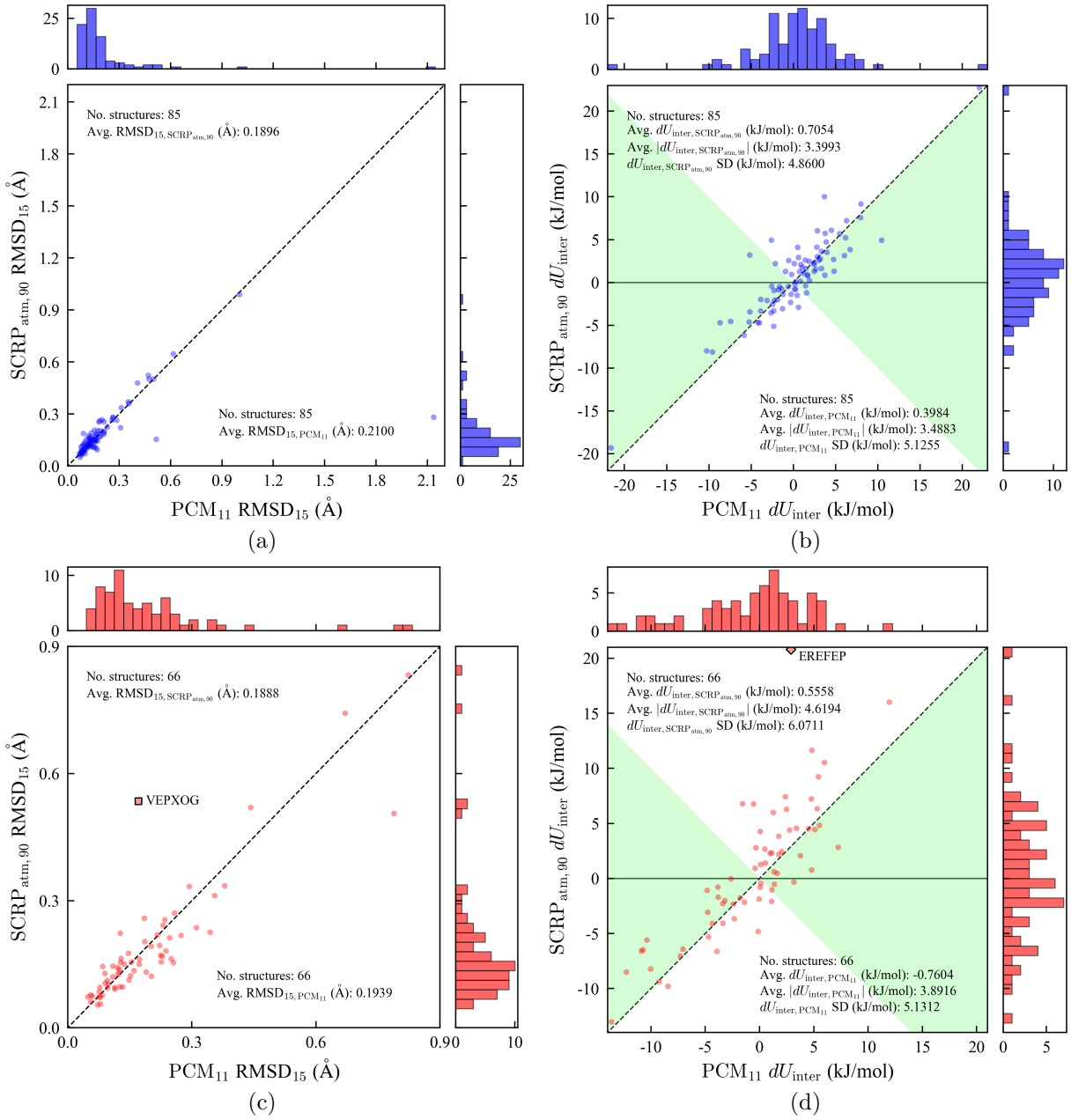


Figure 8.10: CE–755 TST set errors for PCM_{11} and post-CSORM $SCRP_{atm,90}$, using Scheme 7: (a) NHB crystals $RMSD_{15}$, (b) NHB crystals dU_{inter}, (c) HB crystals $RMSD_{15}$, and (d) HB crystals dU_{inter}. Structures with $\Delta RMSD_{15} \geq 0.20$ Å and $|\Delta dU_{inter}| \geq 15.0$ kJ/mol are marked with □ and ◇ respectively. Green areas in (b) and (d) denote where the y-axis model has a lower $|dU_{inter}|$. Histograms of each model's errors are shown on the corresponding axes.

Figure 7.7b, wherein the SCERP model had predicted greater U_{ind} stabilization than the PCM_{11} counterpart (coordinates: (−24.64, −29.80)). Nevertheless, upon full minimization with CSO–RM (and with electrostatic damping), EREFEP is predicted with significantly more positive dU_{inter} when using the $SCRP_{atm,90}$, versus PCM_{11}. Like with the training set aberrations, this anomalous behavior is observed regardless of the fitted FF, suggesting that the model error lies in the electrostatic description, rather than the FF. To avoid skewing the assessment of FF transferability in predicting crystal energies, I have chosen to compare the training and test

Table 8.8: AAD and SD metrics for dU_{inter} in the CE–755 training and test set, using $\text{SCRP}_{\text{atm,90}}$. Outliers have been omitted from the corresponding training (CYCYPR, KUXPUP01, CLTRZL, and GEYSEI) and test set (EREFEP) statistics. $\Delta\text{X} = \text{X}_{\text{TST}} - \text{X}_{\text{TRN}}$ are given in parenthesis.

dU_{inter}	Scheme 5 FF		Scheme 7 FF	
	NHB	HB	NHB	HB
CE–755 Training Set				
AAD (kJ/mol)	2.9710	4.2944	2.8756	4.3519
SD (kJ/mol)	3.8857	5.8262	3.7191	5.7254
CE–755 Test Set				
AAD (kJ/mol)	3.2242 (+0.2532)	4.4127 (+0.1183)	3.3993 (+0.5237)	4.3706 (+0.0187)
SD (kJ/mol)	4.7156 (+0.8299)	5.5146 (−0.3116)	4.8600 (+1.1409)	5.5620 (−0.1634)

set performance in the absence of any dU_{inter} outliers (Table 8.8 and Figure M.4). In general, the change in energy error metrics (ΔX in Table 8.8) between the training and test sets is comparable to that reported for the FFs parametrized in Chapter 6 (*cf.* Figures 6.10). Using the $\text{SCRP}_{\text{atm,90}}$ models, ΔAADs do not exceed 0.52 kJ/mol and ΔSDs do not surpass 1.15 kJ/mol. This suggests that the $\text{SCRP}_{\text{atm,90}}$-compatible FFs ultimately exhibit a similar degree of transferability as the FFs parametrized without SCERP. Comparing the test set performance between the $\text{SCRP}_{\text{atm,90}}$ model and their PCM_{11} counterparts (in the absence of outliers), the

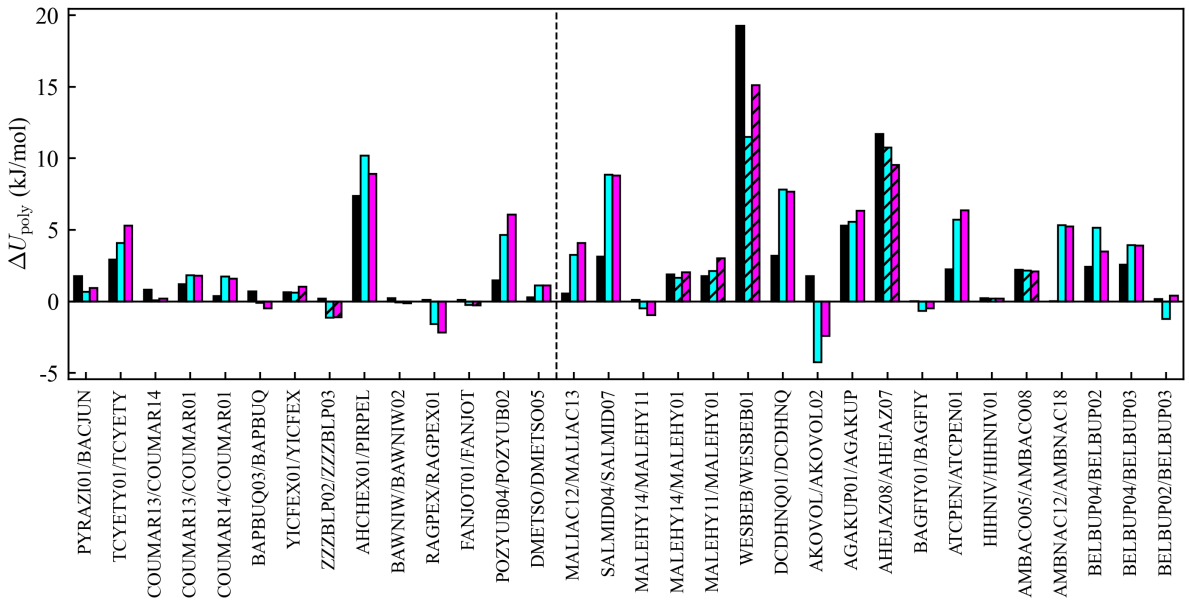


Figure 8.11: ΔU_{poly} (kJ/mol) when using the $\text{SCRP}_{\text{atm,90}}$ multipole model. Each polymorphic pair is labeled with their CSD Refcodes ordered as 'More Stable/Less Stable'. $\Delta U_{\text{poly}}^{\text{ref}}$ values are given in black, while $\Delta U_{\text{poly}}^{\text{pred}}$ obtained using the Scheme 5 and 7 models are in cyan and magenta respectively. Hatched bars indicate that a member of the pair is from the test set. NHB and HB systems are on the left and right of the dashed line respectively. The KUXPUP/KUXPUP01 pair is omitted as KUXPUP01 presents with erroneous behavior with SCERP.

$SCRP_{atm,90}$ model has slightly superior accuracy for the NHB structures (Figures 8.9b and 8.10b), but loses out for the HB structures (Figure M.4).

8.3.4.4 Prediction of Relative Polymorphic Energies

The final test of $SCRP_{atm,90}$ model performance I study is the prediction of relative polymorphic energies, within the CE–755. I follow the test methodology outlined in Section 6.4, and evaluate relative polymorphic energy differences, $\Delta U_{poly,i}$, as defined by Equation 6.4.1. The KUXPUP/KUXPUP01 polymorphic pair is expected to be egregiously in error, due to the spurious predictions of U_{inter} for KUXPUP01. As such, this pair has been omitted from the current results, presented in Figure 8.11 and Table 8.9. The corresponding statistics for the PCM_{11} model, with omission of KUXPUP/KUXPUP01, are also included in Table 8.9.

The results of the TRN and TST did not clearly favor one electrostatic model ($SCRP_{atm,90}$ or PCM_{11}) over the other. However, the predictions of polymorphic energy differences provide additional evidence that the $SCRP_{atm,90}$ model may offer some improvement in model accuracy, for practical CSP. Among the NHB polymorphic pairs, the $SCRP_{atm,90}$ model paired with both the Scheme 5 and 7 fitted-FFs have lower AAD in ΔU_{poly}, versus their PCM_{11} counterparts, by about 0.15 kJ/mol. Among the HB polymorph pairs, there is closer parity in ΔU_{poly} AAD, such

Table 8.9: Summary of the key performance metrics in the prediction of ΔU_{poly} for fitting Schemes 5 and 7 over the different multipole models. These include the rank accuracy ('Rank Acc.', i.e., number of correctly ranked pairs), number of pairs with $|d\Delta U_{poly,i}| \leq 5$ kJ/mol ('Within 5 kJ/mol'), the AD in ΔU_{poly}, AAD in ΔU_{poly}, and the minimum/maximum $d\Delta U_{poly}$ (Min/Max D respectively).

Induction Model	Fitting Scheme	Rank Acc.	Within 5 kJ/mol	AD (kJ/mol)	AAD (kJ/mol)	Min D (kJ/mol)	Max D (kJ/mol)
NHB Pairs (No. of pairs: 14)							
PCM_{11}*	5	10	13	0.4147	1.3062	−1.9593	5.4612
PCM_{11}*	7	10	13	0.4985	1.4442	−2.2296	7.0469
$SCRP_{atm,90}$	5	9	14	0.2582	1.1619	−1.6896	3.1794
$SCRP_{atm,90}$	7	9	14	0.3287	1.3205	−2.2891	4.6005
HB Pairs (No. of pairs: 18)†							
PCM_{11}*	5	13	14	1.1642	2.3964	−5.2278	7.9281
PCM_{11}*	7	15	15	1.3605	2.3019	−4.1980	8.0506
$SCRP_{atm,90}$	5	14	14	0.4858	2.4567	−7.7835	5.7274
$SCRP_{atm,90}$	7	15	16	0.8756	2.2379	−4.1936	5.6644

† The KUXPUP/KUXPUP01 pair has been omitted since KUXPUP01 likely has highly erroneous predictions with the SCERP model.

* The corresponding PCM_{11} multipole results have been reproduced here, from Table 6.4. For the HB pairs, these statistics reflect the omission of KUXPUP/KUXPUP01.

that $\text{SCRP}_{\text{atm,90}}$ slightly outperforms PCM_{11} using the Scheme 7 FF, but conversely PCM_{11} has a lower AAD with the Scheme 5 FF. This observation probably reflects that the PCM_{11} induction model is a poor representation of the crystalline environment for less-polarized NHB crystals, versus HB crystals. $\text{SCRP}_{\text{atm,90}}$ is better able to accommodate both NHB and HB environments, thus improving performance in the former without significant trade-off in the latter. The changes in AAD are further reflected in the Min/Max $d\Delta U_{\text{poly}}$, wherein when using the $\text{SCRP}_{\text{atm,90}}$ model, the range of $d\Delta U_{\text{poly}}$ narrows by about 2.4 kJ/mol, relative to PCM_{11}. Overall, the reduction in ΔU_{poly} errors translates to about 1–2 more structures within the 5 kJ/mol $d\Delta U_{\text{poly}}$ window, and slightly better rank accuracy when using the $\text{SCRP}_{\text{atm,90}}$ model.

8.4 Summary

The central goal of this chapter was to parametrize HAIEFF FFs for use with the SCERP polarization model. In service of this, and based on recommendations from previous authors who investigated induction, I first explored the proposition of employing electrostatic damping within the HAIEFF model. From testing different variations of electrostatic damping, I consistently attain the greatest model performance with just a mild degree of electrostatic damping. Heavy-handed use of damping is prone to subpar predictions of cohesive energies, while a complete neglect of damping is physically and numerically ill-advised. Modulation of damping strength is made possible through the use of $\beta_{\text{atm}}^{tt'}$, as derived via Equation 8.2.4. Unlike $\beta_{\text{mol}}^{\text{AB}}$ [71, 189, 223, 224, 250], atom-pair specific $\beta_{\text{atm}}^{tt'}$ are more chemically intuitive and should be better suited for use in complex chemical systems. Ultimately, I advance the $\text{SCRP}_{\text{atm,90}}$ electrostatic model as the best candidate for use with HAIEFF models.

With a finalized decision on the SCERP model, I progress to parametrize deployment-ready SCERP-compatible FFs (Appendix M.3). Combined with the results from Chapter 6, I now have complete ensembles of HAIEFF models based on several distinct electrostatic models (Static, PCM_{11}, and $\text{SCRP}_{\text{atm,90}}$), paired with their respective complementary FFs (based on Schemes 5 and 7). Employing these HAIEFF models congruously, I compared the SCERP model accuracy against the incumbent induction model, PCM, but found only tenuous evidence that the former leads to significant enhancement of modeling performance. In particular, for predictions of absolute energies, the two models largely perform on par with each other and it is not immediately obvious which induction model is superior. However, for predictions of *relative* energies between polymorphs, there is evidence that the SCERP-embedded HAIEFF model outperforms its PCM counterparts. This result aligns with the intuition established regarding the fidelity of the SCERP model: the PCM polarization model cannot discern between different packing arrangements, whereas the SCERP model can (*cf.* Figure 7.8). This would manifest more blatantly when comparing predictions between polymorphs with distinct crystal packings. Potentially, even greater SCERP model performance could be extracted if a commensurate parameter estimation effort was expended here, as in Chapter 6 (i.e., using more Sobol' seeds). Unfortunately, due to resource limitations, this could not be done. Additionally, strategies to

suppress SCERP performance drift should be further explored.

A 'loose end' that has not been addressed are the outliers in geometry and energy predictions that emerge when using SCERP. It is notable that the same structures are consistently flagged as outliers in the context of dU_{inter}, while the RMSD_{15} outliers vary depending on the FF. This, plus evidence from Figure 7.7, leads me to believe that the flaw in the SCERP model pertains primarily to predicted U_{ind}, while the geometry outliers are simply a knock-on effect of the fitted FF parameters. The origin of this error, as well as further validation of HAIEFF model performance, will be explored in the next chapter.

Chapter 9

CSP Case Studies

Over Chapters 6–8, contributions have been made towards improving the description of intermolecular interactions for the HAIEFF class of CSP energy model. In this final chapter, I employ the augmented models within a series of CSP studies, with two key objectives. First and foremost, CSP studies offer an opportunity to rigorously validate these model developments through the prediction of relative energies across entire polymorphic crystal structure landscapes. In this way, I hope to benchmark the efficacy of the intermolecular energy models towards improving CSP outcomes. Secondly, I hope to further evaluate the algorithmic implementation of SCERP, particularly as it pertains to its utilization for flexible systems. In Section 9.1, I give a brief overview of the CSP methodology which is applied to all target systems. Following this, I conduct CSP studies over a range of systems that vary in flexibility, Z', and molecular chemistry. These systems comprise: a rigid monohydrate (Section 9.2), a flexible neutral molecule (Section 9.3), two flexible amino acids (Sections 9.4 and 9.5), and a flexible cocrystal (Section 9.7).

9.1 General CSP Methodology

The multistage CSP methodology [2] described in Section 2.2 is applied here. Due to various constraints, the CSP strategy employed differs slightly for each system. Here, I provide the high-level methodological details common to all the CSPs, while detailed specifications (e.g., program versions, search settings) can be found in the respective CSP sections in Appendix N.

9.1.1 Global Search

CrystalPredictor (Section 4.2.1) is used to execute the Global Search candidate generation + lattice energy minimization. Candidate generation is conducted in the appropriate Z' for each test system (Z' = 2 for the monohydrate and cocrystal, and Z' = 1 otherwise), and the 61 most common space groups for organic molecules are sampled non-uniformly according to their propensity in the CSD [74]. These include: *P1*, *P-1*, *P21*, *P21/c*, *P212121*, *Pna21*, *Pca21*, *Pbca*, *Pbcn*, *C2/c*, *Cc*, *C2*, *Pc*, *Cm*, *P21/m*, *C2/m*, *P2/c*, *C2221*, *Pmn21*, *Cmc21*, *Aba2*, *Fdd2*, *Iba2*, *Pnna*, *Pccn*, *Pbcm*, *Pnnm*, *Pmmn*, *Pnma*, *Cmcm*, *Cmca*, *Fddd*, *Ibam*, *P41*, *P43*, *I-4*, *P4/n*, *P42/n*, *I4/m*, *I41/a*, *P41212*, *P43212*, *P-421c*, *I-42d*, *P31*, *P32*, *R3*, *P-3*, *R-3*, *P3121*,

P3221, *R3c*, *R-3c*, *P61*, *P63*, *P63/m*, *P213*, *Pa-3*, *P2221*, and *Pba2*. The default CrystalPredictor lattice parameter sampling bounds have been used for all Global Searches.

In CrystalPredictor, U_{latt} is described by a HAIEFF model based on an atomic-charge electrostatic description. Except for the monohydrate, all other systems involve at least one flexible molecule. For these, LAM models are constructed using the adaptive LAM algorithm [98]. All QM calculations to determine conformational energies and charge densities are done in Gaussian, at the PBE0/6-311G(d,p) level of theory. Since all of the test systems have the potential to form intermolecular hydrogen bonds, PCM with ε_{PCM} = 11 is applied in all QM calculations, so as to embed induction into the electrostatic model. HLYGAt-derived atomic-charges are determined as part of the Gaussian calculations. The HAIEFF FF used at this stage is decided from local minimization of the experimentally-reported crystal structures, as described in Section 9.1.5. The default CrystalPredictor values for summation cutoffs, spline settings, and Ewald summation settings have been used in all lattice energy minimizations.

In this work, I focus my analysis on the Refinement stages, while Global Search results are provided in Appendix N. I consider the Global Search to be adequate if good geometric matches to the known forms are present relatively low on the landscape, and there are enough minima to form a comprehensive set of crystals for refinement.

9.1.2 Refinement I

Unique structures low on the Global Search landscape, ~10–20 kJ/mol from the global minimum (GM), are brought forward for refinement. Depending on the flexibility of the structure, either CSO–RM (Section 4.2.2) or CSO–FM (Section 4.2.3) is used. During Refinement I, the intermolecular model is based on rank 4 GDMA-derived multipoles ($r_{\text{H}}^{\text{GDMA}}$ = 0.650 Å), but does not include SCERP. For inputs to GDMA, charge densities are determined at the PBE0/6-311G(d,p) level of theory, using Gaussian. Different intermolecular models are tested at this stage, varying in: (1) their use of PCM, and (2) the FF paired with the electrostatic model. In all intermolecular calculations, I use real-space cutoffs of 15.0 Å and 10.0 Å for the FF and Ewald summations respectively, an Ewald truncation error of 1×10^{-6} kJ/mol, and a spline threshold of 10%. These are consistent with the settings used during the FF parameter estimation (Section 6.1). For flexible molecules, LAMs are constructed on-the-fly through QM calculations in Gaussian. LAM validity ranges are provided in Appendix N. It is important to note that CSO–FM separates the construction of the intramolecular and electrostatic LAMs (Figure 4.2, A1 and A3), hence distinct QM settings (level of theory, PCM, etc.) can be used in each surrogate model. For Refinement I, I use the same settings for both classes of LAM.

9.1.3 Refinement II

From Refinement I, the lowest 100–200 ranked structures undergo a secondary refinement by the SCERP induction model. The calculation and model settings applied are mostly similar to Refinement I, but the Refinement II QM settings require further attention.

In Refinement I, the intramolecular and electrostatic LAMs use identical QM models, but in Refinement II, this is not always the case. In the SCERP induction model, $\Delta\boldsymbol{Q}$ represent the response of the multipole moments to the crystalline electric field, relative to their static (unperturbed) state $\boldsymbol{Q}^{\text{static}}$. If PCM is applied during the electrostatic LAM generation, a portion of $\Delta\boldsymbol{Q}$ is inadvertently passed to $\boldsymbol{Q}^{\text{static}}$. From Equation 7.3.1, a different U_{ind} will be predicted for the interaction involving $(Q^{\text{static}})(\Delta Q)$ versus $(Q^{\text{static}} + dQ^{\text{PCM}})(\Delta Q - dQ^{\text{PCM}})$, where dQ^{PCM} is the implicit induction embedded into Q^{static} (and omitted from ΔQ) by PCM. On this basis, when constructing electrostatic LAMs for use in SCERP-coupled HAIEFF models, a *static QM environment* is required. This further ensures consistency with the electrostatic model definition used during the SCERP-compatible FF parameter estimation (Chapter 8). For neutral molecules, a static environment can be used during the intramolecular LAM construction without issue, and consistency between the intramolecular/electrostatic QM settings can be preserved. However, for amino acids, the zwitterionic state observed in their crystals is stable only within a polarizing environment. In this case, PCM is applied for the intramolecular LAMs, but omitted for the electrostatic LAMs. Other than the use of PCM, identical QM settings are employed for the two LAM surrogates. During Refinement II, I also test the different approximations for the SCERP-coupled flexible gradients (Section 7.4.2 and Table 7.6).

9.1.4 Structure Clustering and Matching

Following lattice energy minimization, minima corresponding to similar structures are clustered together to remove likely duplicates. Two structures are considered for clustering if their $|\Delta U_{\text{latt}}| \leq 1.0$ kJ/mol and $|\Delta\rho| \leq 20.0$ kg/m^3. Pairs that satisfy these cutoffs undergo further comparison by the COMPACK [242] algorithm. If an $\text{RMSD}_{15} \leq 0.20$ Å is achieved, the structures are deemed non-unique, and only the lower energy crystal is retained. The default CCDC Python API COMPACK settings are used during clustering. After clustering, matches to the known forms are identified. Experimental structures are selected from the CSD, with a preference for entries with low r-factors. CSP structures that (using COMPACK) return $\text{RMSD}_{20} \leq 0.80$ Å relative to the reference geometries are accepted as matches, per the threshold used during the Sixth Blind Test [48]. Slightly higher COMPACK distance/angle tolerances are used to facilitate generating RMSD_{20} scores. All COMPACK comparisons ignore hydrogen positions.

9.1.5 Local Minimizations

Specific information may be sought through independent lattice energy minimizations of individual crystal structures. This is done in three main cases:

1. Local minimizations of the experimental structures are conducted using CrystalPredictor, with the Global Search energy model. This supports the selection of the energy model (particularly the FF), by verifying that good RMSD_{20} scores can be achieved.
2. Local minimizations of the experimental structures are conducted with periodic DFT-D to obtain reference relative stabilities of the known forms. While experimental relative

stabilities may be reported in the literature, since the HAIEFF FFs are fit to static DFT-D data, an equivalent ground truth is more appropriate for benchmarking model accuracy.

3. To validate the HAIEFF-predicted relative stabilities of unobserved polymorphic forms, structures from Refinement II may be locally minimized by periodic DFT-D.

Minimizations done with periodic DFT-D are conducted in VASP, as described in Section 5.3.1. The reported energy, U_{vasp}, can be recast in HAIEFF terms according to

$$U_{\text{vasp}} = \sum_{i=1}^{Z}(U_{\text{inter},i} + \Delta U_{\text{intra},i}) + Z \times U_{\text{intra}}^{\text{gp}}, \tag{9.1.1}$$

and the corresponding polymorphic U_{latt} difference for a form n, relative to m, is given by

$$\Delta U_{\text{latt}}^{n,m} = \left(\frac{U_{\text{vasp}}}{Z}\right)_n - \left(\frac{U_{\text{vasp}}}{Z}\right)_m. \tag{9.1.2}$$

For brevity, when comparing a structure n to the global minimum (GM), this is denoted as $\Delta U_{\text{latt}}^{\text{GM}} = U_{\text{latt}}^{n} - U_{\text{latt}}^{\text{GM}}$.

9.2 Rigid Monohydrate: 2,6-Diamino-4(3H)-Pyrimidinone

9.2.1 Background and Early-Stage CSP

The first system I consider is the monohydrate of 2,6-Diamino-4(3H)-Pyrimidinone, shortened as MH-DAP (Figure 9.1). Three known forms [372–374] are reported for this system, which I denote MH-I, MH-II, and MH-III. Of the three forms, MH-I presents as a channel hydrate. Relative stabilities (experimental and DFT-derived) for MH-DAP were reported by Zhang [71]. Slurry experiments agree with static DFT-D predictions that MH-II should be the most stable form. However, at ambient conditions, experimental results suggest that MH-III should be more stable than MH-I, while static DFT-D computations reverse this stability order. Given that $\Delta U_{\text{latt}}^{\text{MH-III,MH-I}}$ from DFT-D is only about +2.4 kJ/mol, the differences in MH-I/MH-III stability could likely be attributed to temperature effects.

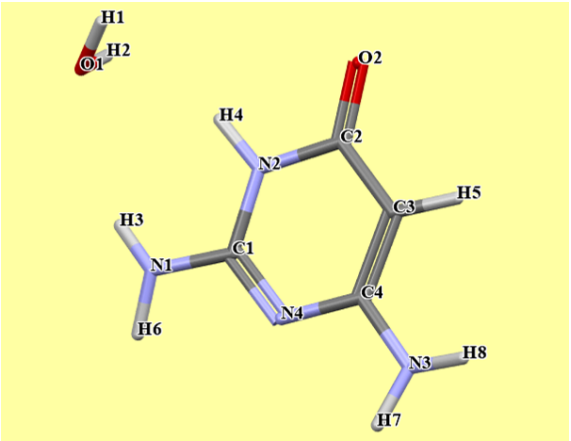


Figure 9.1: MH-DAP ($C_4H_6N_4O + H_2O$, 15 + 3 atoms). Both molecules are treated as rigid.

Table 9.1: Polymorphs of MH-DAP. *Z'/Z* include the active ingredient and water molecule. $RMSD_{15}$ values are obtained by comparing the starting and optimized crystals, and $\Delta U_{\text{latt}}^{\text{GM}}$ values are in kJ/mol.

Form	Refcode	Space group	*Z'/Z*	Relative Stability		
				Exp. [71]	Periodic DFT-D [71]	
					$RMSD_{15}$ (Å)	Ranking [$\Delta U_{\text{latt}}^{\text{GM}}$]
MH-I	SEYDIJ [372]	*Pbca*	2/16	Amb. 3rd	0.17	2nd [+0.70]
MH-II	SEYDIJ01 [373]	*P-1*	2/4	Amb. 1st	0.09	1st [GM]
MH-III	SEYDIJ03 [374]	*C2/c*	2/16	Amb. 2nd	0.14	3rd [+3.09]

Here, I opt to treat the MH-DAP system as rigid. In principle, the $-NH_2$ groups on the DAP molecule are not sterically hindered from rotating, relative to the aromatic ring. However, such rotations disrupt π-delocalization over the planar conformation, making them unfavorable. Comparing the conformation of DAP in MH-I through III, $RMSD_1$ values (without hydrogens) do not exceed 0.020 Å, suggesting that a rigid treatment of the DAP molecule is sufficient to find all forms of MH-DAP. Local minimization of the experimental forms under the rigid molecular model corroborates this (Table N.2). As such, I progress the CSP study of MH-DAP under the rigid-molecule approximation. The Global Search landscape (Figure N.1) comprises 890 structures within $\Delta U_{\text{latt}}^{\text{GM}} \leq 10.0$ kJ/mol. Within this landscape, relatively good geometry matches to all three MH-DAP forms are found, albeit with incorrect relative rankings.

9.2.2 Refinement Stage

For the MH-DAP system, I choose to conduct only a single refinement stage (rather than sequential refinements), i.e., *all* structures are refined with the SCERP induction model. This is possible given the relatively small number of Global Search structures, and the fact that the

Table 9.2: Models used in the refinement of MH-DAP. All QM calculations are done at the PBE0/6-311G(d,p) level of theory.

Model	Induction Model	HAIEFF FF	FF Parameters
1	None (Static)	FIT	Table I.1
2	PCM_{11}	FIT	Table I.1
3	None (Static)	$DB1_{\text{mp}}$	Table I.1 of Ref [2]
4	PCM_{11}	$DB14_{\text{mp}}$	Table I.14 of Ref [2]
5	None (Static)	$BT_{\text{dBhp,mp}}^{\text{Static}}$	Table J.18
6	PCM_{11}	$BT_{\text{dBhp,mp}}^{\text{PCM}}$	Table J.25
7	SCRP[a]	$BT_{\text{dBhp,mp}}^{\text{SCRP}}$	Table M.3
8	None (Static)	$BT_{\text{Mie,mp}}^{\text{Static}}$	Table J.20
9	PCM_{11}	$BT_{\text{Mie,mp}}^{\text{PCM}}$	Table J.27
10	SCRP[a]	$BT_{\text{Mie,mp}}^{\text{SCRP}}$	Table M.4

[a] The SCERP model uses: $\varepsilon_{\text{scrp}} = 1 \times 10^{-3}$ a.u., $r_{\text{scrp}} = 40.0$ Å, $t_{\text{slow}} = 6$, $t_{\text{max}} = 15$, $\varepsilon_{\text{scrpRM}} = 1 \times 10^{-3}$ kJ/mol, and $k_{\text{max}} = 10$, with electrostatic damping employing $\beta_{atm}^{tt'}$ derived from $\%f_6 = 90\%$.

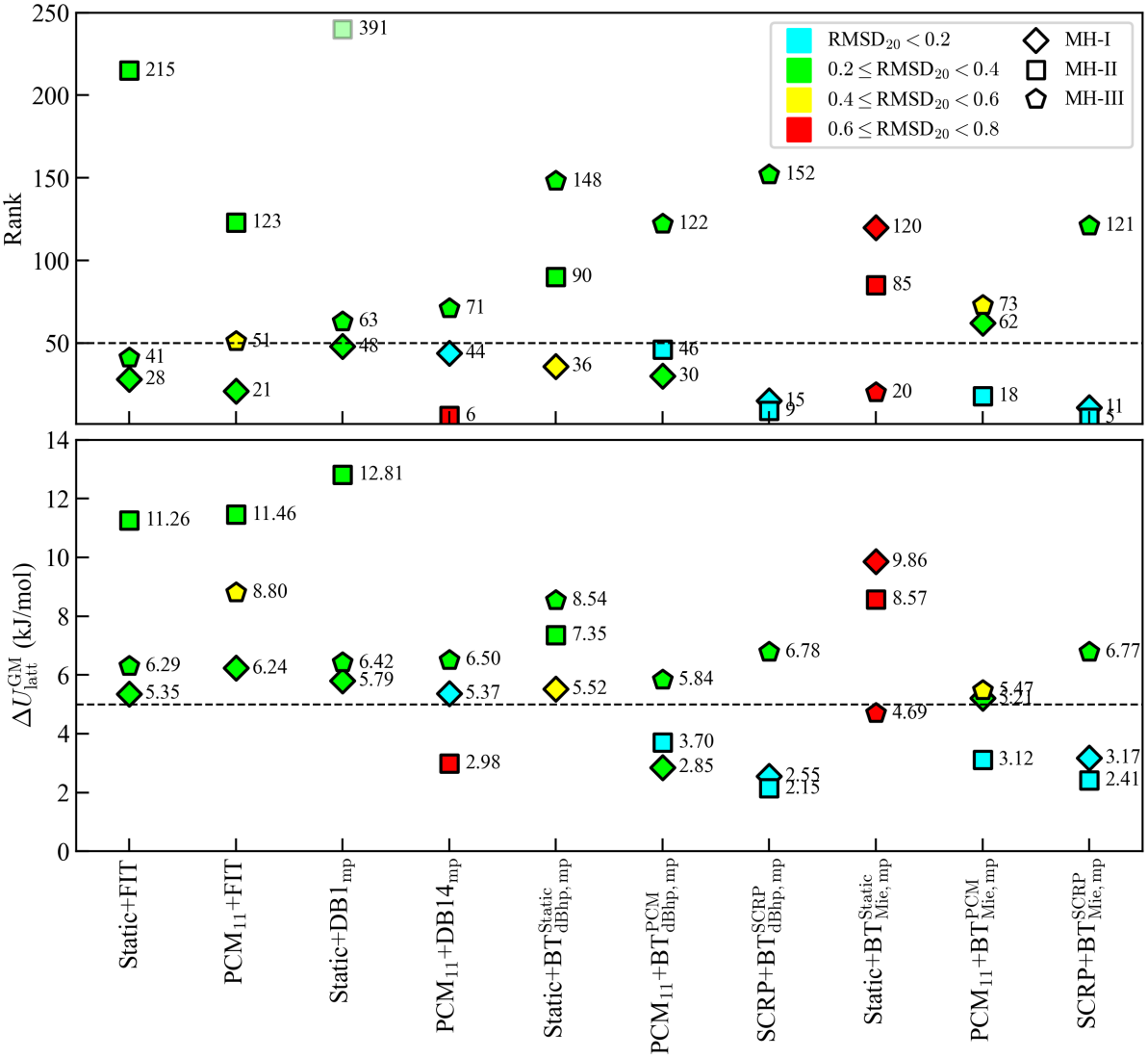


Figure 9.2: Summary of rank (top) and $\Delta U_{\text{latt}}^{\text{GM}}$ (bottom) for the MH-DAP matches using the refinement models in Table 9.2. MH-I, II, and III are given by ◇, □, and ⬠ respectively, while the marker fill colors reflect their RMSD_{20} scores (in Å).

MH-DAP system is treated rigidly. The 890 crystal structures from the Global Search stage undergo (independent) refinement with the models shown in Table 9.2. Following clustering and structure matching, the lowest-energy matches for each form are presented in Figure 9.2. Detailed refinement landscapes can be found in Appendix N.1.4

HAIEFF models based on the FIT FF [166, 198] are 'staples' of CSP [47–49], despite this FF being relatively dated. When applied to MH-DAP, the Static + FIT model finds all three polymorphic forms, with MH-I and MH-III ranked relatively low on the landscape ($\leq 50^{\text{th}}$). However, a glaring issue is that the most stable form, MH-II, is high on the landscape: ranked 215th with $\Delta U_{\text{latt}}^{\text{GM}} \approx 11.3$ kJ/mol. Adding PCM (PCM_{11} + FIT) does little to improve these predictions. While MH-II's rank decreases, its $\Delta U_{\text{latt}}^{\text{GM}}$ is mostly unchanged; moreover, results for MH-I and MH-III worsen. This contradicts results from other CSP studies [48, 188], and from this work, that suggest using PCM benefits hydrogen bonding systems. A likely explanation is

that the FIT FF was not parametrized with PCM [166, 198], hence induction-effects are already implicit in its FF parameters. Applying PCM double-counts these effects, leading to spurious errors in the model predictions. In any case, using either FIT-based model, it would be unlikely for MH-II to progress to the final re-ranking by periodic DFT-D.

Bowskill [2] had parametrized FFs (denoted DBx_{mp}) complementary to a range of QM models (including the Static and PCM_{11} model) under Scheme 1 of Table 6.2. His FFs are tested here, as a closer analogue to the FFs fit in the present work. Without PCM, the Static + $\mathrm{DB1}_{\mathrm{mp}}$ model performs poorly, predicting relative stabilities of MH-I, II, and III that are reminiscent of Static + FIT, but systematically worse. Notably, MH-II fails to be identified as the most stable form, and is extremely high on the landscape (391th and $\Delta U_{\mathrm{latt}}^{\mathrm{GM}} \approx 12.8$ kJ/mol). This is unsurprising: intermolecular energies for HB crystal structures were omitted from the fitting of the $\mathrm{DB1}_{\mathrm{mp}}$ FF. As a result, the Static + $\mathrm{DB1}_{\mathrm{mp}}$ model completely neglects all effects of polarization, which play an important role in hydrate systems [48, 71]. With the addition of PCM (PCM_{11} + $\mathrm{DB14}_{\mathrm{mp}}$), the model predictions improve considerably. The three MH-DAP forms exhibit correct relative stabilities, and MH-I and II are ranked relatively low on the landscape ($\leq 50^{\mathrm{th}}$). However, it is notable that the MH-II match only just makes the geometry cutoff (RMSD_{20} = 0.7881 Å). Advancing all three forms for periodic DFT-D would entail refining (minimally) 71 structures in a $\Delta U_{\mathrm{latt}}^{\mathrm{GM}}$ window of 6.50 kJ/mol, which are reasonable cutoffs for a CSP study.

Finally, I evaluate the models developed in the present work. Without induction, the use of these models ($\mathrm{BT}_{\mathrm{dBhp,mp}}^{\mathrm{Static}}$ and $\mathrm{BT}_{\mathrm{Mie,mp}}^{\mathrm{Static}}$) leads to matches for all three forms, but with erroneous relative stabilities. For both models, $\Delta U_{\mathrm{latt}}^{\mathrm{GM}} > 5$ kJ/mol for almost all matches, and MH-II is (wrongly) ranked 2nd among the three forms. While these models perform better than Static + $\mathrm{DB1}_{\mathrm{mp}}$, they are still unlikely to provide meaningful CSP results. Clearly, absorbing induction-errors into the FF parameters has improved model accuracy, but not to an extent that it eliminates the need to describe induction via U_{elec}. In that vein, the results demonstrate that using PCM ($\mathrm{BT}_{\mathrm{dBhp,mp}}^{\mathrm{PCM}}$ and $\mathrm{BT}_{\mathrm{Mie,mp}}^{\mathrm{PCM}}$) significantly enhances CSP outcomes. With the implicit induction model, MH-III consistently emerges as less stable than MH-I and II, although I and II are not necessarily correctly ranked (*cf.* $\mathrm{BT}_{\mathrm{dBhp,mp}}^{\mathrm{PCM}}$). Crucially, compared to their Static counterparts, the energy window containing all three forms decreases substantially when PCM is employed, e.g., $\Delta U_{\mathrm{latt}}^{\mathrm{MH\text{-}III,GM}}$ decreases from 9.86 → 5.47 kJ/mol, between (Static + $\mathrm{BT}_{\mathrm{Mie,mp}}^{\mathrm{Static}}$) → ($\mathrm{PCM}_{11}$ + $\mathrm{BT}_{\mathrm{Mie,mp}}^{\mathrm{PCM}}$). This corresponds to a decrease in the rank window from $122^{\mathrm{nd}} \rightarrow 73^{\mathrm{rd}}$. Overall, the PCM_{11} + BT models provide sound cutoffs for advancing towards periodic DFT-D refinement, with slightly lower $\Delta U_{\mathrm{latt}}^{\mathrm{GM}}$ than PCM_{11} + $\mathrm{DB14}_{\mathrm{mp}}$.

Substituting PCM with SCERP, the last two models ($\mathrm{BT}_{\mathrm{dBhp,mp}}^{\mathrm{SCRP}}$ and $\mathrm{BT}_{\mathrm{Mie,mp}}^{\mathrm{SCRP}}$) consistently predict accurate relative rankings for MH-I through III, with relative energy gaps in agreement with periodic DFT-D (Table 9.1). Compared to their PCM_{11} counterparts, the two more-stable forms, MH-I and II, fall much lower on the landscape (rank $\leq 15^{\mathrm{th}}$, $\Delta U_{\mathrm{latt}}^{\mathrm{GM}} \leq 3.2$ kJ/mol), but MH-III ascends slightly ($\Delta U_{\mathrm{latt}}^{\mathrm{GM}} \approx 6.8$ kJ/mol). Good geometry predictions are achieved using both PCM_{11} and SCERP, but the latter yields slightly lower RMSD_{20}, particularly for MH-I and III. Like PCM_{11}, progressing to periodic DFT-D from either SCERP-landscape would

realistically encapsulate all three MH-DAP.

Summarizing, models reliant on the FIT FF generally perform poorly and do not reliably improve through the use of PCM. Conversely, models that preserve complementary-FF deployment more reliably exhibit improvements in accuracy commensurate with the fidelity of the *ab initio* HAIEFF-component. For the DB and BT FFs, this means that applying PCM consistently gives superior results relative to the Static model. This is especially acute for models using the DB FFs, which completely omit induction without PCM. Progressing to SCERP, there appears to be diminishing returns from using the more-sophisticated induction model: while MH-I and II predictions improve, MH-III results deteriorate slightly. Ultimately, depending on the assessment criteria, one could make good arguments for using either SCERP or PCM_{11}.

9.2.3 SCERP-Embedded Algorithmic Performance

Before closing the study of MH-DAP, I analyze the algorithmic performance of the SCERP-embedded calculations, using the computations done with the Mie-FFs.

I first study the convergence behavior of CSO–RM minimizations with SCERP (Figure N.12). Over the 890 structures refined with SCRP + $BT^{SCRP}_{Mie,mp}$, 13 fail to reach the 1×10^{-3} kJ/mol convergence threshold within $k_{max} = 10$. Of the remaining 877 structures, an average of 6.35 CSO–RM minimization iterations are spent to reach convergence, including the $k = -1$ iteration. These statistics align with the results from Section 8.3.4.1 and Ref [71]. From $k = 0$ to 10, the average number of QM computations to converge the SCERP process also decays quickly. This reiterates the reduction in QM expenditure that stems from re-using $\boldsymbol{Q}^{total,pc}_{I,field,k}$. Among the 13 structures that fail to converge, 7 are likely to converge given additional CSO–RM iterations, while 6 present with oscillatory non-convergence (Figure N.13). Overall, the convergence of the rigid minimization appears adequately robust, with < 2% of structures failing to converge.

Next, I benchmark the computational cost of the energy minimizations (Table 9.3). For CSO–RM calculations, this entails two stages: (1) single-point QM calculations to determine SCF charge densities, and (2) the minimization of U_{inter} using the derived multipoles. For the Static and PCM_{11} models, each stage is done only once. In contrast, the SCERP model relies on multiple QM evaluations and U_{inter} minimizations. By the current protocol, the total cost

Table 9.3: Total CPU time (in secs) at each stage of a lattice energy minimization, averaged over all crystals minimized (3 experimental structures for periodic DFT-D, and 890 otherwise).

Stage	Static	PCM_{11}	SCERP	Periodic DFT-D
Minimization[a]	139.19	174.56	1,404.83	N/A
QM[b]	303.01	299.58	11,558.73	80,194.14[c]

[a] This includes the time for E04UFF minimization + Newton refinement in CSO–RM.

[b] This includes the time for QM calculations done *in vacuo* (static or PCM calculations), and QM calculations done in the SCERP field.

[c] Periodic DFT-D minimizations in VASP do not distinguish between the QM and optimization costs.

of a SCERP minimization is ~30x that of a Static or PCM_{11} evaluation, but about 15% of a minimization by periodic DFT-D.

This cost analysis should be caveated by a few things. Firstly, periodic DFT-D minimizations are atomistic, hence a fairer comparison would be against CSO–FM. From Table 9.3, one could infer that SCERP-embedded CSO–FM minimizations, which rely on recurrent CSO–RM calls, are likely to be more costly than periodic DFT-D. However, in my estimation of computing resources utilized, this seems far from true, leading me to believe that these CPU times are inaccurate, particularly for SCERP. Secondly, periodic DFT-D is used for minimization of only the experimental structures. If less favorable starting geometries were provided (e.g., the Global Search landscape), its CPU times would likely increase. Finally, I firmly believe that the high cost of the SCERP method is tied to the suboptimal use of external Gaussian calls. Each call requires Gaussian to read and write long I/O files, which is a highly inefficient approach to update the values of $\boldsymbol{Q}_{\text{field}}^{\text{total,pc}}$. A similar self-consistent protocol (SCRF) is applied for PCM, yet it is considerably more efficient, since its execution is self-contained to Gaussian.

9.3 Flexible Z' = 1 Neutral Molecule: Urea

9.3.1 Background and Early-Stage CSP

The second system I consider is the neutral organic molecule, urea (Figure 9.3), which has five reported forms [28]. F-I [375] is the stable form under ambient conditions, while F-III [376] and F-IV [377] form sequentially as pressure is increased. F-I was thought to convert into F-II at elevated temperature and pressure [378]. However Dziubek *et al.* [28] assert that this was F-IV mis-reported as a distinct phase and in the CSD there are no entries for F-II. F-V forms at even higher pressures (> 7 GPa [379]), but currently has no resolved crystal structure.

By local minimization of F-I, III, and IV with periodic DFT-D, F-III emerges as most stable at static conditions. This concurs with experimental reports suggesting F-III is stable at low temperatures [380], and could also be surmised by extrapolating the urea phase diagram proposed in Ref [28]. Notably, under static conditions, F-IV is unstable and interconverts into F-I (Figure N.14). The minimized F-IV crystal cannot match 20 molecules with its initial structure, but gives $RMSD_{20}$ = 0.216 Å when compared to F-I. Olejniczak *et al.* [381] note that

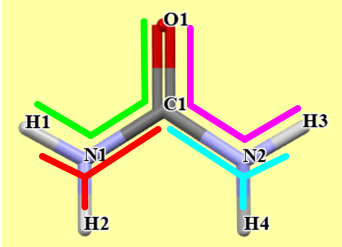


Figure 9.3: Urea (CH_4N_2O, 8 atoms). Four flexible torsions are defined: two proper (green and magenta), and two improper torsions (red and cyan).

Table 9.4: Polymorphs of Urea. $RMSD_{20}$ values are obtained by comparing the starting and optimized crystals, and $\Delta U_{\text{latt}}^{\text{GM}}$ values are in kJ/mol.

Form	Refcode	Space group	Z'/Z	Relative Stability		
				Exp. [28]	Periodic DFT-D	
					$RMSD_{20}$ (Å)	Ranking [$\Delta U_{\text{latt}}^{\text{GM}}$]
F-I	UREAXX25 [375]	*P-421m*	0.25/2	Amb. form	0.060	2nd [+2.65]
F-III	UREAXX33 [376]	*P212121*	1/4	> 0.48 GPa	0.212	1st [GM]
F-IV	UREAXX26 [377]	*P21212*	0.50/2	> 2.80 GPa	Interconverts to F-I	

F-I and IV are closely related, with similar hydrogen bonding networks. It is thus reasonable that without high pressure, the distorted packing of F-IV relaxes back to F-I.

Given the subtle differences in hydrogen bonding among F-I, III, and IV, I permit flexibility of the amine hydrogens via four flexible degrees of freedom (Figure 9.3). The flexible improper torsions help mitigate numerical issues in the LAM model that can arise from locally minimum positions of the hydrogens. Details of the LAM construction, and Global Search are summarized in Appendix N.2. The final Global Search landscape (Figure N.16) spans 1,909 structures in $\Delta U_{\text{latt}}^{\text{GM}} \leq 16.67$ kJ/mol, with good geometry matches to all three urea forms. Among these matches, F-IV is identified at the highest energy (1,079th, $\Delta U_{\text{latt}}^{\text{GM}} = 14.11$ kJ/mol).

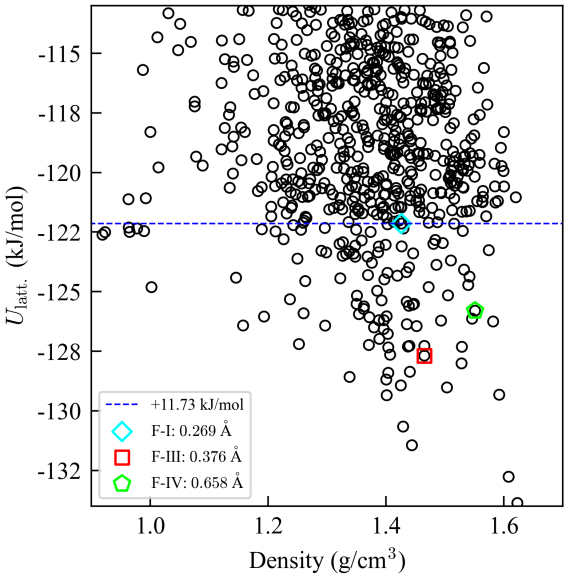
9.3.2 Refinement Stage

Refinement of the urea Global Search landscape is done in two stages. In Refinement I, the bottom 1,342 structures from the Global Search ($\Delta U_{\text{latt}}^{\text{GM}}$ = +15.0 kJ/mol) are independently evaluated using three energy models (Table 9.5, Models 1–3). After clustering, the bottom 200 structures from Model 3 advance to Refinement II for evaluation by the SCERP model (Table 9.5, Model 4). For all refinements, the same flexible definition as the Global Search (four flexible torsions) is used, and in Refinement II rigorous CSO–FM SCERP gradients (i.e., Table 7.6, Mode 4) are employed. Detailed summaries of the refinement calculations are tabulated in Appendix N.2.5, and the (post-clustering) refinement landscapes are shown in Figure 9.4.

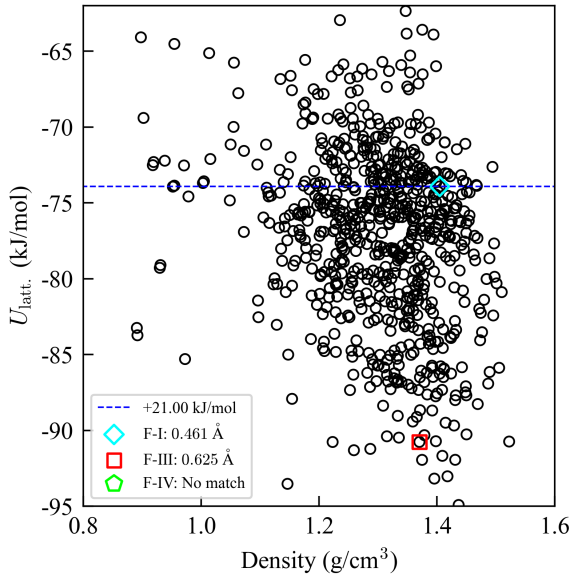
All Refinement I models (Models 1–3) correctly identify F-III as the most stable form among the known polymorphs, and it is well ranked on all three landscapes (< 20th). For the models utilizing the BT FFs (Models 2 and 3), $\Delta U_{\text{latt}}^{\text{GM}}$ for F-III are less than 5.0 kJ/mol, while for

Table 9.5: Models used in the refinement of urea. All QM calculations are done at the PBE0/6-311G(d,p) level of theory.

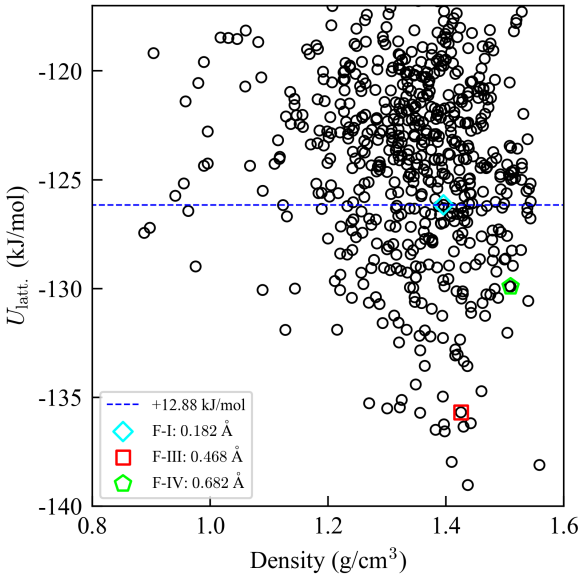
Model	Intramolecular Model	Intermolecular Model		
		Induction Model	HAIEFF FF	FF Parameters
1	PBE0/6-311G(d,p) + PCM_{11}	PCM_{11}	$DB14_{mp}$	Table I.14 of Ref [2]
2	PBE0/6-311G(d,p) + Static	None (Static)	$BT_{Mie,mp}^{Static}$	Table J.20
3	PBE0/6-311G(d,p) + PCM_{11}	PCM_{11}	$BT_{Mie,mp}^{PCM}$	Table J.27
4	PBE0/6-311G(d,p) + Static	SCRP	$BT_{Mie,mp}^{SCRP}$	Table M.4



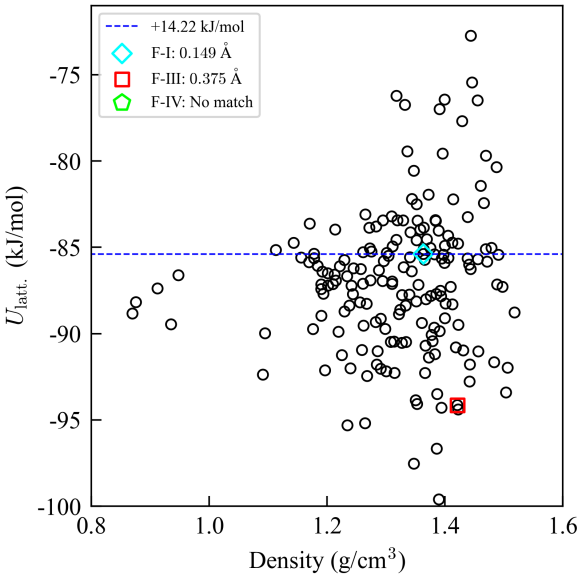
Form	$\mathbf{RMSD_{20}}$ (Å)	$\Delta\boldsymbol{U}_{\text{latt}}^{\text{GM}}$ (kJ/mol)	Rank
F-I	0.2688	+11.7251	111
F-III	0.3761	+6.1849	14
F-IV	0.6583	+8.0688	41



Form	$\mathbf{RMSD_{20}}$ (Å)	$\Delta\boldsymbol{U}_{\text{latt}}^{\text{GM}}$ (kJ/mol)	Rank
F-I	0.4605	+21.0021	505
F-III	0.6255	+4.1592	11
F-IV	No match found		



Form	$\mathbf{RMSD_{20}}$ (Å)	$\Delta\boldsymbol{U}_{\text{latt}}^{\text{GM}}$ (kJ/mol)	Rank
F-I	0.1817	+12.8766	182
F-III	0.4677	+3.3429	10
F-IV	0.6823	+9.1204	65



Form	$\mathbf{RMSD_{20}}$ (Å)	$\Delta\boldsymbol{U}_{\text{latt}}^{\text{GM}}$ (kJ/mol)	Rank
F-I	0.1490	+14.2193	124
F-III	0.3751	+5.4640	8
F-IV	No match found		

Figure 9.4: Refined landscape of urea using Models 1 (top-left), 2 (top-right), 3 (bottom-left), and 4 (bottom-right). The lowest-energy match to F-I, F-III, and F-IV are labeled as ◇, □, and ⬠ respectively, with their corresponding RMSD_{20} scores.

Model 1, which employs the DB FF, this gap is slightly larger at 6.1 kJ/mol. Given its excellent ranking on all three landscapes, there is a high likelihood that F-III would advance to a final re-ranking by periodic DFT-D using any of the Refinement I models.

In contrast, while F-I is found with good geometry accuracy by all three Refinement I models, its relative stability is poorly predicted. For the models employing PCM, F-I's $\Delta U_{\text{latt}}^{\text{GM}}$ is between 11.7–12.9 kJ/mol, corresponding to a rank of 111th and 182nd (Models 1 and 3 respectively). Omitting PCM, this worsen considerably, with the Model 2 F-I match exhibiting a $\Delta U_{\text{latt}}^{\text{GM}}$ of 21.0 kJ/mol and a rank of 505th. The abysmal result for F-I without PCM aligns with findings from the MH-DAP study. HAIEFF models that neglect induction (e.g., DB FFs without PCM), or rely solely on the FF to capture induction effects (e.g., BT FFs without PCM/SCERP) are inferior to models that embed induction into the electrostatic term. This is rational given that the formulation of the repulsion–dispersion FF (r^{-6}) is suboptimal for describing electrostatic interactions (r^{-n}, with $n < 6$). The F-I results stress that (implicit) induction models, like PCM, are crucial for achieving more-reasonable predictions in HB crystal structures. Moreover, crystal geometries tend to be more accurate with the PCM models (Models 1 and 3).

With that said, one is reminded that PCM is imperfect, as demonstrated by F-IV. Models 1 and 3 both incorrectly identify F-IV as a stable minimum on the static landscape, and even rank it as more-stable than F-I. This error may stem from PCM modeling polarization indifferently to the crystal's packing. Using PCM, the urea molecule in F-IV is likely over-polarized relative to what it should be given its packing geometry, and over-stabilization from induction is allowing F-IV to persist on the static landscape. Model 2, which omits PCM, avoids this over-polarization problem and extinguishes F-IV as a stable minimum.

The bottom 200 structures from Model 3, which include the three experimental matches, undergo refinement by Model 4. F-III is again well predicted with excellent relative rank (8th) and a reasonable $\Delta U_{\text{latt}}^{\text{GM}}$ (5.5 kJ/mol). Conversely, F-I continues to be predicted high on the landscape, ranked 124th with a slightly worse $\Delta U_{\text{latt}}^{\text{GM}}$ of 14.2 kJ/mol. Whether this is 'better' than the PCM models from Refinement I depends on the cutoff criteria: on one hand rankings improve over Model 3, yet $\Delta U_{\text{latt}}^{\text{GM}}$ values increase. Critically, F-IV is correctly eliminated as a stable minimum when using Model 4. With COMPACK, I study Cluster 65 from Model 3 and its corresponding cluster (Cluster 79) following refinement by Model 4 (Table 9.6). While the former matches F-IV, following Model 4 refinement Cluster 79 is more closely matched to F-I. This mirrors the behavior seen during periodic DFT-D local minimizations, and supports that packing-dependent polarization from SCERP avoids the over-stabilization of F-IV observed

Table 9.6: COMPACK comparison of refined crystals from Model 3 and 4, against F-I and F-IV.

Structure	$\text{RMSD}_{20,\text{F-I}}$ (Å)	$\text{RMSD}_{20,\text{F-IV}}$ (Å)
Model 3, Cluster 65[a]	1.020	0.682
Model 4, Cluster 79[b]	0.820	0.900

[a] This corresponds to the F-IV match found using Model 3.

[b] This corresponds to the F-IV match from Model 3, after refinement by Model 4.

with PCM. Likely, using a smaller ε_{PCM} would also eliminate F-IV from the landscape, but the benefit of SCERP is its ability to adapt the polarizing environment to the crystal packing.

9.3.3 Final Periodic DFT-D Re-ranking

The preceding results highlight the need to embed induction into the HAIEFF model. The PCM (Model 3) and SCERP (Model 4) refinements yield mostly similar CSP outcomes, but the latter has subtle indications of being more physically robust (removal of F-IV). As a final test of accuracy, the bottom 10 structures from the Model 4 landscape, plus structure 124 (the Model 4 F-I match), are re-ranked by periodic DFT-D. Each DFT-D optimized structure gives a small RMSD_{20} relative to its initial Model 4 structure (Table N.10), indicating that the SCERP model is predicting accurate stable geometries. Structures 8 and 124 from Model 4, the F-III and I match respectively, optimize to the same DFT-D minima as the calculations initialized from the experimental structures of F-III and I. In Figure 9.5a, relative lattice energies with respect to the F-III match ($\Delta U_{\text{latt}}^{\text{F-III}}$) are presented for periodic DFT-D, Model 4, and the corresponding Model 3 structures. In Figure 9.5b, the relative ranks among these 11 structures is also shown. From the $\Delta U_{\text{latt}}^{\text{F-III}}$ error metrics (AD, AAD, and RMSD), the PCM model (Model 3) exhibits slightly better agreement with the DFT-D relative energies. Conversely, the Pearson-r and Spearman-ρ scores both suggest better correlation with DFT-D is achieved by the SCERP model (Model 4). Given the small sample size, one should be wary about drawing strong inferences from these statistics. Visually, the parity plots do not clearly favor either HAIEFF model.

Summarizing, mixed success is achieved in the urea CSP study, for all HAIEFF models. Without PCM/SCERP, extremely poor predictions are obtained for F-I, while F-IV is correctly

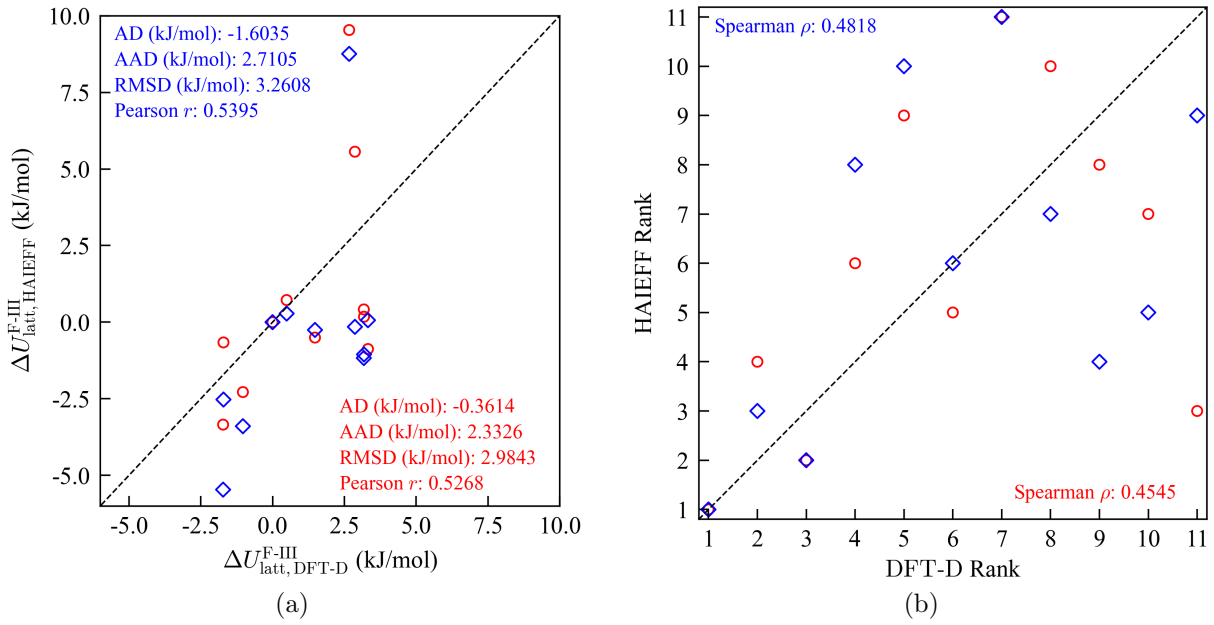


Figure 9.5: Comparison of (a) $\Delta U_{\text{latt}}^{\text{F-III}}$, and (b) ranks, predicted by the HAIEFF models versus periodic DFT-D. The bottom 10 structures + structure 124 from Model 4 are denoted by blue ◇, while their corresponding Model 3 results are denoted by red ○.

de-stabilized; conversely with PCM, F-I has more-reasonable predictions (albeit still poor quality), but F-IV is wrongly stabilized. The SCERP model seems to offer a preferable middle ground, wherein the performance for F-I is similar to the PCM models, while F-IV is correctly de-stabilized. Overall, however, the average accuracy of the SCERP model (relative to periodic DFT-D) appears to be (at best) only slightly superior to PCM, despite its much higher cost.

9.3.4 Assessment of SCERP Gradient Approximations

It is expensive to employ rigorous CSO–FM finite-difference (FD) gradients with Model 4, as done in the preceding refinement. To lower this cost, I consider the efficacy of the FD approximations proposed in Table 7.6. Local minimization of F-I, III, and IV is done using the four FD modes, with SCERP settings: $r_{\text{scrp}} = 40.0$ Å, $t_{\max} = 30$, $t_{\text{slow}} = 6$, $k_{\max} = 20$, $\varepsilon_{\text{scrp}} = 1 \times 10^{-5}$ a.u., and $\varepsilon_{\text{scrpRM}} = 1 \times 10^{-5}$ kJ/mol. CSO–FMv1.4.0.1 is employed, such that early-stopping is disabled; the default NAG optimality tolerance is used. Minimization results are summarized in Table 9.7, while minimization trajectories are presented in Appendix N.2.7.1.

In most cases, the final $F(\boldsymbol{x})$ of Modes 1–3 deviate from Mode 4 by ≤ 0.5 kJ/mol, suggesting that the approximate FD modes agree with the rigorous gradients. However, the dof values reveal that Modes 1–3 do not actually converge to an identical conformation as Mode 4, with x_n deviating by $\geq 5°$ in some cases. For Mode 1, this is observed in F-I, while Modes 2 and 3 exhibit this in F-IV. It is notable that Modes 2 and 3 share similar minimization trajectories and often converge in close proximity. This may be because both approximations differ only subtly in how they update $\Delta\boldsymbol{Q}$, but are alike otherwise.

At the zeroth iteration, the conformation is unaltered and $\|\boldsymbol{g}\|_0$ can be compared. Approxi-

Table 9.7: Minimization[a] of urea by FD gradient Modes 1, 2, 3, and 4. Converged runs are denoted by '*'; other runs terminate from walltime or E04UFF failing to progress. All values are from the last E04UFF major iteration, except $\|\boldsymbol{g}(\boldsymbol{x})\|_0$ which is from the zeroth iteration.

Form	Mode	$F(\boldsymbol{x})$ (kJ/mol)	$\|\boldsymbol{g}(\boldsymbol{x})\|_0$ (kJ/mol)	$\|\boldsymbol{g}(\boldsymbol{x})\|$ (kJ/mol)	x_1 (°)	x_2 (°)	x_3 (°)	x_4 (°)
F-I	1	−83.46	476.87	40.06	−7.69	−169.97	4.66	181.48
	2	−83.59	339.18	0.29*	−1.98	−168.33	−0.21	177.11
	3	−83.59	338.82	4.30	−1.23	−168.38	−1.27	176.04
	4	−83.51	368.99	1.75	−1.56	−168.39	−1.83	174.88
F-III	1	−78.21	341.39	3.91	10.45	174.20	−11.71	−165.07
	2	−77.76	199.38	21.21	11.74	172.07	−9.55	−167.43
	3	−77.88	202.35	15.49	10.61	173.75	−11.05	−166.11
	4	−77.92	247.21	3.22	11.16	173.16	−10.06	−166.23
F-IV	1	−84.19	476.23	5.62	−6.14	−166.38	7.14	−175.04
	2	−83.79	338.89	23.11	−1.49	−167.67	−0.38	176.71
	3	−83.23	339.53	91.98	−0.25	−167.83	−0.27	177.75
	4	−84.16	366.96	36.17	−7.29	−167.86	5.10	−178.69

[a] The minima found here may not match those found in Figure 9.4, likely due to different initialization.

mate gradients from Modes 1–3 are clearly distinct to Mode 4, with Mode 1 being most in error. Modes 2 and 3 predict very similar initial gradients, explaining their common trajectories. $F(\boldsymbol{x})$ reduces rapidly early in all optimizations, albeit by distinct $\boldsymbol{x}$ trajectories (e.g., Figure N.17b). In all but one case (Mode 2, F-I), progress plateaus before convergence can be reached, likely due to the precision limit of the SCERP-embedded CSO–RM minimization preventing the detection of a decreasing step. The converged case is likely coincidental, and not a validation of Mode 2. Using Mode 4, CSO–RM is also employed when evaluating $\boldsymbol{g}(\boldsymbol{x})$, hence these effects extend to the elements of $\boldsymbol{g}$ and noise emerges near the end of the $\|\boldsymbol{g}\|$ trajectory (e.g., Figure N.17a). Despite progress stalling, E04UFF may persist in making futile linesearches: this is most acute for Mode 1, which has markedly prolonged trajectories for F-I and IV. Ultimately, these observations justify the various strategies imposed in CSO–FMv1.4.0.2 (Section 7.4.3).

To study relative costs, I consider the metrics in Table 9.8. Firstly, the average number of linesearch attempts used per major iteration ($\frac{N_{\text{line}}}{N_{\text{maj}}}$) provides insight into the linesearch efficiency and gradient accuracy. Since each linesearch attempt entails a $F(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x})$ evaluation, lowering this average also reduces cost. Secondly, I consider the average number of CSO–RM and SCERP QM iterations expended per linesearch attempt (N_{CSORM} and N_{SCRP} respectively). For $\boldsymbol{g}(\boldsymbol{x})$ evaluations, N_{CSORM} and N_{SCRP} are averaged by the number of flexible dofs. Lastly, $\frac{N_{\text{SCRP}}}{N_{\text{CSORM}}}$ is the number of SCERP QM computations per CSO–RM iteration.

The accuracy of the FD gradients is expected to improve from Modes 1 to 4, and the $\frac{N_{\text{line}}}{N_{\text{maj}}}$ values reflect this. More-accurate gradients should enhance linesearch efficiency (i.e., a decreasing step is found in fewer search iterations), and correspondingly $\frac{N_{\text{line}}}{N_{\text{maj}}}$ tends to decrease from Modes 1 to 4. Unfortunately, the need to perform a minimization and/or SCERP process during the $\boldsymbol{g}(\boldsymbol{x})$ evaluation still scales more poorly than the linesearch efficiency, since the minimizations/QM calculations are executed in *both* the forward/backward step of *every* flexible dof. As such, Modes 3 and 4 remain more expensive than Modes 1 and 2. Nevertheless, all of

Table 9.8: Cost metrics of the urea local minimization by FD gradient Modes 1, 2, 3, and 4. Definitions for each metric are given in the main text.

Form	Mode	$\frac{N_{\text{line}}}{N_{\text{maj}}}$	$F(\boldsymbol{x})$ Evaluation			$\boldsymbol{g}(\boldsymbol{x})$ Evaluation		
			N_{CSORM}	N_{SCRP}	$\frac{N_{\text{SCRP}}}{N_{\text{CSORM}}}$	N_{CSORM}	N_{SCRP}	$\frac{N_{\text{SCRP}}}{N_{\text{CSORM}}}$
F-I	1	7.74	6.57	26.88	4.09	-	-	-
	2	2.06	6.86	27.26	3.97	-	-	-
	3	2.33	6.12	23.81	3.89	-	4.46	-
	4	2.65	5.27	15.14	2.87	6.43	11.38	1.77
F-III	1	2.54	9.93	37.08	3.73	-	-	-
	2	4.17	9.07	41.27	4.55	-	-	-
	3	3.57	6.80	28.31	4.16	-	4.82	-
	4	1.74	6.95	29.57	4.25	9.06	24.32	2.68
F-IV	1	6.39	6.36	24.49	3.85	-	-	-
	2	5.33	5.07	18.83	3.71	-	-	-
	3	3.19	6.60	26.31	3.99	-	4.51	-
	4	2.86	5.62	21.95	3.91	8.14	18.26	2.24

Table 9.9: Convergence behavior during urea Refinement II, using FD Modes 2–4.

Mode	Average N_{maj}	Average N_{line}	Average $\frac{N_{\text{line}}}{N_{\text{maj}}}$	No. Converged [%]
2	7.6	30.4	4.7	21 [10.5%]
3	10.7	28.8	3.2	61 [30.5%]
4	15.5	29.4	2.1	150 [75.0%]

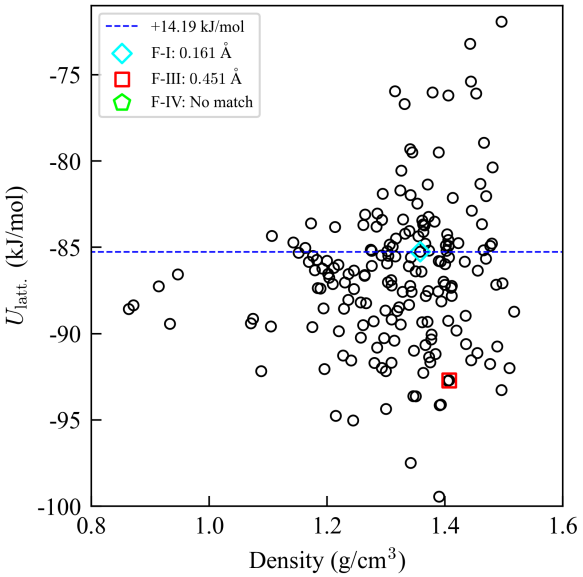
the FD approaches prove to be practicable on the scale of structures refined here. From $\frac{N_{\text{SCRP}}}{N_{\text{CSORM}}}$, SCERP process convergence is reached as quickly as (if not faster than) in previous tests of CSO–RM and the SCERP process. This is despite the tighter $\varepsilon_{\text{scrp}}$ (1×10^{-5} a.u.) used in tandem with CSO–FM. Most likely, the reuse of $\boldsymbol{Q}_{I,\text{field},*}^{\text{total,pc}}$ between CSO–RM minimizations (Section 7.4.1) is a catalyst for enabling SCERP-embedded CSO–FM, especially Modes 3 and 4, which rely heavily on QM computation. Overall, these results suggest no reason to use Mode 1: it has comparable costs to Mode 2, but is considerably less accurate (based on $\|\boldsymbol{g}(\boldsymbol{x})\|_0$). Modes 2 and 3 exhibit similar accuracies despite their distinct approaches to update $\boldsymbol{\Delta Q}$. Costs rise from Modes 2 to 4, as expected, but remain practicable due to the strategies implemented in CSO–FM.

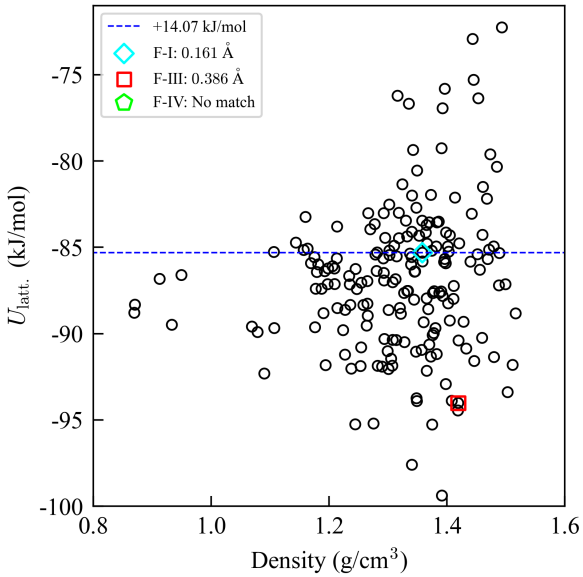
From local minimizations alone, it is hard to separate the performance of Modes 2–4. As such, I redo Refinement II using Modes 2 and 3, and compare their predictions to Mode 4 (Figure 9.4, Model 4). Unlike the local minimizations, early-stopping is applied together with a more-sensible τ_{nag} (Table N.9). The convergence behavior of the three FD modes is given in Table 9.9 (and Figures N.20–N.22). From the refinement, clearer distinctions in performance emerge. As before, average $\frac{N_{\text{line}}}{N_{\text{maj}}}$ decreases from Mode 2 to 4, consistent with the intuited gradient accuracies. Both N_{maj} and convergence rates are the lowest when using Mode 2, suggesting that its less-accurate gradients are leading to earlier termination on the basis of early-stopping. Modes 3 and 4 progressively exhibit better convergence behavior.

Finally, I assess the ability of Modes 2 and 3 to match the Mode 4 landscape (Table 9.10 and Figure N.23). Mode 3 should have more-accurate gradients than Mode 2, hence the former tends to make more progress in the minimization and is less susceptible to early-stopping (per Table 9.9). As such, compared to Mode 2, Mode 3 has smaller energy/geometry error metrics and closer alignment with Mode 4. The tendency for Mode 3 to advance further in the minimization is also evidenced by its less positive U_{latt} AD (relative to $U_{\text{latt}}^{\text{Mode 4}}$). While the clustered landscapes of Modes 2 and 3 (Figure 9.6) resemble Mode 4 (Figure 9.4, bottom-right), the experimental matches reveal subtle discrepancies: the Mode 2 F-III match deviates by > 1.4

Table 9.10: Optimized U_{latt} and geometries for urea predicted by Modes 2 and 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$, while RMSD_{20} scores are generated by the COMPACK algorithm. All comparisons are done on a structure-by-structure basis.

Mode	AD U_{latt} (kJ/mol)	AAD U_{latt} (kJ/mol)	No. $\lvert\Delta U_{\text{latt}}\rvert$ > 0.5 kJ/mol	Avg. RMSD_{20} (Å)	Max. RMSD_{20} (Å)	No. RMSD_{20} > 0.5 Å
2	0.281	0.431	47	0.062	1.031	2
3	0.262	0.373	31	0.050	0.733	2





Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
F-I	0.1613	+14.1912	121
F-III	0.4511	+6.7441	12
F-IV	No match found		

Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
F-I	0.1612	+14.0699	127
F-III	0.3856	+5.3631	7
F-IV	No match found		

Figure 9.6: Refined landscape of urea using Model 4, with Mode 2 (left) and Mode 3 (right) FD gradient approximations. The lowest-energy match to F-I, F-III, and F-IV are labeled as ◇, □, and ⬠ respectively, with their corresponding $RMSD_{20}$ scores.

kJ/mol in U_{latt} from the Mode 3 and 4 match. Modes 3 and 4 have only minor variation between their experimental matches.

Ultimately, to reduce the SCERP model's cost, the Mode 3 FD gradient approximation is a promising choice, since it most closely reproduces the rigorous gradient results; Mode 2, whilst cheaper, lacks reliability. These FD approximations continue to be tested below.

9.4 Flexible Z' = 1 Zwitterion: Glycine

9.4.1 Background and Early-Stage CSP

The third system I study is the zwitterionic amino acid, glycine (Figure 9.7), for which 6 experimental forms have been reported [27]. At ambient conditions, α [382], β [383], and γ [384] are observed, with relative stabilities [385–389]: γ (most stable) > α > β. α/γ are purportedly enantiotropically related, with α favored at elevated temperatures (> 400 K) [385–389], suggesting that the two forms are thermodynamically competitive. At elevated pressures, δ (> 0.8 GPa) and ε (> 1.9 GPa) are observed [390]. A sixth form, β' [391], was also reported at > 0.8 GPa, however COMPACK comparison with δ suggests that the two forms are identical ($RMSD_{50}$ = 0.1296 Å). As such, β' is not considered here.

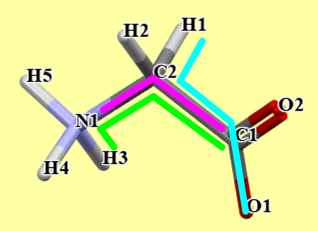


Figure 9.7: Glycine ($C_2H_5NO_2$, 10 atoms). Three flexible dofs are defined: two proper torsions (green and cyan), and one bond angle (magenta).

By periodic DFT-D local minimization, relative stabilities at static conditions are determined as: α (most stable) > γ > β > δ > ε. Xavier *et al.* [392] performed similar DFT-D calculations on the ambient forms and also reported a reversed α/γ stability order under static conditions. The $\Delta U_{\text{latt}}^{\gamma,\alpha}$ found here (0.89 kJ/mol) and in Ref [392] (0.81 kJ/mol) concur, supporting the possibility of enantiotropic behavior between α/γ. Lastly, the relative stabilities of the high pressure forms conform to the extent of pressurization required to observe them.

Because CrystalPredictor permits only torsions to be flexible, during the Global Search I treat two torsions flexibly, to allow the rotation of the hydrogen bonding moieties: the carboxylate and amino group. Details of the LAM construction, and Global Search are summarized in Appendix N.3. When studying zwitterionic systems, like amino acids, the use of PCM (ε_{PCM} = 11) is necessary to preserve the zwitterionic state during isolated-molecule LAM QM calculations. The final Global Search landscape (Figure N.25) spans 4,702 structures in $\Delta U_{\text{latt}}^{\text{GM}} \leq 19.85$ kJ/mol, with good geometry matches to all five glycine forms. Among these matches, the highest-energy match corresponds to ε-glycine (701st, $\Delta U_{\text{latt}}^{\text{GM}}$ = 9.86 kJ/mol).

9.4.2 Refinement Stage

Refinement of the glycine Global Search landscape is done in two stages. In Refinement I, the bottom 3,000 structures from the Global Search ($\Delta U_{\text{latt}}^{\text{GM}}$ = +16.67 kJ/mol) are independently evaluated by three energy models (Table 9.12, Models 1–3). After clustering, the bottom 200 structures from Model 3 advance to Refinement II for evaluation by the SCERP model (Table 9.12, Model 4). In both refinement stages, a bond angle (N_1−C_2−C_1) is added to the set of

Table 9.11: Polymorphs of Glycine. $RMSD_{20}$ values are obtained by comparing the starting and optimized crystals, and $\Delta U_{\text{latt}}^{\text{GM}}$ values are in kJ/mol.

Form	Refcode	Space group	Z'/Z	Relative Stability		
				Exp.	Periodic DFT-D	
				[385–389]	$RMSD_{20}$ (Å)	Ranking [$\Delta U_{\text{latt}}^{\text{GM}}$]
α	GLYCIN86 [382]	*P21/n*	1/4	Amb. 2nd	0.066	1st [GM]
β	GLYCIN32 [383]	*P21*	1/2	Amb. 3nd	0.037	3rd [+2.21]
γ	GLYCIN16 [384]	*P32*	1/3	Amb. 1st	0.063	2nd [+0.89]
δ	GLYCIN67 [390]	*P21/a*	1/4	> 0.80 GPa	0.179	4th [+2.56]
ε	GLYCIN68 [390]	*Pn*	1/2	> 1.90 GPa	0.256	5th [+3.45]

$\boldsymbol{\theta}^{\text{ind}}$ used during the Global Search (i.e., a total of two flexible torsions and one flexible bond angle). In Refinement II, rigorous CSO–FM SCERP gradients (i.e., Table 7.6, Mode 4) are employed. Detailed summaries of the refinement calculations are tabulated in Appendix N.3.4 and the (post-clustering) refinement landscapes are shown in Figure 9.8.

Using the Refinement I models, all forms of glycine are consistently found with good geometry accuracy: RMSD_{20} scores do not exceed 0.46 Å for any match to the known forms. On average, the BT FFs do predict better geometries than the DB FF, but the difference is marginal. In terms of predicted U_{latt}, the α, β, and δ forms are strongly favored by all three models, such that they are within 4 kJ/mol of the GM and ranked below 30th. For these three forms, Model 1 provides the most accurate predictions: the three forms are in the bottom 10 ranked structures and α is correctly ranked as the most-stable of the three. Although with Model 1 the stabilities of β and δ are reversed, ΔU_{latt} values from periodic DFT-D (Table 9.11) indicate that they are extremely close in energy, so this is reasonable. Models 2 and 3 both suggest a relative stability of $\beta > \alpha > \delta$, and in fact β is the GM of the Model 3 landscape. These models use the dBHP and MIP FFs parametrized in the present work, which share a similar fitting approach, hence it is unsurprising that their CSP results are alike. Ultimately, as a precursor to re-ranking by DFT-D, any of these Refinement I models is likely to find and advance the α, β, and δ forms.

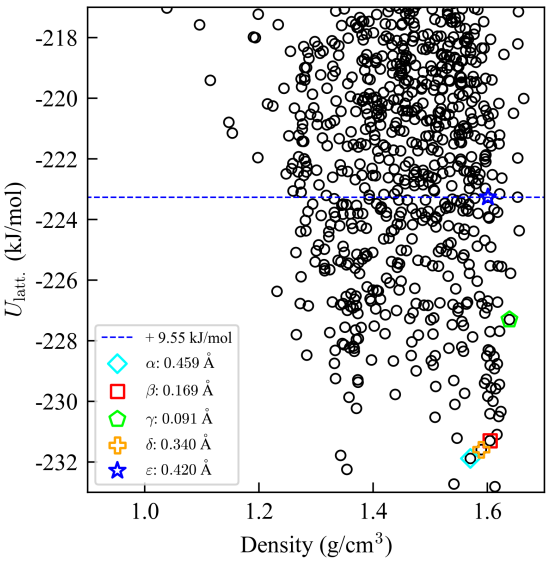
Unfortunately, the Refinement I models' predictions for the other two forms are below par: γ and ε are separated from α, β, and δ by ≥ 4 kJ/mol, and exhibit $\Delta U_{\text{latt}}^{\text{GM}} \geq 5.5$ kJ/mol. While Model 1's γ match is less distal from α, β, and δ, in any case, upwards of 187 crystals (Model 3) must advance to periodic DFT-D re-ranking for all five forms to be present.

Advancing to Refinement II, the bottom 200 structures from Model 3 are refined by Model 4. With the higher-fidelity induction model, it was hoped that the errors pertaining to the γ and ε forms could be resolved; from Figure 9.8, this is not the case. On one hand, the α, β, and δ forms continue to be well-predicted, with $\Delta U_{\text{latt}}^{\text{GM}} \leq 5.5$ kJ/mol and relative ranks $< 30^{\text{th}}$. Model 4 does rectify the mis-ranking of α by Model 3 and (like Model 1) predicts $\alpha > \delta > \beta$. δ and β are incorrectly ranked, but are energetically competitive. On the other hand, predictions for γ- and ε-glycine severely degrade, with their $\Delta U_{\text{latt}}^{\text{GM}}$ ballooning to > 61 kJ/mol!

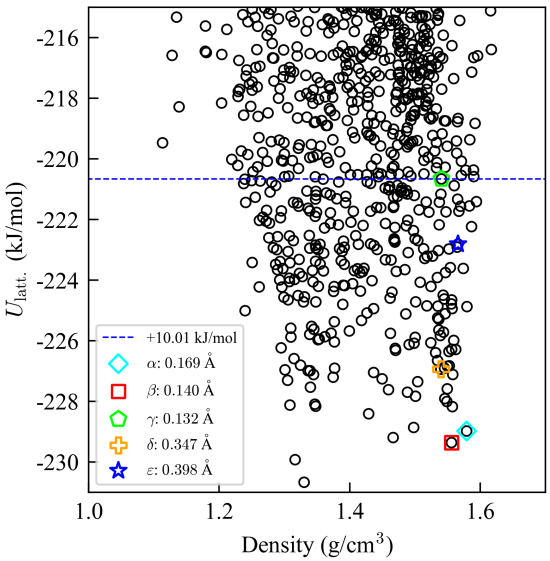
The systematic failure of all four models to describe γ and ε suggest a more fundamental error with the physical model. Clarity is offered by Hartman *et al.* [226], who employed a charge-embedding protocol (similar to SCERP) to model polarization in the prediction of NMR

Table 9.12: Models used in the refinement of glycine. All QM calculations are done at the PBE0/6-311G(d,p) level of theory.

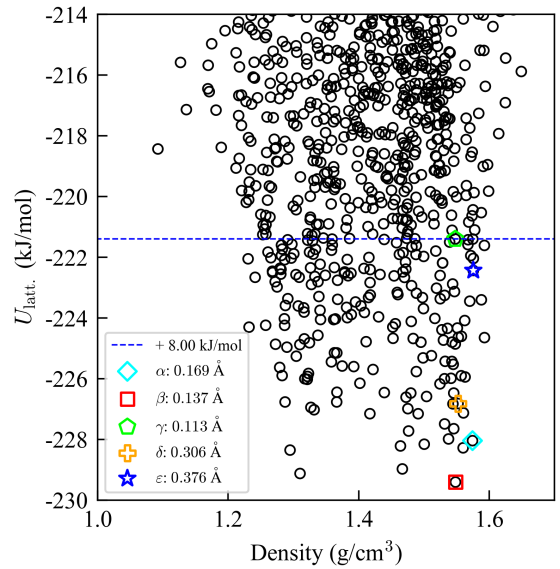
Model	Intramolecular Model	Intermolecular Model		
		Induction Model	HAIEFF FF	FF Parameters
1	PBE0/6-311G(d,p) + PCM_{11}	PCM_{11}	DB14_{mp}	Table I.14 of Ref [2]
2	PBE0/6-311G(d,p) + PCM_{11}	PCM_{11}	$\text{BT}_{\text{dBHp,mp}}^{\text{PCM}}$	Table J.25
3	PBE0/6-311G(d,p) + PCM_{11}	PCM_{11}	$\text{BT}_{\text{Mie,mp}}^{\text{PCM}}$	Table J.27
4	PBE0/6-311G(d,p) + PCM_{11}	SCRP	$\text{BT}_{\text{Mie,mp}}^{\text{SCRP}}$	Table M.4



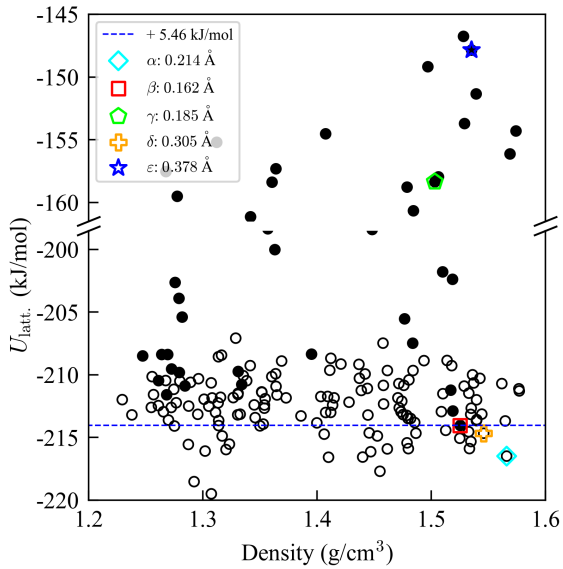
Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
α	0.4592	+0.9309	4
β	0.1688	+1.5169	7
γ	0.0909	+5.5102	76
δ	0.3399	+1.2220	6
ε	0.4196	+9.5491	260



Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
α	0.1686	+1.6905	5
β	0.1401	+1.3084	3
γ	0.1316	+10.0052	254
δ	0.3468	+3.7285	29
ε	0.3975	+7.8561	159



Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
α	0.1686	+1.3629	8
β	0.1366	+0.0000	1
γ	0.1125	+8.0017	187
δ	0.3062	+2.5853	20
ε	0.3759	+6.9676	138



Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
α	0.2135	+3.0070	6
β	0.1622	+5.4583	29
γ	0.1848	+61.1333	174
δ	0.3049	+4.8305	21
ε	0.3783	+71.6423	185

Figure 9.8: Refined landscape of glycine using Models 1 (top-left), 2 (top-right), 3 (bottom-left), and 4 (bottom-right). The lowest-energy match to α, β, γ, δ, and ε are labeled as ◇, □, ⬠, ✙, and ☆ respectively, with their corresponding $RMSD_{20}$ scores. In the Model 4 plot, crystals belonging to polar space groups are denoted by •.

chemical shifts of γ-glycine. The authors stressed that a simple charge-embedding scheme based on a finite-embedding radius is inadequate to describe γ-glycine, as it contains highly-charged zwitterionic molecules in a polar unit cell. In non-centrosymmetric unit cells, molecular dipole moments do not cancel out, resulting in the unit cell carrying a non-zero dipole moment — these are termed 'polar unit cells' [228–230]. In the infinite lattice, the dipole moments of polar unit cells can align and act cooperatively to incite extremely long-ranged electrostatic effects, like induction. To account for this long-ranged polarization, Hartman *et al.* [226] proposed a more sophisticated embedding scheme, termed SCRMP, that combines long-range summation techniques (i.e., Ewald sums) with auxiliary charges (on top of the SCERP atomic-charges).

The current SCERP model does not account for this form of long-range polarization, explaining the marked shortfall in stabilization observed for γ-glycine when using Model 4. In fact, from Figure 9.8, it can be seen that this erroneous behavior extends to *all* glycine structures belonging to polar unit cells (including ε-glycine), and structures with $\Delta U_{\text{latt}}^{\text{GM}} \geq 15$ kJ/mol are exclusively polar crystals. Going further, I return to the outliers noted throughout Chapters 7 and 8. From Table 9.13, these outliers all belong to polar space groups, and in cases where a related form is non-polar (e.g., KUXPUP/KUXPUP01), only the non-polar form exhibits reasonable behavior using SCERP. As a final verification of this fault, I modify CSO–RMv1.5.2.3 to output the ESP associated with the real-space term of the Ewald sum at atomic sites. To mimic the SCERP-field, I use only an atomic-charge model during these evaluations. I present the ESPs for the ambient forms of glycine in Figure 9.9. For the non-polar form (α) the ESP plateaus within the range of r considered, whilst for the polar forms (β and γ) the ESPs continue to diverge even up to r = 150 Å. This supports the idea that with the current SCERP-embedding protocol, polarization in polar unit cells is likely under-represented. β diverges less aggressively than γ, which could explain why it exhibits only minor error with SCERP, despite also belonging to a polar unit cell. Based on whether the unit cell is polar or not, this divergent/convergent trend is also observed for δ-/ε-glycine (Figure N.26) and KUXPUP/KUXPUP01 (Figure N.27).

Altogether, there is compelling evidence that the current SCERP model has a serious deficiency in describing these types of systems, particularly if a polar unit cell contains highly-polarized/charged molecules. Without more-elaborate charge-embedding (e.g., SCRMP [226]), polarization from the crystal environment is spuriously represented by SCERP. Ironically, the

Table 9.13: Polar unit cell classification of several crystal structures.

Structure	Space Group	Polar?
KUXPUP	*P21/c*	×
KUXPUP01	*Pna21*	✓
CYCYPR	*R3c*	✓
CLTRZL	*Pca21*	✓
GEYSEI	*Pca21*	✓
EREFEP	*Cc*	✓

Structure	Space Group	Polar?
GLYCIN86 (α)	*P21/n*	×
GLYCIN32 (β)	*P21*	✓
GLYCIN16 (γ)	*P32*	✓
GLYCIN67 (δ)	*P21/a*	×
GLYCIN68 (ε)	*Pn*	✓

[a] None of the known forms of MH-DAP or urea exist in polar unit cells, hence the analysis done in their respective CSPs is unaffected by this phenomenon.

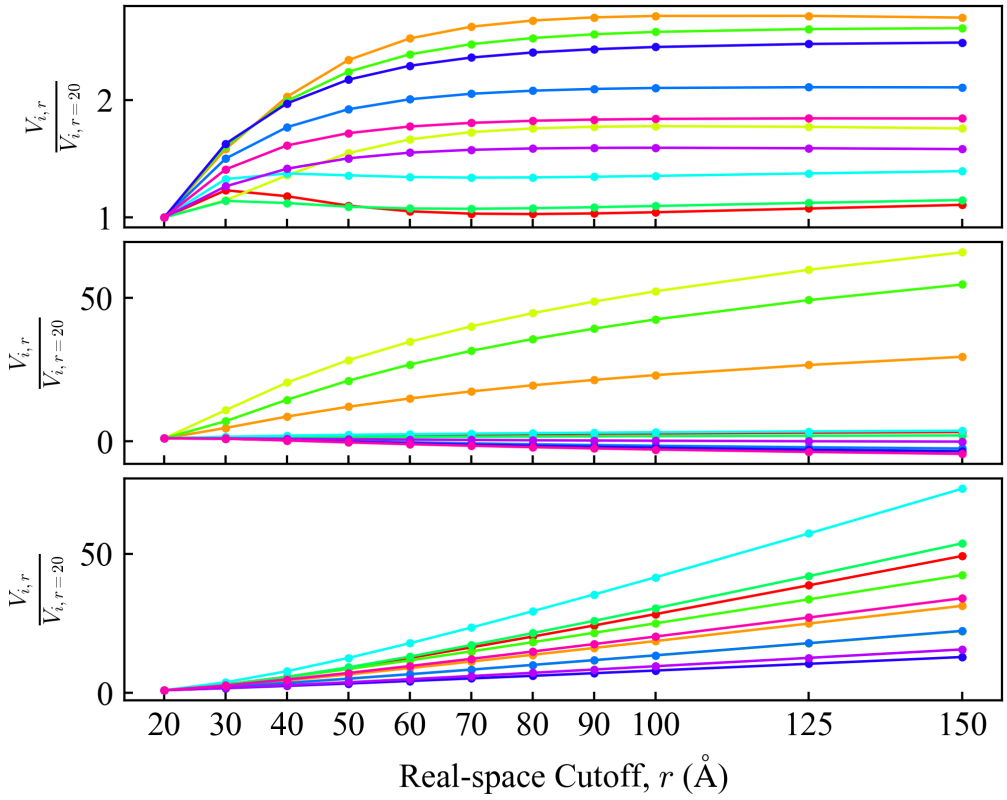


Figure 9.9: ESP exerted on atom i of α- (top), β- (middle), and γ-glycine (bottom) at different real-space cutoffs. All ESPs are reported as a ratio with respect to the ESP at r = 20 Å.

PCM model benefits from ignoring the crystalline lattice when describing polarization, such that there is a smaller relative error between polar and non-polar crystals. Nevertheless, predictions for the polar forms of glycine are inadequate with both induction models, but only the SCERP methodology offers a systematic route to improve model fidelity.

9.4.3 Final Periodic DFT-D Re-ranking

As before, I compare the accuracy of the PCM (Model 3) and SCERP (Model 4) models, to periodic DFT-D. The bottom 6 structures from Model 4 (6th: α match), plus structures 21 (δ match), 29 (β match), 174 (γ match), and 185 (ε match), are re-ranked by periodic DFT-D.

The DFT-D optimized structures continue to give good RMSD_{20} scores when compared to their initial Model 4 crystal structures (Table N.18). Also, the five experimental matches from the Model 4 landscape 'return' to the optima identified during the DFT-D minimization of the experimental structures. Overall, the SCERP model is at least capable of reliably predicting stable crystal geometries. This also confirms that structures 174 and 185 from Model 4, γ and ε respectively, should have much lower U_{latt} and the SCERP model is deficient. In Figure 9.10, relative lattice energies with respect to α ($\Delta U^{\alpha}_{\text{latt}}$) and relative ranks are presented for all structures, except 174 and 185. The latter two are omitted to avoid skewing the analysis. By DFT-D, the α form is more stable than the other 7 forms by > 1.8 kJ/mol. Conversely, Models 3 and 4 predict at least 5 structures as more stable than α-glycine. Whilst the error

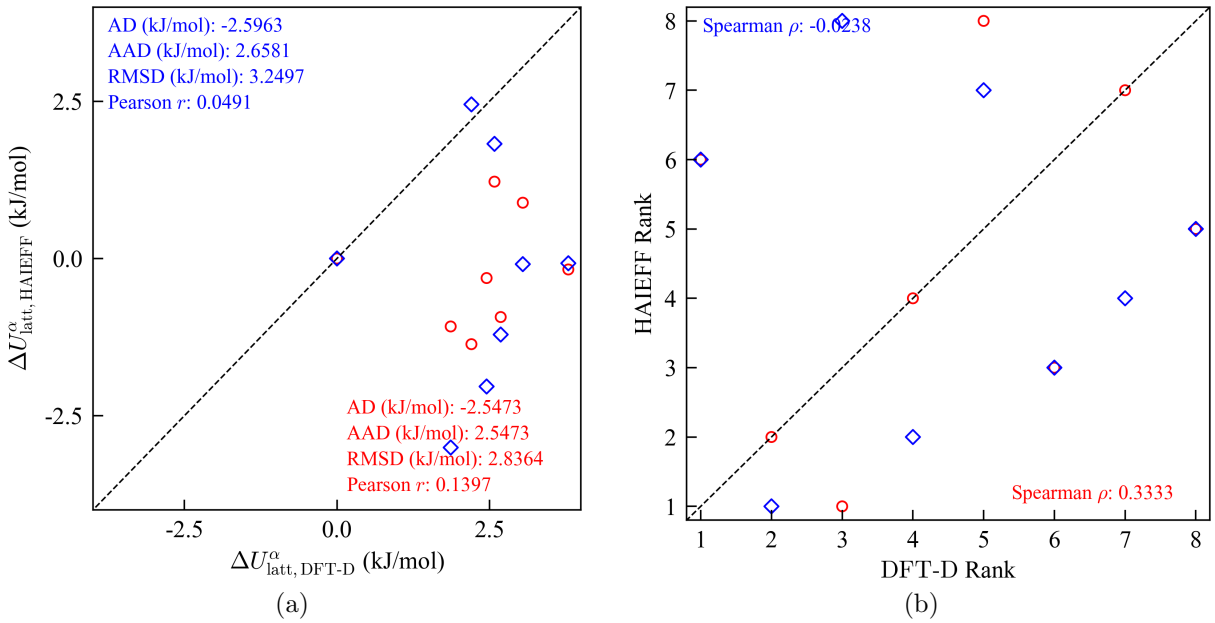


Figure 9.10: Comparison of (a) $\Delta U^{\alpha}_{\text{latt}}$, and (b) ranks, predicted by the HAIEFF models versus periodic DFT-D. The bottom 6 structures + structures 21 and 29 from Model 4 are denoted by blue ◇, while their corresponding Model 3 results are denoted by red ○.

metrics and correlation scores favor the PCM model (Model 3), visually, the parity plots reveal no meaningful distinction in prediction quality between the two models.

To summarize, both the urea and glycine CSP studies reiterate that HAIEFF models are adequate precursors to DFT-D re-ranking, but fall short of DFT-D accuracy. Even with the more-sophisticated SCERP model, predictions seem to improve only on a case-by-case basis, while overall model accuracy is mostly on par with PCM. This may indicate that the bottleneck to HAIEFF accuracy remains elsewhere in the model (e.g., the transferable FF). Importantly, the glycine CSP reveals a critical flaw in the SCERP (and to a lesser extent PCM) model: polar unit cells. Development of the SCERP approach according to Ref [226] may resolve this, while the PCM model lacks an analogous developmental 'roadmap'.

9.4.4 Assessment of SCERP Gradient Approximations

Building on the analysis done for urea, Refinement II is repeated for glycine using Model 4 (Table N.17) and CSO–FMv1.4.0.2. This is done with FD Modes 2 and 3, and compared to the results obtained with Mode 4. The convergence behavior of the three approaches is tabulated in Table 9.14, with minimization trajectories presented in Figures N.28–N.30. From Mode 2 to 4, as FD gradient fidelity improves, the propensity for early-stopping falls (N_{maj} increases), linesearch efficiency improves ($\frac{N_{\text{line}}}{N_{\text{maj}}}$ decreases), and rate of convergence increases. Whilst these trends are consistent with results from the urea CSP, the convergence rates observed for glycine are significantly lower than for urea. There are a few possible reasons for this.

Firstly, the E04UFF optimality tolerance used in Refinement II is decided by taking a margin of error from $\tau_{\text{nag,min}}$ (Equation 7.4.14). This approach is highly approximate, and may suggest

Table 9.14: Convergence behavior during glycine Refinement II, using FD Modes 2–4.

Mode	Average N_{maj}	Average N_{line}	Average $\frac{N_{\text{line}}}{N_{\text{maj}}}$	No. Converged [%]
2	3.8	19.6	6.1	8 [4.0%]
3	5.9	21.4	4.7	36 [18.0%]
4	9.3	26.4	3.1	99 [49.5%]

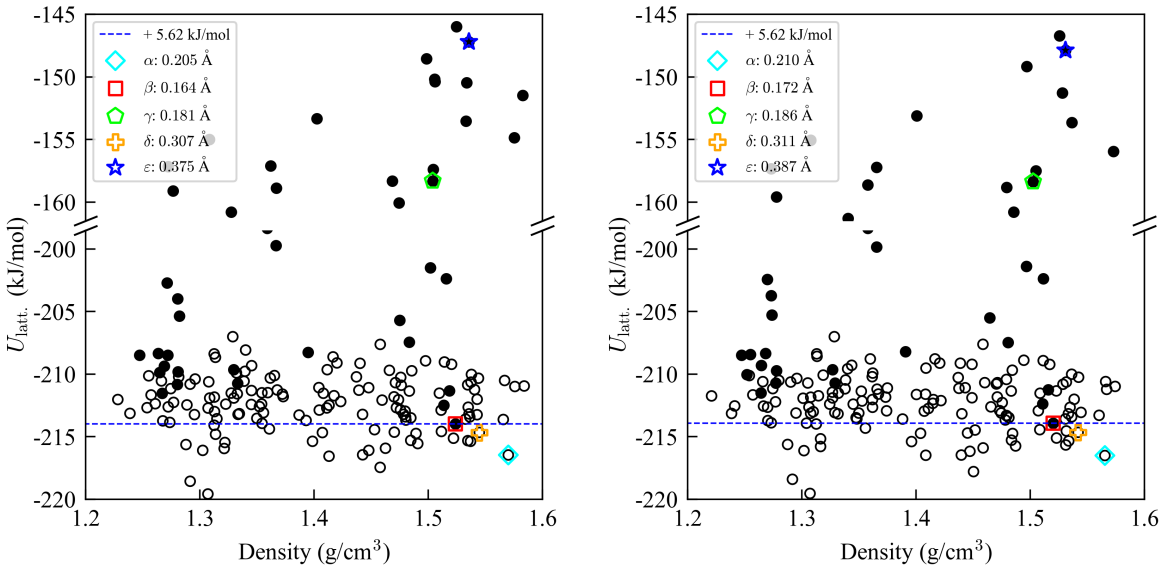
suboptimal τ_{nag} that lead to poor convergence rates. Moreover, due to the polar cell problem, $F(\boldsymbol{x})$ for glycine span a much wider range (~75 kJ/mol) than for urea (~25 kJ/mol), and this exacerbates the convergence issue for glycine. Since the E04UFF subroutine's convergence criterion is scaled by $F(\boldsymbol{x})$ (Equation 7.4.11), it is more challenging to select a 'one-size-fits-all' optimality tolerance suitable for a wider $F(\boldsymbol{x})$ range. With Mode 4, 16 unconverged structures have $|F(\boldsymbol{x})| < 195$ kJ/mol, while $\tau_{\text{nag,min}}$ was derived using $|F(\boldsymbol{x})| = 200$ kJ/mol.

Aside from the choice of τ_{nag}, errors in $\boldsymbol{g}(\boldsymbol{x})$ also contribute to poor CSO–FM convergence. Using Mode 4, CSO–FM FD computations involve CSO–RM *minimizations*. When these fail to converge tightly to $\varepsilon_{\text{scrpRM}}$, spurious errors in the CSO–FM gradients emerge, e.g., $\|\boldsymbol{g}(\boldsymbol{x})\|$ 'spikes' in Figure N.30. This prompts either immediate termination of E04UFF, or imprecise linesearches that force early-stopping. From the Mode 4 trajectories for urea (Figure N.22) and glycine (Figure N.30), the latter seems to exhibit more frequent $\|\boldsymbol{g}(\boldsymbol{x})\|$ 'spikes', which widens the gap in convergence rates. Modes 2 and 3 use only single-point energy evaluations during their FD calculations, hence these gradient errors are absent from their $\|\boldsymbol{g}(\boldsymbol{x})\|$ trajectories.

Poor convergence notwithstanding, I compare the landscapes obtained by Modes 2 and 3, against the Mode 4 results (Table 9.15 and Figure N.31). Over the Refinement II landscape, both approximate gradient methods achieve excellent agreement with the rigorous FDs, in terms of geometry and energy. Again, Mode 3 achieves better agreement with Mode 4, versus Mode 2, but this is marginal. Parity between the FD modes is further shown in Figure 9.11, wherein the clustered landscapes for Modes 2 and 3 exhibit almost no discrepancy compared to the Mode 4 clustered landscape (Figure 9.8, bottom-right). The matches identified for all five glycine forms are also very similar among the three FD modes. Relative to urea, Refinement II of glycine suggests a smaller discrepancy between Modes 2–4. However, this may be biased by the glycine minimizations expending fewer major iterations (around half) versus urea. Over a smaller number of minimization steps, there is less opportunity for the FD methods to express their differences. In any case, these results reiterate that between Modes 2 and 3, the latter is

Table 9.15: Optimized U_{latt} and geometries for glycine predicted by Modes 2 and 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$, while RMSD_{20} scores are generated by the COMPACK algorithm. All comparisons are done on a structure-by-structure basis.

Mode	AD U_{latt} (kJ/mol)	AAD U_{latt} (kJ/mol)	No. $\lvert\Delta U_{\text{latt}}\rvert$ > 0.5 kJ/mol	Avg. RMSD_{20} (Å)	Max. RMSD_{20} (Å)	No. RMSD_{20} > 0.5 Å
2	0.137	0.198	19	0.026	0.148	0
3	0.085	0.165	10	0.024	0.165	0



Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
α	0.2047	+3.1191	5
β	0.1645	+5.6158	28
γ	0.1813	+61.2629	172
δ	0.3067	+4.9199	22
ε	0.3745	+72.3748	185

Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
α	0.2097	+3.0300	4
β	0.1719	+5.6234	29
γ	0.1855	+61.1592	175
δ	0.3113	+4.8832	22
ε	0.3869	+71.6501	185

Figure 9.11: Refined landscape of glycine using Model 4, with Mode 2 (left) and Mode 3 (right) FD gradient approximations. The lowest-energy match to α, β, γ, δ, and ε are labeled as ◇, □, ⬠, ✙, and ☆ respectively, with their corresponding $RMSD_{20}$ scores. Crystals belonging to polar space groups are denoted by •.

a better approximation for the rigorous gradients.

9.5 Flexible Z' = 1 Zwitterion: Proline

9.5.1 Background and Early-Stage CSP

The fourth system investigated is the zwitterionic amino acid, proline (Figure 9.12). Unlike glycine, proline has a chiral centre on its α-carbon and exhibits stereoisomerism. As enantiomers, L- and D-proline exhibit identical thermodynamic behavior, but due to their preponderance in biological systems, the L-forms of amino acids are of greater practical interest. Two forms of L-proline have been observed experimentally [393], denoted L-I and L-II here. L-I is stable at ambient conditions, whilst L-II was resolved only at elevated temperatures (> 400 K) [393]. Because L-II has Z' > 1, I do not expect to find it during this Z' = 1 CSP. In addition to enantiopure crystals of proline, DL-proline crystals can also be found from a Z' = 1 CSP, through space groups with inversion centres (e.g., *P21/c*). As such, I also consider the racemic proline crystal [394], denoted DL-I. To my knowledge, no experimental assessment of L-/DL-proline relative stability has been conducted, but it is often asserted that racemic crystals are thermodynamically favored over their chiral counterparts [395].

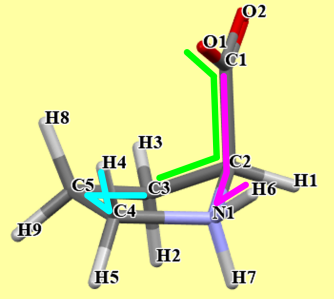


Figure 9.12: L-Proline ($C_5H_9NO_2$, 17 atoms). Three flexible proper torsions are defined (green, cyan, and magenta).

The three experimental forms of proline undergo local minimization by periodic DFT-D, from which their relative stability at static conditions emerges as: L-I (most stable) > DL-I > L-II. L-II is observed at elevated temperatures, hence under static 0 K conditions it is reasonable that its optimized structure from DFT-D gives a poor match to the initial structure ($RMSD_{20}$ > 0.1 Å) and is thermodynamically un-favored. L-I and DL-I, post-optimization, are still close geometric matches with their experimental counterparts. The preference for enantiopure L-I, over DL-I, contradicts conventional wisdom [395], but $\Delta U_{\text{latt}}^{\text{DL-I,L-I}}$ is small enough that DL-I/L-I can be considered practically iso-energetic.

During the Global Search, I employ a flexible molecular definition that allows rotation of the hydrogen bonding moieties. As shown in Figure 9.12, this corresponds to the carboxylate (green torsion) and amino (magenta torsion) group rotation. Details of the LAM construction, and Global Search are summarized in Appendix N.4. As before, PCM (ε_{PCM} = 11) is used to preserve the zwitterionic state of the proline molecule during isolated-molecule LAM QM calculations. The final Global Search landscape (Figure N.33) spans 2,739 structures in $\Delta U_{\text{latt}}^{\text{GM}} \leq 17.56$ kJ/mol. Good geometry matches to L-I and DL-I are identified, with the highest-energy match being DL-I (315th, $\Delta U_{\text{latt}}^{\text{GM}}$ = 8.80 kJ/mol). L-II was not searched for during the Global Search.

9.5.2 Refinement Stage

Refinement of the proline Global Search landscape is conducted in two stages. In Refinement I, the bottom 2,471 structures from the Global Search ($\Delta U_{\text{latt}}^{\text{GM}}$ = +17.00 kJ/mol) are indepen-

Table 9.16: Polymorphs of Proline. $RMSD_{20}$ values are obtained by comparing the starting and optimized crystals, and $\Delta U_{\text{latt}}^{\text{GM}}$ values are in kJ/mol. 'L' denotes the enantiopure Laevus form; 'DL' denotes the racemic form.

Form	Refcode	Space group	Z'/Z	Relative Stability		
				Exp. [393]	Periodic DFT-D	
					$RMSD_{20}$ (Å)	Ranking [$\Delta U_{\text{latt}}^{\text{GM}}$]
L-I	PROLIN03 [393]	*P212121*	1/4	L, Amb.	0.038	1st [GM]
L-II	PROLIN04 [393]	*P1*	2/2	L, 413 K	0.424	3rd [+2.36]
DL-I	QANRUT [394]	*P21/c*	1/4	DL, Amb.	0.043	2nd [+0.27]

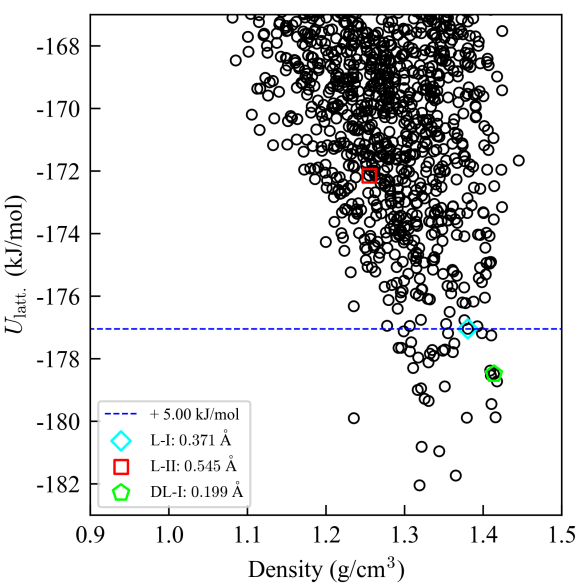
Table 9.17: Models used in the refinement of proline. All QM calculations are done at the PBE0/6-311G(d,p) level of theory.

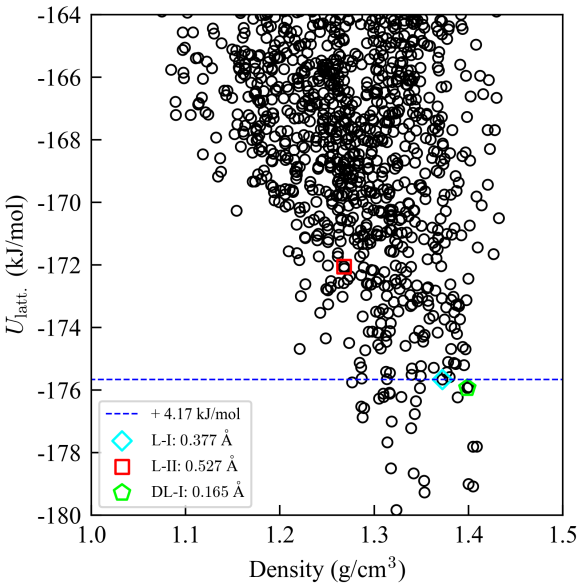
Model	Intramolecular Model	Intermolecular Model		
		Induction Model	HAIEFF FF	FF Parameters
1	PBE0/6-311G(d,p) + PCM_{11}	PCM_{11}	$DB14_{mp}$	Table I.14 of Ref [2]
2	PBE0/6-311G(d,p) + PCM_{11}	PCM_{11}	$BT^{PCM}_{dBHp,mp}$	Table J.25
3	PBE0/6-311G(d,p) + PCM_{11}	PCM_{11}	$BT^{PCM}_{Mie,mp}$	Table J.27
4	PBE0/6-311G(d,p) + PCM_{11}	SCRP	$BT^{SCRP}_{Mie,mp}$	Table M.4

dently evaluated by three energy models (Table 9.17, Models 1–3). After clustering, the bottom 200 structures from Model 3 advance to Refinement II for evaluation by the SCERP model (Table 9.17, Model 4). This was initially attempted with an identical flexible definition as the Global Search (only two flexible torsions), but poor numerical behavior surfaced due to the five-membered ring puckering in different orientations. For this reason, Refinement I and II are conducted with an additional flexible torsion (H_4−C_4−C_5−C_3) to reflect the ring flexion. In Refinement II, rigorous CSO–FM SCERP gradients (i.e., Table 7.6, Mode 4) are employed. Detailed summaries of the refinement calculations are tabulated in Appendix N.4.4 and the (post-clustering) refinement landscapes are shown in Figure 9.13.

For the Z' = 1 forms (L-I and DL-I), all three Refinement I models yield excellent geometry and energy predictions. In all cases, the geometry match to the known forms have $RMSD_{20}$ < 0.38 Å, with DL-I's geometry consistently being better-predicted than L-I's. Between Models 1–3, there are only marginal differences in the quality of geometry prediction for the known forms. In terms of relative stabilities, all Refinement I models indicate that DL-I is more stable than L-I. Whilst this contradicts the DFT-D results (Table 9.16), the $\Delta U_{latt}^{L\text{-}I,DL\text{-}I}$ derived from DFT-D is likely too small for a HAIEFF model to reliably resolve, particularly when using a transferable FF. Setting aside unrealistic expectations, the $\Delta U_{latt}^{L\text{-}I,DL\text{-}I}$ predicted by Models 1–3 are small enough (< 1.45 kJ/mol) to consider this an acceptable level of accuracy, given the HAIEFF model's fidelity. Relative to the rest of the structures, these two forms place low on all three landscapes: with $\Delta U_{latt}^{GM} \leq 5.0$ kJ/mol and ranks < 40th. As with glycine, Models 2 and 3 present with very similar results, since they employ BT FFs derived through a common fitting approach. Given results from any of these three Refinement I models, it is likely that both L-I and DL-I would be successfully advanced to re-ranking by periodic DFT-D.

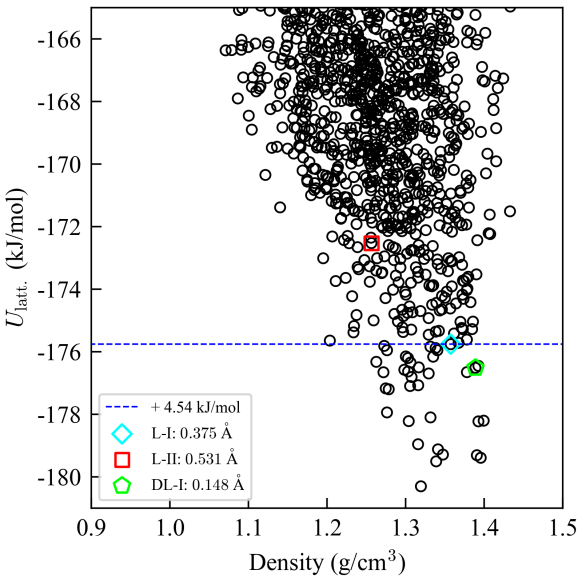
Due to its Z' number, L-II is not expected to be found from the refinement of a Z' = 1 Global Search. Nevertheless, screening of the Refinement I landscapes reveals that each model predicts a modest match to L-II ($RMSD_{20}$ ~ 0.5 Å), in the *P21* space group, with ΔU_{latt}^{GM} > 7.7 kJ/mol and rank > 140th. Although L-II should be the least stable among the three proline forms (under static conditions), one would still expect it to be in closer thermodynamic competition with L-I and DL-I. There are two potential factors contributing to the inferior predictions for L-II (versus L-I/DL-I). Firstly, the Z' = 1 model over-constrains the symmetry of L-II, preventing its lattice from relaxing further than it would if an appropriate Z' was employed. Secondly, L-II

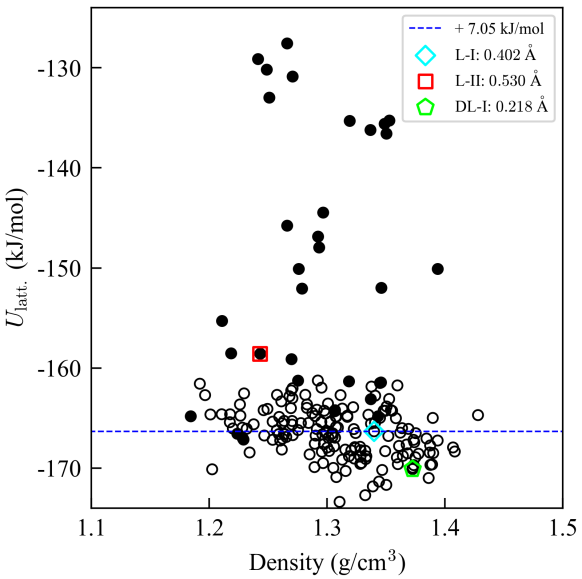




Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
L-I	0.3709	+5.0037	36
L-II	0.5454	+9.9038	266
DL-I	0.1985	+3.5616	16

Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
L-I	0.3772	+4.1734	29
L-II	0.5267	+7.7868	147
DL-I	0.1650	+3.9040	26





Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
L-I	0.3747	+4.5427	36
L-II	0.5306	+7.7750	140
DL-I	0.1477	+3.7878	23

Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
L-I	0.4017	+7.0452	86
L-II	0.5304	+14.8047	171
DL-I	0.2184	+3.2776	11

Figure 9.13: Refined landscape of proline using Models 1 (top-left), 2 (top-right), 3 (bottom-left), and 4 (bottom-right). The lowest-energy match to L-I, L-II, and DL-I are labeled as ◇, □, and ⬠ respectively, with their corresponding $RMSD_{20}$ scores. In the Model 4 plot, crystals belonging to polar space groups are denoted by •.

is a polar crystal and, as demonstrated earlier, the existing induction models (PCM/SCERP) falter in describing these structures accurately, especially when charged molecules are involved. Altogether, one cannot glean too much about the quality of the Refinement I models from L-II, other than reiterating that PCM is likely deficient for describing polar cells.

The bottom 200 structures from Model 3 undergo further refinement by Model 4, and the results from this Refinement II stage align with findings from the glycine CSP. Due to the polar nature of the L-II unit cell, Refinement II energy predictions for this structure worsen considerably ($\Delta U_{\text{latt}}^{\text{GM}}$ > 14.8 kJ/mol), relative to the Refinement I PCM models. The SCERP model's polarization predictions are more severely hampered due to its attempt to explicitly represent the lattice, versus PCM which substitutes the crystal for a continuum. For the non-polar forms of proline, DL-I falls lower on the landscape, both in terms of rank (11th) and $\Delta U_{\text{latt}}^{\text{GM}}$ (3.3 kJ/mol), while L-I ascends to rank 86th and $\Delta U_{\text{latt}}^{\text{GM}}$ = 7.0 kJ/mol. Overall, the Refinement II model seems to perform poorer than its Refinement I counterpart.

9.5.3 Final Periodic DFT-D Re-ranking

Finally, I compare the accuracy of the PCM (Model 3) and SCERP (Model 4) models, to periodic DFT-D. The bottom 18 structures from Model 4 (11th: DL-I match), plus structure 86 (L-I match) are re-ranked by periodic DFT-D. Structure 171 (L-II match) is excluded from this analysis since the Z' = 1 HAIEFF model may bias it against matching the periodic DFT-D result. From Table N.26, all of the DFT-D optimized crystal structures return good RMSD_{20} scores when compared to their original Model 4 structures, corroborating that the SCERP model is accurate at reproducing DFT-D crystal geometries. For structures 86 and 11 (Model 4's L-I and DL-I match, respectively), the DFT-D optimized structures match the corresponding known forms of proline, and align with earlier local minimizations of L-I and DL-I.

In Figure 9.14a, I study the relative lattice energies, with respect to L-I ($\Delta U_{\text{latt}}^{\text{L-I}}$), predicted for the 19 proline crystal structures. In the CSPs for glycine and urea, similar analyzes of ΔU_{latt} suggested that Models 3 and 4 reproduce DFT-D relative stabilities to comparable accuracy, and the superior model was ambiguous. For proline, there is strong evidence that Model 3 has better agreement (than Model 4) with DFT-D. Model 4 not only has a significantly larger AAD and RMSD than Model 3, Model 4 also predicts a *negative* correlation with DFT-D (whereas Model 3 is positively correlated). The negative correlation is particularly concerning, as it suggests that Model 4 predicts relative stabilities in opposition to DFT-D. Overall, from visual inspection of the parity plot, I would assert that Model 4 is inferior to Model 3 for predicting $\Delta U_{\text{latt}}^{\text{L-I}}$. Relative ranks for the 19 crystal structures are presented in Figure 9.14b. Admittedly, the distribution of points on the parity plot appear almost random for both Models 3 and 4. However, like with $\Delta U_{\text{latt}}^{\text{L-I}}$, Model 4 again exhibits a slight negative rank correlation with DFT-D, whilst Model 3 at least preserves a positive rank correlation.

In summary, proline is the only system for which the Refinement I PCM models convincingly outperform the Refinement II SCERP model. As precursors to periodic DFT-D re-ranking, the

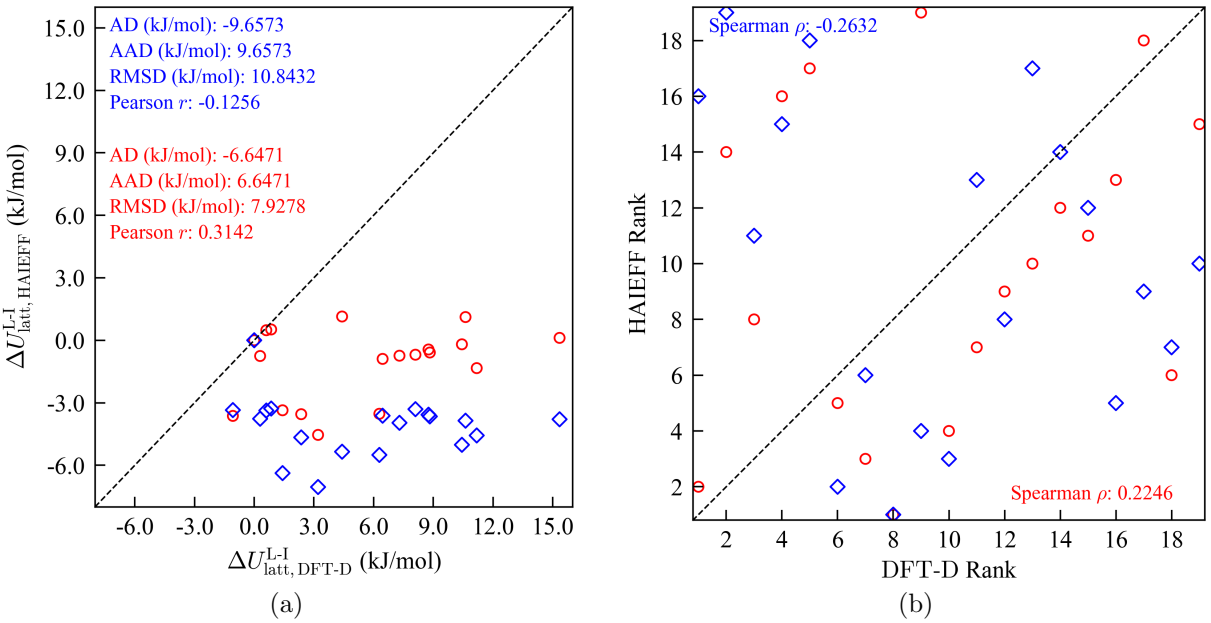


Figure 9.14: Comparison of (a) $\Delta U_{\text{latt}}^{\text{L-I}}$, and (b) ranks, predicted by the HAIEFF models versus periodic DFT-D. The bottom 18 structures + structures 86 from Model 4 are denoted by blue $\diamond$, while their corresponding Model 3 results are denoted by red $\circ$.

Refinement I landscapes require lower cutoff ranks and lower $\Delta U_{\text{latt}}^{\text{GM}}$ to advance both L-I and DL-I, versus Refinement II. With reference to periodic DFT-D results, the SCERP model (Model 4) exhibits greater errors in $\Delta U_{\text{latt}}^{\text{L-I}}$, compared to Model 3. Granted, the AAD and RMSD metrics are dependent on the chosen reference structure, L-I in this case. However, the correlation scores are agnostic to the reference, and they provide stronger evidence that Model 4 is in poorer agreement with DFT-D than Model 3. Model 4's negative correlation with DFT-D could indicate a more fundamental flaw in the description of proline using SCERP. Further study of proline, and other highly charged species, may be necessary to gain more insight.

9.5.4 Assessment of SCERP Gradient Approximations

As a final appraisal of the SCERP FD gradients, I repeat the gradient tests once more. Refinement II is repeated for proline using Model 4 (Table N.25) and CSO–FMv1.4.0.2. Results obtained by FD Modes 2 and 3 are compared to previous results determined by Mode 4.

The convergence of the three FD modes are compared in Table 9.18, with minimization trajectories presented in Figures N.34–N.36. The trends for proline align with findings from urea and glycine: enhanced minimization performance is observed according to the improving accuracy of the FD approximation between Modes 2–4. The Mode 4 convergence rate for proline is similar to urea (~75%), both of which are higher than glycine (~50%). Like urea, proline's Mode 4 minimization trajectories (Figure N.36) present with fewer spikes in $\|\boldsymbol{g}(\boldsymbol{x})\|$, compared to glycine. This suggests that (SCERP-embedded) CSO–RM convergence is better in urea and proline, relative to glycine, such that their gradients experience fewer spurious errors. It is unclear whether a particular trait of the molecule/crystal encourages a proclivity for this

Table 9.18: Convergence behavior during proline Refinement II, using FD Modes 2–4.

Mode	Average N_{maj}	Average N_{line}	Average $\frac{N_{\text{line}}}{N_{\text{maj}}}$	No. Converged [%]
2	3.6	18.2	6.0	13 [6.5%]
3	5.0	18.6	4.7	42 [21.1%]
4	8.0	19.5	2.4	153 [76.5%]

Table 9.19: Optimized U_{latt} and geometries for proline predicted by Modes 2 and 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$, while RMSD_{20} scores are generated by the COMPACK algorithm. All comparisons are done on a structure-by-structure basis.

Mode	AD U_{latt} (kJ/mol)	AAD U_{latt} (kJ/mol)	No. $\lvert\Delta U_{\text{latt}}\rvert$ > 0.5 kJ/mol	Avg. RMSD_{20} (Å)	Max. RMSD_{20} (Å)	No. RMSD_{20} > 0.5 Å
2	0.119	0.139	3	0.034	1.801	1
3	0.079	0.127	12	0.030	0.184	0

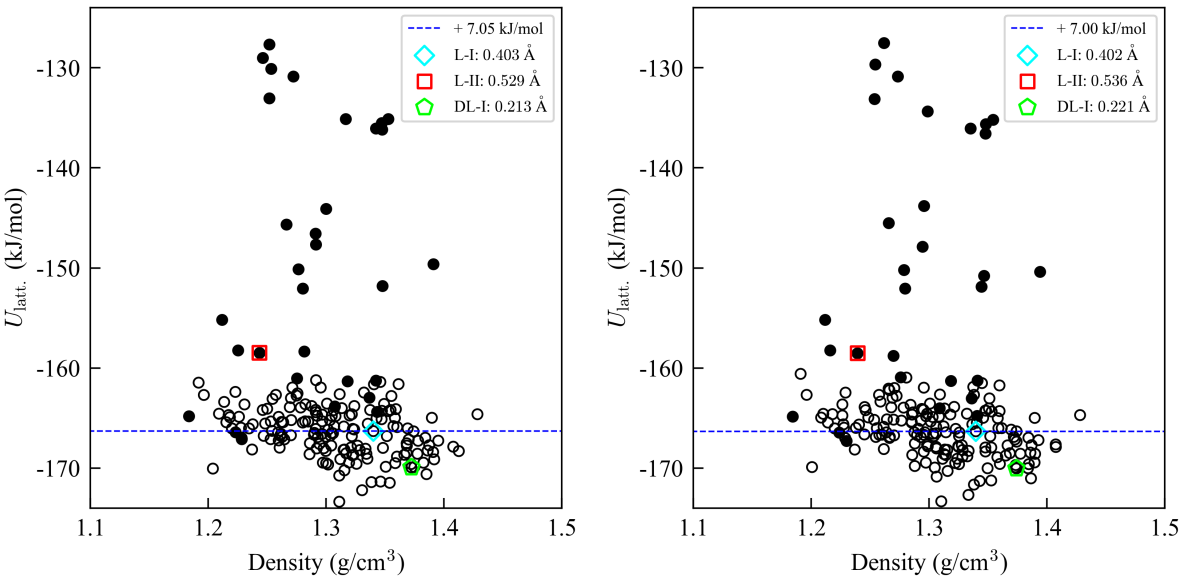
behavior in the CSO–RM minimization, or if it occurs arbitrarily. Like glycine, proline's Mode 2 and 3 minimizations also exhibit lower convergence rates than their urea counterparts. This may be due to the charged-molecule-in-polar-cell problem, which spreads proline's U_{latt} over a broad range (~50 kJ/mol), complicating the choice of τ_{nag}.

Despite their dissimilar convergence behavior, Modes 2/3 exhibit almost no disparity with Mode 4 in optimized crystal geometries and energies, over the Refinement II landscape (Table 9.19 and Figure N.37). Like glycine, this is explained by the relatively few steps (N_{maj}) expended to optimize the proline crystals. The exception to this is Structure 145 from Mode 2, which has poor agreement in geometry (RMSD_{20} = 1.801 Å) and energy ($\Delta U_{\text{latt}}^{\text{Mode 2,Mode 4}} \approx 5$ kJ/mol) with its Mode 4 counterpart. Ultimately, Mode 3 still exhibits closer agreement with Mode 4, than Mode 2. While Mode 3 has more structures than Mode 2 with $|\Delta U_{\text{latt}}| > 0.5$ kJ/mol, these 'outliers' should only be marginally beyond the 0.5 kJ/mol threshold, since Mode 3's U_{latt} AAD is lower than Mode 2's. Finally, comparing the clustered landscapes, there is no obvious discrepancy between Modes 2/3 (Figure 9.15) and Mode 4 (Figure 9.13, bottom-right) and matches to L-I, L-II, and DL-I are consistent between the three FD approaches.

9.6 SCERP Gradient Approximations: Discussion

To conclude my assessment of the SCERP-embedded CSO–FM FD gradients, I highlight the key findings from the urea, glycine, and proline CSP studies.

In terms of gradient accuracy, there is a clear hierarchy among the four FD Modes: 4 (most accurate) > 3 > 2 > 1. Modes 1 and 2 have equivalent costs, as neither employ the SCERP process or CSO–RM minimization in the FD, hence Mode 1 is unambiguously inferior to Mode 2. Between Modes 2–4, gradient accuracies are distinct enough to incite tangible differences in the minimization behavior. In particular, Model 3 consistently progresses further in the minimization (compared to Mode 2), such that it ultimately agrees more closely with the



Form	$\mathbf{RMSD_{20}}$ (Å)	$\boldsymbol{\Delta U_{\text{latt}}^{\text{GM}}}$ (kJ/mol)	Rank
L-I	0.4027	+7.0527	82
L-II	0.5287	+14.8809	169
DL-I	0.2132	+3.4248	10

Form	$\mathbf{RMSD_{20}}$ (Å)	$\boldsymbol{\Delta U_{\text{latt}}^{\text{GM}}}$ (kJ/mol)	Rank
L-I	0.4025	+6.9976	87
L-II	0.5360	+14.8074	172
DL-I	0.2210	+3.2878	9

Figure 9.15: Refined landscape of proline using Model 4, with Mode 2 (left) and Mode 3 (right) FD gradient approximations. The lowest-energy match to L-I, L-II, and DL-I are labeled as ◇, □, and ⬠ respectively, with their corresponding RMSD_{20} scores. Crystals belonging to polar space groups are denoted by •.

rigorous Mode 4 FD gradients. With that said, disparate minimization performance does not necessarily translate to *practical* repercussions on CSP outcomes. The Refinement II landscapes for proline and glycine are virtually unaffected by the FD mode, while the urea landscape does exhibit noticeable differences using Mode 2. Clearly, if the minimization starts near the optima, it is inconsequential if premature termination occurs from using Modes 2/3, since only minor adjustments are made to the crystal structure. For glycine and proline, structures that converged using Mode 4 exhibit an average change in U_{latt} of 0.27 and 0.21 kJ/mol (respectively) over the course of the minimization, while urea structures exhibit an average change of 2.52 kJ/mol. This is also reflected in the smaller N_{maj} expended for glycine and proline versus urea. One factor that could be contributing to this behavior is the choice of LAM validity ranges. For glycine and proline, I employed tighter validity ranges in Refinement I/II (Tables N.16, N.17, N.24, and N.25), compared to the urea Refinement (Tables N.8 and N.9). As such, in the case of glycine and proline, the optimized structures from Model 3 (which serve as starting points for the SCERP minimizations) may be better converged with respect to the conformational dofs, reducing the effort that is required during the SCERP minimization.

Speaking on costs, I have chosen not to compare compute times for a few reasons. Firstly, as suggested from MH-DAP, I believe CPU times are inconsistently reported between the different

programs (CSO–RM/FM, Gaussian, VASP) and this is leading to false inferences about the relative costs of these methods. In my experience, SCERP-embedded CSO–FM minimizations (using Mode 4) are slightly more than an order of magnitude more expensive than 'regular' CSO–FM minimizations, while DFT-D minimizations in VASP are even more costly. This aligns with the number of structures treated in Refinement I and II, and finally by periodic DFT-D. Comparing between the FD Modes, CPU times bias the less accurate FD approximations that are more prone to early-stopping. Overall, multiple competing factors contribute to the cost of the FD methods (linesearch efficiency, early-stopping, SCERP process convergence, and CSO–RM convergence), but on net Mode 2 is cheapest since it avoids SCERP/CSO–RM during the FD. For the systems tested in the present work, Modes 3 and 4 are still practicable due to the cost reduction strategies embedded in CSO–FM, CSO–RM, and the SCERP process. However, Modes 3 and 4 scale poorly with Z' and/or number of conformational dofs, since these compound the number of QM computations required per CSO–FM iteration. This is especially acute for Mode 4, since its use of CSO–RM minimizations adds another multiplicative factor.

Taking a pragmatic stance, I recommend the use of Mode 3 FD gradients, since this offers a good balance of cost and accuracy. If U_{latt} values do not change appreciably over the minimization (< 1 kJ/mol), treatment by Mode 3 is likely adequate. Conversely, if U_{latt} undergoes significant variation, the Mode 3 results should be refined with Mode 4. This strategy mitigates errors in Mode 3 that amass over long minimization trajectories, whilst reducing the number of steps that must be taken with Mode 4. Mode 2 should be used only if costs become prohibitive.

9.7 Flexible Cocrystal: Caffeine + m-Nitrobenzoic Acid

9.7.1 Background and Early-Stage CSP

Cocrystals offer novel avenues to tune the physicochemical and thermodynamic properties of solid state products [12–15], making them an important class of multicomponent crystal. To that end, the final system I consider is the 1:1 caffeine + m-nitrobenzoic acid cocrystal, abbreviated as CAFF-mNBA (Figure 9.16). Two forms of CAFF-mNBA are reported in the CSD, denoted here as F-I [397] and F-II [396]. From characterization done by Singaraju *et al.* [396], slurry experiments indicate F-II is stable at ambient conditions, while DSC measurements sug-

Table 9.20: Polymorphs of the 1:1 CAFF + mNBA cocrystal. *Z'/Z* include both molecular species. RMSD_{20} values are obtained by comparing the starting and optimized crystals, and $\Delta U_{\text{latt}}^{\text{GM}}$ values are in kJ/mol.

Form	Refcode	Space group	Z'/Z	Relative Stability		
				Exp. [396]	Periodic DFT-D	
					RMSD_{20} (Å)	Ranking [$\Delta U_{\text{latt}}^{\text{GM}}$]
F-I	DIJVUN [397]	*P21/a*	2/8	High-T	0.112	1st [GM]
F-II	DIJVUN01 [396]	*P21/c*	2/8	Amb. form	0.062	2nd [+0.30]

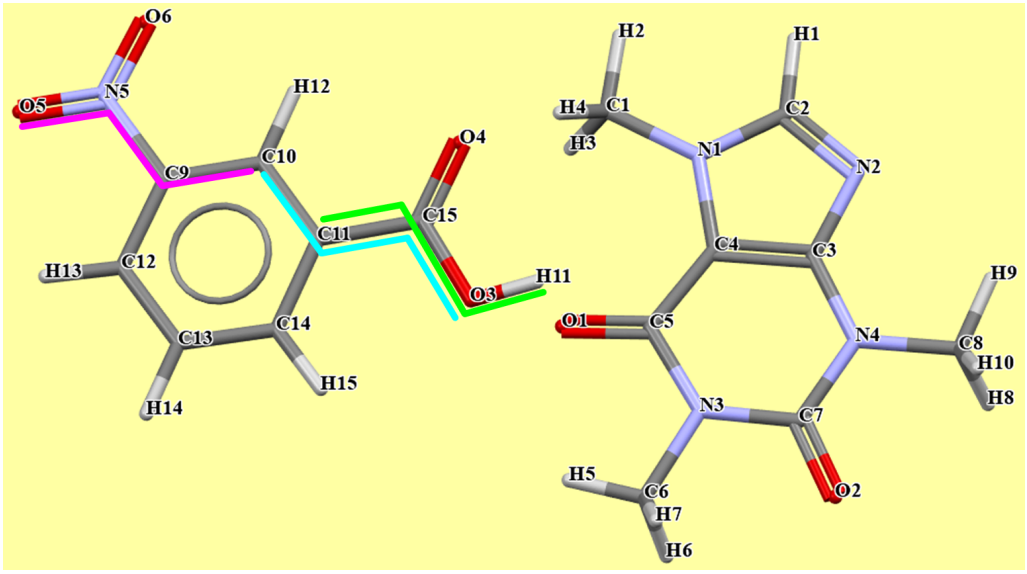


Figure 9.16: CAFF-mNBA ($C_8H_{10}N_4O_2$ + $C_7H_5NO_4$, 24 + 17 atoms). Three flexible torsions (green, cyan, and magenta) are defined for mNBA (left), while CAFF (right) is treated as rigid.

gest F-I is favored at elevated temperature (~433 K). When subject to grinding, F-II → F-I interconversion is also observed [396]. Altogether, the two forms of CAFF-mNBA are likely in close thermodynamic competition. This is confirmed by local minimization of F-I and II using periodic DFT-D. Under static conditions, F-I emerges as more stable than F-II. Whilst this contradicts the experimental relative stabilities, $\Delta U_{\text{latt}}^{\text{F-II,F-I}}$ is small enough that the inclusion of vibrational effects could reverse the stability order [24], aligning it with experiment.

During the Global Search of CAFF-mNBA, I treat caffeine as rigid, while *m*-nitrobenzoic acid is considered flexibly. In CAFF, the fused ring moiety is unlikely to exhibit significant conformational change and flexibility is limited to methyl group rotations. These rotations usually do not impact CSP results and can be neglected. In mNBA, I permit flexibility about the hydrogen bonding carboxylic acid group (green and cyan torsions) and the nitro-group (magenta torsion). Rotations of $-NO_2$ and $-COOH$, relative to the aromatic ring, may be unfavorable since they disrupt π-delocalization, but the LAM model should inherently recognize and penalize these conformations. Details of the LAM construction, and Global Search are summarized in Appendix N.5. The final Global Search landscape (Figure N.39) spans 52,666 unique clusters in $\Delta U_{\text{latt}}^{\text{GM}} \leq 15.00$ kJ/mol. Good geometry matches to F-I and F-II are identified, with the highest-energy match being F-II (203rd, $\Delta U_{\text{latt}}^{\text{GM}} = 3.94$ kJ/mol).

9.7.2 Refinement Stage

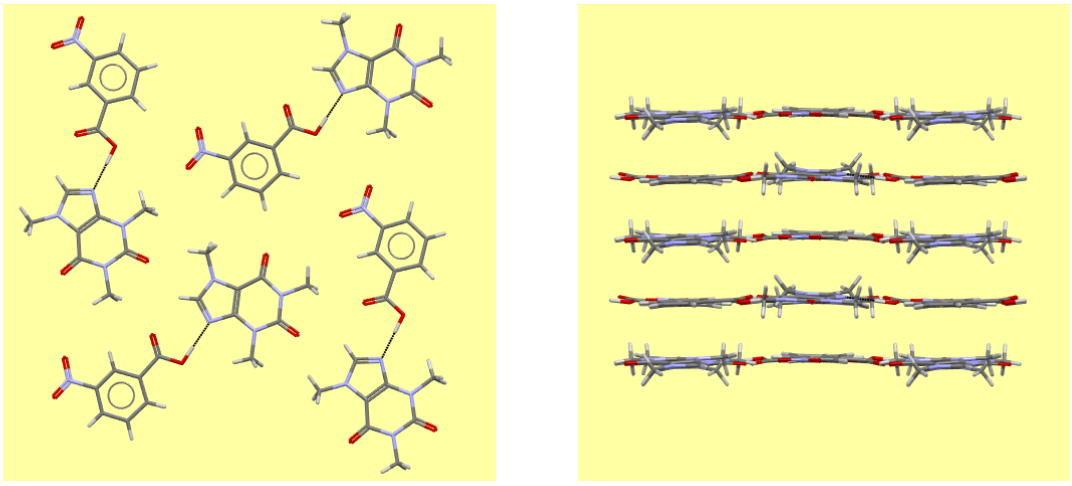
Refinement of the CAFF-mNBA Global Search landscape is done in two stages. In Refinement I, the bottom 3,000 structures from the Global Search ($\Delta U_{\text{latt}}^{\text{GM}} = +7.82$ kJ/mol) are independently evaluated using three energy models (Table 9.21, Models 1–3). After clustering, the bottom 120 structures from Model 3 advance to Refinement II for evaluation by the SCERP model (Table 9.21, Model 4). For all refinements, the same flexible definition as the Global Search

Table 9.21: Models used in the refinement of CAFF-mNBA. All QM calculations are done at the PBE0/6-311G(d,p) level of theory.

Model	Intramolecular Model	Intermolecular Model		
		Induction Model	HAIEFF FF	FF Parameters
1	PBE0/6-311G(d,p) + PCM_{11}	PCM_{11}	$DB14_{mp}$	Table I.14 of Ref [2]
2	PBE0/6-311G(d,p) + Static	None (Static)	$BT^{Static}_{Mie,mp}$	Table J.20
3	PBE0/6-311G(d,p) + PCM_{11}	PCM_{11}	$BT^{PCM}_{Mie,mp}$	Table J.27
4	PBE0/6-311G(d,p) + Static	SCRP	$BT^{SCRP}_{Mie,mp}$	Table M.4

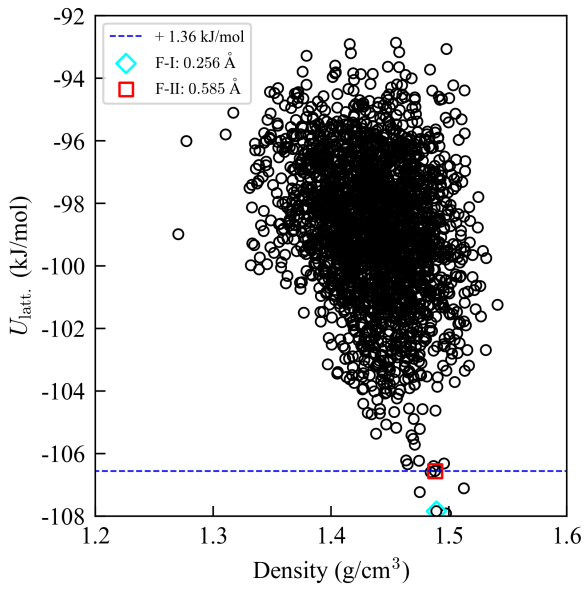
(CAFF: rigid; mNBA: three flexible torsions) is used, and in Refinement II, rigorous CSO–FM SCERP gradients (i.e., Table 7.6, Mode 4) are employed. Detailed summaries of the refinement calculations are tabulated in Appendix N.5.4, and the (post-clustering) refinement landscapes are shown in Figure 9.18.

In Refinement I, excellent results are achieved by all three models. For both forms, geometries are predicted with $RMSD_{20} \leq 0.60$ Å, and the predicted crystal geometries for F-I are consistently of higher quality ($RMSD_{20} \leq 0.26$ Å) than F-II. Between the three models, those employing the BT FFs (Models 2 and 3) also have marginally more-accurate predictions of geometry, versus Model 1, which uses the DB FF. In terms of relative stabilities, all three models correctly identify F-I as more stable than F-II, by 1.2–1.4 kJ/mol. This energy gap is several times larger than determined by periodic DFT-D (~0.30 kJ/mol), however, sub-kJ/mol accuracy is not expected for HAIEFF models employing transferable FFs. Relative to the other predicted structures, F-I and F-II are exceptionally low on all three landscapes, with ranks $< 10^{th}$ and $\Delta U^{GM}_{latt} < 2.0$ kJ/mol. In fact F-I is ranked 2nd, 1st, and 3rd using Models 1–3 respectively. Overall, all three Refinement I models produce highly satisfactory results and, given any of their landscapes, both forms are very likely to advance for re-ranking by periodic DFT-D.

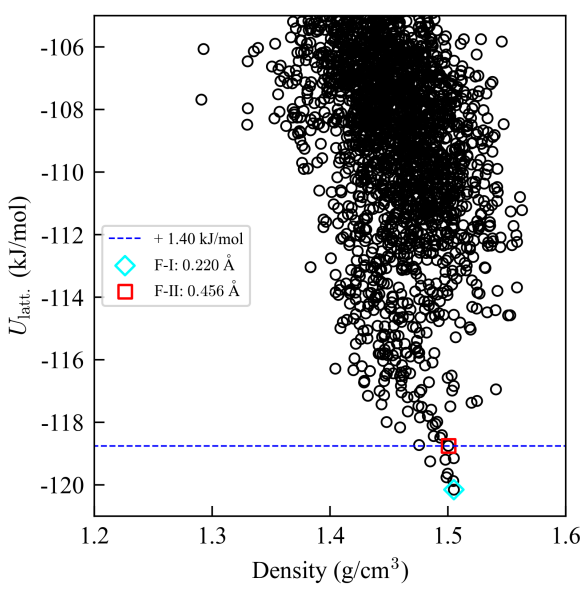


(a) (b)

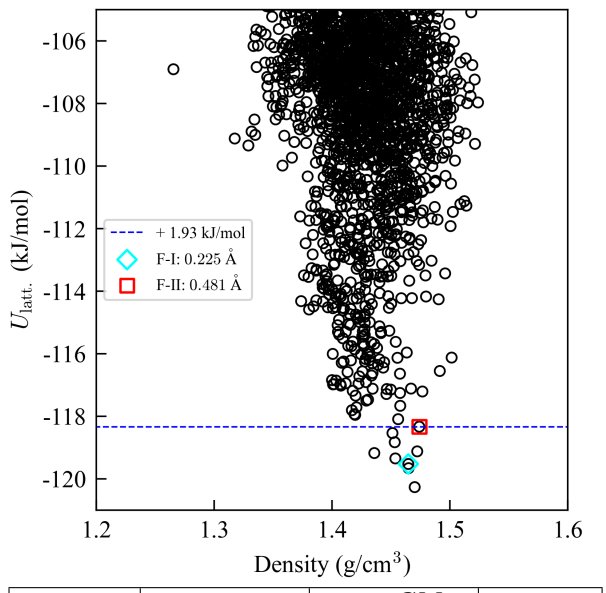
Figure 9.17: Crystal packing of CAFF-mNBA F-I: (a) face-on view of a single packing plane, (b) side-on view along the packing plane. HB contacts are given by the black dotted lines.



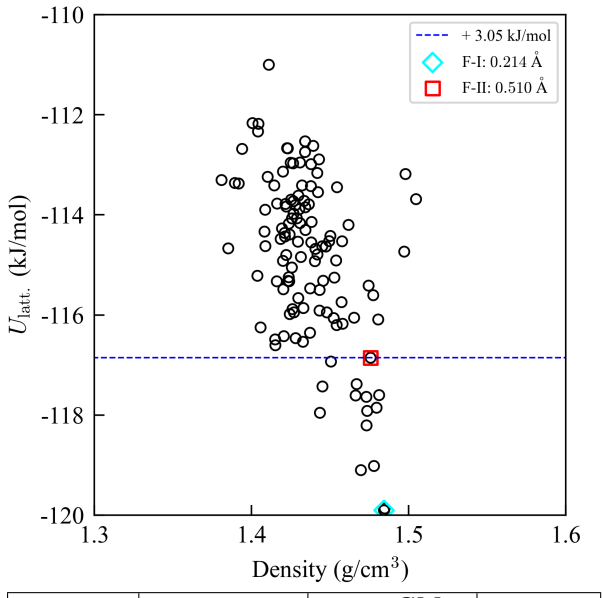
Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
F-I	0.2563	+0.0750	2
F-II	0.5850	+1.3592	6



Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
F-I	0.2204	+0.0000	1
F-II	0.4560	+1.4012	8



Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
F-I	0.2255	+0.7397	3
F-II	0.4814	+1.9328	9



Form	$RMSD_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
F-I	0.2137	+0.0000	1
F-II	0.5101	+3.0480	15

Figure 9.18: Refined landscape of CAFF-mNBA using Models 1 (top-left), 2 (top-right), 3 (bottom-left), and 4 (bottom-right). The lowest-energy match to F-I and F-II are labeled as ◇ and □ respectively, with their corresponding $RMSD_{20}$ scores.

It is notable that Model 2, which does not use PCM, yields predictions on par with the PCM-embedded models. This is at odds with earlier findings, that suggest an induction-free model would give inferior results for hydrogen bonding systems, like CAFF-mNBA. Examination of CAFF-mNBA's crystal structures provides insight into this behavior. From graph set analysis [263], intermolecular hydrogen bonding in F-I and F-II is limited to dimers between CAFF/mNBA pairs, while extended networks of hydrogen bonding chains/rings are absent. These CAFF/mNBA dimers arrange into planes, which then stack via vdW and/or π-π interactions. This is depicted for F-I in Figure 9.17; unlike F-I, F-II's planes alternate in a herringbone fashion. Accordingly, only a small proportion of cohesive energy arises from hydrogen bonds and Model 2 is capable of reasonable CSP predictions with the limited description of induction contained within the $\mathrm{BT}^{\mathrm{Static}}_{\mathrm{Mie,mp}}$ FF. Nevertheless, without *a priori* knowledge of the crystal packing, equally good results are obtained with PCM (Models 1 and 3).

During Refinement II, the bottom 120 structures from Model 3 are refined by Model 4. In general, Refinement II preserves the successes of Refinement I. With Model 4, F-I and F-II are predicted with comparable geometry accuracy to Model 3: F-I improves marginally, while F-II deteriorates slightly. Additionally, F-I is predicted as the global minimum structure, while F-II remains in close competition ($\Delta U^{\mathrm{GM}}_{\mathrm{latt}} \approx 3.0$ kJ/mol). While the energy gap between F-I/F-II widens slightly, they are both still extremely well-ranked on the Refinement II landscape ($\leq 15^{\mathrm{th}}$). Considering the stellar results from Refinement I, one cannot expect much improvement in the CSP predictions, and ultimately progression to periodic DFT-D re-ranking is equally feasible from the Refinement II landscape.

Final re-ranking by periodic DFT-D was attempted for the bottom 20 structures from Model 4, however these calculations were ultimately abandoned as convergence could not be reached in a reasonable amount of time. Whilst only anecdotal in nature, this does support the previous claim that SCERP-embedded CSO–FM is more cost-efficient than periodic DFT-D. Refinement II for CAFF-mNBA was also re-attempted with the approximate SCERP FD gradients. The findings from these calculations (Appendix N.5.5) are consistent with conclusions made in Section 9.6, hence I do not discuss them further.

9.8 HAIEFF Model Accuracy: Discussion

Over these five CSPs, I have benchmarked the accuracy for a range of HAIEFF models over a myriad of systems. Below, I distill some pertinent findings.

Firstly, FFs like FIT [4, 166, 198] continue to find use in modern CSP due to their positive track record throughout the development of CSP. Over time, the HAIEFF model has continued to evolve, primarily through using more-sophisticated models in its *ab initio* component (e.g., the use of PCM/SCERP, AIM-multipole methods). Unfortunately, the present results suggest that pairing modern *ab initio* models with incompatible empirical FFs may ultimately undo any expected improvement in CSP outcomes. This view aligns with findings from Section 6.5 and has been echoed by other authors [2, 169]. On this basis, comparisons of HAIEFF model

accuracy are meaningful only when the *ab initio* and empirical FF components are congruent. Calculations involving the DB and BT FFs have stayed true to this philosophy.

To appraise the various model developments, models using the DB FFs [2] were chosen as a yardstick for accuracy. Not only do these FFs honor FF-compatibility, but they were also fit under a similar framework as the BT FFs. The DB and BT FFs mainly differ in their fitting scheme (in Table 6.2, DB uses Scheme 1, while BT employs Schemes 5/7) and their reference data (DB–445 versus CE–755). The DB FFs have been benchmarked against FIT in the past [2], however the systems studied were mostly non-hydrogen bonding and rigid in nature. From Chapter 6, models using the BT FFs should exhibit superior energy accuracy in HB systems like those studied here. Unlike the DB FFs, the BT FFs provide some implicit description of induction, which enhances CSP results when polarization (PCM/SCERP) is completely neglected in the electrostatic term. With that said, models employing wholly static electrostatic descriptions are consistently outperformed by their PCM-embedded counterparts. With PCM, similar model accuracy is achieved using both the DB and BT FFs across all CSPs. Altogether, successful CSP with HAIEFF models appears to rely heavily on fortuitous error-cancellation between crystal structures of the *same* molecule. As such, reducing parameter estimation errors over a *diverse* set of molecules may provide only tenuous insight into the ultimate CSP accuracy.

For a given electrostatic model, the dBHP and MIP BT FFs tend to predict similar landscapes. While this is reasonable, as they are trained on the same reference data, it also suggests that model accuracy is insensitive to the form of the empirical FF. Nevertheless, there is a subtle benefit to using the MIP BT FFs. During lattice energy minimization with the BT dBHP FFs (and the DB BHP FFs), structures are susceptible to fusing, whereas these occurrences are eliminated when using the BT MIP FFs. Although the proportion of the landscape affected by fusing ($< 1\%$) is unlikely to influence CSP outcomes, it is still preferable to suppress this using the MIP. Overall, the CSPs support that FFs based on the MIP can provide results as good as, if not better, than the BHP/dBHP, whilst eliminating unphysical behavior.

To study the electrostatic model's effect on CSP accuracy, the SCERP-embedded HAIEFF model has been explored alongside PCM. As suggested in Chapter 7, SCERP should offer a higher-fidelity description of polarization than PCM, yet the SCERP-embedded models do

Table 9.22: Summary of the case study results, ignoring forms that present with the "polar cell problem". Models 3 and 4 correspond to the the HAIEFF models employing the $\mathrm{BT}^{\mathrm{PCM}}_{\mathrm{Mie,mp}}$ and $\mathrm{BT}^{\mathrm{SCRP}}_{\mathrm{Mie,mp}}$ FF respectively.

System	Forms Considered	Highest $\Delta U^{\mathrm{GM}}_{\mathrm{latt}}$ (kJ/mol)		Highest Rank	
		Model 3	Model 4	Model 3	Model 4
MH-DAP	MH-I, MH-II, MH-III	5.47	6.77	73	121
Urea	F-I, F-III	12.88	14.22	182	124
Glycine	α, β, δ	2.59	5.46	20	29
Proline	L-I, DL-I	4.54	7.05	36	86
CAFF-mNBA	F-I, F-II	1.93	3.05	9	15

not systematically improve CSP results over their PCM counterparts. Even when excluding structures afflicted by the polar cell problem, using SCERP broadens the range of U_{latt} over which experimental matches are observed (Table 9.22). As a result, $\Delta U_{\text{latt}}^{\text{GM}}$ for these matches tend to increase, even if their ranks are more favorable. On one hand, this aligns with the intent of SCERP, wherein each crystal is offered a bespoke description of polarization that emphasizes their differences. On the other hand, the tailored model may not be equally accurate across all structures, potentially disrupting the systematic error-cancellation that benefits simpler models like PCM. Charged molecules in polar cells are an extreme example of this, for which SCERP's predictions are of much lower quality than for the same molecule in non-polar cells.

With this in mind, is a model development like SCERP still necessary? I would argue that it is. First and foremost, the fundamental basis of SCERP is more robust than that of PCM, and this has been observed on a case-by-case basis (e.g., F-IV of urea). SCERP may currently lack some large scale reliability, but its theoretical underpinnings offer clearer routes for model development that can rectify this. For instance, in Appendix K, I discuss the use of off-site charges in the SCERP-field; approaches like SCRMP [226] are also natural evolutions of the SCERP method that address the polar cell problem. In contrast, PCM can only be modulated through the choice of ε_{PCM}. This is a coarse instrument that may enhance the description of polarization in some structures, at the expense of others. Finally, I note that parameter estimates for the SCERP-compatible FFs may be suboptimal for two reasons. First, these FFs were parametrized with polar cell outliers present in the reference data, and as such, the current FF parameters must contain some undesirable artefact of error-correction for these structures. Second, less effort was also expended on their parametrization, relative to the Static- and PCM-compatible FFs. In short, SCERP-compatible FF parametrization should be re-attempted in earnest once the physical deficiencies of the SCERP model have been remedied.

Lastly, looking at the CSP problem as a whole, the present CSP results reiterate the notion that HAIEFF models still fall short of periodic DFT-D accuracy. Predictions of relative stability from PCM- and SCERP-embedded HAIEFF models exhibit little to no correlation with results from periodic DFT-D, suggesting an excess of random error in the HAIEFF models. This is consistent with the magnitude of SDs reported in Chapters 6 and 8, as well as benchmarking done by Nyman *et al.* [66]. Without significant reduction of this random error, HAIEFF model predictions can only continue being precursors to final re-ranking by periodic DFT-D.

9.9 Summary

In this chapter, I have investigated the model developments proposed in the current work, through five CSP case studies. The chosen systems span molecules with different flexibility and chemistry, as well as different classes of crystal (neat, hydrate, and cocrystal), all with a proclivity for forming intermolecular hydrogen bonds. Model accuracy at the Refinement stage has been analyzed across different pairings of empirical FFs and electrostatic induction models. Relative to the FFs proposed by Bowskill [2], the FFs parametrized in the present work do contribute

to better predictions in the absence of induction in the electrostatic model. When induction is introduced through PCM, the two sets of FFs perform similarly, and in most cases significantly outperform their PCM-free counterparts. Overall, the PCM-compatible FFs parametrized in the current work do generally give good predictions for hydrogen bonding systems, with most of the known forms found relatively low in energy ($\Delta U_{\text{latt}}^{\text{GM}} \leq 5$ kJ/mol) and rank ($\leq 50^{\text{th}}$) on their respective CSP landscapes.

When induction is introduced through SCERP, there is only tenuous improvement in the CSP results, mostly on a case-by-case basis. This is despite the much higher cost associated with conducting lattice energy minimizations within the SCERP paradigm. To reduce the cost of these SCERP-embedded calculations, I have compared the use of various approximate finite-difference gradients (Table 7.6). From this, relationships that aligns with intuition emerge between the fidelity of these FD approaches, their relative costs, and the minimization behavior. A good balance in cost and accuracy is struck by Mode 3, that should significantly reduce the cost of SCERP, whilst preserving agreement with the more-rigorous FD gradients.

A critical finding is the fact that SCERP (and PCM to a lesser extent) is deficient in modeling crystals with polar unit cells. Addressing this issue is paramount for reliable CSP predictions with either model. For PCM, it is unclear how this resolution may be achieved, whereas there are clearer avenues for model development in SCERP (e.g., Ref [226]). As such, I believe that further development of SCERP is warranted, despite its 'lukewarm' results here.

Chapter 10

Conclusions & Future Work

In this thesis, I have investigated the description of intermolecular interactions within the hybrid *ab initio*/empirical force-field class of CSP energy model. There are two key intermolecular contributions within the HAIEFF formulation: (1) the empirical (repulsion–dispersion) force-field term, and (2) the *ab initio*-derived electrostatic term. Through this work, I have aimed to address limitations in both terms, thereby holistically enhancing the intermolecular model.

I first contend with the parametrization of the empirical FF, which has arguably received less attention than the rest of the HAIEFF model. Fitting of accurate, transferable HAIEFF FFs has historically been hindered by the availability of reliable (and diverse) reference data. To address this, Bowskill [2] laid a foundation of high-quality, synthetic, periodic DFT-D data, to which I have systematically added complementary data to curate the CE–755 dataset [1, 256]. As a result, the CE–755 encapsulates more-comprehensively the industrially relevant chemical space of atomic environments and intermolecular hydrogen bonding motifs. The reliability of the data was also verified by comparison with available experimental data.

With access to high-quality and diverse data in the CE–755, I was able to extensively study HAIEFF FF parametrization. The focal point of this investigation was whether departing from the 'standard' HAIEFF FF formulation (that uses the Buckingham potential) could enable meaningful improvement in model accuracy. To that end, various fitting approaches were explored, encompassing parametrization strategies applied by other authors [2, 169, 198] and novel methodologies, like using the damped-Buckingham and Mie potentials. Ultimately, I demonstrated that the latter 'more-complex' FFs (Table 6.2, Schemes 5 and 7) provide effective descriptions of crystal geometry and energy, while 'conventional' parametrizations are hindered by the unphysical attributes of the BHP. Critically, further testing of these new FFs validated that they are adequately transferable, provided that the *ab initio* (electrostatic) HAIEFF component is defined congruently with the model used during FF training. Barring this, FF performance is liable to deteriorate significantly in the event of *ad hoc* FF-deployment. This is an important finding that must be recognized in all developments of the HAIEFF model.

To facilitate complementary-FF utilization, FF fitting was performed for electrostatic models that are relevant to different stages of the CSP workflow (point-charge/multipoles) and that are representative of distinct crystalline environments (Static/PCM). Model accuracy was consistently enhanced by implicitly embedding polarization into the electrostatic term, through the

use of PCM. This aligns with evidence indicating the importance of induction in organic systems [71, 183, 189] and motivates the second half of this work: the accurate modeling of polarization.

To this end, the SCERP model [71, 183, 189] was implemented into CSO–RM and CSO–FM. In the process, several decisions made by Zhang [71] were revisited. These affect algorithmic efficiency, precision of evaluated energies, and the limits of SCERP-embedded minimizations. With this revised implementation, I showed that using SCERP should yield more-reliable predictions of induction, relative to using PCM. Before its deployment into 'real' CSPs, and in adherence to the principle of complementary-FF utilization, I fitted SCERP-compatible HAIEFF FFs. Concurrently, I studied the need for electrostatic damping and the best approach to its utilization. From both numerical and physical perspectives, electrostatic damping is justified. However, I found that the prevailing method for determining damping parameters [71, 189, 250] is suboptimal, and better results are achieved with intuition-derived, atom-pair specific parameters.

Finally, these model developments were benchmarked through a series of CSP studies on hydrogen bonding systems. When used in tandem with PCM, the new FFs achieved good predictions in most cases. Geometries were predicted to excellent fidelity and the predicted relative stabilities were adequate precursors for re-ranking by periodic DFT-D, in the majority of known forms. When SCERP is applied, there are instances in which its physical-validity led to better predictions over PCM. Moreover, the high cost of SCERP-embedded CSO–FM can likely be managed through judicious use of approximate gradients.

Altogether, the current work influences several facets of state-of-the-art multistage CSP. First, the 'conventional' CSP stages based on atomic-charge/distributed-multipole HAIEFF models continue to offer competitive balances of cost and accuracy, making them desirable prescreening stages before re-ranking with more-accurate models. Importantly, the present results indicate that implicit induction models, like PCM, should be more routinely used in these stages. While the neglect of induction can be partially absorbed by the empirical HAIEFF component, leveraging error assimilation remains an inferior strategy compared to simply using PCM, especially since PCM contributes negligibly to computational cost. Critically, I have also shown that the benefits of employing PCM may be relatively insensitive towards the specific choice of dielectric constant, provided that a PCM-compatible HAIEFF parametrization is used. In light of this, I would urge practitioners not only to embrace modern HAIEFF parametrizations (like those developed here), but also to remain mindful of the continued need for routine parameterization in all future model developments. Broaching the task of HAIEFF parametrization, I would put forward the Mie potential as an effective substitute for the Buckingham potential. The former offers more modeling degrees of freedom and less unphysical behavior, which together facilitate parameter estimation and greater model performance.

Compared to their closest analogues (Bowskill's potentials [2]), the PCM-compatible Mie-potential HAIEFF models developed here demonstrate an equal or superior ability to prescreen the CSP landscape prior to final re-ranking. Considering their lower absolute accuracy during FF training, the competitive results achieved with Bowskill's potentials likely rely on fortuitous error cancellation more so than the newly parametrized models. In systems with stronger hydrogen

bonding and/or more diverse functionalities, I expect that the benefits of the new parametrization may become more apparent. This is supported by the amino acid examples, wherein the new Mie potentials predict tighter energy and rank cutoffs for the experimental forms, compared to Bowskill's potentials. Additionally, an understated benefit of the Mie parameterizations are their greater resilience against unphysical molecular fusing behavior. Due to their compatibility with PCM, the HAIEFF models developed here seem to offer better prescreening capabilities than incongruous HAIEFF models employing the FIT potential.

Finally, this work examines the prospect of introducing an additional prescreening step prior to final re-ranking, in the form of SCERP-embedded HAIEFF models, but the merits of this Refinement II stage remain somewhat uncertain. On one hand, the SCERP model has good theoretical underpinnings supporting its validity over PCM. On the other hand, the current implementation is shown to have deficiencies that obfuscate the appraisal of its prescreening capabilities. In particular, it is unclear whether improvements in the relative ranking of known polymorphs reflect genuine corrections or are merely artefacts of polar crystals rising in the landscape. Optimizations performed with SCERP-coupled models cost about one order of magnitude more than SCERP-free HAIEFF models, but one order of magnitude less than periodic DFT-D. As such, unless a significant enhancement in prescreening-capability can be achieved with the SCERP models, the inclusion of Refinement II may not be well justified. Ultimately, I believe there is strong physical justification to continue development of the SCERP approach, though its practicality in CSP will remain limited until its cost and accuracy are more robustly addressed. One cost-saving strategy that should be explored is the effectiveness of performing only *single-point* energy re-ranking with the SCERP-embedded model, as opposed to complete geometry optimization. Avenues to enhance model accuracy are discussed below.

Looking forward, the present work has also highlighted some critical limitations in the current suite of HAIEFF models. First and foremost is the treatment of polar unit cells, particularly when highly charged/polarized molecules are involved and when induction is closely tied to the details of the crystalline lattice. Due to the unique, long-ranged polarization that can emerge from polar unit cells, both PCM and SCERP provide subpar predictions of induction in these crystals. This error is magnified on a CSP landscape when the polar cells are compared to their less-problematic, non-polar counterparts. To bolster CSP reliability, it is critical that strategies for managing polar cells are developed. The relative simplicity of PCM likely limits this to managing relative error between polar/non-polar cells via judicious selection of $\varepsilon_{\mathrm{PCM}}$. Whilst only a band-aid solution, this is still important in the context of the hierarchical CSP workflow. To rigorously address the underlying model deficiency, more-sophisticated charge-embedding methods, like SCRMP [226], should be considered for SCERP's development. This would enable the SCERP methodology to more faithfully represent the crystalline electric field and enhance model accuracy in both polar and non-polar crystals.

Secondly, I echo the view of other authors [2, 66]: the empirical FF will continue to be a major bottleneck for HAIEFF model accuracy, as long as random error associated with its predictions remain in large excess of expected polymorphic energy differences. Taking inspiration from

molecular mechanics FFs, I believe significant diversification of atom types would be necessary for transferable HAIEFF FFs to attain this level of accuracy. However, this poses several challenges: (1) the rational design of atom types is not well defined, (2) further expansion of the CE–755 may prove necessary where data is currently scarce (e.g., sulphur and the halogens), and (3) this may ultimately be impractical for 'one-shot' parameter estimation. In lieu of this, *molecule-specific* HAIEFF FFs could prove to be a more prudent strategy. By removing the requirement for generality, a tailored FF should outperform even the best transferable FF, whilst only needing data to parametrize a limited set of atom types relevant to the molecule. However, given the need to generate molecule-specific data for this parametrization, such a bespoke model would be most appropriate in a secondary refinement (e.g., Refinement II in tandem with SCERP). To limit the need for synthesizing new data in each CSP, molecule-specific data could be supplemented by 'general' data [70] of chemically similar molecules from the CE–755 dataset. Ultimately, I believe HAIEFF models will remain precursors for periodic DFT-D re-ranking, irrespective of developments in the electrostatic model, until HAIEFF FF accuracy can be properly addressed.

Bibliography

[1] Bowskill, D. H.; Tan, B. I.; Keates, A.; Sugden, I. J.; Adjiman, C. S.; Pantelides, C. C. Large-Scale Parameter Estimation for Crystal Structure Prediction. Part 1: Dataset, Methodology, and Implementation. *Journal of Chemical Theory and Computation* **2024**, *20*, 10288–10315.

[2] Bowskill, D. H. Reliable and Efficient Parameter Estimation Methodologies for Crystal Structure Prediction. Ph.D. thesis, Imperial College London, 2021.

[3] Bondi, A. van der Waals Volumes and Radii. *The Journal of Physical Chemistry* **1964**, *68*, 441–451.

[4] London, U. C. The FIT and WILL01 empirical "repulsion-dispersion" potentials. 2017; `http://www.chem.ucl.ac.uk/cposs/dmacrys/fit_and_will01_empirical_potentials.htm`, Accessed: 2023-04-18.

[5] Variankaval, N.; Cote, A. S.; Doherty, M. F. From form to function: Crystallization of active pharmaceutical ingredients. *AIChE Journal* **2008**, *54*, 1682–1688.

[6] Olenik, B.; Keil, B.; Jeschke, P. Importance of chemical polymorphism in modern crop protection. *Pest Management Science* **2022**, *78*, 2746–2758.

[7] Panina, N. Crystal Structure and Morphology Prediction of Organic Pigments. Ph.D. thesis, Radboud University Nijmegen, 2009.

[8] Hunger, K. The effect of crystal structure on colour application properties of organic pigments*. *Review of Progress in Coloration and Related Topics* **1999**, *29*, 71–84.

[9] Aina, A. A.; Misquitta, A. J.; Phipps, M. J. S.; Price, S. L. Charge Distributions of Nitro Groups Within Organic Explosive Crystals: Effects on Sensitivity and Modeling. *ACS Omega* **2019**, *4*, 8614–8625.

[10] Song, Y.; Wang, Y.; Xu, R.; Zhang, Q. Research progress and prospect of explosive crystallization (2022-present). *Energetic Materials Frontiers* **2024**, *5*, 147–157, Advanced Technologies for Preparing Energetic Materials.

[11] Wang, C.; Dong, H.; Jiang, L.; Hu, W. Organic semiconductor crystals. *Chem. Soc. Rev.* **2018**, *47*, 422–500.

[12] Berry, D. J.; Steed, J. W. Pharmaceutical cocrystals, salts and multicomponent systems; intermolecular interactions and property based design. *Advanced Drug Delivery Reviews* **2017**, *117*, 3–24, Engineering of pharmaceutical cocrystals, salts and polymorphs: Advances and Challenges.

[13] Karagianni, A.; Malamatari, M.; Kachrimanis, K. Pharmaceutical Cocrystals: New Solid Phase Modification Approaches for the Formulation of APIs. *Pharmaceutics* **2018**, *10*.

[14] López-Cedrún, J.; Videla, S.; Burgueño, M.; Juárez, I.; Aboul-Hosn, S.; Martín-Granizo, R.; Grau, J.; Puche, M.; Gil-Diez, J.-L.; Hueto, J.-A.; Vaqué, A.; Sust, M.; Plata-Salamán, C.; Monner, A. Co-crystal of Tramadol-Celecoxib in Patients with Moderate to Severe Acute Post-surgical Oral Pain: A Dose-Finding, Randomised, Double-Blind, Placebo- and Active-Controlled, Multicentre, Phase II Trial. *Drugs R D* **2018**, *18*.

[15] Thakur, T. S.; Thakuria, R. Crystalline Multicomponent Solids: An Alternative for Addressing the Hygroscopicity Issue in Pharmaceutical Materials. *Crystal Growth & Design* **2020**, *20*, 6245–6265.

[16] Price, S. L.; Braun, D. E.; Reutzel-Edens, S. M. Can computed crystal energy landscapes help understand pharmaceutical solids? *Chem. Commun.* **2016**, *52*, 7065–7077.

[17] Brittain, H. G. *Polymorphism in Pharmaceutical Solids: Second edition*; CRC Press, 2009; pp 1–640.

[18] Stahly, G. P. Diversity in Single- and Multiple-Component Crystals. The Search for and Prevalence of Polymorphs and Cocrystals. *Crystal Growth & Design* **2007**, *7*, 1007–1026.

[19] Lee, A. Y.; Erdemir, D.; Myerson, A. S. Crystal Polymorphism in Chemical Process Development. *Annual Review of Chemical and Biomolecular Engineering* **2011**, *2*, 259–280, PMID: 22432619.

[20] Aitipamula, S.; Chow, P. S.; Tan, R. B. H. Polymorphism in cocrystals: a review and assessment of its significance. *CrystEngComm* **2014**, *16*, 3451–3465.

[21] Cruz-Cabeza, A. J.; Reutzel-Edens, S. M.; Bernstein, J. Facts and fictions about polymorphism. *Chem. Soc. Rev.* **2015**, *44*, 8619–8635.

[22] Kersten, K.; Kaur, R.; Matzger, A. Survey and analysis of crystal polymorphism in organic structures. *IUCrJ* **2018**, *5*, 124–129.

[23] McCrone, W. *Physics and Chemistry of the Organic Solid State*; Wiley-Interscience, 1965; Chapter 8, p 725–767.

[24] Nyman, J.; Day, G. M. Static and lattice vibrational energy differences between polymorphs. *CrystEngComm* **2015**, *17*, 5154–5165.

[25] Maher, A.; Rasmuson, Å. C.; Croker, D. M.; Hodnett, B. K. Solubility of the Metastable Polymorph of Piracetam (Form II) in a Range of Solvents. *Journal of Chemical & Engineering Data* **2012**, *57*, 3525–3531.

[26] Sacchetti, M. Thermodynamic Analysis of DSC Data for Acetaminophen Polymorphs. *Journal of Thermal Analysis and Calorimetry* **2000**, *63*, 345–350.

[27] Görbitz, C. H. Crystal structures of amino acids: from bond lengths in glycine to metal complexes and high-pressure polymorphs. *Crystallography Reviews* **2015**, *21*, 160–212.

[28] Dziubek, K.; Citroni, M.; Fanetti, S.; Cairns, A. B.; Bini, R. High-Pressure High-Temperature Structural Properties of Urea. *The Journal of Physical Chemistry C* **2017**, *121*, 2380–2387.

[29] Issa, N.; Karamertzanis, P. G.; Welch, G. W. A.; Price, S. L. Can the Formation of Pharmaceutical Cocrystals Be Computationally Predicted? I. Comparison of Lattice Energies. *Crystal Growth & Design* **2009**, *9*, 442–453.

[30] Karamertzanis, P. G.; Kazantsev, A. V.; Issa, N.; Welch, G. W.; Adjiman, C. S.; Pantelides, C. C.; Price, S. L. Can the Formation of Pharmaceutical Cocrystals Be Computationally Predicted? 2. Crystal Structure Prediction. *Journal of Chemical Theory and Computation* **2009**, *5*, 1432–1448, PMID: 26609729.

[31] Taylor, C. R.; Day, G. M. Evaluating the Energetic Driving Force for Cocrystal Formation. *Crystal Growth & Design* **2018**, *18*, 892–904.

[32] Sugden, I. J.; Braun, D. E.; Bowskill, D. H.; Adjiman, C. S.; Pantelides, C. C. Efficient Screening of Coformers for Active Pharmaceutical Ingredient Cocrystallization. *Crystal Growth & Design* **2022**, *22*, 4513–4527.

[33] Mullin, J. In *Crystallization (Fourth Edition)*, fourth edition ed.; Mullin, J., Ed.; Butterworth-Heinemann: Oxford, 2001; pp 1–31.

[34] Bučar, D.-K.; Lancaster, R. W.; Bernstein, J. Disappearing Polymorphs Revisited. *Angewandte Chemie International Edition* **2015**, *54*, 6972–6993.

[35] Chemburkar, S. R. et al. Dealing with the Impact of Ritonavir Polymorphs on the Late Stages of Bulk Drug Process Development. *Organic Process Research & Development* **2000**, *4*, 413–417.

[36] Karpinski, P. H. Polymorphism of Active Pharmaceutical Ingredients. *Chemical Engineering & Technology* **2006**, *29*, 233–237.

[37] Morissette, S. L.; Soukasene, S.; Levinson, D.; Cima, M. J.; Almarsson, Ö. Elucidation of crystal form diversity of the HIV protease inhibitor ritonavir by high-throughput crystallization. *Proceedings of the National Academy of Sciences* **2003**, *100*, 2180–2184.

[38] Mortazavi, M.; Hoja, J.; Aerts, L.; Quéré, L.; van de Streek, J.; Neumann, M.; Tkatchenko, A. Computational polymorph screening reveals late-appearing and poorly-soluble form of rotigotine. *Communications Chemistry* **2019**, *2*.

[39] Rietveld, I. B.; Céolin, R. Rotigotine: Unexpected Polymorphism with Predictable Overall Monotropic Behavior. *Journal of pharmaceutical sciences* **2015**, *104 12*, 4117–4122.

[40] Braga, D.; Casali, L.; Grepioni, F. The Relevance of Crystal Forms in the Pharmaceutical Field: Sword of Damocles or Innovation Tools? *International Journal of Molecular Sciences* **2022**, *23*.

[41] Nyman, J.; Reutzel-Edens, S. M. Crystal structure prediction is changing from basic science to applied technology. *Faraday Discuss.* **2018**, *211*, 459–476.

[42] Pantelides, C. C.; Adjiman, C. S.; Kazantsev, A. V. *Prediction and Calculation of Crystal Structures: Methods and Applications*; Springer International Publishing: Cham, 2014; pp 25–58.

[43] Lommerse, J. P. M.; Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Mooij, W. T. M.; Price, S. L.; Schweizer, B.; Schmidt, M. U.; van Eijck, B. P.; Verwer, P.; Williams, D. E. A test of crystal structure prediction of small organic molecules. *Acta Crystallographica Section B* **2000**, *56*, 697–714.

[44] Motherwell, W. D. S. et al. Crystal structure prediction of small organic molecules: a second blind test. *Acta Crystallographica Section B* **2002**, *58*, 647–661.

[45] Day, G. M. et al. A third blind test of crystal structure prediction. *Acta Crystallographica Section B* **2005**, *61*, 511–527.

[46] Day, G. M. et al. Significant progress in predicting the crystal structures of small organic molecules – a report on the fourth blind test. *Acta Crystallographica Section B* **2009**, *65*, 107–125.

[47] Bardwell, D. A. et al. Towards crystal structure prediction of complex organic compounds – a report on the fifth blind test. *Acta Crystallographica Section B* **2011**, *67*, 535–551.

[48] Reilly, A. M. et al. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallographica Section B* **2016**, *72*, 439–459.

[49] Hunnisett, L. M. et al. The seventh blind test of crystal structure prediction: structure generation methods. *Acta Crystallographica Section B* **2024**, *80*, 517–547.

[50] Bowskill, D. H.; Sugden, I. J.; Konstantinopoulos, S.; Adjiman, C. S.; Pantelides, C. C. Crystal Structure Prediction Methods for Organic Molecules: State of the Art. *Annual Review of Chemical and Biomolecular Engineering* **2021**, *12*, 593–623.

[51] Beran, G. J. O. Frontiers of molecular crystal structure prediction for pharmaceuticals and functional organic materials. *Chem. Sci.* **2023**, *14*, 13290–13312.

[52] Price, S. L. Is zeroth order crystal structure prediction (CSP_0) coming to maturity? What should we aim for in an ideal crystal structure prediction code? *Faraday Discuss.* **2018**, *211*, 9–30.

[53] Braun, D. E.; Orlova, M.; Griesser, U. J. Creatine: Polymorphs Predicted and Found. *Crystal Growth & Design* **2014**, *14*, 4895–4900.

[54] Braun, D. E.; Oberacher, H.; Arnhard, K.; Orlova, M.; Griesser, U. J. 4-Aminoquinaldine monohydrate polymorphism: prediction and impurity aided discovery of a difficult to access stable form. *CrystEngComm* **2016**, *18*, 4053–4067.

[55] Vasileiadis, M.; Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C. The polymorphs of ROY: application of a systematic crystal structure prediction technique. *Acta Crystallographica Section B* **2012**, *68*, 677–685.

[56] Beran, G. J. O.; Sugden, I. J.; Greenwell, C.; Bowskill, D. H.; Pantelides, C. C.; Adjiman, C. S. How many more polymorphs of ROY remain undiscovered. *Chem. Sci.* **2022**, *13*, 1288–1297.

[57] Taylor, C. R.; Mulvee, M. T.; Perenyi, D. S.; Probert, M. R.; Day, G. M.; Steed, J. W. Minimizing Polymorphic Risk through Cooperative Computational and Experimental Exploration. *Journal of the American Chemical Society* **2020**, *142*, 16668–16680.

[58] Arlin, J.-B.; Price, L. S.; Price, S. L.; Florence, A. J. A strategy for producing predicted polymorphs: catemeric carbamazepine form V. *Chem. Commun.* **2011**, *47*, 7074–7076.

[59] Mayo, R. A.; Marczenko, K. M.; Johnson, E. R. Quantitative matching of crystal structures to experimental powder diffractograms. *Chem. Sci.* **2023**, *14*, 4777–4785.

[60] Sparrow, G. M.; Mayo, R. A.; Johnson, E. R. Matching ROY crystal structures to high-throughput PXRD. *CrystEngComm* **2024**, –.

[61] Pulido, A. et al. Functional materials discovery using energy–structure–function maps. *Nature* **2017**, *543*, 657 – 664.

[62] Zhao, C.; Chen, L.; Che, Y.; Pang, Z.; Wu, X.; Lu, Y.; Liu, H.; Day, G. M.; Cooper, A. I. Digital navigation of energy–structure–function maps for hydrogen-bonded porous molecular crystals. *Nature Communications* **2021**, *12*.

[63] Aakeröy, C. B.; Salmon, D. J. Building co-crystals with molecular sense and supramolecular sensibility. *CrystEngComm* **2005**, *7*, 439–448.

[64] Saha, S.; Desiraju, G. R. Acid···Amide Supramolecular Synthon in Cocrystals: From Spectroscopic Detection to Property Engineering. *Journal of the American Chemical Society* **2018**, *140*, 6361–6373, PMID: 29697258.

[65] Taylor, C. R.; Butler, P. W. V.; Day, G. M. Predictive crystallography at scale: mapping, validating, and learning from 1000 crystal energy landscapes. *Faraday Discuss.* **2024**, –.

[66] Nyman, J.; Pundyke, O. S.; Day, G. M. Accurate force fields and methods for modelling organic molecular crystals at finite temperatures. *Phys. Chem. Chem. Phys.* **2016**, *18*, 15828–15837.

[67] Konstantinopoulos, S. Free Energy calculations for Crystal Structure Prediction studies. Ph.D. thesis, Imperial College London, 2023.

[68] Price, S. L. Why don't we find more polymorphs? *Acta Crystallographica Section B* **2013**, *69*, 313–328.

[69] Dybeck, E. C.; McMahon, D. P.; Day, G. M.; Shirts, M. R. Exploring the Multi-minima Behavior of Small Molecule Crystal Polymorphs at Finite Temperature. *Crystal Growth & Design* **2019**, *19*, 5568–5580.

[70] Bowskill, D. H.; Sugden, I. J.; George, N.; Keates, A.; Webb, J.; Pantelides, C. C.; Adjiman, C. S. In *30th European Symposium on Computer Aided Process Engineering*; Pierucci, S., Manenti, F., Bozzano, G. L., Manca, D., Eds.; Computer Aided Chemical Engineering; Elsevier, 2020; Vol. 48; pp 493–498.

[71] Zhang, Y. Crystal Structure Prediction for Cocrystals and Hydrates: Energy Models and Structure Generation. Ph.D. thesis, Imperial College London, 2023.

[72] Nyman, J.; Day, G. M. Modelling temperature-dependent properties of polymorphic organic molecular crystals. *Phys. Chem. Chem. Phys.* **2016**, *18*, 31132–31143.

[73] Price, S. L. Predicting crystal structures of organic compounds. *Chem. Soc. Rev.* **2014**, *43*, 2098–2111.

[74] Cambridge Crystallographic Data Centre *CSD Space Group Statistics – Space Group Frequency Ordering*; 2024.

[75] Cruz-Cabeza, A. J.; Bernstein, J. Conformational Polymorphism. *Chemical Reviews* **2014**, *114*, 2170–2191.

[76] Thompson, H. P. G.; Day, G. M. Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape. *Chem. Sci.* **2014**, *5*, 3173–3182.

[77] Dudek, M. K.; Drużbicki, K. Along the road to crystal structure prediction (CSP) of pharmaceutical-like molecules. *CrystEngComm* **2022**, *24*, 1665–1678.

[78] Nowell, H.; Price, S. L. Validation of a search technique for crystal structure prediction of flexible molecules by application to piracetam. *Acta Crystallographica Section B* **2005**, *61*, 558–568.

[79] Day, G. M.; S. Motherwell, W. D.; Jones, W. A strategy for predicting the crystal structures of flexible molecules: the polymorphism of phenobarbital. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1693–1704.

[80] Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864–B871.

[81] Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.

[82] Johnson, E. R.; Mackie, I. D.; DiLabio, G. A. Dispersion interactions in density-functional theory. *Journal of Physical Organic Chemistry* **2009**, *22*, 1127–1135.

[83] Hait, D.; Head-Gordon, M. Delocalization Errors in Density Functional Theory Are Essentially Quadratic in Fractional Occupation Number. *The Journal of Physical Chemistry Letters* **2018**, *9*, 6280–6288.

[84] Bryenton, K. R.; Adeleke, A. A.; Dale, S. G.; Johnson, E. R. Delocalization error: The greatest outstanding challenge in density-functional theory. *WIREs Computational Molecular Science* **2023**, *13*, e1631.

[85] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* **2010**, *132*, 154104.

[86] Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry* **2011**, *32*, 1456–1465.

[87] Tkatchenko, A.; Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **2009**, *102*, 073005.

[88] Becke, A. D.; Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction. *The Journal of Chemical Physics* **2005**, *122*, 154104.

[89] Becke, A. D.; Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction revisited. *The Journal of Chemical Physics* **2007**, *127*, 154108.

[90] Greenwell, C.; McKinley, J. L.; Zhang, P.; Zeng, Q.; Sun, G.; Li, B.; Wen, S.; Beran, G. J. O. Overcoming the difficulties of predicting conformational polymorph energetics in molecular crystals via correlated wavefunction methods. *Chem. Sci.* **2020**, *11*, 2200–2214.

[91] Greenwell, C.; Beran, G. J. O. Inaccurate Conformational Energies Still Hinder Crystal Structure Prediction in Flexible Organic Molecules. *Crystal Growth & Design* **2020**, *20*, 4875–4881.

[92] Aina, A. A.; Misquitta, A. J.; Price, S. L. From dimers to the solid-state: Distributed intermolecular force-fields for pyridine. *The Journal of Chemical Physics* **2017**, *147*, 161722.

[93] Karamertzanis, P. G.; Pantelides, C. C. Ab initio crystal structure prediction—I. Rigid molecules. *Journal of Computational Chemistry* **2005**, *26*, 304–324.

[94] Karamertzanis, P. G.; Pantelides, C. C. Ab initio crystal structure prediction. II. Flexible molecules. *Molecular Physics* **2007**, *105*, 273–291.

[95] Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C. Efficient Handling of Molecular Flexibility in Lattice Energy Minimization of Organic Crystals. *Journal of Chemical Theory and Computation* **2011**, *7*, 1998–2016.

[96] Habgood, M.; Sugden, I. J.; Kazantsev, A. V.; Adjiman, C. S.; Pantelides, C. C. Efficient Handling of Molecular Flexibility in Ab Initio Generation of Crystal Structures. *Journal of Chemical Theory and Computation* **2015**, *11*, 1957–1969.

[97] Sugden, I.; Adjiman, C. S.; Pantelides, C. C. Accurate and efficient representation of intramolecular energy in *ab initio* generation of crystal structures. I. Adaptive local approximate models. *Acta Crystallographica Section B* **2016**, *72*, 864–874.

[98] Sugden, I. J.; Adjiman, C. S.; Pantelides, C. C. Accurate and efficient representation of intramolecular energy in *ab initio* generation of crystal structures. II. Smoothed intramolecular potentials. *Acta Crystallographica Section B* **2019**, *75*, 423–433.

[99] Neumann, M. A. Tailor-Made Force Fields for Crystal-Structure Prediction. *The Journal of Physical Chemistry B* **2008**, *112*, 9810–9829.

[100] Neumann, M. A.; Leusen, F. J. J.; Kendrick, J. A Major Advance in Crystal Structure Prediction. *Angewandte Chemie International Edition* **2008**, *47*, 2427–2430.

[101] Mattei, A.; Hong, R. S.; Dietrich, H.; Firaha, D.; Helfferich, J.; Liu, Y. M.; Sasikumar, K.; Abraham, N. S.; Miglani Bhardwaj, R.; Neumann, M. A.; Sheikh, A. Y. Efficient Crystal Structure Prediction for Structurally Related Molecules with Accurate and Transferable Tailor-Made Force Fields. *Journal of Chemical Theory and Computation* **2022**, *18*, 5725–5738.

[102] Fredericks, S.; Parrish, K.; Sayre, D.; Zhu, Q. PyXtal: A Python library for crystal structure generation and symmetry analysis. *Computer Physics Communications* **2021**, *261*, 107810.

[103] Sobol', I. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics* **1967**, *7*, 86–112.

[104] Case, D. H.; Campbell, J. E.; Bygrave, P. J.; Day, G. M. Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling. *Journal of Chemical Theory and Computation* **2016**, *12*, 910–924.

[105] Avant-Garde Materials Simulations GRACE. 2024; `https://www.avmatsim.eu/`, Accessed: 2024-10-01.

[106] Curtis, F.; Li, X.; Rose, T.; Vázquez-Mayagoitia, A.; Bhattacharya, S.; Ghiringhelli, L. M.; Marom, N. GAtor: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction. *Journal of Chemical Theory and Computation* **2018**, *14*, 2246–2264.

[107] Curtis, F.; Rose, T.; Marom, N. Evolutionary niching in the GAtor genetic algorithm for molecular crystal structure prediction. *Faraday Discuss.* **2018**, *211*, 61–77.

[108] XTalPi XTalPi: Drug Design Partner. 2024; `https://www.xtalpi.com/en`, Accessed: 2024-10-01.

[109] Zhang, P.; Wood, G. P. F.; Ma, J.; Yang, M.; Liu, Y.; Sun, G.; Jiang, Y. A.; Hancock, B. C.; Wen, S. Harnessing Cloud Architecture for Crystal Structure Prediction Calculations. *Crystal Growth & Design* **2018**, *18*, 6891–6900.

[110] Cambridge Crystallographic Data Centre CSD Landscape Generator. 2024; `https://www.ccdc.cam.ac.uk/solutions/software/csd-landscape-generator/`, Accessed: 2024-10-01.

[111] Jensen, F. *Introduction to Computational Chemistry*; Wiley, 2007.

[112] Sherrill, C. D. An Introduction to Hartree-Fock Molecular Orbital Theory. 2000.

[113] Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618–622.

[114] Maschio, L.; Usvyat, D.; Manby, F. R.; Casassa, S.; Pisani, C.; Schütz, M. Fast local-MP2 method with density-fitting for crystals. I. Theory and algorithms. *Phys. Rev. B* **2007**, *76*, 075101.

[115] Usvyat, D.; Maschio, L.; Manby, F. R.; Casassa, S.; Schütz, M.; Pisani, C. Fast local-MP2 method with density-fitting for crystals. II. Test calculations and application to the carbon dioxide crystal. *Phys. Rev. B* **2007**, *76*, 075102.

[116] Maschio, L. Local MP2 with Density Fitting for Periodic Systems: A Parallel Implementation. *Journal of Chemical Theory and Computation* **2011**, *7*, 2818–2830.

[117] Cremer, D. Møller–Plesset perturbation theory: from small molecule methods to methods for thousands of atoms. *WIREs Computational Molecular Science* **2011**, *1*, 509–530.

[118] Beran, G. J. O. Modeling Polymorphic Molecular Crystals with Electronic Structure Theory. *Chemical Reviews* **2016**, *116*, 5567–5613.

[119] Riley, K. E.; Pitoňák, M.; Jurečka, P.; Hobza, P. Stabilization and Structure Calculations for Noncovalent Interactions in Extended Molecular Systems Based on Wave Function and Density Functional Theories. *Chemical Reviews* **2010**, *110*, 5023–5063.

[120] Rappoport, D.; Crawford, N. R. M.; Furche, F.; Burke, K. *Encyclopedia of Inorganic Chemistry*; John Wiley & Sons, Ltd, 2009.

[121] Burns, L. A.; Vázquez-Mayagoitia, Å.; Sumpter, B. G.; Sherrill, C. D. Density-functional approaches to noncovalent interactions: A comparison of dispersion corrections (DFT-D), exchange-hole dipole moment (XDM) theory, and specialized functionals. *The Journal of Chemical Physics* **2011**, *134*, 084107.

[122] Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *The Journal of Chemical Physics* **2005**, *123*.

[123] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

[124] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Phys. Rev. Lett.* **1997**, *78*, 1396–1396.

[125] Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of chemical physics* **1999**, *110*, 6158–6170.

[126] Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **1993**, *98*, 5648–5652.

[127] Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789.

[128] Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. Results obtained with the correlation energy density functionals of becke and Lee, Yang and Parr. *Chemical Physics Letters* **1989**, *157*, 200–206.

[129] Perdew, J. P.; Ruzsinszky, A.; Constantin, L. A.; Sun, J.; Csonka, G. I. Some Fundamental Issues in Ground-State Density Functional Theory: A Guide for the Perplexed. *Journal of Chemical Theory and Computation* **2009**, *5*, 902–908.

[130] Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta–Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.

[131] Otero-de-la Roza, A.; Johnson, E. R. A benchmark for non-covalent interactions in solids. *The Journal of Chemical Physics* **2012**, *137*, 054103.

[132] Brandenburg, J. G.; Grimme, S. Organic crystal polymorphism: a benchmark for dispersion-corrected mean-field electronic structure methods. *Acta Crystallographica Section B* **2016**, *72*, 502–513.

[133] Otero-de-la Roza, A.; Johnson, E. R. Many-body dispersion interactions from the exchange-hole dipole moment model. *The Journal of Chemical Physics* **2013**, *138*, 054103.

[134] Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. Van der Waals Density Functional for General Geometries. *Phys. Rev. Lett.* **2004**, *92*, 246401.

[135] Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C. Dispersion-Corrected Mean-Field Electronic Structure Methods. *Chemical Reviews* **2016**, *116*, 5105–5154.

[136] Otero-de-la Roza, A.; LeBlanc, L. M.; Johnson, E. R. What is "many-body" dispersion and should I worry about it? *Phys. Chem. Chem. Phys.* **2020**, *22*, 8266–8276.

[137] Dobson, J. F. Beyond pairwise additivity in London dispersion interactions. *International Journal of Quantum Chemistry* **2014**, *114*, 1157–1161.

[138] Axilrod, B. M.; Teller, E. Interaction of the van der Waals Type Between Three Atoms. *The Journal of Chemical Physics* **1943**, *11*, 299–300.

[139] Muto, Y. Force between nonpolar molecules. *J. Phys. Math. Soc. Jpn* **1943**, *17*, 629–631.

[140] Tkatchenko, A.; DiStasio, R. A.; Car, R.; Scheffler, M. Accurate and Efficient Method for Many-Body van der Waals Interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.

[141] Ambrosetti, A.; Reilly, A. M.; DiStasio, J., Robert A.; Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. *The Journal of Chemical Physics* **2014**, *140*, 18A508.

[142] Reilly, A. M.; Tkatchenko, A. Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals. *The Journal of Chemical Physics* **2013**, *139*, 024705.

[143] Moellmann, J.; Grimme, S. DFT-D3 Study of Some Molecular Crystals. *The Journal of Physical Chemistry C* **2014**, *118*, 7615–7621.

[144] Bhaskar Rana, G. J. O. B.; Herbert, J. M. Correcting π-delocalisation errors in conformational energies using density-corrected DFT, with application to crystal polymorphs. *Molecular Physics* **2023**, *121*, e2138789.

[145] Wen, S.; Nanda, K.; Huang, Y.; Beran, G. J. O. Practical quantum mechanics-based fragment methods for predicting molecular crystal properties. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7578–7590.

[146] Chung, L. W.; Sameera, W. M. C.; Ramozzi, R.; Page, A. J.; Hatanaka, M.; Petrova, G. P.; Harris, T. V.; Li, X.; Ke, Z.; Liu, F.; Li, H.-B.; Ding, L.; Morokuma, K. The ONIOM Method and Its Applications. *Chemical Reviews* **2015**, *115*, 5678–5796.

[147] Beran, G. J. O.; Nanda, K. Predicting Organic Crystal Lattice Energies with Chemical Accuracy. *The Journal of Physical Chemistry Letters* **2010**, *1*, 3480–3487.

[148] Wen, S.; Beran, G. J. O. Accurate Molecular Crystal Lattice Energies from a Fragment QM/MM Approach with On-the-Fly Ab Initio Force Field Parametrization. *Journal of Chemical Theory and Computation* **2011**, *7*, 3733–3742.

[149] Wen, S.; Beran, G. J. O. Accidental Degeneracy in Crystalline Aspirin: New Insights from High-Level ab Initio Calculations. *Crystal Growth & Design* **2012**, *12*, 2169–2172.

[150] Misquitta, A. J.; Szalewicz, K. Symmetry-adapted perturbation-theory calculations of intermolecular forces employing density-functional description of monomers. *The Journal of Chemical Physics* **2005**, *122*, 214109.

[151] Szalewicz, K. Symmetry-adapted perturbation theory of intermolecular forces. *WIREs Computational Molecular Science* **2012**, *2*, 254–272.

[152] Schriber, J.; Cheney, D.; Sherrill, C. D. Levels of symmetry adapted perturbation theory (SAPT). II. Convergence of interaction energy components. *ChemRxiv* **2023**,

[153] Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: a generic force field for molecular simulations. *The Journal of Physical Chemistry* **1990**, *94*, 8897–8909.

[154] Maple, J. R.; Dinur, U.; Hagler, A. T. Derivation of force fields for molecular mechanics and dynamics from ¡i¿ab initio¡/i¿ energy surfaces. *Proceedings of the National Academy of Sciences* **1988**, *85*, 5350–5354.

[155] Sun, H. COMPASS: An ab Initio Force-Field Optimized for Condensed-Phase ApplicationsOverview with Details on Alkane and Benzene Compounds. *The Journal of Physical Chemistry B* **1998**, *102*, 7338–7364.

[156] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* **1996**, *118*, 11225–11236.

[157] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

[158] Marchese Robinson, R. L.; Geatches, D.; Morris, C.; Mackenzie, R.; Maloney, A. G. P.; Roberts, K. J.; Moldovan, A.; Chow, E.; Pencheva, K.; Vatvani, D. R. M. Evaluation of Force-Field Calculations of Lattice Energies on a Large Public Dataset, Assessment of Pharmaceutical Relevance, and Comparison to Density Functional Theory. *Journal of Chemical Information and Modeling* **2019**, *59*, 4778–4792.

[159] Sun, H.; Jin, Z.; Yang, C.; Akkermans, R. L. C.; Robertson, S. H.; Spenley, N. A.; Miller, S.; Todd, S. M. COMPASS II: extended coverage for polymer and drug-like molecule databases. *Journal of Molecular Modeling* **2016**, *22*, 47.

[160] Neumann, M. A.; Perrin, M.-A. Energy Ranking of Molecular Crystals Using Density Functional Theory Calculations and an Empirical van der Waals Correction. *The Journal of Physical Chemistry B* **2005**, *109*, 15531–15541.

[161] Yao, C.; Guzei, I. A.; Jin, Y.; Ruan, S.; Sun, G.; Gui, Y.; Wang, L.; Yu, L. Polymorphism of Piroxicam: New Polymorphs by Melt Crystallization and Crystal Structure Prediction. *Crystal Growth & Design* **2020**, *20*, 7874–7881.

[162] Li, X.; Ou, X.; Wang, B.; Rong, H.; Wang, B.; Chang, C.; Shi, B.; Yu, L.; Lu, M. Rich polymorphism in nicotinamide revealed by melt crystallization and crystal structure prediction. *Communications Chemistry* **2020**, *3*, 152.

[163] Sun, G.; Liu, X.; Abramov, Y. A.; Nilsson Lill, S. O.; Chang, C.; Burger, V.; Broo, A. Current State-of-the-art In-house and Cloud-Based Applications of Virtual Polymorph Screening of Pharmaceutical Compounds: A Challenging Case of AZD1305. *Crystal Growth & Design* **2021**, *21*, 1972–1983.

[164] Sun, G.; Jin, Y.; Li, S.; Yang, Z.; Shi, B.; Chang, C.; Abramov, Y. A. Virtual Coformer Screening by Crystal Structure Predictions: Crucial Role of Crystallinity in Pharmaceutical Cocrystallization. *The Journal of Physical Chemistry Letters* **2020**, *11*, 8832–8838.

[165] Aina, A. A.; Misquitta, A. J.; Price, S. L. A non-empirical intermolecular force-field for trinitrobenzene and its application in crystal structure prediction. *The Journal of Chemical Physics* **2021**, *154*, 094123.

[166] Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M. Role of Electrostatic Interactions in Determining the Crystal Structures of Polar Organic Molecules. A Distributed Multipole Study. *The Journal of Physical Chemistry* **1996**, *100*, 7352–7360.

[167] Day, G. M.; Motherwell, W. D. S.; Jones, W. Beyond the Isotropic Atom Model in Crystal Structure Prediction of Rigid Molecules: Atomic Multipoles versus Point Charges. *Crystal Growth & Design* **2005**, *5*, 1023–1033.

[168] Kramer, C.; Spinn, A.; Liedl, K. R. Charge Anisotropy: Where Atomic Multipoles Matter Most. *Journal of Chemical Theory and Computation* **2014**, *10*, 4488–4496.

[169] Pyzer-Knapp, E.; Thompson, H.; Day, G. An optimized intermolecular force field for hydrogen-bonded organic molecular crystals using atomic multipole electrostatics. *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials* **2016**, *72*, 477 – 487.

[170] Manz, T. A.; Limas, N. G. Introducing DDEC6 atomic population analysis: part 1. Charge partitioning theory and methodology. *RSC Adv.* **2016**, *6*, 47771–47801.

[171] Campañá, C.; Mussard, B.; Woo, T. K. Electrostatic Potential Derived Atomic Charges for Periodic Systems Using a Modified Error Functional. *Journal of Chemical Theory and Computation* **2009**, *5*, 2866–2878.

[172] Golze, D.; Hutter, J.; Iannuzzi, M. Wetting of water on hexagonal boron nitride@Rh(111): a QM/MM model based on atomic charges derived for nano-structured substrates. *Phys. Chem. Chem. Phys.* **2015**, *17*, 14307–14316.

[173] Hu, H.; Lu, Z.; Yang, W. Fitting molecular electrostatic potentials from quantum mechanical calculations. *Journal of chemical theory and computation* **2007**, *3*, 1004–1013.

[174] Breneman, C. M.; Wiberg, K. B. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *Journal of Computational Chemistry* **1990**, *11*, 361–373.

[175] Singh, U. C.; Kollman, P. A. An approach to computing electrostatic charges for molecules. *Journal of Computational Chemistry* **1984**, *5*, 129–145.

[176] Besler, B. H.; Merz Jr., K. M.; Kollman, P. A. Atomic charges derived from semiempirical methods. *Journal of Computational Chemistry* **1990**, *11*, 431–439.

[177] Winn, P. J.; Ferenczy, G. G.; Reynolds, C. A. Toward Improved Force Fields. 1. Multipole-Derived Atomic Charges. *The Journal of Physical Chemistry A* **1997**, *101*, 5437–5445.

[178] Stone, A. Distributed multipole analysis, or how to describe a molecular charge distribution. *Chemical Physics Letters* **1981**, *83*, 233–239.

[179] Stone, A. J.; Alderton, M. Distributed multipole analysis. *Molecular Physics* **1985**, *56*, 1047–1064.

[180] Stone, A. J.; Alderton, M. Distributed multipole analysis Methods and applications. *Molecular Physics* **2002**, *100*, 221–233.

[181] Stone, A. J. Distributed Multipole Analysis: Stability for Large Basis Sets. *Journal of Chemical Theory and Computation* **2005**, *1*, 1128–1132.

[182] Stone, A. J. Distributed Multipole Analysis of Gaussian wavefunctions.

[183] Welch, G. The induction contribution to the lattice energy of organic crystals. Ph.D. thesis, University College London, 2010.

[184] Unke, O. T.; Devereux, M.; Meuwly, M. Minimal distributed charges: Multipolar quality at the cost of point charge electrostatics. *The Journal of Chemical Physics* **2017**, *147*, 161712.

[185] Bultinck, P.; Van Alsenoy, C.; Ayers, P. W.; Carbó-Dorca, R. Critical analysis and extension of the Hirshfeld atoms in molecules. *The Journal of Chemical Physics* **2007**, *126*, 144111.

[186] Lillestolen, T. C.; Wheatley, R. J. Redefining the atom: atomic charge densities produced by an iterative stockholder approach. *Chem. Commun.* **2008**, 5909–5911.

[187] Cooper, T. G.; Jones, W.; Motherwell, W. D. S.; Day, G. M. Database guided conformation selection in crystal structure prediction of alanine. *CrystEngComm* **2007**, *9*, 595–602.

[188] Cooper, T. G.; Hejczyk, K. E.; Jones, W.; Day, G. M. Molecular Polarization Effects on the Relative Energies of the Real and Putative Crystal Structures of Valine. *Journal of Chemical Theory and Computation* **2008**, *4*, 1795–1805.

[189] Welch, G.; Karamertzanis, P.; Misquitta, A.; Stone, A.; Price, S. Is the Induction Energy Important for Modeling Organic Crystals? *Journal of Chemical Theory and Computation - J CHEM THEORY COMPUT* **2008**, *4*, 522–532.

[190] Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chemical Reviews* **2005**, *105*, 2999–3094.

[191] Day, G. M.; Cooper, T. G. Crystal packing predictions of the alpha-amino acids: methods assessment and structural observations. *CrystEngComm* **2010**, *12*, 2443–2453.

[192] Williams, D. Improved intermolecular force field for crystalline hydrocarbons containing four- or three-coordinated carbon. *Journal of Molecular Structure* **1999**, *485-486*, 321–347.

[193] Williams, D. Improved Intermolecular Force Field for Crystalline Oxohydrocarbons Including O–H ... O Hydrogen Bonding. *Journal of Computational Chemistry* **2001**, *22*, 1–20.

[194] Williams, D. E.; Cox, S. R. Nonbonded potentials for azahydrocarbons: the importance of the Coulombic interaction. *Acta Crystallographica Section B* **1984**, *40*, 404–417.

[195] Cox, S. R.; Hsu, L. Y.; Williams, D. E. Nonbonded Potential Function Models for Crystalline Oxohydrocarbons. *Acta Crystallographica Section A* **1981**, *37*, 293–301.

[196] Williams, D. E.; Houpt, D. J. Fluorine nonbonded potential parameters derived from crystalline perfluorocarbons. *Acta Crystallographica Section B* **1986**, *42*, 286–295.

[197] Hsu, L. Y.; Williams, D. E. G. Intermolecular potential-function models for crystalline perchlorohydrocarbons. *Acta Crystallographica Section A* **1980**, *36*, 277–281.

[198] Beyer, T.; Price, S. L. Dimer or Catemer? Low-Energy Crystal Packings for Small Carboxylic Acids. *The Journal of Physical Chemistry B* **2000**, *104*, 2647–2655.

[199] Williams, D. Improved intermolecular force field for molecules containing H, C, N, and O atoms, with application to nucleoside and peptide crystals. *Journal of Computational Chemistry* **2001**, *22*, 1154–1166.

[200] Stone, A. J.; Misquitta, A. J. Atom–atom potentials from ab initio calculations. *International Reviews in Physical Chemistry* **2007**, *26*, 193–222.

[201] Misquitta, A. J.; Welch, G. W. A.; Stone, A. J.; Price, S. L. A first principles prediction of the crystal structure of C6Br2ClFH2. *Chemical Physics Letters* **2008**, *456*, 105–109.

[202] Misquitta, A. J.; Stone, A. J. Ab Initio Atom–Atom Potentials Using CamCASP: Theory and Application to Many-Body Models for the Pyridine Dimer. *Journal of Chemical Theory and Computation* **2016**, *12*, 4184–4208.

[203] Podeszwa, R.; Bukowski, R.; Rice, B. M.; Szalewicz, K. Potential energy surface for cyclotrimethylene trinitramine dimer from symmetry-adapted perturbation theory. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5561–5569.

[204] Podeszwa, R.; Rice, B. M.; Szalewicz, K. Predicting Structure of Molecular Crystals from First Principles. *Phys. Rev. Lett.* **2008**, *101*, 115503.

[205] Podeszwa, R.; Rice, B. M.; Szalewicz, K. Crystal structure prediction for cyclotrimethylene trinitramine (RDX) from first principles. *Phys. Chem. Chem. Phys.* **2009**, *11*, 5512–5518.

[206] Chickos, J. S. Enthalpies of Sublimation after a Century of Measurement: A View as Seen through the Eyes of a Collector. *Netsu Sokutei* **2003**, *30*, 116–124.

[207] Kulik, H. J. et al. Roadmap on Machine learning in electronic structure. *Electronic Structure* **2022**, *4*, 023004.

[208] McDonagh, D.; Skylaris, C.-K.; Day, G. M. Machine-Learned Fragment-Based Energies for Crystal Structure Prediction. *Journal of Chemical Theory and Computation* **2019**, *15*, 2743–2758.

[209] Egorova, O.; Hafizi, R.; Woods, D. C.; Day, G. M. Multi-fidelity Statistical Machine Learning for Molecular Crystal Structure Prediction. *The Journal of Physical Chemistry A* **2020**, *124*, 8065–8078.

[210] Wengert, S.; Csányi, G.; Reuter, K.; Margraf, J. T. Data-efficient machine learning for molecular crystal structure prediction. *Chem. Sci.* **2021**, *12*, 4536–4546.

[211] Butler, P. W. V.; Hafizi, R.; Day, G. M. Machine-Learned Potentials by Active Learning from Organic Crystal Structure Prediction Landscapes. *The Journal of Physical Chemistry A* **2024**, *128*, 945–957.

[212] Kapil, V.; Engel, E. A. A complete description of thermodynamic stabilities of molecular crystals. *Proceedings of the National Academy of Sciences* **2022**, *119*, e2111769119.

[213] Gatsiou, C. A.; Adjiman, C. S.; Pantelides, C. C. Repulsion–dispersion parameters for the modelling of organic molecular crystals containing N, O, S and Cl. *Faraday Discuss.* **2018**, *211*, 297–323.

[214] Chickos, J. S.; Gavezzotti, A. Sublimation Enthalpies of Organic Compounds: A Very Large Database with a Match to Crystal Structure Determinations and a Comparison with Lattice Energies. *Crystal Growth & Design* **2019**, *19*, 6566–6576.

[215] Heit, Y. N.; Beran, G. J. O. How important is thermal expansion for predicting molecular crystal structures and thermochemistry at finite temperatures? *Acta Crystallographica Section B* **2016**, *72*, 514–529.

[216] Mayo, R. A.; Johnson, E. R. Improved quantitative crystal-structure comparison using powder diffractograms via anisotropic volume correction. *CrystEngComm* **2021**, *23*, 7118–7131.

[217] Johnson, E. R. Dependence of dispersion coefficients on atomic environment. *The Journal of Chemical Physics* **2011**, *135*, 234109.

[218] Tulip, P. R.; Clark, S. J. Lattice dynamical and dielectric properties of L-amino acids. *Phys. Rev. B* **2006**, *74*, 064301.

[219] Schatschneider, B.; Liang, J.-J.; Reilly, A. M.; Marom, N.; Zhang, G.-X.; Tkatchenko, A. Electrodynamic response and stability of molecular crystals. *Phys. Rev. B* **2013**, *87*, 060104.

[220] Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; Wang, J. Polarization effects in molecular mechanical force fields. *Journal of Physics: Condensed Matter* **2009**, *21*, 333102.

[221] Stone, A. *The Theory of Intermolecular Forces*; Oxford University Press, 2016.

[222] Misquitta, A. J.; Stone, A. J. Distributed polarizabilities obtained using a constrained density-fitting algorithm. *The Journal of Chemical Physics* **2006**, *124*, 024111.

[223] Misquitta, A. J.; Stone, A. J. Accurate Induction Energies for Small Organic Molecules: 1. Theory. *Journal of Chemical Theory and Computation* **2008**, *4*, 7–18.

[224] Misquitta, A. J.; Stone, A. J.; Price, S. L. Accurate Induction Energies for Small Organic Molecules. 2. Development and Testing of Distributed Polarizability Models against SAPT(DFT) Energies. *Journal of Chemical Theory and Computation* **2008**, *4*, 19–32.

[225] Misquitta, A. J.; Stone, A. J. ISA-Pol: distributed polarizabilities and dispersion models from a basis-space implementation of the iterated stockholder atoms procedure. *Theoretical Chemistry Accounts* **2018**, *137*, 153.

[226] Hartman, J. D.; Balaji, A.; Beran, G. J. O. Improved Electrostatic Embedding for Fragment-Based Chemical Shift Calculations in Molecular Crystals. *Journal of Chemical Theory and Computation* **2017**, *13*, 6043–6051.

[227] Frisch, M. J. et al. Gaussian 09, Revision D.01. 2010; Gaussian Inc. Wallingford CT.

[228] van Eijck, B. P.; Kroon, J. Coulomb Energy of Polar Crystals. *The Journal of Physical Chemistry B* **1997**, *101*, 1096–1100.

[229] van Eijck, B. P.; Kroon, J. Comment on "Crystal Structure Prediction by Global Optimization as a Tool for Evaluating Potentials: Role of the Dipole Moment Correction Term in Successful Predictions". *The Journal of Physical Chemistry B* **2000**, *104*, 8089–8089.

[230] Wedemeyer, W. J.; Arnautova, Y. A.; Pillardy, J.; Wawak, R. J.; Czaplewski, C.; Scheraga, H. A. Reply to "Comment on 'Crystal Structure Prediction by Global Optimization as a Tool for Evaluating Potentials: Role of the Dipole Moment Correction Term in Successful Predictions'" by B. P. van Eijck and J. Kroon. *The Journal of Physical Chemistry B* **2000**, *104*, 8090–8092.

[231] Nijboer, B.; De Wette, F. On the calculation of lattice sums. *Physica* **1957**, *23*, 309–321.

[232] Vasileiadis, M. Calculation of the free energy of crystalline solids. Ph.D. thesis, Imperial College London, 2013.

[233] Karamertzanis, P. Prediction of Crystal Structure of Molecular Solids. Ph.D. thesis, Imperial College London, 2004.

[234] Hirata, S. Fast electron-correlation methods for molecular crystals: An application to the α, β_1, and β_2 modifications of solid formic acid. *The Journal of Chemical Physics* **2008**, *129*, 204104.

[235] Blow, D. *Outline of Crystallography for Biologists*; Oxford University Press, 2002; Chapter 02.

[236] Aroyo, M. I., Ed. *International Tables for Crystallography, Volume A: Space-Group Symmetry*, 6th ed.; International Tables for Crystallography; Wiley: Chichester, UK, 2016; p 896.

[237] Macrae, C. F.; Sovago, I.; Cottrell, S. J.; Galek, P. T. A.; McCabe, P.; Pidcock, E.; Platings, M.; Shields, G. P.; Stevens, J. S.; Towler, M.; Wood, P. A. *Mercury 4.0*: from visualization to analysis, design and prediction. *Journal of Applied Crystallography* **2020**, *53*, 226–235.

[238] Frenkel, D.; Smit, B. In *Understanding Molecular Simulation (Second Edition)*, second edition ed.; Frenkel, D., Smit, B., Eds.; Academic Press: San Diego, 2002; pp 291–320.

[239] Numerical Algorithms Group The NAG Library. `https://www.nag.com/`.

[240] The Numerical Algorithms Group (NAG) The NAG Library for Fortran. `https://www.nag.com`, Last accessed on 16.03.2021.

[241] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallographica Section B* **2016**, *72*, 171–179.

[242] Chisholm, J. A.; Motherwell, S. *COMPACK*: a program for identifying crystal structure similarity using distances. *Journal of Applied Crystallography* **2005**, *38*, 228–231.

[243] Mayo, R. A.; Otero-de-la Roza, A.; Johnson, E. R. Development and assessment of an improved powder-diffraction-based method for molecular crystal structure similarity. *CrystEngComm* **2022**, *24*, 8326–8338.

[244] Frisch, M. J. et al. Gaussian16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

[245] Kresse, G.; Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **1993**, *47*, 558–561.

[246] Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science* **1996**, *6*, 15–50.

[247] Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **1996**, *54*, 11169–11186.

[248] Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **1999**, *59*, 1758–1775.

[249] Hafner, J. *Ab-initio* simulations of materials using VASP: Density-functional theory and beyond. *Journal of Computational Chemistry* **2008**, *29*, 2044–2078.

[250] Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M. Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. *Phys. Chem. Chem. Phys.* **2010**, *12*, 8478–8490.

[251] Gatsiou, C.-A. Improving the accuracy of lattice energy calculations in crystal structure prediction using experimental data. Ph.D. thesis, Imperial College London, 2016.

[252] Desiraju, G. R. Supramolecular Synthons in Crystal Engineering—A New Organic Synthesis. *Angewandte Chemie International Edition in English* **1995**, *34*, 2311–2327.

[253] Bernstein, J.; Davis, R. E.; Shimoni, L.; Chang, N.-L. Patterns in Hydrogen Bonding: Functionality and Graph Set Analysis in Crystals. *Angewandte Chemie International Edition in English* **1995**, *34*, 1555–1573.

[254] Spek, A. L. Single-crystal structure validation with the program PLATON. *Journal of Applied Crystallography* **2003**, *36*, 7–13.

[255] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallographica Section B* **2016**, *72*, 171–179.

[256] Bowskill, D. H.; Tan, B. I.; Keates, A.; Sugden, I. J.; Adjiman, C. S.; Pantelides, C. C. CE-755 DFT-D3 Optimised Crystal Structure Database. 2024; `https://zenodo.org/records/7813566`, Version 1.0.

[257] Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *npj Computational Materials* **2015**, *1*, 15010.

[258] Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *Journal of Metals* **2013**, *65*, 1501–1509.

[259] van de Streek, J.; Neumann, M. A. Validation of experimental molecular crystal structures with dispersion-corrected density functional theory calculations. *Acta Crystallographica Section B* **2010**, *66*, 544–558.

[260] van de Streek, J.; Neumann, M. A. Validation of molecular crystal structures from powder diffraction data with dispersion-corrected density functional theory (DFT-D). *Acta Crystallographica Section B* **2014**, *70*, 1020–1032.

[261] Chatterjee, S.; Deb, U.; Datta, S.; Walther, C.; Gupta, D. K. Common explosives (TNT, RDX, HMX) and their fate in the environment: Emphasizing bioremediation. *Chemosphere* **2017**, *184*, 438–451.

[262] Pfaff, R. A.; Beltz, J.; King, E.; Ercal, N. Medicinal Thiols: Current Status and New Perspectives. *Mini-Reviews in Medicinal Chemistry* **2020**, *20*, 513–529.

[263] Etter, M. C.; MacDonald, J. C.; Bernstein, J. Graph-set analysis of hydrogen-bond patterns in organic crystals. *Acta Crystallographica Section B* **1990**, *46*, 256–262.

[264] Stephenson, R. M.; Malanowski, S. K. *Handbook of the thermodynamics of organic compounds*; Springer, 1987.

[265] Hirt, R. C.; Steger, J. E.; Simard, G. L. Vapor pressure of 2,4,6-triamino-s-triazine (melamine). *Journal of Polymer Science* **1960**, *43*, 319–323.

[266] Ribeiro da Silva, M. A. V.; Monte, M. J. S.; Matos, M. A. R. Enthalpies of combustion, vapour pressures, and enthalpies of sublimation of 8-hydroxyquinoline, 5-nitro-8-hydroxyquinoline, and 2-methyl-8-hydroxyquinoline. *The Journal of Chemical Thermodynamics* **1989**, *21*, 159–166.

[267] Horton, G. R.; Wendlandt, W. W. The heats of dissociation of the 8-quinolinol and substituted 8-quinolinol metal chelates of uranium (VI), thorium (IV) and scandium (III). *Journal of Inorganic and Nuclear Chemistry* **1963**, *25*, 241–245.

[268] Sakiyama, M.; Nakano, T.; Seki, S. Enthalpies of Combustion of Organic Compounds. I. 8-Quinolinol. *Bulletin of the Chemical Society of Japan* **1975**, *48*, 1705–1708.

[269] Oja, V.; Chen, X.; Hajaligol, M. R.; Chan, W. G. Sublimation Thermodynamic Parameters for Cholesterol, Ergosterol, β-Sitosterol, and Stigmasterol. *Journal of Chemical & Engineering Data* **2009**, *54*, 730–734.

[270] De Kruif, C. G. Enthalpies of sublimation and vapour pressures of 11 polycyclic hydrocarbons. *The Journal of Chemical Thermodynamics* **1980**, *12*, 243–248.

[271] Ribeiro da Silva, M. A. V.; Morais, V. M. F.; Matos, M. A. R.; Rio, C. M. A. Thermochemical and Theoretical Studies of Some Bipyridines. *The Journal of Organic Chemistry* **1995**, *60*, 5291–5294.

[272] Faour, M.; Akasheh, T. S. Heat of combustion of some N-heterocycle compounds. Part 1. *J. Chem. Soc., Perkin Trans. 2* **1985**, 811–813.

[273] Emmenegger, F. The Complex of Co(2,2,6,6-tetramethyl-3,5-heptanedionate)2 with 2,2‘-Bipyridine. Its Formation in the Gas Phase and in Solution. *Inorganic Chemistry* **1996**, *35*, 5931–5934.

[274] Rojas-Aguilar, A.; Flores-Lara, H.; Martinez-Herrera, M.; Ginez-Carbajal, F. Thermochemistry of benzoquinones. *The Journal of Chemical Thermodynamics* **2004**, *36*, 453–463.

[275] Magnus, A. Die Resonanzenergien der Parachinone. *Zeitschrift für Physikalische Chemie* **1956**, *9*, 141–161.

[276] Coolidge, A. S.; Coolidge, M. S. The Sublimation Pressures of Substituted Quinones and Hydroquinones. *Journal of the American Chemical Society* **1927**, *49*, 100–104.

[277] Wakayama, N.; Inokuchi, H. Heats of Sublimation of Polycyclic Aromatic Hydrocarbons and Their Molecular Packings. *Bulletin of the Chemical Society of Japan* **1967**, *40*, 2267–2271.

[278] Domalski, E. S.; Hearing, E. D. Heat Capacities and Entropies of Organic Compounds in the Condensed Phase. Volume III. *Journal of Physical and Chemical Reference Data* **1996**, *25*, 1–1.

[279] Roux, M. V.; Temprado, M.; Chickos, J. S.; Nagano, Y. Critically Evaluated Thermochemical Properties of Polycyclic Aromatic Hydrocarbons. *Journal of Physical and Chemical Reference Data - J Phys Chem Ref Data* **2008**, *37*, 1855–1996.

[280] Bashir-Hashemi, A.; Chickos, J. S.; Hanshaw, W.; Zhao, H.; Farivar, B. S.; Liebman, J. F. The enthalpy of sublimation of cubane. *Thermochimica Acta* **2004**, *424*, 91–97.

[281] Stull, D. R. Vapor Pressure of Pure Substances. Organic and Inorganic Compounds. *Industrial & Engineering Chemistry* **1947**, *39*, 517–540.

[282] Pilcher, G.; Parchment, O. G.; Hillier, I. H.; Heatley, F.; Fletcher, D.; Ribeiro da Silva, M. A. V.; Ferrao, M. L. C. C. H.; Monte, M. J. S.; Fang, J. Thermochemical and theoretical studies on cyclohexanediones. *The Journal of Physical Chemistry* **1993**, *97*, 243–247.

[283] Sabbah, R.; An, X. W. Etude thermodynamique des chlorobenzenes. *Thermochimica Acta* **1991**, *179*, 81–88.

[284] Dávalos, J. Z.; Flores, H.; Jiménez, P.; Notario, R.; Roux, M. V.; Juaristi, E.; Hosmane, R. S.; Liebman, J. F. Calorimetric, Computational (G2(MP2) and G3) and Conceptual Study of the Energetics of the Isomeric 1,3- and 1,4-Dithianes. *The Journal of Organic Chemistry* **1999**, *64*, 9328–9336.

[285] Acree, J., W.; Chickos, J. S. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; National Institute of Standards and Technology: Gaithersburg MD, 2023.

[286] Busfield, W. K.; Mackle, H.; O'Hare, P. A. G. Studies in the thermochemistry of sulphones. Part 2. — The standard heats of formation of sulphones of the type RSO2CH3·. *Trans. Faraday Soc.* **1961**, *57*, 1054–1057.

[287] Ferro, D.; Piacente, V.; Gigli, R.; D'Ascenzo, G. Determination of the vapour pressures of o-, m-, and p-dinitrobenzene by the torsion-effusion method. *The Journal of Chemical Thermodynamics* **1976**, *8*, 1137–1143.

[288] Ribeiro da Silva, M. A. V.; Santos, A. F. L. O. M.; Gomes, J. R. B.; Roux, M. V.; Temprado, M.; Jiménez, P.; Notario, R. Thermochemistry of Bithiophenes and Thienyl Radicals. A Calorimetric and Computational Study. *The Journal of Physical Chemistry A* **2009**, *113*, 11042–11050.

[289] Geiseler, G.; Quitzsch, K.; Rauh, H.-J.; Schaffernicht, H.; Walther, H.-J. Bildungsenthalpien und Mesomerieenergien von π-Bindungssystemen. 1. Mitteilung: Bildungsenthalpien und Mesomerieenergien einiger mehrkerniger Aromaten und verschiedener Pseudoazulene. *Berichte der Bunsengesellschaft für physikalische Chemie* **1966**, *70*, 551–556.

[290] Bondi, A. Heat of Sublimation of Molecular Crystals: A Catalog of Molecular Structure Increments. *Journal of Chemical & Engineering Data* **1963**, *8*, 371–381.

[291] Verevkin, S. P. Relationships among strain energies of mono- and poly-cyclic cyclohexanoid molecules and strain of their component rings. *The Journal of Chemical Thermodynamics* **2002**, *34*, 263–275.

[292] Wada, T.; Kishida, E.; Tomiie, Y.; Suga, H.; Seki, S.; Nitta, I. Crystal Structure and Thermodynamical Investigations of Triethylenediamine, C6H12N2. *Bulletin of the Chemical Society of Japan* **1960**, *33*, 1317–1318.

[293] Chirico, R. D.; Knipmeyer, S. E.; Nguyen, A.; Steele, W. V. The thermodynamic properties to the temperature 700 K of naphthalene and of 2,7-dimethylnaphthalene. *The Journal of Chemical Thermodynamics* **1993**, *25*, 1461–1494.

[294] Speros, D. M.; Rossini, F. D. Heats of combustion and formation of naphthalene, the two methylnaphthalenes, cis and trans- decahydronaphthalene, and related compounds. *The Journal of Physical Chemistry* **1960**, *64*, 1723–1727.

[295] Irving, R. J. The standard enthalpy of sublimation of naphthalene. *The Journal of Chemical Thermodynamics* **1972**, *4*, 793–794.

[296] Coleman, D. J.; Pilcher, G. Heats of combustion of biphenyl, bibenzyl, naphthalene, anthracene and phenanthrene. *Trans. Faraday Soc.* **1966**, *62*, 821–827.

[297] Miller, G. A. Vapor Pressure of Naphthalene. Thermodynamic Consistency with Proposed Frequency Assignments. *Journal of Chemical & Engineering Data* **1963**, *8*, 69–72.

[298] Van Ekeren, P. J.; Jacobs, M. H. G.; Offringa, J. C. A.; De Kruif, C. Vapour-pressure measurements on trans-diphenylethene and naphthalene using a spinning-rotor friction gauge. *The Journal of Chemical Thermodynamics* **1983**, *15*, 409–417.

[299] Sinke, G. C. A method for measurement of vapor pressures of organic compounds below 0.1 Torr Naphthalene as a reference substance. *The Journal of Chemical Thermodynamics* **1974**, *6*, 311–316.

[300] Murata, S.; Sakiyama, M.; Seki, S. Construction and testing of a sublimation calorimetric system using a Calvet microcalorimeter. *The Journal of Chemical Thermodynamics* **1982**, *14*, 707–721.

[301] Torres-Gómez, L. A.; Barreiro-Rodríguez, G.; Galarza-Mondragón, A. A new method for the measurement of enthalpies of sublimation using differential scanning calorimetry. *Thermochimica Acta* **1988**, *124*, 229–233.

[302] Morawetz, E. Enthalpies of vaporization for a number of aromatic compounds. *The Journal of Chemical Thermodynamics* **1972**, *4*, 455–460.

[303] Chickos, J.; Hesse, D.; Hosseini, S.; Nichols, G.; Webb, P. Sublimation enthalpies at 298.15K using correlation gas chromatography and differential scanning calorimetry measurements. *Thermochimica Acta* **1998**, *313*, 101–110.

[304] Wolf, K. L.; Weghofer, H. Über Sublimationswärmen. *Zeitschrift für Physikalische Chemie* **1938**, *39B*, 194–208.

[305] Inokuchi, H.; Shiba, S.; Handa, T.; Akamatu, H. Heats of Sublimation of Condensed Polynuclear Aromatic Hydrocarbons. *Bulletin of the Chemical Society of Japan* **1952**, *25*, 299–302.

[306] Chirico, R. D.; Kazakov, A. F.; Steele, W. V. Thermodynamic properties of three-ring aza-aromatics. 1. Experimental results for phenazine and acridine, and mutual validation of experiments and computational methods. *The Journal of Chemical Thermodynamics* **2010**, *42*, 571–580.

[307] Månsson, M. Non-bonded Oxygen-Oxygen Interactions in 2,4,10-Trioxa-adamantane and 1,3,5,7,9-Pentoxecane. *Acta Chemica Scandinavica* **1974**, *28b*, 895–899.

[308] Sakoguchi, A.; Ueoka, R.; Kato, Y.; Arai, Y. Vapor Pressure of Alkylpyridines and Alkylpyrazines. *Journal of Chemical Engineering* **1995**, *21*, 219–223.

[309] Meng-Yan, Y.; Pilcher, G. Enthalpies of combustion of succinic anhydride, glutaric anhydride, and glutarimide. *The Journal of Chemical Thermodynamics* **1990**, *22*, 893–898.

[310] Boyd, R. H. Thermochemistry of Cyanocarbons. *The Journal of Chemical Physics* **1963**, *38*, 2529–2535.

[311] Westrum, E. F. J.; Rapport, N. J.; Andrews, J. T. S. Enthalpies of formation for globular molecules III. Succinonitrile and triethylenediamine. *Journal of the American Chemical Society* **1971**, *93*, 4363–4365.

[312] Månsson, M.; Nakase, Y.; Sunner, S. The Enthalpies of Formation of Trioxane and Tetroxane. *Acta Chemica Scandinavica* **1969**, *23*, 56–60.

[313] Roux, M. V.; Jiménez, P.; Dávalos, J. Z.; Notario, R.; Juaristi, E. Calorimetric and Computational Study of 1,3,5-Trithiane. *The Journal of Organic Chemistry* **2001**, *66*, 5343–5351.

[314] Byström, K. The stabilization energy of 1,3,5-triazine derived from measurements of the enthalpies of combustion and sublimation. *The Journal of Chemical Thermodynamics* **1982**, *14*, 865–870.

[315] Bogdanova, K. A.; Berlin, A. A.; Kompaniets, V. Z.; Rakova, G. V.; Miroshnichenko, Y. A.; Lebedev, Y. A.; Yenikolopyan, N. S. Enthalpies of the homogeneous polymerization of trioxane and tetroxane. *Polymer Science U.S.S.R.* **1975**, *17*, 759–766.

[316] Calis-Van Ginkel, C. H. D.; Calis, G. H. M.; Timmermans, C. W. M.; de Kruif, C. G.; Oonk, H. A. J. Enthalpies of sublimation and dimerization in the vapour phase of formic, acetic, propanoic, and butanoic acids. *The Journal of Chemical Thermodynamics* **1978**, *10*, 1083–1088.

[317] Sabbah, R.; Chastel, R.; Laffitte, M. Thermodynamique de composés azotés. I. Etude calorimétrique des enthalpies de sublimation des trois acides aminobenzoïques. *Canadian Journal of Chemistry* **1974**, *52*, 2201–2205.

[318] Nabavian, M.; Sabbah, R.; Chastel, R.; Laffitte, M. Thermodynamique de composés azotés - II. - Étude thermochimique des acides aminobenzoïques, de la pyrimidine, de l'uracile et de la thymine. *J. Chim. Phys.* **1977**, *74*, 115–126.

[319] Sabbah, R.; da Silva Eusébio, M. E. Energétique des liaisons inter- et intramoléculaires dans les trois isomères de l'aminopyridine. *Canadian Journal of Chemistry* **1998**, *76*, 18–24.

[320] Bickerton, J.; Pilcher, G.; Al-Takhin, G. Enthalpies of combustion of the three aminopyridines and the three cyanopyridines. *The Journal of Chemical Thermodynamics* **1984**, *16*, 373–378.

[321] Sabbah, R.; Perez, L. Energétique des liaisons inter- et intramoléculaires dans les trois isomères du benzènediamine. *Canadian Journal of Chemistry* **1997**, *75*, 357–364.

[322] Linstrom, P. J.; Mallard, W. G. The NIST Chemistry WebBook: A Chemical Data Resource on the Internet. *Journal of Chemical & Engineering Data* **2001**, *46*, 1059–1063.

[323] Monte, M. J. S.; Gonçalves, M. V.; Ribeiro da Silva, M. D. M. C. Vapor Pressures and Enthalpies of Combustion of the Dihydroxybenzoic Acid Isomers. *Journal of Chemical & Engineering Data* **2010**, *55*, 2246–2251.

[324] Jiménez, P.; Roux, M. V.; Turrión, C.; Gomis, F. Thermochemical properties of N-heterocyclic compounds I. Enthalpies of combustion, vapour pressures and enthalpies of sublimation, and enthalpies of formation of pyrazole, imidazole, indazole, and benzimidazole. *The Journal of Chemical Thermodynamics* **1987**, *19*, 985–992.

[325] Zimmermann, H.; Geisenfelder, H. Über die Mesomerieenergie von Azolen. *Zeitschrift für Elektrochemie, Berichte der Bunsengesellschaft für physikalische Chemie* **1961**, *65*, 368–371.

[326] Lehman, R. L.; Gentry, J. S.; Glumac, N. G. Thermal stability of potassium carbonate near its melting point. *Thermochimica Acta* **1998**, *316*, 1–9.

[327] Ribeiro da Silva, M. A. V.; Fonseca, J. M. S.; Carvalho, R. P. B. M.; Monte, M. J. S. Thermodynamic study of the sublimation of six halobenzoic acids. *The Journal of Chemical Thermodynamics* **2005**, *37*, 271–279.

[328] Sabbah, R.; Aguilar, A. R. Étude thermodynamique des trois isomères de l'acide chlorobenzoïque. Partie II. *Canadian Journal of Chemistry* **1995**, *73*, 1538–1545.

[329] Jiménez, P.; Roux, M. V.; Turrión, C. Thermochemical properties of N-heterocyclic compounds IV. Enthalpies of combustion, vapour pressures and enthalpies of sublimation, and enthalpies of formation of 2-methylimidazole and 2-ethylimidazole. *The Journal of Chemical Thermodynamics* **1992**, *24*, 1145–1149.

[330] Arshadi, M. R. Determination of heats of sublimation of organic compounds by a mass spectrometric–knudsen effusion method. *J. Chem. Soc., Faraday Trans. 1* **1974**, *70*, 1569–1571.

[331] Sabbah, R.; Gilbert, M.; Julg, A. Détermination de l'énergie de conjugaison des acides naphtoïques. *Thermochimica Acta* **1974**, *10*, 345–351.

[332] Cox, J. D.; Gundry, H. A.; Harrop, D.; Head, A. J. Thermodynamic properties of fluorine compounds 9. Enthalpies of formation of some compounds containing the pentafluorophenyl group. *The Journal of Chemical Thermodynamics* **1969**, *1*, 77–87.

[333] Bernardes, C. E. S.; Minas da Piedade, M. E. Energetics of the O–H Bond and of Intramolecular Hydrogen Bonding in HOC6H4C(O)Y (Y =, CH3, CH2CHCH2, CCH, CH2F, NH2, NHCH3, NO2, OH, OCH3, OCN, CN, F, Cl, SH, and SCH3) Compounds. *The Journal of Physical Chemistry A* **2008**, *112*, 10029–10039.

[334] Ribeiro da Silva, M. D. M. C.; Araújo, N. R. M. Thermochemical studies on salicylaldehyde and salicylamide. *The Journal of Chemical Thermodynamics* **2007**, *39*, 1372–1376, Honour issue Professor J. Simões Redinha on the occasion of his 80th birthday.

[335] De Wit, H. G. M.; Van Miltenburg, J. C.; De Kruif, C. G. Thermodynamic properties of molecular organic crystals containing nitrogen, oxygen, and sulphur 1. Vapour pressures and enthalpies of sublimation. *The Journal of Chemical Thermodynamics* **1983**, *15*, 651–663.

[336] Jiménez, P.; Roux, M. V.; Turrión, C. Thermochemical properties of N-heterocyclic compounds II. Enthalpies of combustion, vapour pressures, enthalpies of sublimation, and enthalpies of formation of 1,2,4-triazole and benzotriazole. *The Journal of Chemical Thermodynamics* **1989**, *21*, 759–764.

[337] Santos, A. F. L. O. M.; Ribeiro da Silva, M. A. V. Experimental and Computational Study on the Molecular Energetics of 2-Pyrrolecarboxylic Acid and 1-Methyl-2-pyrrolecarboxylic Acid. *The Journal of Physical Chemistry A* **2009**, *113*, 9741–9750.

[338] Maria, T. M. R.; Costa, F. S.; Leitão, M. L. P.; Redinha, J. S. A calorimetric study of phase transitions for some cyclohexanediols. *Thermochimica Acta* **1995**, *269-270*, 405–413, Recent Advances in Thermal Analysis and Calorimetry.

[339] McLean, A. D.; Chandler, G. S. Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z=11–18. *The Journal of Chemical Physics* **1980**, *72*, 5639–5648.

[340] Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *The Journal of Chemical Physics* **1980**, *72*, 650–654.

[341] Frisch, M. J.; Pople, J. A.; Binkley, J. S. Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets. *The Journal of Chemical Physics* **1984**, *80*, 3265–3269.

[342] Tang, K. T.; Toennies, J. P. An improved simple model for the van der Waals potential based on universal damping functions for the dispersion coefficients. *The Journal of Chemical Physics* **1984**, *80*, 3726–3741.

[343] Gavezzotti, A.; Filippini, G. Geometry of the Intermolecular X-H···Y (X, Y = N, O) Hydrogen Bond and the Calibration of Empirical Hydrogen-Bond Potentials. *The Journal of Physical Chemistry* **1994**, *98*, 4831–4837.

[344] Mann, H. B.; Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **1947**, *18*, 50 – 60.

[345] Hunter, C. A.; Sanders, J. K. M. The nature of .pi.-.pi. interactions. *Journal of the American Chemical Society* **1990**, *112*, 5525–5534.

[346] Totton, T. S.; Misquitta, A. J.; Kraft, M. A transferable electrostatic model for intermolecular interactions between polycyclic aromatic hydrocarbons. *Chemical Physics Letters* **2011**, *510*, 154–160.

[347] Martinez, C. R.; Iverson, B. L. Rethinking the term "pi-stacking". *Chem. Sci.* **2012**, *3*, 2191–2201.

[348] Roothaan, C. C. J. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.* **1951**, *23*, 69–89.

[349] Frisch, M. J.; Head-Gordon, M.; Pople, J. A. A direct MP2 gradient method. *Chemical Physics Letters* **1990**, *166*, 275–280.

[350] Frisch, M. J.; Head-Gordon, M.; Pople, J. A. Semi-direct algorithms for the MP2 energy and gradient. *Chemical Physics Letters* **1990**, *166*, 281–289.

[351] Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 energy evaluation by direct methods. *Chemical Physics Letters* **1988**, *153*, 503–506.

[352] Sæbø, S.; Almlöf, J. Avoiding the integral storage bottleneck in LCAO calculations of electron correlation. *Chemical Physics Letters* **1989**, *154*, 83–89.

[353] Head-Gordon, M.; Head-Gordon, T. Analytic MP2 frequencies without fifth-order storage. Theory and application to bifurcated hydrogen bonds in the water hexamer. *Chemical Physics Letters* **1994**, *220*, 122–128.

[354] Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *The Journal of Chemical Physics* **1971**, *54*, 724–728.

[355] Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *The Journal of Chemical Physics* **1972**, *56*, 2257–2261.

[356] Hariharan, P. C.; Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theoretical Chemistry Accounts* **1973**, *28*, 213–222.

[357] Hariharan, P.; Pople, J. Accuracy of AH n equilibrium geometries by single determinant molecular orbital theory. *Molecular Physics* **1974**, *27*, 209–214.

[358] Gordon, M. S. The isomers of silacyclopropane. *Chemical Physics Letters* **1980**, *76*, 163–168.

[359] Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *The Journal of Chemical Physics* **1982**, *77*, 3654–3665.

[360] Binning Jr., R. C.; Curtiss, L. A. Compact contracted basis sets for third-row atoms: Ga–Kr. *Journal of Computational Chemistry* **1990**, *11*, 1206–1216.

[361] Blaudeau, J.-P.; McGrath, M. P.; Curtiss, L. A.; Radom, L. Extension of Gaussian-2 (G2) theory to molecules containing third-row atoms K and Ca. *The Journal of Chemical Physics* **1997**, *107*, 5016–5021.

[362] Rassolov, V. A.; Pople, J. A.; Ratner, M. A.; Windus, T. L. 6-31G* basis set for atoms K through Zn. *The Journal of Chemical Physics* **1998**, *109*, 1223–1229.

[363] Rassolov, V. A.; Ratner, M. A.; Pople, J. A.; Redfern, P. C.; Curtiss, L. A. 6-31G* basis set for third-row atoms. *Journal of Computational Chemistry* **2001**, *22*, 976–984.

[364] Binkley, J. S.; Pople, J. A.; Hehre, W. J. Self-consistent molecular orbital methods. 21. Small split-valence basis sets for first-row elements. *Journal of the American Chemical Society* **1980**, *102*, 939–947.

[365] Gordon, M. S.; Binkley, J. S.; Pople, J. A.; Pietro, W. J.; Hehre, W. J. Self-consistent molecular-orbital methods. 22. Small split-valence basis sets for second-row elements. *Journal of the American Chemical Society* **1982**, *104*, 2797–2803.

[366] Pietro, W. J.; Francl, M. M.; Hehre, W. J.; DeFrees, D. J.; Pople, J. A.; Binkley, J. S. Self-consistent molecular orbital methods. 24. Supplemented small split-valence basis sets for second-row elements. *Journal of the American Chemical Society* **1982**, *104*, 5039–5048.

[367] Dobbs, K. D.; Hehre, W. J. Molecular orbital theory of the properties of inorganic and organometallic compounds 4. Extended basis sets for third-and fourth-row, main-group elements. *Journal of Computational Chemistry* **1986**, *7*, 359–378.

[368] Dobbs, K. D.; Hehre, W. J. Molecular orbital theory of the properties of inorganic and organometallic compounds 5. Extended basis sets for first-row transition metals. *Journal of Computational Chemistry* **1987**, *8*, 861–879.

[369] Dobbs, K. D.; Hehre, W. J. Molecular orbital theory of the properties of inorganic and organometallic compounds. 6. Extended basis sets for second-row transition metals. *Journal of Computational Chemistry* **1987**, *8*, 880–893.

[370] Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. R. Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements, Li–F. *Journal of Computational Chemistry* **1983**, *4*, 294–301.

[371] Totton, T. S.; Misquitta, A. J.; Kraft, M. A First Principles Development of a General Anisotropic Potential for Polycyclic Aromatic Hydrocarbons. *Journal of Chemical Theory and Computation* **2010**, *6*, 683–695.

[372] Skoweranda, J.; Bukowska-Strzyzewska, M.; Bartnik, R.; Strzyżewski, W. *Mercury 4.0*: from visualization to analysis, design and prediction. *Journal of Crystallographic and Spectroscopic Research* **1990**, *20*, 117–121.

[373] Suleiman Gwaram, N.; Khaledi, H.; Mohd Ali, H.; Olmstead, M. M. Two new polymorphs of 2,6-diaminopyrimidin-4(3*H*)-one monohydrate. *Acta Crystallographica Section C* **2011**, *67*, o6–o9.

[374] Nasiri, H.; Bolte, M. CCDC 740440: Experimental Crystal Structure Determination. *CSD Communication* **2009**,

[375] Birkedal, H.; Madsen, D.; Mathiesen, R. H.; Knudsen, K.; Weber, H.-P.; Pattison, P.; Schwarzenbach, D. The charge density of urea from synchrotron diffraction data. *Acta Crystallographica Section A* **2004**, *60*, 371–381.

[376] Roszak, K.; Katrusiak, A. Giant Anomalous Strain between High-Pressure Phases and the Mesomers of Urea. *The Journal of Physical Chemistry C* **2017**, *121*, 778–784.

[377] Olejniczak, A.; Ostrowska, K.; Katrusiak, A. H-Bond Breaking in High-Pressure Urea. *The Journal of Physical Chemistry C* **2009**, *113*, 15761–15767.

[378] Bridgman, P. W. Polymorphism at High Pressures. *Proceedings of the American Academy of Arts and Sciences* **1916**, *52*, 91–187.

[379] Weber, H. P.; Marshall, W. G.; Dmitriev, V. High-pressure polymorphism in deuterated urea. *Acta Crystallographica Section A: Foundations and Advances* **2002**,

[380] Andersson, O.; Ross, R. G. Phase behavior and thermal conductivity of urea at pressures up to 1 GPa and at temperatures in the range 50–370 K. *International Journal of Thermophysics* **1994**, *15*, 513–524.

[381] Olejniczak, A.; Ostrowska, K.; Katrusiak, A. H-Bond Breaking in High-Pressure Urea. *The Journal of Physical Chemistry C* **2009**, *113*, 15761–15767.

[382] Mondal, S.; Prathapa, S. J.; van Smaalen, S. Experimental dynamic electron densities of multipole models at different temperatures. *Acta Crystallographica Section A* **2012**, *68*, 568–581.

[383] Boldyreva, E. V.; Drebushchak, T. N.; Shutova, E. S. Structural distortion of the α, β, and γ polymorphs of glycine on cooling. *Zeitschrift für Kristallographie - Crystalline Materials* **2003**, *218*, 366–376.

[384] Kvick, Å.; Canning, W. M.; Koetzle, T. F.; Williams, G. J. B. An experimental study of the influence of temperature on a hydrogen-bonded system: the crystal structure of γ-glycine at 83 K and 298 K by neutron diffraction. *Acta Crystallographica Section B* **1980**, *36*, 115–120.

[385] Han, G.; Chow, P. S.; Tan, R. B. H. Understanding the Salt-Dependent Outcome of Glycine Polymorphic Nucleation. *Pharmaceutics* **2021**, *13*.

[386] Perlovich, G. L.; Hansen, L. K.; Bauer-Brandl, A. The Polymorphism of Glycine. Thermochemical and structural aspects. *Journal of Thermal Analysis and Calorimetry* **2001**, *66*, 699–715.

[387] Srinivasan, K. Crystal growth of α and γ glycine polymorphs and their polymorphic phase transformations. *Journal of Crystal Growth* **2008**, *311*, 156–162.

[388] Boldyreva, E. V.; Drebushchak, V. A.; Drebushchak, T. N.; Paukov, I. E.; Kovalevskaya, Y. A.; Shutova, E. S. Polymorphism of glycine, Part I. *Journal of Thermal Analysis and Calorimetry* **2003**, *73*, 409–418.

[389] Boldyreva, E. V.; Drebushchak, V. A.; Drebushchak, T. N.; Paukov, I. E.; Kovalevskaya, Y. A.; Shutova, E. S. Polymorphism of glycine, Part II. *Journal of Thermal Analysis and Calorimetry* **2003**, *73*, 419–428.

[390] Dawson, A.; Allan, D. R.; Belmonte, S. A.; Clark, S. J.; David, W. I. F.; McGregor, P. A.; Parsons, S.; Pulham, C. R.; Sawyer, L. Effect of High Pressure on the Crystal Structures of Polymorphs of Glycine. *Crystal Growth & Design* **2005**, *5*, 1415–1427.

[391] Tumanov, N. A.; Bolldyreva, E. V.; Ahsbahs, H. CCDC 682610: Experimental Crystal Structure Determination. *CSD Communication* **2010**,

[392] Xavier, N. F. J.; da Silva, A. M. J.; Bauerfeldt, G. F. What Rules the Relative Stability of α-, β-, and γ-Glycine Polymorphs? *Crystal Growth & Design* **2020**, *20*, 4695–4706.

[393] Tumanova, N.; Tumanov, N.; Robeyns, K.; Fischer, F.; Fusaro, L.; Morelle, F.; Ban, V.; Hautier, G.; Filinchuk, Y.; Wouters, J.; Leyssens, T.; Emmerling, F. Opening Pandora's Box: Chirality, Polymorphism, and Stoichiometric Diversity in Flurbiprofen/Proline Cocrystals. *Crystal Growth & Design* **2018**, *18*, 954–961.

[394] Myung, S.; Pink, M.; Baik, M.-H.; Clemmer, D. E. DL-Proline. *Acta Crystallographica Section C* **2005**, *61*, o506–o508.

[395] Brock, C. P.; Schweizer, W. B.; Dunitz, J. D. On the validity of Wallach's rule: on the density and stability of racemic crystals compared with their chiral counterparts. *Journal of the American Chemical Society* **1991**, *113*, 9811–9820.

[396] Singaraju, A. B.; Bahl, D.; Wang, C.; Swenson, D. C.; Sun, C. C.; Stevens, L. L. Molecular Interpretation of the Compaction Performance and Mechanical Properties of Caffeine Cocrystals: A Polymorphic Study. *Molecular Pharmaceutics* **2020**, *17*, 21–31.

[397] Martin, R.; Lilley, T. H.; Bailey, N. A.; Falshaw, C. P.; Haslam, E.; Magnolato, D.; Begley, M. J. Polyphenol–caffeine complexation. *J. Chem. Soc., Chem. Commun.* **1986**, 105–106.

[398] Devereux, M.; Raghunathan, S.; Fedorov, D. G.; Meuwly, M. A Novel, Computationally Efficient Multipolar Model Employing Distributed Charges for Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation* **2014**, *10*, 4229–4241.

Appendices

Appendix A

Gamma Functions and their Derivatives

Gamma functions are used in the generalized Ewald summation described in Section 3.2.1.1. Because we are only interested in multipole interactions which have integer ranks, it is possible to derive analytical expressions for the gamma function across all sum-of-multipole-ranks of interest. This has done below, up to $l_1 + l_2 = 4$ (Table A.1). Importantly, we leverage the recursive relationship between gamma functions, $\Gamma(z+1) = z\Gamma(z)$, to expedite this derivation.

Table A.1: Analytical solutions for $\Gamma(z)$ at different values of $l_1 + l_2$.

$\boldsymbol{l_1 + l_2}$	$\boldsymbol{\Gamma(\frac{1}{2} + l_1 + l_2)}$
0	$\sqrt{\pi}$
1	$\frac{1}{2}\sqrt{\pi}$
2	$\frac{3}{4}\sqrt{\pi}$
3	$\frac{15}{8}\sqrt{\pi}$
4	$\frac{105}{16}\sqrt{\pi}$

Similarly, analytical expressions for the upper and lower incomplete gamma functions are also necessary to evaluate the Ewald sum. For the case of charge–charge interactions (i.e., $l_1 + l_2 = 0$), the upper and lower incomplete gamma functions are given by

$$\Gamma\left(\frac{1}{2}, x\right) = \sqrt{\pi}\text{erfc}(\sqrt{x}), \tag{A.0.1}$$

and

$$\gamma\left(\frac{1}{2}, x\right) = \sqrt{\pi}\text{erf}(\sqrt{x}), \tag{A.0.2}$$

respectively. 'erf' and 'erfc' are the error and complimentary error function respectively. Again, recursive gamma function relationships can be leveraged to get the corresponding upper and lower incomplete gamma functions for higher-ranked multipole interactions. These recursive relationships are

$$\Gamma(z+1, x) = z\Gamma(z, x) + x^z e^{-x}, \tag{A.0.3}$$

and

$$\gamma(z+1,x) = z\gamma(z,x) - x^z e^{-x}, \tag{A.0.4}$$

for the upper and lower incomplete gamma functions respectively.

During lattice energy minimization, first- and second-derivatives of these incomplete gamma functions will be needed. For the upper incomplete gamma function, these can be obtained by differentiation of Equation 3.2.11:

$$\frac{\partial\Gamma(z,x)}{\partial x} = -x^{z-1}e^{-x}, \tag{A.0.5}$$

$$\frac{\partial^2\Gamma(z,x)}{\partial x^2} = -(z-1)x^{z-2}e^{-x} + x^{z-1}e^{-x}. \tag{A.0.6}$$

By definition, $\Gamma(z) = \Gamma(z,x) + \gamma(z,x)$. Hence from here, it is trivial to obtain the relationship,

$$\frac{\partial^n\gamma(z,x)}{\partial x^n} = -\frac{\partial^n\Gamma(z,x)}{\partial x^n}. \tag{A.0.7}$$

Appendix B

Trigonometric Identities used in Ewald Sums

In the derivation of the reciprocal term of the generalized Ewald sum, Equation 3.2.19 (re-written below),

$$
\begin{aligned}
U^{\mathrm{rec}}_{\boldsymbol{l}_k} = & \sum_{l_1,l_2} \frac{2^{-l_1-l_2}\sqrt{\pi}}{\Gamma(\frac{1}{2}+l_1+l_2)} \frac{1}{\|\boldsymbol{k}\|^{2-l_1-l_2}} \exp\left(\frac{-\|\boldsymbol{k}\|^2}{4\cdot\pi\cdot\alpha_{l_1+l_2}}\right) \\
& \cdot \sum_{m_1,m_2} \sum_{I=1}^{Z}\sum_{J=1}^{Z} \widetilde{T}^{m_1,m_2}_{l_1,l_2,l_1+l_2}(\boldsymbol{A}^T_{\mathrm{rot},I}, \boldsymbol{A}^T_{\mathrm{rot},J}, \hat{\boldsymbol{k}}) \\
& \cdot \sum_{i=1}^{\eta_I}\sum_{j=1}^{\eta_J} Q^{l_1m_1}_{Ii} Q^{l_2m_2}_{Jj} \{ i^{l_1+l_2} \cdot \cos\left[\boldsymbol{k}\cdot(\boldsymbol{r}_{\boldsymbol{0}Jj} - \boldsymbol{r}_{\boldsymbol{0}Ii})\right] - i^{l_1+l_2+1} \cdot \sin\left[\boldsymbol{k}\cdot(\boldsymbol{r}_{\boldsymbol{0}Jj} - \boldsymbol{r}_{\boldsymbol{0}Ii})\right] \},
\end{aligned}
\tag{B.0.1}
$$

is simplified using trigonometric identities. For clarity, I will neglect parts of the summation that are unimportant to the trigonometric manipulation.

For even values of $l_1 + l_2$, the imaginary sine term is eliminated and only the cosine term remains in Equation B.0.1,

$$
\sigma_{\mathrm{even}} = \sum_{i=1}^{\eta_I}\sum_{j=1}^{\eta_J} Q_{Ii} Q_{Jj} \cdot \cos[\boldsymbol{k}\cdot(\boldsymbol{r}_{\boldsymbol{0}Jj} - \boldsymbol{r}_{\boldsymbol{0}Ii})]. \tag{B.0.2}
$$

Using the trigonometric identity, $\cos(A-B) = \cos(A)\cos(B) + \sin(A)\sin(B)$, this can be re-written as

$$
\sigma_{\mathrm{even}} = \sum_{i=1}^{\eta_I}\sum_{j=1}^{\eta_J} Q_{Ii} Q_{Jj} \cdot \big[\cos(\boldsymbol{k}\cdot\boldsymbol{r}_{\boldsymbol{0}Jj})\cos(\boldsymbol{k}\cdot\boldsymbol{r}_{\boldsymbol{0}Ii}) + \sin(\boldsymbol{k}\cdot\boldsymbol{r}_{\boldsymbol{0}Jj})\sin(\boldsymbol{k}\cdot\boldsymbol{r}_{\boldsymbol{0}Ii}) \big]. \tag{B.0.3}
$$

Thereafter, it is trivial to separate the i and j indexed terms,

$$
\begin{aligned}
\sigma_{\mathrm{even}} = & \sum_{i=1}^{\eta_I} Q_{Ii}\cos(\boldsymbol{k}\cdot\boldsymbol{r}_{\boldsymbol{0}Ii}) \sum_{j=1}^{\eta_J} Q_{Jj}\cos(\boldsymbol{k}\cdot\boldsymbol{r}_{\boldsymbol{0}Jj}) \\
& + \sum_{i=1}^{\eta_I} Q_{Ii}\sin(\boldsymbol{k}\cdot\boldsymbol{r}_{\boldsymbol{0}Ii}) \sum_{j=1}^{\eta_J} Q_{Jj}\sin(\boldsymbol{k}\cdot\boldsymbol{r}_{\boldsymbol{0}Jj}).
\end{aligned}
\tag{B.0.4}
$$

These are exactly the expressions given in Equations 3.2.20–3.2.22.

Similarly, for odd values of $l_1 + l_2$, the imaginary cosine term is eliminated and only the sine term remains in Equation B.0.1,

$$\sigma_{\text{odd}} = \sum_{i=1}^{\eta_I} \sum_{j=1}^{\eta_J} Q_{Ii} Q_{Jj} \cdot \sin[\boldsymbol{k} \cdot (\boldsymbol{r}_{\mathbf{0}Jj} - \boldsymbol{r}_{\mathbf{0}Ii})]. \tag{B.0.5}$$

Using the trigonometric identity, $\sin(A - B) = \sin(A)\cos(B) - \cos(A)\sin(B)$, this can be re-written as

$$\sigma_{\text{odd}} = \sum_{i=1}^{\eta_I} \sum_{j=1}^{\eta_J} Q_{Ii} Q_{Jj} \cdot \big[\sin(\boldsymbol{k} \cdot \boldsymbol{r}_{\mathbf{0}Jj}) \cos(\boldsymbol{k} \cdot \boldsymbol{r}_{\mathbf{0}Ii}) - \cos(\boldsymbol{k} \cdot \boldsymbol{r}_{\mathbf{0}Jj}) \sin(\boldsymbol{k} \cdot \boldsymbol{r}_{\mathbf{0}Ii}) \big], \tag{B.0.6}$$

and separating the indices, we again obtain

$$\begin{aligned} \sigma_{\text{odd}} = &\sum_{i=1}^{\eta_I} Q_{Ii} \cos(\boldsymbol{k} \cdot \boldsymbol{r}_{\mathbf{0}Ii}) \sum_{j=1}^{\eta_J} Q_{Jj} \sin(\boldsymbol{k} \cdot \boldsymbol{r}_{\mathbf{0}Jj}) \\ &- \sum_{i=1}^{\eta_I} Q_{Ii} \sin(\boldsymbol{k} \cdot \boldsymbol{r}_{\mathbf{0}Ii}) \sum_{j=1}^{\eta_J} Q_{Jj} \cos(\boldsymbol{k} \cdot \boldsymbol{r}_{\mathbf{0}Jj}). \end{aligned} \tag{B.0.7}$$

This is exactly the expression given in Equation 3.2.23.

Appendix C

Derivatives of (Reciprocal) Unit Cell Matrix with respect to Lattice Parameters

C.1 First-Derivatives

To evaluate the first-derivatives of $\boldsymbol{R}$ and $\boldsymbol{R}^{-1}$ with respect to $\boldsymbol{X}$, I define the following:

$$W = 1 - \sum_{\zeta=\alpha,\beta,\gamma} \cos^2(\zeta) + 2 \prod_{\zeta=\alpha,\beta,\gamma} \cos(\zeta), \tag{C.1.1}$$

$$\frac{\partial W}{\partial \zeta} = 2\cos(\zeta)\sin(\zeta) - 2\sin(\zeta) \prod_{\zeta'=\alpha,\beta,\gamma,\zeta'\neq\zeta} \cos(\zeta') \qquad \forall \quad \zeta \in \{\alpha,\beta,\gamma\}. \tag{C.1.2}$$

Table C.1: Analytical expressions for $\nabla_X \boldsymbol{R}$.

X	$\nabla_X \boldsymbol{R}$
a	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
b	$\begin{pmatrix} 0 & \cos(\gamma) & 0 \\ 0 & \sin(\gamma) & 0 \\ 0 & 0 & 0 \end{pmatrix}$
c	$\begin{pmatrix} 0 & 0 & \cos(\beta) \\ 0 & 0 & \frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{\sin(\gamma)} \\ 0 & 0 & \frac{\sqrt{W}}{\sin(\gamma)} \end{pmatrix}$
α	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\frac{c\sin(\alpha)}{\sin(\gamma)} \\ 0 & 0 & \frac{c}{2\sqrt{W}\sin(\gamma)}\frac{\partial W}{\partial \alpha} \end{pmatrix}$
β	$\begin{pmatrix} 0 & 0 & -c\sin(\beta) \\ 0 & 0 & \frac{c\sin(\beta)\cos(\gamma)}{\sin(\gamma)} \\ 0 & 0 & \frac{c}{2\sqrt{W}\sin(\gamma)}\frac{\partial W}{\partial \beta} \end{pmatrix}$
γ	$\begin{pmatrix} 0 & -b\sin(\gamma) & 0 \\ 0 & b\cos(\gamma) & c\frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{\sin^2(\gamma)} \\ 0 & 0 & c\left[\frac{1}{2\sqrt{W}\sin(\gamma)}\frac{\partial W}{\partial \gamma} - \frac{\sqrt{W}\cos(\gamma)}{\sin^2(\gamma)}\right] \end{pmatrix}$

Table C.2: Analytical expressions for $\nabla_X \boldsymbol{R}^{-1}$.

X	$\nabla_X \boldsymbol{R}^{-1}$
a	$\begin{pmatrix} -\frac{1}{a^2} & \frac{\cos(\gamma)}{a^2\sin(\gamma)} & \frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{a^2\sin(\gamma)\sqrt{W}} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
b	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & -\frac{1}{b^2\sin(\gamma)} & \frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{b^2\sin(\gamma)\sqrt{W}} \\ 0 & 0 & 0 \end{pmatrix}$
c	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\frac{\sin(\gamma)}{c^2\sqrt{W}} \end{pmatrix}$
α	$\begin{pmatrix} 0 & 0 & \frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{2a\sin(\gamma)W^{3/2}}\frac{\partial W}{\partial\alpha} - \frac{\sin(\alpha)\cos(\gamma)}{a\sin(\gamma)\sqrt{W}} \\ 0 & 0 & \frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{2b\sin(\gamma)W^{3/2}}\frac{\partial W}{\partial\alpha} + \frac{\sin(\alpha)}{b\sin(\gamma)\sqrt{W}} \\ 0 & 0 & -\frac{\sin(\gamma)}{2cW^{3/2}}\frac{\partial W}{\partial\alpha} \end{pmatrix}$
β	$\begin{pmatrix} 0 & 0 & \frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{2a\sin(\gamma)W^{3/2}}\frac{\partial W}{\partial\beta} + \frac{\sin(\beta)}{a\sin(\gamma)\sqrt{W}} \\ 0 & 0 & \frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{2b\sin(\gamma)W^{3/2}}\frac{\partial W}{\partial\beta} - \frac{\sin(\beta)\cos(\gamma)}{b\sin(\gamma)\sqrt{W}} \\ 0 & 0 & -\frac{\sin(\gamma)}{2cW^{3/2}}\frac{\partial W}{\partial\beta} \end{pmatrix}$
γ	$\begin{pmatrix} 0 & \frac{1}{a\sin^2(\gamma)} & \frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{2a\sin(\gamma)W^{3/2}}\frac{\partial W}{\partial\gamma} - \frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{a\sin^2(\gamma)\sqrt{W}} \\ 0 & -\frac{\cos(\gamma)}{b\sin^2(\gamma)} & \frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{2b\sin(\gamma)W^{3/2}}\frac{\partial W}{\partial\gamma} - \frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{b\sin^2(\gamma)\sqrt{W}} \\ 0 & 0 & \frac{\cos(\gamma)}{c\sqrt{W}} - \frac{\sin(\gamma)}{2cW^{3/2}}\frac{\partial W}{\partial\gamma} \end{pmatrix}$

C.2 Double-Derivatives

To evaluate the double-derivatives of $\boldsymbol{R}$ and $\boldsymbol{R}^{-1}$ with respect to $\boldsymbol{X}$, I define the following::

$$W = 1 - \sum_{\zeta=\alpha,\beta,\gamma} \cos^2(\zeta) + 2 \prod_{\zeta=\alpha,\beta,\gamma} \cos(\zeta), \tag{C.2.1}$$

$$\frac{\partial W}{\partial \zeta} = 2\cos(\zeta)\sin(\zeta) - 2\sin(\zeta) \prod_{\zeta'=\alpha,\beta,\gamma,\zeta'\neq\zeta} \cos(\zeta') \qquad \forall \quad \zeta \in \{\alpha,\beta,\gamma\}, \tag{C.2.2}$$

$$\frac{\partial^2 W}{\partial \zeta' \partial \zeta} = 2\sin(\zeta)\sin(\zeta')\cos(\zeta'') \qquad \forall \quad \zeta,\zeta',\zeta'' \in \{\alpha,\beta,\gamma\} \quad \zeta \neq \zeta' \neq \zeta'', \tag{C.2.3}$$

$$\frac{\partial^2 W}{\partial \zeta^2} = 2[\cos^2(\zeta) - \sin^2(\zeta)] - 2 \prod_{\zeta'=\alpha,\beta,\gamma} \cos(\zeta') \qquad \forall \quad \zeta \in \{\alpha,\beta,\gamma\}. \tag{C.2.4}$$

Table C.3: Analytical expressions for $\nabla_X \nabla_{X'} \boldsymbol{R} \; \forall \; X, X' \in \{a,b,c,\alpha,\beta,\gamma\}$. Note that $\nabla_X \nabla_{X'} \boldsymbol{R} = \nabla_{X'} \nabla_X \boldsymbol{R}$.

X	X'	$\nabla_X \nabla_{X'} \boldsymbol{R}$
a	a	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
a	b	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
a	c	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
a	α	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
a	β	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
a	γ	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
b	b	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
b	c	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

b	α	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
b	β	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
b	γ	$\begin{pmatrix} 0 & -\sin(\gamma) & 0 \\ 0 & \cos(\gamma) & 0 \\ 0 & 0 & 0 \end{pmatrix}$
c	c	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
c	α	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\frac{\sin(\alpha)}{\sin(\gamma)} \\ 0 & 0 & \frac{1}{2\sqrt{W}\sin(\gamma)}\frac{\partial W}{\partial \alpha} \end{pmatrix}$
c	β	$\begin{pmatrix} 0 & 0 & -\sin(\beta) \\ 0 & 0 & \frac{\sin(\beta)\cos(\gamma)}{\sin(\gamma)} \\ 0 & 0 & \frac{1}{2\sqrt{W}\sin(\gamma)}\frac{\partial W}{\partial \beta} \end{pmatrix}$
c	γ	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{\sin^2(\gamma)} \\ 0 & 0 & \left[\frac{1}{2\sqrt{W}\sin(\gamma)}\frac{\partial W}{\partial \gamma} - \frac{\sqrt{W}\cos(\gamma)}{\sin^2(\gamma)}\right] \end{pmatrix}$
α	α	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\frac{c\cos(\alpha)}{\sin(\gamma)} \\ 0 & 0 & \frac{c}{2\sqrt{W}\sin(\gamma)}\left[\frac{\partial^2 W}{\partial \alpha^2} - \frac{1}{2W}\left(\frac{\partial W}{\partial \alpha}\right)^2\right] \end{pmatrix}$
α	β	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{c}{2\sqrt{W}\sin(\gamma)}\left[\frac{\partial^2 W}{\partial \alpha \partial \beta} - \frac{1}{2W}\frac{\partial W}{\partial \alpha}\frac{\partial W}{\partial \beta}\right] \end{pmatrix}$
α	γ	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{c\sin(\alpha)\cos(\gamma)}{\sin^2(\gamma)} \\ 0 & 0 & \frac{c}{2\sqrt{W}\sin(\gamma)}\left[\frac{\partial^2 W}{\partial \alpha \partial \gamma} - \frac{1}{2W}\frac{\partial W}{\partial \alpha}\frac{\partial W}{\partial \gamma} - \frac{\cos(\gamma)}{\sin(\gamma)}\frac{\partial W}{\partial \alpha}\right] \end{pmatrix}$
β	β	$\begin{pmatrix} 0 & 0 & -c\cos(\beta) \\ 0 & 0 & \frac{c\cos(\beta)\cos(\gamma)}{\sin(\gamma)} \\ 0 & 0 & \frac{c}{2\sqrt{W}\sin(\gamma)}\left[\frac{\partial^2 W}{\partial \beta^2} - \frac{1}{2W}\left(\frac{\partial W}{\partial \beta}\right)^2\right] \end{pmatrix}$
β	γ	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -c\frac{\sin(\beta)}{\sin^2(\gamma)} \\ 0 & 0 & \frac{c}{2\sqrt{W}\sin(\gamma)}\left[\frac{\partial^2 W}{\partial \beta \partial \gamma} - \frac{1}{2W}\frac{\partial W}{\partial \beta}\frac{\partial W}{\partial \gamma} - \frac{\cos(\gamma)}{\sin(\gamma)}\frac{\partial W}{\partial \beta}\right] \end{pmatrix}$
γ	γ	$\begin{pmatrix} 0 & -b\cos(\gamma) & 0 \\ 0 & -b\sin(\gamma) & c\left[\frac{\cos(\alpha)\sin(\gamma)[\cos^2(\gamma)+1]-2\cos(\beta)\cos(\gamma)\sin(\gamma)}{\sin^4(\gamma)}\right] \\ 0 & 0 & \frac{c}{2\sqrt{W}\sin(\gamma)}\left[\frac{\partial^2 W}{\partial \gamma^2} - \frac{1}{2W}\left(\frac{\partial W}{\partial \gamma}\right)^2 - \frac{2\cos(\gamma)}{\sin(\gamma)}\frac{\partial W}{\partial \gamma} + \frac{2W[\cos^2(\gamma)+1]}{\sin^2(\gamma)}\right] \end{pmatrix}$

Table C.4: Analytical expressions for $\nabla_X \nabla_{X'} \boldsymbol{R}^{-1} \;\forall\; X, X' \in \{a, b, c, \alpha, \beta, \gamma\}$. Note that $\nabla_X \nabla_{X'} \boldsymbol{R}^{-1} = \nabla_{X'} \nabla_X \boldsymbol{R}^{-1}$.

X	X'	$\nabla_X \nabla_{X'} \boldsymbol{R}^{-1}$
a	a	$\begin{pmatrix} \frac{2}{a^3} & -\frac{2\cos(\gamma)}{a^3\sin(\gamma)} & -2\frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{a^3\sin(\gamma)\sqrt{W}} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
a	b	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
a	c	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
a	α	$\begin{pmatrix} 0 & 0 & \frac{1}{a^2\sin(\gamma)\sqrt{W}}\left[\sin(\alpha)\cos(\gamma) - \frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{2W}\frac{\partial W}{\partial \alpha}\right] \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
a	β	$\begin{pmatrix} 0 & 0 & \frac{1}{a^2\sin(\gamma)\sqrt{W}}\left[-\sin(\beta) - \frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{2W}\frac{\partial W}{\partial \beta}\right] \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
a	γ	$\begin{pmatrix} 0 & -\frac{1}{a^2\sin^2(\gamma)} & \frac{1}{a^2\sin(\gamma)\sqrt{W}}\left[\frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{\sin(\gamma)} - \frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{2W}\frac{\partial W}{\partial \gamma}\right] \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
b	b	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{2}{b^3\sin(\gamma)} & -2\frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{b^3\sin(\gamma)\sqrt{W}} \\ 0 & 0 & 0 \end{pmatrix}$
b	c	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
b	α	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{1}{b^2\sin(\gamma)\sqrt{W}}\left[-\sin(\alpha) - \frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{2W}\frac{\partial W}{\partial \alpha}\right] \\ 0 & 0 & 0 \end{pmatrix}$
b	β	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{1}{b^2\sin(\gamma)\sqrt{W}}\left[\sin(\beta)\cos(\gamma) - \frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{2W}\frac{\partial W}{\partial \beta}\right] \\ 0 & 0 & 0 \end{pmatrix}$
b	γ	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{\cos(\gamma)}{b^2\sin^2(\gamma)} & \frac{1}{b^2\sin(\gamma)\sqrt{W}}\left[\frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{\sin(\gamma)} - \frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{2W}\frac{\partial W}{\partial \gamma}\right] \\ 0 & 0 & 0 \end{pmatrix}$
c	c	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{2\sin(\gamma)}{c^3\sqrt{W}} \end{pmatrix}$
c	α	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{\sin(\gamma)}{2c^2W^{3/2}}\frac{\partial W}{\partial \alpha} \end{pmatrix}$

c	β	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{\sin(\gamma)}{2c^2W^{3/2}}\frac{\partial W}{\partial \beta} \end{pmatrix}$
c	γ	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{\sin(\gamma)}{2c^2W^{3/2}}\frac{\partial W}{\partial \gamma} - \frac{\cos(\gamma)}{c^2\sqrt{W}} \end{pmatrix}$
α	α	$\begin{pmatrix} 0 & 0 & \frac{1}{a\sin(\gamma)\sqrt{W}}\left[-\cos(\alpha)\cos(\gamma) + \frac{\sin(\alpha)\cos(\gamma)}{W}\frac{\partial W}{\partial \alpha} - \frac{3[\cos(\beta)-\cos(\alpha)\cos(\gamma)]}{4W^2}\left(\frac{\partial W}{\partial \alpha}\right)^2 + \frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{2W}\frac{\partial^2 W}{\partial \alpha^2}\right] \\ 0 & 0 & \frac{1}{b\sin(\gamma)\sqrt{W}}\left[\cos(\alpha) - \frac{\sin(\alpha)}{W}\frac{\partial W}{\partial \alpha} - \frac{3[\cos(\alpha)-\cos(\beta)\cos(\gamma)]}{4W^2}\left(\frac{\partial W}{\partial \alpha}\right)^2 + \frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{2W}\frac{\partial^2 W}{\partial \alpha^2}\right] \\ 0 & 0 & \frac{\sin(\gamma)}{4cW^{3/2}}\left[\frac{3}{W}\left(\frac{\partial W}{\partial \alpha}\right)^2 - 2\frac{\partial^2 W}{\partial \alpha^2}\right] \end{pmatrix}$
α	β	$\begin{pmatrix} 0 & 0 & \frac{1}{a\sin(\gamma)\sqrt{W}}\left[-\frac{\sin(\beta)}{2W}\frac{\partial W}{\partial \alpha} + \frac{\sin(\alpha)\cos(\gamma)}{2W}\frac{\partial W}{\partial \beta} - \frac{3[\cos(\beta)-\cos(\alpha)\cos(\gamma)]}{4W^2}\frac{\partial W}{\partial \alpha}\frac{\partial W}{\partial \beta} + \frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{2W}\frac{\partial^2 W}{\partial \alpha \partial \beta}\right] \\ 0 & 0 & \frac{1}{b\sin(\gamma)\sqrt{W}}\left[\frac{\sin(\beta)\cos(\gamma)}{2W}\frac{\partial W}{\partial \alpha} - \frac{\sin(\alpha)}{2W}\frac{\partial W}{\partial \beta} - \frac{3[\cos(\alpha)-\cos(\beta)\cos(\gamma)]}{4W^2}\frac{\partial W}{\partial \alpha}\frac{\partial W}{\partial \beta} + \frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{2W}\frac{\partial^2 W}{\partial \alpha \partial \beta}\right] \\ 0 & 0 & \frac{\sin(\gamma)}{4cW^{3/2}}\left[\frac{3}{W}\frac{\partial W}{\partial \alpha}\frac{\partial W}{\partial \beta} - 2\frac{\partial^2 W}{\partial \alpha \partial \beta}\right] \end{pmatrix}$
α	γ	$\begin{pmatrix} 0 & 0 & \frac{1}{a\sin(\gamma)\sqrt{W}}\left[\frac{\sin(\alpha)}{\sin(\gamma)} + \frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{2\sin(\gamma)W}\frac{\partial W}{\partial \alpha} + \frac{\sin(\alpha)\cos(\gamma)}{2W}\frac{\partial W}{\partial \gamma} - \frac{3[\cos(\beta)-\cos(\alpha)\cos(\gamma)]}{4W^2}\frac{\partial W}{\partial \alpha}\frac{\partial W}{\partial \gamma} + \frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{2W}\frac{\partial^2 W}{\partial \alpha \partial \gamma}\right] \\ 0 & 0 & \frac{1}{b\sin(\gamma)\sqrt{W}}\left[-\frac{\sin(\alpha)\cos(\gamma)}{\sin(\gamma)} + \frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{2\sin(\gamma)W}\frac{\partial W}{\partial \alpha} - \frac{\sin(\alpha)}{2W}\frac{\partial W}{\partial \gamma} - \frac{3[\cos(\alpha)-\cos(\beta)\cos(\gamma)]}{4W^2}\frac{\partial W}{\partial \alpha}\frac{\partial W}{\partial \gamma} + \frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{2W}\frac{\partial^2 W}{\partial \alpha \partial \gamma}\right] \\ 0 & 0 & \frac{1}{2cW^{3/2}}\left[-\cos(\gamma)\frac{\partial W}{\partial \alpha} + \frac{3\sin(\gamma)}{2W}\frac{\partial W}{\partial \alpha}\frac{\partial W}{\partial \gamma} - \sin(\gamma)\frac{\partial^2 W}{\partial \alpha \partial \gamma}\right] \end{pmatrix}$
β	β	$\begin{pmatrix} 0 & 0 & \frac{1}{a\sin(\gamma)\sqrt{W}}\left[\cos(\beta) - \frac{\sin(\beta)}{W}\frac{\partial W}{\partial \beta} - \frac{3[\cos(\beta)-\cos(\alpha)\cos(\gamma)]}{4W^2}\left(\frac{\partial W}{\partial \beta}\right)^2 + \frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{2W}\frac{\partial^2 W}{\partial \beta^2}\right] \\ 0 & 0 & \frac{1}{b\sin(\gamma)\sqrt{W}}\left[-\cos(\beta)\cos(\gamma) + \frac{\sin(\beta)\cos(\gamma)}{W}\frac{\partial W}{\partial \beta} - \frac{3[\cos(\alpha)-\cos(\beta)\cos(\gamma)]}{4W^2}\left(\frac{\partial W}{\partial \beta}\right)^2 + \frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{2W}\frac{\partial^2 W}{\partial \beta^2}\right] \\ 0 & 0 & \frac{\sin(\gamma)}{2cW^{3/2}}\left[\frac{3}{2W}\left(\frac{\partial W}{\partial \beta}\right)^2 - \frac{\partial^2 W}{\partial \beta^2}\right] \end{pmatrix}$

β	γ	$\begin{pmatrix} & & \frac{1}{a\sin(\gamma)\sqrt{W}}\left[-\frac{\sin(\beta)\cos(\gamma)}{\sin(\gamma)}+\frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{2\sin(\gamma)W}\frac{\partial W}{\partial\beta}-\frac{\sin(\beta)}{2W}\frac{\partial W}{\partial\gamma}\right. \\ 0 & 0 & \left.-\frac{3[\cos(\beta)-\cos(\alpha)\cos(\gamma)]}{4W^2}\frac{\partial W}{\partial\beta}\frac{\partial W}{\partial\gamma}+\frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{2W}\frac{\partial^2 W}{\partial\beta\partial\gamma}\right] \\ & & \frac{1}{b\sin(\gamma)\sqrt{W}}\left[\frac{\sin(\beta)}{\sin(\gamma)}+\frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{2\sin(\gamma)W}\frac{\partial W}{\partial\beta}+\frac{\sin(\beta)\cos(\gamma)}{2W}\frac{\partial W}{\partial\gamma}\right. \\ 0 & 0 & \left.-\frac{3[\cos(\alpha)-\cos(\beta)\cos(\gamma)]}{4W^2}\frac{\partial W}{\partial\beta}\frac{\partial W}{\partial\gamma}+\frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{2W}\frac{\partial^2 W}{\partial\beta\partial\gamma}\right] \\ 0 & 0 & \frac{1}{2cW^{3/2}}\left[-\cos(\gamma)\frac{\partial W}{\partial\beta}+\frac{3\sin(\gamma)}{2W}\frac{\partial W}{\partial\beta}\frac{\partial W}{\partial\gamma}-\sin(\gamma)\frac{\partial^2 W}{\partial\beta\partial\gamma}\right] \end{pmatrix}$
γ	γ	$\begin{pmatrix} & & \frac{1}{a\sin(\gamma)\sqrt{W}}\left[\frac{2\cos(\alpha)\cos(\gamma)-\cos(\beta)[\cos^2(\gamma)+1]}{\sin^2(\gamma)}+\frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{\sin(\gamma)W}\frac{\partial W}{\partial\gamma}\right. \\ 0 & \frac{-2\cos(\gamma)}{a\sin^3(\gamma)} & \left.-\frac{3[\cos(\beta)-\cos(\alpha)\cos(\gamma)]}{4W^2}\left(\frac{\partial W}{\partial\gamma}\right)^2\right. \\ & & \left.+\frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{2W}\frac{\partial^2 W}{\partial\gamma^2}\right] \\ & & \frac{1}{b\sin(\gamma)\sqrt{W}}\left[\frac{2\cos(\beta)\cos(\gamma)-\cos(\alpha)[\cos^2(\gamma)+1]}{\sin^2(\gamma)}+\frac{\cos(\beta)-\cos(\alpha)\cos(\gamma)}{\sin(\gamma)W}\frac{\partial W}{\partial\gamma}\right. \\ 0 & \frac{\cos^2(\gamma)+1}{b\sin^3(\gamma)} & \left.-\frac{3[\cos(\alpha)-\cos(\beta)\cos(\gamma)]}{4W^2}\left(\frac{\partial W}{\partial\gamma}\right)^2\right. \\ & & \left.+\frac{\cos(\alpha)-\cos(\beta)\cos(\gamma)}{2W}\frac{\partial^2 W}{\partial\gamma^2}\right] \\ 0 & 0 & \frac{1}{c\sqrt{W}}\left[-\sin(\gamma)-\frac{\cos(\gamma)}{W}\frac{\partial W}{\partial\gamma}+\frac{3\sin(\gamma)}{4W^2}\left(\frac{\partial W}{\partial\gamma}\right)^2-\frac{\sin(\gamma)}{2W}\frac{\partial^2 W}{\partial\gamma^2}\right] \end{pmatrix}$

Appendix D

Derivation of CP2 Flexible Gradients

It was observed that small errors in the analytical gradients with respect to the independent dofs were resulting in poor performance during the CP2 Global Search. These errors were traced back to intermediate gradient quantities related to the change in atomic positions with respect to the conformation dofs. Below, I derive the analytical expressions related to these gradient quantities.

D.1 Constructing Molecules

D.1.1 Background Information

When constructing a molecule, it is assumed that we have a full Zmatrix definition of the molecule. This encompasses,

1. The connectivity of all atoms in the molecule.
2. The corresponding values for each conformational degree of freedom (dof): bond length, bond angle, and dihedral/torsion angle.

The full Zmatrix should then contain $(N_{\text{atm}} \times 3 - 6)$ elements, where N_{atm} is the number of atoms in the molecule. This is obvious when one considers that every atom site will have a triplet of conformational dofs except atom 1 (no bonds, angles, or dihedrals), atom 2 (only a bond), and atom 3 (no dihedral). With this information, the molecule can be constructed.

Henceforth, all bonds, angles, and dihedrals will be identified as $b_{i,j}$, $a_{i,j,k}$, and $d_{i,j,k,l}$ respectively, where the subscripts refer to the participating atoms. Note that this initial section is irrespective of whether the dof is independent/dependent, and where/how the dof value might be obtained (optimizer-specified, or LAM-derived).

D.1.2 Atoms 1, 2, and 3

To begin, the first three atoms which have no dihedral information can be easily placed. Atom 1 of each atom will always be defined as the origin of the coordinate system. Thereafter, atom 2 is displaced by $b_{1,2}$ in the positive direction along the x-axis. Atom 3 can be defined in two

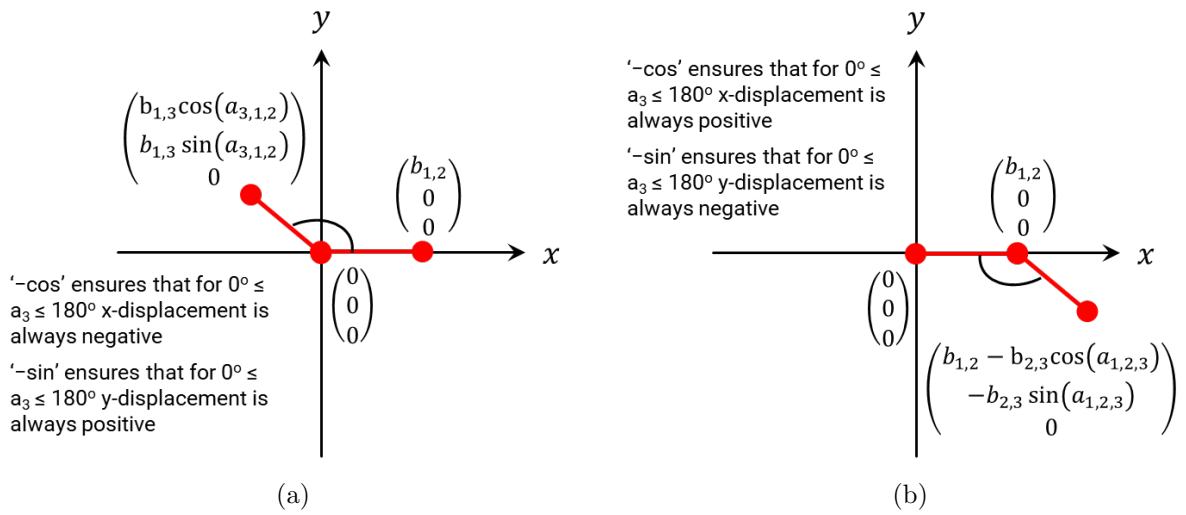


Figure D.1: Cartesian coordinates of the first 3 atoms in a molecule where (a) atom 3 is bonded to atom 1, or (b) atom 3 is bonded to atom 2.

ways, either bonded to atom 1 or bonded to atom 2. In either case, atom 3's position on the xy-plane can be determined by simple trigonometric operations, as shown in Figure D.1.

D.1.3 Atoms 4+

Once the first 3 atoms are placed, each subsequent atom can be placed using the bond, angle, and dihedral information from preceding atoms. For simplicity, I will index these preceding atoms as 1, 2, and 3, and the atom-of-interest as 4. In practice the preceding atoms can be of any combination of indices below the atom-of-interest. From the Zmatrix, knowledge of $b_{4,3}$, $a_{4,3,2}$, and $d_{4,3,2,1}$ is given. Moreover, the Cartesian coordinates of atom 1–3 ($\mathbf{r}_{1-3}$) are also known (since, by definition, they precede 4). This is illustrated in Figure D.2.

To place atom 4, a local coordinate system centered about atom 2 can be defined, as shown in Figure D.3. The axis of this local system may be obtained as follows. From the (known) Cartesian coordinates of the preceding atoms, two directional unit-vectors can be defined as

$$\mathbf{tvec1} = \mathbf{r}_{2\to3} = \frac{\mathbf{r}_3 - \mathbf{r}_2}{|\mathbf{r}_3 - \mathbf{r}_2|}, \tag{D.1.1}$$

and

$$\mathbf{tvec2} = \mathbf{r}_{2\to1} = \frac{\mathbf{r}_1 - \mathbf{r}_2}{|\mathbf{r}_1 - \mathbf{r}_2|}. \tag{D.1.2}$$

From these two vectors, an orthonormal vector is obtained by normalizing their cross product,

$$\mathbf{tcrp} = \frac{\mathbf{tvec1} \times \mathbf{tvec2}}{|\mathbf{tvec1} \times \mathbf{tvec2}|}. \tag{D.1.3}$$

Finally, the last axis of the coordinate system may be obtained by conducting an additional cross product,

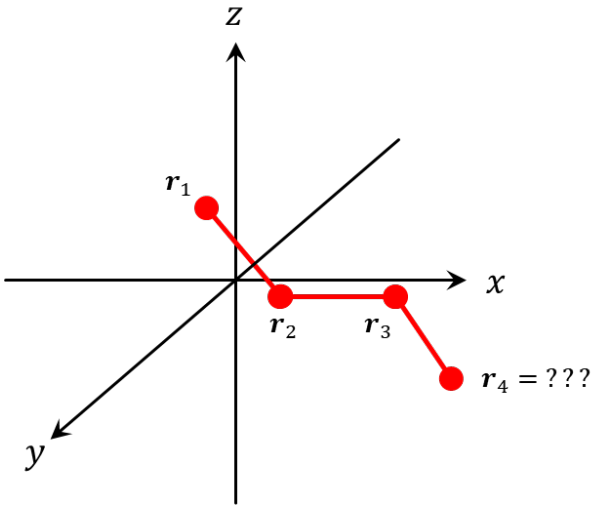


Figure D.2: The key problem we hope to solve is determine the position of atom 4 given knowledge of the connectivity relationship between 1, 2, 3, and 4.

$$\mathbf{ncrp} = \mathbf{tcrp} \times \mathbf{tvec1}. \tag{D.1.4}$$

It is worth noting that since **tcrp** and **tvec1** are orthonormal, **ncrp**, as defined in Equation D.1.4, is also a unit vector. Because atom 2 and atom 3 are aligned along the **tvec1** direction, the local system can be equivalently defined with atom 3 as its origin.

Using this re-centered local system, it is more convenient to perceive the system in spherical coordinates as shown in Figure D.4. In this system, atom 2 and atom 3 are placed along the **tvec1** vector (equivalent to the z-axis). Atom 1 lies on the **ncrp**/**tvec1**-plane. The angle, $a_{4,3,2}$, defines the polar angle ($0° \leq \theta \leq 180°$) between the 'z-axis' and the vector $\mathbf{r}_{3\to4}$. The dihedral, $d_{4,3,2,1}$, then defines the azimuthal angle ($0° \leq \phi \leq 360°$) from the **ncrp** axis to the projection of

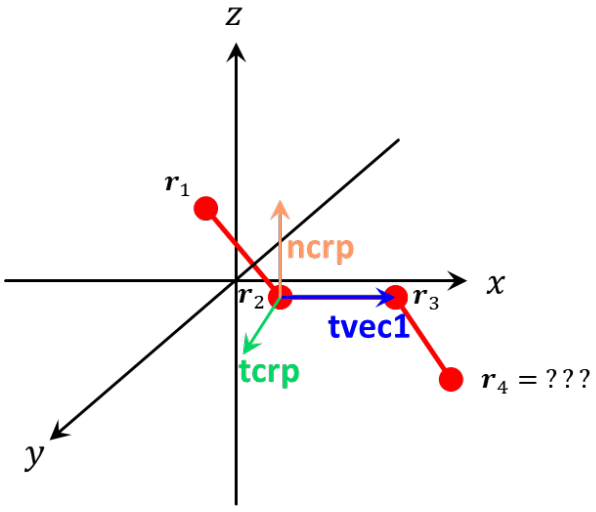


Figure D.3: A local coordinate system is defined around atom 2.

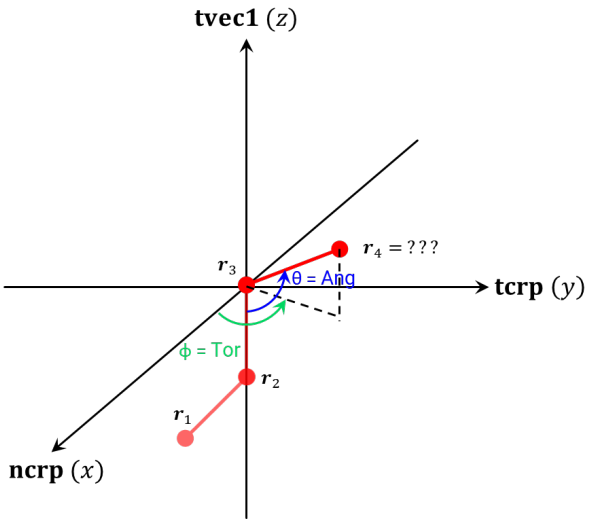


Figure D.4: A spherical coordinate system is defined around atom 3.

$\mathbf{r}_{3\to4}$ onto the **ncrp**/**tcrp**-plane. With these definitions, the directional vector $\mathbf{r}_{3\to4}$ is given by

$$\mathbf{r}_{3\to4} = \begin{pmatrix} b_{4,3}\cdot\sin(a_{4,3,2})\cdot\cos(d_{4,3,2,1}) \\ b_{4,3}\cdot\sin(a_{4,3,2})\cdot\sin(d_{4,3,2,1}) \\ b_{4,3}\cdot\cos(a_{4,3,2}) \end{pmatrix}. \tag{D.1.5}$$

Since this vector is currently defined in terms of the spherical coordinate system (and its corresponding Cartesian axes), it is necessary to 'rotate' the vector back to the original coordinate system, as given by

$$\begin{aligned} \mathbf{r}_{3\to4} = {} & \mathbf{ncrp}\cdot b_{4,3}\cdot\sin(a_{4,3,2})\cdot\cos(d_{4,3,2,1}) \\ & + \mathbf{tcrp}\cdot b_{4,3}\cdot\sin(a_{4,3,2})\cdot\sin(d_{4,3,2,1}) \\ & - \mathbf{tvec1}\cdot b_{4,3}\cdot\cos(a_{4,3,2}). \end{aligned} \tag{D.1.6}$$

From here, $\mathbf{r}_4$ is trivially obtained from $\mathbf{r}_3$ by,

$$\mathbf{r}_4 = \mathbf{r}_3 + \mathbf{r}_{3\to4}. \tag{D.1.7}$$

All atom sites defined with torsions will follow these steps to obtain their corresponding Cartesian positions.

D.2 Derivatives of Atomic Cartesian Coordinates

Derivatives of these Cartesian coordinates, with respect to the independent dofs, are used to calculate $\frac{\partial U_{\text{inter}}}{\partial\theta^{\text{ind}}}$. In this section, I derive their analytical expressions with judicious application of the chain and product rules.

In practice, values of $b_{i,j}$, $a_{i,j,k}$, and $d_{i,j,k,l}$, as well as their derivatives with respect to the independent dofs, $\boldsymbol{\theta}^{\text{ind}}$, will be necessary to evaluate these analytical expressions. Within the LAM framework described in Section 3.2.2, obtaining these values is trivial. The LAM model defines the value of all dofs (independent or dependent) generally as functions of $\boldsymbol{\theta}^{\text{ind}}$,

$$b_{i,j} = b_{i,j}(\boldsymbol{\theta}^{\text{ind}}), \tag{D.2.1}$$

$$a_{i,j,k} = a_{i,j,k}(\boldsymbol{\theta}^{\text{ind}}), \tag{D.2.2}$$

$$d_{i,j,k,l} = d_{i,j,k,l}(\boldsymbol{\theta}^{\text{ind}}). \tag{D.2.3}$$

Likewise, the derivatives of all dofs with respect to $\boldsymbol{\theta}^{\text{ind}}$ are obtainable from the LAM surrogate model,

$$\frac{\partial b_{i,j}}{\partial \theta^{\text{ind}}} = \frac{\partial b_{i,j}}{\partial \theta^{\text{ind}}}(\boldsymbol{\theta}^{\text{ind}}), \tag{D.2.4}$$

$$\frac{\partial a_{i,j,k}}{\partial \theta^{\text{ind}}} = \frac{\partial a_{i,j,k}}{\partial \theta^{\text{ind}}}(\boldsymbol{\theta}^{\text{ind}}), \tag{D.2.5}$$

$$\frac{\partial d_{i,j,k,l}}{\partial \theta^{\text{ind}}} = \frac{\partial d_{i,j,k,l}}{\partial \theta^{\text{ind}}}(\boldsymbol{\theta}^{\text{ind}}). \tag{D.2.6}$$

A special case arises when the dof is also an independent dof. In CP2, only torsions ($d_{i,j,k,l}$) may be treated as independent dofs. If

$$d_{i,j,k,l} = \theta_1^{\text{ind}}, \tag{D.2.7}$$

then

$$\frac{\partial d_{i,j,k,l}}{\partial \theta_1^{\text{ind}}} = \frac{\partial \theta_1^{\text{ind}}}{\partial \theta_1^{\text{ind}}} = 1, \tag{D.2.8}$$

and

$$\frac{\partial d_{i,j,k,l}}{\partial \theta_{i\neq 1}^{\text{ind}}} = \frac{\partial \theta_1^{\text{ind}}}{\partial \theta_{i\neq 1}^{\text{ind}}} = 0, \tag{D.2.9}$$

where $\theta_{i\neq 1}^{\text{ind}}$ is any independent dof other than θ_i^{ind}. The last equation must be true given that the independent dofs, by definition, have no dependence on each other. With this in mind, evaluation of the analytical gradient expressions is wholly possible from just the LAMs and the optimizer-specified $\boldsymbol{\theta}^{\text{ind}}$.

D.2.1 Derivatives of Atoms 1, 2, and 3

As before, I will treat the first 3 atoms separately. For clarity, any derivative that reduces to zero has been omitted from the respective equations. For atom 1, which is fixed to the origin:

$$\mathbf{r}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \rightarrow \frac{\partial \mathbf{r}_1}{\partial \theta^{\text{ind}}} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{D.2.10}$$

For atom 2, which is displaced from atom 1 in only the x-direction:

$$\mathbf{r}_2 = \begin{pmatrix} x_1 + b_{1,2} \\ 0 \\ 0 \end{pmatrix} \rightarrow \frac{\partial \mathbf{r}_2}{\partial \theta^{\text{ind}}} = \begin{pmatrix} \frac{\partial b_{1,2}}{\partial \theta^{\text{ind}}} \\ 0 \\ 0 \end{pmatrix}. \tag{D.2.11}$$

And lastly for atom 3, both the case where atom 3 is bonded to atom 1,

$$\mathbf{r}_3 = \begin{pmatrix} x_1 + b_{1,3}\cos\left(a_{3,1,2}\right) \\ y_1 + b_{1,3}\sin\left(a_{3,1,2}\right) \\ 0 \end{pmatrix} \rightarrow \frac{\partial \mathbf{r}_3}{\partial \theta^{\text{ind}}} = \begin{pmatrix} \frac{\partial b_{1,3}}{\partial \theta^{\text{ind}}}\cos\left(a_{3,1,2}\right) - b_{1,3}\frac{\partial a_{3,1,2}}{\partial \theta^{\text{ind}}}\sin\left(a_{3,1,2}\right) \\ \frac{\partial b_{1,3}}{\partial \theta^{\text{ind}}}\sin\left(a_{3,1,2}\right) + b_{1,3}\frac{\partial a_{3,1,2}}{\partial \theta^{\text{ind}}}\cos\left(a_{3,1,2}\right) \\ 0 \end{pmatrix}, \tag{D.2.12}$$

as well as the case where atom 3 is bonded to atom 2,

$$\mathbf{r}_3 = \begin{pmatrix} x_2 - b_{2,3}\cos\left(a_{1,2,3}\right) \\ y_2 - b_{2,3}\sin\left(a_{1,2,3}\right) \\ 0 \end{pmatrix} \rightarrow \frac{\partial \mathbf{r}_3}{\partial \theta^{\text{ind}}} = \begin{pmatrix} \frac{\partial x_2}{\partial \theta^{\text{ind}}} - \frac{\partial b_{2,3}}{\partial \theta^{\text{ind}}}\cos\left(a_{1,2,3}\right) + b_{2,3}\frac{\partial a_{1,2,3}}{\partial \theta^{\text{ind}}}\sin\left(a_{1,2,3}\right) \\ -\frac{\partial b_{2,3}}{\partial \theta^{\text{ind}}}\sin\left(a_{1,2,3}\right) - b_{2,3}\frac{\partial a_{1,2,3}}{\partial \theta^{\text{ind}}}\cos\left(a_{1,2,3}\right) \\ 0 \end{pmatrix}, \tag{D.2.13}$$

are considered.

D.2.2 Derivatives of Atoms 4+

As derived in the preceding section, the Cartesian coordinates of an atom defined by its bond, angle, and dihedral value are obtained through a series of vector operations. To derive corresponding derivatives, I will begin by re-expressing the results of these vector operations explicitly using the (known) Cartesian coordinates of atoms 1, 2, and 3. Thereafter, the corresponding derivative quantities may be expressed in terms of the the (known) Cartesian coordinates of atoms 1, 2, and 3 as well as their derivatives with respect to θ^{ind}. It must be stressed that due to the sequential nature of the molecule's construction, all $\frac{\partial \mathbf{r}_n}{\partial \theta^{\text{ind}}}$ values for n lower than the current atom's index would be known when assessing $\frac{\partial \mathbf{r}}{\partial \theta^{\text{ind}}}$ for the current atom. With this in mind, I first define the following auxiliary variables:

$$\Delta x_{i,j} = x_j - x_i, \tag{D.2.14}$$

$$\Delta y_{i,j} = y_j - y_i, \tag{D.2.15}$$

$$\Delta z_{i,j} = z_j - z_i, \tag{D.2.16}$$

and their corresponding derivative values:

$$\frac{\partial \Delta x_{i,j}}{\partial \theta^{\text{ind}}} = \frac{\partial x_j}{\partial \theta^{\text{ind}}} - \frac{\partial x_i}{\partial \theta^{\text{ind}}}, \tag{D.2.17}$$

$$\frac{\partial \Delta y_{i,j}}{\partial \theta^{\text{ind}}} = \frac{\partial y_j}{\partial \theta^{\text{ind}}} - \frac{\partial y_i}{\partial \theta^{\text{ind}}}, \tag{D.2.18}$$

$$\frac{\partial \Delta z_{i,j}}{\partial \theta^{\text{ind}}} = \frac{\partial z_j}{\partial \theta^{\text{ind}}} - \frac{\partial z_i}{\partial \theta^{\text{ind}}}. \tag{D.2.19}$$

For the first two vector quantities defined earlier:

$$\mathbf{tvec1} = \frac{\mathbf{r}_3 - \mathbf{r}_2}{|\mathbf{r}_3 - \mathbf{r}_2|} = \frac{1}{\sqrt{(\Delta x_{2,3})^2 + (\Delta y_{2,3})^2 + (\Delta z_{2,3})^2}} \begin{pmatrix} \Delta x_{2,3} \\ \Delta y_{2,3} \\ \Delta z_{2,3} \end{pmatrix}, \tag{D.2.20}$$

$$\mathbf{tvec2} = \frac{\mathbf{r}_1 - \mathbf{r}_2}{|\mathbf{r}_1 - \mathbf{r}_2|} = \frac{1}{\sqrt{(\Delta x_{2,1})^2 + (\Delta y_{2,1})^2 + (\Delta z_{2,1})^2}} \begin{pmatrix} \Delta x_{2,1} \\ \Delta y_{2,1} \\ \Delta z_{2,1} \end{pmatrix}. \tag{D.2.21}$$

The vector **tcrp** should be orthonormal to **tvec1** and **tvec2**. An orthogonal vector can be found by taking the cross product of **tvec1** and **tvec2**:

$$\mathbf{tvec1} \times \mathbf{tvec2} = \frac{\begin{pmatrix} \Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3} \\ \Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3} \\ \Delta x_{2,3} \cdot \Delta y_{2,1} - \Delta x_{2,1} \cdot \Delta y_{2,3} \end{pmatrix}}{\sqrt{[(\Delta x_{2,3})^2 + (\Delta y_{2,3})^2 + (\Delta z_{2,3})^2] \cdot [(\Delta x_{2,1})^2 + (\Delta y_{2,1})^2 + (\Delta z_{2,1})^2]}}, \tag{D.2.22}$$

and normalizing this gives **tcrp**:

$$\mathbf{tcrp} = \frac{\begin{pmatrix} \Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3} \\ \Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3} \\ \Delta x_{2,3} \cdot \Delta y_{2,1} - \Delta x_{2,1} \cdot \Delta y_{2,3} \end{pmatrix}}{\sqrt{\begin{array}{l} (\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3})^2 \\ + (\Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3})^2 \\ + (\Delta x_{2,3} \cdot \Delta y_{2,1} - \Delta x_{2,1} \cdot \Delta y_{2,3})^2 \end{array}}}. \tag{D.2.23}$$

For brevity, I will re-express the denominators using these variables:

$$\Psi_{2,1} = (\Delta x_{2,1})^2 + (\Delta y_{2,1})^2 + (\Delta z_{2,1})^2, \tag{D.2.24}$$

$$\Psi_{2,3} = (\Delta x_{2,3})^2 + (\Delta y_{2,3})^2 + (\Delta z_{2,3})^2, \tag{D.2.25}$$

$$\begin{aligned} \Psi_{2,3\times 2,1} =& (\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3})^2 \\ &+ (\Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3})^2 \\ &+ (\Delta x_{2,3} \cdot \Delta y_{2,1} - \Delta x_{2,1} \cdot \Delta y_{2,3})^2, \end{aligned} \tag{D.2.26}$$

$$\Psi_{(2,3)\times(2,3\times 2,1)} = \Psi_{2,3} \times \Psi_{2,3\times 2,1}, \tag{D.2.27}$$

such that:

$$\mathbf{tvec1} = \frac{\mathbf{r}_3 - \mathbf{r}_2}{|\mathbf{r}_3 - \mathbf{r}_2|} = \frac{1}{\sqrt{\Psi_{2,3}}} \begin{pmatrix} \Delta x_{2,3} \\ \Delta y_{2,3} \\ \Delta z_{2,3} \end{pmatrix}, \tag{D.2.28}$$

$$\mathbf{tvec2} = \frac{\mathbf{r}_1 - \mathbf{r}_2}{|\mathbf{r}_1 - \mathbf{r}_2|} = \frac{1}{\sqrt{\Psi_{2,1}}} \begin{pmatrix} \Delta x_{2,1} \\ \Delta y_{2,1} \\ \Delta z_{2,1} \end{pmatrix}, \tag{D.2.29}$$

$$\mathbf{tcrp} = \frac{1}{\sqrt{\Psi_{2,3\times 2,1}}} \begin{pmatrix} \Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3} \\ \Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3} \\ \Delta x_{2,3} \cdot \Delta y_{2,1} - \Delta x_{2,1} \cdot \Delta y_{2,3} \end{pmatrix}. \tag{D.2.30}$$

Finally, we can obtain **ncrp** by taking the corresponding cross product of the **tcrp** and **tvec1**:

$$\mathbf{ncrp} = \mathbf{tcrp} \times \mathbf{tvec1} = \frac{\begin{pmatrix} \Delta z_{2,3} \cdot (\Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3}) \\ -\Delta y_{2,3} \cdot (\Delta x_{2,3} \cdot \Delta y_{2,1} - \Delta x_{2,1} \cdot \Delta y_{2,3}) \\ \Delta x_{2,3} \cdot (\Delta x_{2,3} \cdot \Delta y_{2,1} - \Delta x_{2,1} \cdot \Delta y_{2,3}) \\ -\Delta z_{2,3} \cdot (\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3}) \\ \Delta y_{2,3} \cdot (\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3}) \\ -\Delta x_{2,3} \cdot (\Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3}) \end{pmatrix}}{\sqrt{\Psi_{(2,3)\times(2,3\times2,1)}}}. \tag{D.2.31}$$

Note that the denominator normalization factor is just the scalar multiplication between the normalization factors of the original two vectors ($\Psi_{2,3}$ and $\Psi_{2,3\times2,1}$).

Finally, I can explicitly express the Cartesian coordinates of atom 4 as:

$$\begin{aligned} r_4 = r_3 &+ r_{\mathbf{ncrp}} \cdot b_{4,3} \cdot \sin(a_{4,3,2}) \cdot \cos(d_{4,3,2,1}) \\ &+ r_{\mathbf{tcrp}} \cdot b_{4,3} \cdot \sin(a_{4,3,2}) \cdot \sin(d_{4,3,2,1}) \\ &- r_{\mathbf{tvec1}} \cdot b_{4,3} \cdot \cos(a_{4,3,2}), \end{aligned} \tag{D.2.32}$$

where r_{xxx} is just a general variable representing the x, y, and z components of the vector, $\mathbf{r}_{\mathrm{xxx}}$. By extension, the corresponding gradient expression will be:

$$\begin{aligned} \frac{\partial r_4}{\partial \theta^{\mathrm{ind}}} = &\frac{\partial r_3}{\partial \theta^{\mathrm{ind}}} \\ &+ \frac{\partial r_{\mathbf{ncrp}}}{\partial \theta^{\mathrm{ind}}} \cdot [b_{4,3} \cdot \sin(a_{4,3,2}) \cdot \cos(d_{4,3,2,1})] \\ &+ r_{\mathbf{ncrp}} \cdot \left\{ \frac{\partial b_{4,3}}{\partial \theta^{\mathrm{ind}}} \cdot [\sin(a_{4,3,2}) \cdot \cos(d_{4,3,2,1})] \right\} \\ &+ r_{\mathbf{ncrp}} \cdot \left\{ b_{4,3} \cdot \left[\frac{\partial a_{4,3,2}}{\partial \theta^{\mathrm{ind}}} \cos(a_{4,3,2}) \cos(d_{4,3,2,1}) - \frac{\partial d_{4,3,2,1}}{\partial \theta^{\mathrm{ind}}} \sin(a_{4,3,2}) \sin(d_{4,3,2,1}) \right] \right\} \\ &+ \frac{\partial r_{\mathbf{tcrp}}}{\partial \theta^{\mathrm{ind}}} \cdot [b_{4,3} \cdot \sin(a_{4,3,2}) \cdot \sin(d_{4,3,2,1})] \\ &+ r_{\mathbf{tcrp}} \cdot \left\{ \frac{\partial b_{4,3}}{\partial \theta^{\mathrm{ind}}} \cdot [\sin(a_{4,3,2}) \cdot \sin(d_{4,3,2,1})] \right\} \\ &+ r_{\mathbf{tcrp}} \cdot \left\{ b_{4,3} \cdot \left[\frac{\partial a_{4,3,2}}{\partial \theta^{\mathrm{ind}}} \cos(a_{4,3,2}) \sin(d_{4,3,2,1}) + \frac{\partial d_{4,3,2,1}}{\partial \theta^{\mathrm{ind}}} \sin(a_{4,3,2}) \cos(d_{4,3,2,1}) \right] \right\} \\ &- \frac{\partial r_{\mathbf{tvec1}}}{\partial \theta^{\mathrm{ind}}} \cdot [b_{4,3} \cdot \cos(a_{4,3,2})] \\ &- r_{\mathbf{tvec1}} \cdot \left[\frac{\partial b_{4,3}}{\partial \theta^{\mathrm{ind}}} \cos(a_{4,3,2}) - b_{4,3} \frac{\partial a_{4,3,2}}{\partial \theta^{\mathrm{ind}}} \sin(a_{4,3,2}) \right]. \end{aligned} \tag{D.2.33}$$

As described earlier, all quantities within this expression are given from the LAM model, except for $\frac{\partial r_{\mathbf{ncrp}}}{\partial \theta^{\text{ind}}}$, $\frac{\partial r_{\mathbf{tcrp}}}{\partial \theta^{\text{ind}}}$, and $\frac{\partial r_{\mathbf{tvec1}}}{\partial \theta^{\text{ind}}}$. Below, I define these values explicitly in terms of the Cartesian coordinate variables.

D.2.2.1 Derivatives of $\sqrt{\Psi}$ Variables

To facilitate expressing the derivatives of the vector components in a compact fashion, I begin by determining the derivatives of $\sqrt{\Psi}$. I choose to find the derivatives of the *square-root* since this is how the Ψ variables appear in all the vector expressions.
Firstly,

$$\sqrt{\Psi_{2,3}} = \sqrt{(\Delta x_{2,3})^2 + (\Delta y_{2,3})^2 + (\Delta z_{2,3})^2}, \tag{D.2.34}$$

$$\frac{\partial \sqrt{\Psi_{2,3}}}{\partial \theta^{\text{ind}}} = \frac{\Delta x_{2,3} \frac{\partial \Delta x_{2,3}}{\partial \theta^{\text{ind}}} + \Delta y_{2,3} \frac{\partial \Delta y_{2,3}}{\partial \theta^{\text{ind}}} + \Delta z_{2,3} \frac{\partial \Delta z_{2,3}}{\partial \theta^{\text{ind}}}}{\sqrt{(\Delta x_{2,3})^2 + (\Delta y_{2,3})^2 + (\Delta z_{2,3})^2}}. \tag{D.2.35}$$

Secondly,

$$\sqrt{\Psi_{2,3\times 2,1}} = \sqrt{\begin{array}{l}(\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3})^2 \\ + (\Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3})^2, \\ + (\Delta x_{2,3} \cdot \Delta y_{2,1} - \Delta x_{2,1} \cdot \Delta y_{2,3})^2\end{array}} \tag{D.2.36}$$

$$\begin{aligned}\frac{\partial \sqrt{\Psi_{2,3\times 2,1}}}{\partial \theta^{\text{ind}}} = & \\ & \frac{\begin{array}{l}(\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3}) \\ \cdot (\Delta z_{2,1} \frac{\partial \Delta y_{2,3}}{\partial \theta^{\text{ind}}} + \Delta y_{2,3} \frac{\partial \Delta z_{2,1}}{\partial \theta^{\text{ind}}} - \Delta z_{2,3} \frac{\partial \Delta y_{2,1}}{\partial \theta^{\text{ind}}} - \Delta y_{2,1} \frac{\partial \Delta z_{2,3}}{\partial \theta^{\text{ind}}})\end{array}}{\sqrt{\begin{array}{l}(\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3})^2 \\ + (\Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3})^2 \\ + (\Delta x_{2,3} \cdot \Delta y_{2,1} - \Delta x_{2,1} \cdot \Delta y_{2,3})^2\end{array}}} \\ & + \frac{\begin{array}{l}(\Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3}) \\ \cdot (\Delta x_{2,1} \frac{\partial \Delta z_{2,3}}{\partial \theta^{\text{ind}}} + \Delta z_{2,3} \frac{\partial \Delta x_{2,1}}{\partial \theta^{\text{ind}}} - \Delta x_{2,3} \frac{\partial \Delta z_{2,1}}{\partial \theta^{\text{ind}}} - \Delta z_{2,1} \frac{\partial \Delta x_{2,3}}{\partial \theta^{\text{ind}}})\end{array}}{\sqrt{\begin{array}{l}(\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3})^2 \\ + (\Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3})^2 \\ + (\Delta x_{2,3} \cdot \Delta y_{2,1} - \Delta x_{2,1} \cdot \Delta y_{2,3})^2\end{array}}} \\ & + \frac{\begin{array}{l}(\Delta x_{2,3} \cdot \Delta y_{2,1} - \Delta x_{2,1} \cdot \Delta y_{2,3}) \\ \cdot (\Delta y_{2,1} \frac{\partial \Delta x_{2,3}}{\partial \theta^{\text{ind}}} + \Delta x_{2,3} \frac{\partial \Delta y_{2,1}}{\partial \theta^{\text{ind}}} - \Delta y_{2,3} \frac{\partial \Delta x_{2,1}}{\partial \theta^{\text{ind}}} - \Delta x_{2,1} \frac{\partial \Delta y_{2,3}}{\partial \theta^{\text{ind}}})\end{array}}{\sqrt{\begin{array}{l}(\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3})^2 \\ + (\Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3})^2 \\ + (\Delta x_{2,3} \cdot \Delta y_{2,1} - \Delta x_{2,1} \cdot \Delta y_{2,3})^2\end{array}}}.\end{aligned} \tag{D.2.37}$$

For the last variable, it can be expressed as a function of the previous two variables:

$$\sqrt{\Psi_{(2,3)\times(2,3\times2,1)}} = \sqrt{\Psi_{2,3}} \times \sqrt{\Psi_{2,3\times2,1}}, \tag{D.2.38}$$

$$\frac{\partial\sqrt{\Psi_{(2,3)\times(2,3\times2,1)}}}{\partial\theta^{\text{ind}}} = \sqrt{\Psi_{2,3\times2,1}} \cdot \frac{\partial\sqrt{\Psi_{2,3}}}{\partial\theta^{\text{ind}}} + \sqrt{\Psi_{2,3}} \cdot \frac{\partial\sqrt{\Psi_{2,3\times2,1}}}{\partial\theta^{\text{ind}}}. \tag{D.2.39}$$

With these derivatives defined, the expressions for the required vectors can be obtained.

D.2.2.2 Derivatives of tvec1

For **tvec1**, consider just the x-component:

$$x_{\mathbf{tvec1}} = \frac{\Delta x_{2,3}}{\sqrt{\Psi_{2,3}}}, \tag{D.2.40}$$

$$\frac{\partial x_{\mathbf{tvec1}}}{\partial\theta^{\text{ind}}} = \frac{\frac{\partial\Delta x_{2,3}}{\partial\theta^{\text{ind}}}\sqrt{\Psi_{2,3}} - \Delta x_{2,3} \cdot \frac{\partial\sqrt{\Psi_{2,3}}}{\partial\theta^{\text{ind}}}}{\Psi_{2,3}}. \tag{D.2.41}$$

Since the denominator is shared amongst all x, y, and z-components, the corresponding y- and z-components are simply given as:

$$\frac{\partial y_{\mathbf{tvec1}}}{\partial\theta^{\text{ind}}} = \frac{\frac{\partial\Delta y_{2,3}}{\partial\theta^{\text{ind}}}\sqrt{\Psi_{2,3}} - \Delta y_{2,3} \cdot \frac{\partial\sqrt{\Psi_{2,3}}}{\partial\theta^{\text{ind}}}}{\Psi_{2,3}}, \tag{D.2.42}$$

$$\frac{\partial z_{\mathbf{tvec1}}}{\partial\theta^{\text{ind}}} = \frac{\frac{\partial\Delta z_{2,3}}{\partial\theta^{\text{ind}}}\sqrt{\Psi_{2,3}} - \Delta z_{2,3} \cdot \frac{\partial\sqrt{\Psi_{2,3}}}{\partial\theta^{\text{ind}}}}{\Psi_{2,3}}. \tag{D.2.43}$$

D.2.2.3 Derivatives of tcrp

For **tcrp**, consider just the x-component:

$$x_{\mathbf{tcrp}} = \frac{\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3}}{\sqrt{\Psi_{2,3\times2,1}}}. \tag{D.2.44}$$

Again, by applying chain rule we can get the following:

$$\begin{aligned}\frac{\partial x_{\mathbf{tcrp}}}{\partial\theta^{\text{ind}}} = &\frac{\left(\Delta z_{2,1}\frac{\partial\Delta y_{2,3}}{\partial\theta^{\text{ind}}} + \Delta y_{2,3}\frac{\partial\Delta z_{2,1}}{\partial\theta^{\text{ind}}} - \Delta z_{2,3}\frac{\partial\Delta y_{2,1}}{\partial\theta^{\text{ind}}} - \Delta y_{2,1}\frac{\partial\Delta z_{2,3}}{\partial\theta^{\text{ind}}}\right) \cdot \sqrt{\Psi_{2,3\times2,1}}}{\Psi_{2,3\times2,1}} \\ &- \frac{\left(\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3}\right) \cdot \frac{\partial\sqrt{\Psi_{(2,3\times2,1)}}}{\partial\theta^{\text{ind}}}}{\Psi_{2,3\times2,1}}.\end{aligned} \tag{D.2.45}$$

Analogously, the y- and z-components are defined as:

$$\begin{aligned}\frac{\partial y_{\mathbf{tcrp}}}{\partial\theta^{\text{ind}}} = &\frac{\left(\Delta x_{2,1}\frac{\partial\Delta z_{2,3}}{\partial\theta^{\text{ind}}} + \Delta z_{2,3}\frac{\partial\Delta x_{2,1}}{\partial\theta^{\text{ind}}} - \Delta x_{2,3}\frac{\partial\Delta z_{2,1}}{\partial\theta^{\text{ind}}} - \Delta z_{2,1}\frac{\partial\Delta x_{2,3}}{\partial\theta^{\text{ind}}}\right) \cdot \sqrt{\Psi_{2,3\times2,1}}}{\Psi_{2,3\times2,1}} \\ &- \frac{\left(\Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3}\right) \cdot \frac{\partial\sqrt{\Psi_{(2,3\times2,1)}}}{\partial\theta^{\text{ind}}}}{\Psi_{2,3\times2,1}},\end{aligned} \tag{D.2.46}$$

$$\begin{aligned}\frac{\partial z_{\mathbf{tcrp}}}{\partial \theta^{\mathrm{ind}}} = & \frac{\left(\Delta y_{2,1}\frac{\partial \Delta x_{2,3}}{\partial \theta^{\mathrm{ind}}} + \Delta x_{2,3}\frac{\partial \Delta y_{2,1}}{\partial \theta^{\mathrm{ind}}} - \Delta y_{2,3}\frac{\partial \Delta x_{2,1}}{\partial \theta^{\mathrm{ind}}} - \Delta x_{2,1}\frac{\partial \Delta y_{2,3}}{\partial \theta^{\mathrm{ind}}}\right)\cdot\sqrt{\Psi_{2,3\times 2,1}}}{\Psi_{2,3\times 2,1}} \\ & - \frac{\left(\Delta x_{2,3}\cdot\Delta y_{2,1} - \Delta x_{2,1}\cdot\Delta y_{2,3}\right)\cdot\frac{\partial\sqrt{\Psi_{(2,3\times 2,1)}}}{\partial\theta^{\mathrm{ind}}}}{\Psi_{2,3\times 2,1}}.\end{aligned} \tag{D.2.47}$$

D.2.2.4 Derivatives of ncrp

Finally, for **ncrp**, again consider the x-component:

$$x_{\mathbf{ncrp}} = \frac{\Delta z_{2,3}\cdot\left(\Delta z_{2,3}\cdot\Delta x_{2,1} - \Delta z_{2,1}\cdot\Delta x_{2,3}\right) - \Delta y_{2,3}\cdot\left(\Delta x_{2,3}\cdot\Delta y_{2,1} - \Delta x_{2,1}\cdot\Delta y_{2,3}\right)}{\sqrt{\Psi_{(2,3)\times(2,3\times 2,1)}}}. \tag{D.2.48}$$

The corresponding gradient term can be expressed as,

$$\begin{aligned}\frac{\partial x_{\mathbf{ncrp}}}{\partial \theta^{\mathrm{ind}}} = & \frac{\sqrt{\Psi_{(2,3)\times(2,3\times 2,1)}}}{\Psi_{(2,3)\times(2,3\times 2,1)}} \\ & \cdot\Bigg\{\frac{\partial \Delta z_{2,3}}{\partial\theta^{\mathrm{ind}}}\cdot\left(\Delta z_{2,3}\cdot\Delta x_{2,1} - \Delta z_{2,1}\cdot\Delta x_{2,3}\right) \\ & + \Delta z_{2,3}\cdot\left(\Delta x_{2,1}\frac{\partial\Delta z_{2,3}}{\partial\theta^{\mathrm{ind}}} + \Delta z_{2,3}\frac{\partial\Delta x_{2,1}}{\partial\theta^{\mathrm{ind}}} - \Delta x_{2,3}\frac{\partial\Delta z_{2,1}}{\partial\theta^{\mathrm{ind}}} - \Delta z_{2,1}\frac{\partial\Delta x_{2,3}}{\partial\theta^{\mathrm{ind}}}\right) \\ & - \frac{\partial\Delta y_{2,3}}{\partial\theta^{\mathrm{ind}}}\cdot\left(\Delta x_{2,3}\cdot\Delta y_{2,1} - \Delta x_{2,1}\cdot\Delta y_{2,3}\right) \\ & - \Delta y_{2,3}\cdot\left(\Delta y_{2,1}\frac{\partial\Delta x_{2,3}}{\partial\theta^{\mathrm{ind}}} + \Delta x_{2,3}\frac{\partial\Delta y_{2,1}}{\partial\theta^{\mathrm{ind}}} - \Delta y_{2,3}\frac{\partial\Delta x_{2,1}}{\partial\theta^{\mathrm{ind}}} - \Delta x_{2,1}\frac{\partial\Delta y_{2,3}}{\partial\theta^{\mathrm{ind}}}\right)\Bigg\} \\ & - \frac{\Delta z_{2,3}\cdot\left(\Delta z_{2,3}\cdot\Delta x_{2,1} - \Delta z_{2,1}\cdot\Delta x_{2,3}\right) - \Delta y_{2,3}\cdot\left(\Delta x_{2,3}\cdot\Delta y_{2,1} - \Delta x_{2,1}\cdot\Delta y_{2,3}\right)}{\Psi_{(2,3)\times(2,3\times 2,1)}} \\ & \cdot\frac{\partial\sqrt{\Psi_{(2,3)\times(2,3\times 2,1)}}}{\partial\theta^{\mathrm{ind}}}.\end{aligned} \tag{D.2.49}$$

And lastly, its corresponding y- and z-components:

$$
\begin{aligned}
\frac{\partial y_{\mathbf{ncrp}}}{\partial \theta^{\mathrm{ind}}} = & \frac{\sqrt{\Psi_{(2,3)\times(2,3\times2,1)}}}{\Psi_{(2,3)\times(2,3\times2,1)}} \\
& \cdot \Bigg\{ \frac{\partial \Delta x_{2,3}}{\partial \theta^{\mathrm{ind}}} \cdot (\Delta x_{2,3} \cdot \Delta y_{2,1} - \Delta x_{2,1} \cdot \Delta y_{2,3}) \\
& + \Delta x_{2,3} \cdot \left(\Delta y_{2,1} \frac{\partial \Delta x_{2,3}}{\partial \theta^{\mathrm{ind}}} + \Delta x_{2,3} \frac{\partial \Delta y_{2,1}}{\partial \theta^{\mathrm{ind}}} - \Delta y_{2,3} \frac{\partial \Delta x_{2,1}}{\partial \theta^{\mathrm{ind}}} - \Delta x_{2,1} \frac{\partial \Delta y_{2,3}}{\partial \theta^{\mathrm{ind}}} \right) \\
& - \frac{\partial \Delta z_{2,3}}{\partial \theta^{\mathrm{ind}}} \cdot (\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3}) \\
& - \Delta z_{2,3} \cdot \left(\Delta z_{2,1} \frac{\partial \Delta y_{2,3}}{\partial \theta^{\mathrm{ind}}} + \Delta y_{2,3} \frac{\partial \Delta z_{2,1}}{\partial \theta^{\mathrm{ind}}} - \Delta z_{2,3} \frac{\partial \Delta y_{2,1}}{\partial \theta^{\mathrm{ind}}} - \Delta y_{2,1} \frac{\partial \Delta z_{2,3}}{\partial \theta^{\mathrm{ind}}} \right) \Bigg\} \\
& - \frac{\Delta x_{2,3} \cdot (\Delta x_{2,3} \cdot \Delta y_{2,1} - \Delta x_{2,1} \cdot \Delta y_{2,3}) - \Delta z_{2,3} \cdot (\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3})}{\Psi_{(2,3)\times(2,3\times2,1)}} \\
& \cdot \frac{\partial \sqrt{\Psi_{(2,3)\times(2,3\times2,1)}}}{\partial \theta^{\mathrm{ind}}},
\end{aligned}
\tag{D.2.50}
$$

$$
\begin{aligned}
\frac{\partial z_{\mathbf{ncrp}}}{\partial \theta^{\mathrm{ind}}} = & \frac{\sqrt{\Psi_{(2,3)\times(2,3\times2,1)}}}{\Psi_{(2,3)\times(2,3\times2,1)}} \\
& \cdot \Bigg\{ \frac{\partial \Delta y_{2,3}}{\partial \theta^{\mathrm{ind}}} \cdot (\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3}) \\
& + \Delta y_{2,3} \cdot \left(\Delta z_{2,1} \frac{\partial \Delta y_{2,3}}{\partial \theta^{\mathrm{ind}}} + \Delta y_{2,3} \frac{\partial \Delta z_{2,1}}{\partial \theta^{\mathrm{ind}}} - \Delta z_{2,3} \frac{\partial \Delta y_{2,1}}{\partial \theta^{\mathrm{ind}}} - \Delta y_{2,1} \frac{\partial \Delta z_{2,3}}{\partial \theta^{\mathrm{ind}}} \right) \\
& - \frac{\partial \Delta x_{2,3}}{\partial \theta^{\mathrm{ind}}} \cdot (\Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3}) \\
& - \Delta x_{2,3} \cdot \left(\Delta x_{2,1} \frac{\partial \Delta z_{2,3}}{\partial \theta^{\mathrm{ind}}} + \Delta z_{2,3} \frac{\partial \Delta x_{2,1}}{\partial \theta^{\mathrm{ind}}} - \Delta x_{2,3} \frac{\partial \Delta z_{2,1}}{\partial \theta^{\mathrm{ind}}} - \Delta z_{2,1} \frac{\partial \Delta x_{2,3}}{\partial \theta^{\mathrm{ind}}} \right) \Bigg\} \\
& - \frac{\Delta y_{2,3} \cdot (\Delta y_{2,3} \cdot \Delta z_{2,1} - \Delta y_{2,1} \cdot \Delta z_{2,3}) - \Delta x_{2,3} \cdot (\Delta z_{2,3} \cdot \Delta x_{2,1} - \Delta z_{2,1} \cdot \Delta x_{2,3})}{\Psi_{(2,3)\times(2,3\times2,1)}} \\
& \cdot \frac{\partial \sqrt{\Psi_{(2,3)\times(2,3\times2,1)}}}{\partial \theta^{\mathrm{ind}}}.
\end{aligned}
\tag{D.2.51}
$$

Hence with all of these vector components defined, Equation D.2.33 can be evaluated in terms of known dof quantities. This derivation, as shown above, has been implemented into CrystalPredictor V2.5.2 onwards (within build_molecule.f90).

Appendix E

CP2 Shift Algorithm

E.1 Background

In CrystalPredictor II (CP2), Global Search starting structures are generated by a quasi-random Sobol' sequence. Whilst this is generally successful, it has been observed that for unit cells with high Z' and/or large molecules, the initial decision variable vector, $\boldsymbol{x}$, often results in physically-unfeasible candidates with unsound intermolecular separation (termed 'molecular clashing' here). By default, CP2 decides that such points are 'not worth the effort' to minimize, discards them, and proceeds to generate a new $\boldsymbol{x}$. In extreme cases, this brute force strategy can be highly inefficient: long durations are spent generating invalid $\boldsymbol{x}$ candidates before a valid minimization starting point is found. Examples of this include Molecule XXX from the 7th Blind Test (CANNOL) [49].

This issue motivated the implementation of the Shift Algorithm into CPV2.5. Zhang [71] demonstrated an early implementation of this algorithm, but that implementation was incomplete due to the neglect of space group constraints on the unit cell lengths.

E.2 The Simplest Case: Cubic Shift Algorithm

E.2.1 A Brief Intuition

The problem at hand is best described visually. Consider a 2-dimensional unit cell with cell angles fixed at 90°. The cell contains two molecules (A and B) that are clashing (Figure E.1). Note that clashing does not necessitate that the molecules are overlapping, but rather just implies that they are in unreasonable proximity (r_{lim}). For simplicity, I assume that these hypothetical molecules contain only two atoms each, of which their second atoms are clashing. The Shift Algorithm seeks to separate the two clashing sites by expanding the unit cell through scaling of the cell lengths. Because the *fractional coordinates* of each molecule's first atom are fixed throughout this process, by expanding the cell dimensions the two molecules are consequently moved apart. The appeal of this strategy is best explained by considering an alternative: directly translating/rotating molecules within a unit cell of fixed size.

In this alternative strategy, it is non-trivial to determine the best transformation that: not only separates the clashing sites, but also does not introduce clashing elsewhere. This is es-

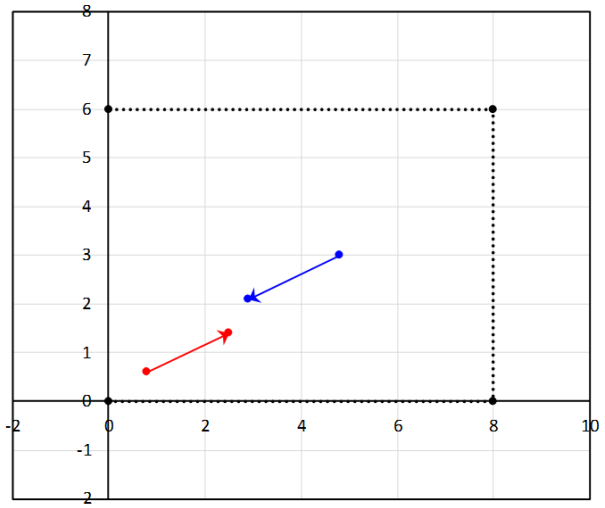


Figure E.1: A simplified representation of the clashing molecule problem. The red and blue structure represent distinct molecules in some arbitrary unit cell. The arrow points from the first-atom of each molecule to their respective 'clash sites'.

pecially so given the periodicity of the crystal: moving a molecule away in one direction risks introducing new clashes not only in the original cell, but also in its neighboring cells. Moreover, the predominant factor controlling the occurrence of clashing is the unit cell. If the unit cell is too small, no amount of shifting molecules around can provide a reasonable separation. Expanding the unit cell bypasses many of these challenges since it is intuitive that if the available space increases, the average separation of molecules in the entire (periodic) system can only increase. As I will outline below, this strategy is a well-defined problem that can be solved analytically.

E.2.2 The Underpinning Mathematics

At the juncture where molecular clashing is assessed, the following information must be available for the two clashing molecules:

1. The (space-fixed) positions of their first atoms ($\boldsymbol{r}_{1,j}$)
2. The (space-fixed) positions of their 'clash sites' ($\boldsymbol{r}_{\mathrm{c},j}$)
3. The vector from atom 1 to their 'clash sites' ($\boldsymbol{r}_{1\to\mathrm{c},j}$)

where j is an index for the clashing molecules.

In practice, the decision variable vector, $\boldsymbol{x}$, defines the space-fixed COM position of each molecule. This COM effectively acts as the 'anchor' for the rest of the molecule. Throughout this derivation I will treat the first atom of each molecule synonymously with the molecule's COM (i.e., the first atom anchors the molecular position in the cell). This substitution has no bearing on the validity of the derivations, and the derived expressions are equally valid when the COM position is used instead. Combining $\boldsymbol{r}_{1,j}$, with the molecular conformation and Euler angles, the space-fixed positions of all other atoms in the molecule are also defined. Assessment

of interatomic distances in the unit cell allows the clash sites to be identified and the second value can be found, The third variable is trivial to obtain from

$$\boldsymbol{r}_{1\to\mathrm{c},j} = \boldsymbol{r}_{\mathrm{c},j} - \boldsymbol{r}_{1,j}. \tag{E.2.1}$$

With this, the key variable of interest, the intersite distance between the clash sites ($|\boldsymbol{\phi}|$) of molecules A and B, can also be defined:

$$\boldsymbol{\phi} = \boldsymbol{r}_{\mathrm{c},A} - \boldsymbol{r}_{\mathrm{c},B}, \tag{E.2.2}$$

$$|\boldsymbol{\phi}| = |\boldsymbol{r}_{\mathrm{c},A} - \boldsymbol{r}_{\mathrm{c},B}|. \tag{E.2.3}$$

Because the first atoms of each molecule are anchored at their fractional coordinates, expanding the cell will translate the first atom and this then propagates to separation of the clash sites. Note that, $\boldsymbol{r}_{1\to\mathrm{c},j}$ remains unchanged since the conformation and Euler angles are fixed throughout. This is given by:

$$\boldsymbol{r}_{\mathrm{c},j} = \boldsymbol{r}_{1,j} + \boldsymbol{r}_{1\to\mathrm{c},j}, \tag{E.2.4}$$

$$\boldsymbol{r}_{\mathrm{c},j} = \boldsymbol{R}\hat{\boldsymbol{r}}_{1,j} + \boldsymbol{r}_{1\to\mathrm{c},j}, \tag{E.2.5}$$

where $\boldsymbol{R}$ is the unit cell matrix and $\hat{\boldsymbol{r}}_{1,j}$ is the fractional coordinates of atom 1 in molecule j. For now, only a cubic cell is considered and the distance equation can be re-expressed according to

$$|\boldsymbol{\phi}| = |\boldsymbol{R}\hat{\boldsymbol{r}}_{1,A} + \boldsymbol{r}_{1\to\mathrm{c},A} - (\boldsymbol{R}\hat{\boldsymbol{r}}_{1,B} + \boldsymbol{r}_{1\to\mathrm{c},B})|, \tag{E.2.6}$$

$$|\boldsymbol{\phi}| = \left| \begin{pmatrix} a\hat{x}_{1,A} \\ b\hat{y}_{1,A} \\ c\hat{z}_{1,A} \end{pmatrix} + \begin{pmatrix} x_{1\to\mathrm{c},A} \\ y_{1\to\mathrm{c},A} \\ z_{1\to\mathrm{c},A} \end{pmatrix} - \begin{pmatrix} a\hat{x}_{1,B} \\ b\hat{y}_{1,B} \\ c\hat{z}_{1,B} \end{pmatrix} - \begin{pmatrix} x_{1\to\mathrm{c},B} \\ y_{1\to\mathrm{c},B} \\ z_{1\to\mathrm{c},B} \end{pmatrix} \right|. \tag{E.2.7}$$

For brevity, I define a Δ operator which indicates the deduction of B from A (i.e., $A - B$), such that

$$\boldsymbol{\phi} = \begin{pmatrix} a\Delta\hat{x}_1 \\ b\Delta\hat{y}_1 \\ c\Delta\hat{z}_1 \end{pmatrix} + \begin{pmatrix} \Delta x_{1\to\mathrm{c}} \\ \Delta y_{1\to\mathrm{c}} \\ \Delta z_{1\to\mathrm{c}} \end{pmatrix}, \tag{E.2.8}$$

$$|\phi| = \left| \begin{pmatrix} a\Delta\hat{x}_1 \\ b\Delta\hat{y}_1 \\ c\Delta\hat{z}_1 \end{pmatrix} + \begin{pmatrix} \Delta x_{1\to c} \\ \Delta y_{1\to c} \\ \Delta z_{1\to c} \end{pmatrix} \right|, \tag{E.2.9}$$

$$|\phi| = \sqrt{(\Delta x_{1\to c} + a\Delta\hat{x}_1)^2 + (\Delta y_{1\to c} + b\Delta\hat{y}_1)^2 + (\Delta z_{1\to c} + c\Delta\hat{z}_1)^2}. \tag{E.2.10}$$

Consider now that an isotropic expansion is applied to the cell, whereby every cell length is multiplied by a factor ε. The vector between the clash sites after expansion ($\mathbf{\Phi}$) and its corresponding distance are given by,

$$\mathbf{\Phi} = \begin{pmatrix} \varepsilon a\Delta\hat{x}_1 \\ \varepsilon b\Delta\hat{y}_1 \\ \varepsilon c\Delta\hat{z}_1 \end{pmatrix} + \begin{pmatrix} \Delta x_{1\to c} \\ \Delta y_{1\to c} \\ \Delta z_{1\to c} \end{pmatrix}, \tag{E.2.11}$$

$$|\mathbf{\Phi}| = \left| \begin{pmatrix} \varepsilon a\Delta\hat{x}_1 \\ \varepsilon b\Delta\hat{y}_1 \\ \varepsilon c\Delta\hat{z}_1 \end{pmatrix} + \begin{pmatrix} \Delta x_{1\to c} \\ \Delta y_{1\to c} \\ \Delta z_{1\to c} \end{pmatrix} \right|, \tag{E.2.12}$$

$$|\mathbf{\Phi}| = \sqrt{(\Delta x_{1\to c} + \varepsilon a\Delta\hat{x}_1)^2 + (\Delta y_{1\to c} + \varepsilon b\Delta\hat{y}_1)^2 + (\Delta z_{1\to c} + \varepsilon c\Delta\hat{z}_1)^2}. \tag{E.2.13}$$

It should be apparent that if the expanded distance ($|\mathbf{\Phi}|$) is equated to a minimum tolerable distance (r_{tol}), the problem is a fully defined system of one unknown (ε) and one equation (Equation E.2.13). Taking the square of Equation E.2.13 gives a quadratic equation in terms of ε which is trivial to solve. This isotropic expansion is the most basic implementation of the Shift Algorithm.

E.2.3 Anisotropic Expansion

Based on the analytical derivation, the isotropic expansion does satisfy our intentions. However in practice, the fact that it offers only one degree of freedom limits its usefulness. Intuitively, if the expansion pulls the clash sites away from each other *along* the direction of the vector connecting the two sites, this should represent the fastest/most-efficient mode of separation. The isotropic expansion fails to do this and hence is liable to over-expanding the unit cell. The output of the Shift Algorithm is subsequently subjected to a local minimization and if the unit cell is expanded excessively, the minimization will inadvertently take more time. Attaining this most-efficient expansion is thus in the best interest of the overall Global Search efficiency.

For this reason, we instead consider the possibility of independent expansion factors for each cell dimension,

$$\boldsymbol{\Phi} = \begin{pmatrix} \varepsilon_x a\Delta\hat{x}_1 \\ \varepsilon_y b\Delta\hat{y}_1 \\ \varepsilon_z c\Delta\hat{z}_1 \end{pmatrix} + \begin{pmatrix} \Delta x_{1\to c} \\ \Delta y_{1\to c} \\ \Delta z_{1\to c} \end{pmatrix}, \tag{E.2.14}$$

$$|\boldsymbol{\Phi}| = \left| \begin{pmatrix} \varepsilon_x a\Delta\hat{x}_1 \\ \varepsilon_y b\Delta\hat{y}_1 \\ \varepsilon_z c\Delta\hat{z}_1 \end{pmatrix} + \begin{pmatrix} \Delta x_{1\to c} \\ \Delta y_{1\to c} \\ \Delta z_{1\to c} \end{pmatrix} \right|, \tag{E.2.15}$$

$$|\boldsymbol{\Phi}| = \sqrt{(\Delta x_{1\to c} + \varepsilon_x a\Delta\hat{x}_1)^2 + (\Delta y_{1\to c} + \varepsilon_y b\Delta\hat{y}_1)^2 + (\Delta z_{1\to c} + \varepsilon_z c\Delta\hat{z}_1)^2}, \tag{E.2.16}$$

where $\varepsilon_{x,y,z}$ are the respective expansion factors in each dimension. With this, we now have three unknowns ($\varepsilon_{x,y,z}$) but only one equation (Equation E.2.16). As stated before, to minimize the overall cell expansion, one would prefer to separate the clash sites along the direction of their connecting vector, as given by Equations E.2.2 and E.2.8. To enforce this, the *unit vectors* connecting the clash sites should be the same before and after the expansion, as given by,

$$\frac{\boldsymbol{\phi}}{|\boldsymbol{\phi}|} = \frac{\boldsymbol{\Phi}}{|\boldsymbol{\Phi}|}. \tag{E.2.17}$$

This actually simplifies the implementation considerable since three equations are obtained without ever having to explicitly evaluate Equation E.2.16:

$$\frac{|\boldsymbol{\Phi}|}{|\boldsymbol{\phi}|}(\Delta x_{1\to c} + a\Delta\hat{x}_1) = \Delta x_{1\to c} + \varepsilon_x a\Delta\hat{x}_1, \tag{E.2.18}$$

$$\frac{|\boldsymbol{\Phi}|}{|\boldsymbol{\phi}|}(\Delta y_{1\to c} + b\Delta\hat{y}_1) = \Delta y_{1\to c} + \varepsilon_y b\Delta\hat{y}_1, \tag{E.2.19}$$

$$\frac{|\boldsymbol{\Phi}|}{|\boldsymbol{\phi}|}(\Delta z_{1\to c} + c\Delta\hat{z}_1) = \Delta z_{1\to c} + \varepsilon_z c\Delta\hat{z}_1. \tag{E.2.20}$$

Once again, if $|\boldsymbol{\Phi}|$ is set to the target value, r_{tol}, then there are exactly three (linear) equations and three unknowns which can be trivially solved for.

E.2.4 Just a System of Linear Equations?

At first glance, it seems the problem has been simplified to solving a set of three linear equations. However, consideration of some 'non-ideal' situations will quickly reveal that this is incorrect. In short, is the solution of this trio of linear equations (Equations E.2.18–E.2.20) synonymous with a solution involving Equation E.2.16?

E.2.4.1 Coincident First Atoms

First, consider if the first atoms of the two molecules are exactly coincident (i.e., $\Delta\hat{r}_1 = 0$ for $r = x, y, z$). In this scenario, both the linear equations as well as the quadratic equation point to the same outcome: no solution of ε_r can be determined since the two molecules are anchored at the same point and expansion of the cell has no effect on the intersite distance:

$$|\mathbf{\Phi}| = \sqrt{(\Delta x_{1\to c})^2 + (\Delta y_{1\to c})^2 + (\Delta z_{1\to c})^2}, \tag{E.2.21}$$

$$\frac{|\mathbf{\Phi}|}{|\boldsymbol{\phi}|}(\Delta x_{1\to c}) = \Delta x_{1\to c}, \tag{E.2.22}$$

$$\frac{|\mathbf{\Phi}|}{|\boldsymbol{\phi}|}(\Delta y_{1\to c}) = \Delta y_{1\to c}, \tag{E.2.23}$$

$$\frac{|\mathbf{\Phi}|}{|\boldsymbol{\phi}|}(\Delta z_{1\to c}) = \Delta z_{1\to c}. \tag{E.2.24}$$

From this, it can established that, **Rule 1:** if clashing molecules have coincident first atoms ($\Delta\hat{\boldsymbol{r}}_1 = \mathbf{0}$), abandon the algorithm.

E.2.4.2 Coincident Clash Sites

Next, considering the opposing extreme whereby the clash sites are coincident (i.e., $\boldsymbol{\phi} = \mathbf{0}$ and $|\boldsymbol{\phi}| = 0$). Under such circumstances, the set of linear equations are eliminated. An interpretation of this is that because the clash sites are overlapping, there is no 'most efficient' direction to expand in and as such the linear equations are meaningless. However, this does not indicate an expansion cannot separate the sites, since Equation E.2.16 is not invalidated in any way. Unfortunately, given that only one valid equation remains, there is no well-defined solution for three expansion factors, and the only option is to settle for the isotropic expansion with a uniform expansion factor. In short, **Rule 2:** if the two clash sites are coincident ($|\boldsymbol{\phi}| = \mathbf{0}$), use the isotropic expansion (Equation E.2.13).

E.2.4.3 Co-linear First Atoms

It is in the intermediate cases, where the first atoms/clash sites are co-linear in only some dimensions (but not completely coincident) that things become more complicated. Consider the case where it is the first atom positions which are co-linear. If only $\Delta\hat{x}_1 = 0$, the post-expansion intersite distance is now given by

$$|\mathbf{\Phi}| = \sqrt{(\Delta x_{1\to c})^2 + (\Delta y_{1\to c} + \varepsilon_y b\Delta\hat{y}_1)^2 + (\Delta z_{1\to c} + \varepsilon_z c\Delta\hat{z}_1)^2}, \tag{E.2.25}$$

which is independent of ε_x. It follows that the linear equation describing x (Equation E.2.22) is no longer 'useful'. There is no reason to contract a as it may cause more clashing, and conversely there is also no reason to expand this dimension to avoid increased minimization time. Hence,

the logical strategy is to set $\varepsilon_x = 1$. This logic extends naturally to the case where the first atoms are co-linear in two dimensions, and Rule 1 follows in the extreme case of all dimensions being co-linear. As such, I suggest **Rule 3:** if the molecules' first atoms are co-linear in a direction ($\Delta\hat{r}_1 = 0$), constrain the corresponding expansion factor to 1.

Returning to the hypothetical case where only the x-dimension is eliminated. This leaves two unknowns (ε_y and ε_z) but *three* equations to choose from (Equations E.2.25, E.2.19, and E.2.20). In the ideal case, where there is no co-linearity, substituting the linear expressions into the quadratic one gives

$$|\boldsymbol{\Phi}| = \sqrt{\left[\frac{|\boldsymbol{\Phi}|}{|\phi|}(\Delta x_{1\to c} + a\Delta\hat{x}_1)\right]^2 + \left[\frac{|\boldsymbol{\Phi}|}{|\phi|}(\Delta y_{1\to c} + b\Delta\hat{y}_1)\right]^2 + \left[\frac{|\boldsymbol{\Phi}|}{|\phi|}(\Delta z_{1\to c} + c\Delta\hat{z}_1)\right]^2} \tag{E.2.26}$$

$$= \frac{|\boldsymbol{\Phi}|}{|\phi|} \cdot \sqrt{(\Delta x_{1\to c} + \varepsilon_x a\Delta\hat{x}_1)^2 + (\Delta y_{1\to c} + \varepsilon_y b\Delta\hat{y}_1)^2 + (\Delta z_{1\to c} + \varepsilon_z c\Delta\hat{z}_1)^2}.$$

This can be easily rearranged to obtain a valid expression for the un-expanded clash distance $|\phi|$ proving that the solution of *all three* linear equations is synonymous to Equation E.2.16. However, the same is untrue in the non-ideal case with co-linearity:

$$|\boldsymbol{\Phi}| = \sqrt{(\Delta x_{1\to c})^2 + \left[\frac{|\boldsymbol{\Phi}|}{|\phi|}(\Delta y_{1\to c} + b\Delta\hat{y}_1)\right]^2 + \left[\frac{|\boldsymbol{\Phi}|}{|\phi|}(\Delta z_{1\to c} + c\Delta\hat{z}_1)\right]^2}, \tag{E.2.27}$$

which if solved would erroneously indicate that $\Delta x_{1\to c}$ has changed. In other words, choosing to solve the two linear equations *does not* necessarily satisfy the distance criteria for a fixed $\Delta x_{1\to c}$. In this scenario (and, by extension, the scenario whereby two $\Delta\hat{r}_1 = 0$), a strategy centered on solving only the linear equations is inadequate for achieving $|\boldsymbol{\Phi}| = r_{tol}$. Thusly, **Rule 4:** if the molecules' first atoms are co-linear in either one or two dimensions, the linear equations will not give the correct separation distance and a different analytical solution must be used. If all three dimensions are co-linear, then Rule 1 takes precedence.

E.2.4.4 Co-linear Clash Sites

I now consider the other intermediate case whereby the clash sites are co-linear in only some dimensions (i.e., $\Delta r_{1\to c} + \boldsymbol{R}\Delta\hat{r}_1 = 0$). Once again, I illustrate with the x-dimension:

$$\Delta x_{1\to c} + a\Delta\hat{x}_1 = 0, \tag{E.2.28}$$

from which it follows that Equation E.2.18 becomes:

$$\Delta x_{1\to c} + \varepsilon_x a\Delta\hat{x}_1 = 0. \tag{E.2.29}$$

Solving these equations simultaneously gives the trivial solution $\varepsilon_x = 1$. This is wholly congruent with the intuition gained from the earlier case where the clash sites were coincident: any co-linear dimension provides no information about the most-efficient expansion direction, i.e., **Rule 5:** If

the molecules' clash sites are co-linear in either one or two dimensions, constrain their expansion factors to 1 (this arises naturally from solving the linear equation). If all three dimensions are co-linear (i.e., coincident clash sites), then Rule 2 takes precedence.

Again, we must consider how the remaining ε_r may be determined. In this case, the distance equation reduces to

$$\begin{aligned} |\mathbf{\Phi}| &= \sqrt{(\Delta x_{1\to c} + \varepsilon_x a\Delta\hat{x}_1)^2 + (\Delta y_{1\to c} + \varepsilon_y b\Delta\hat{y}_1)^2 + (\Delta z_{1\to c} + \varepsilon_z c\Delta\hat{z}_1)^2} \\ &= \sqrt{(\Delta y_{1\to c} + \varepsilon_y b\Delta\hat{y}_1)^2 + (\Delta z_{1\to c} + \varepsilon_z c\Delta\hat{z}_1)^2}, \end{aligned} \tag{E.2.30}$$

where substituting the linear equations (Equation E.2.19 and E.2.20) gives

$$|\mathbf{\Phi}| = \sqrt{\left[\frac{|\mathbf{\Phi}|}{|\phi|}(\Delta y_{1\to c} + b\Delta\hat{y}_1)\right]^2 + \left[\frac{|\mathbf{\Phi}|}{|\phi|}(\Delta z_{1\to c} + c\Delta\hat{z}_1)\right]^2}. \tag{E.2.31}$$

Unlike the case involving co-linear first atoms, the linear equations continue to be congruent with the distance equation! It can be seen that this is true even if two dimensions are co-linear. In other words, if the molecules' clash sites are co-linear in either one or two dimensions, the remaining expansion factors can be solved as the solution of the remaining linear equations. If all three dimensions are co-linear (i.e., coincident clash sites), then Rule 2 takes precedence.

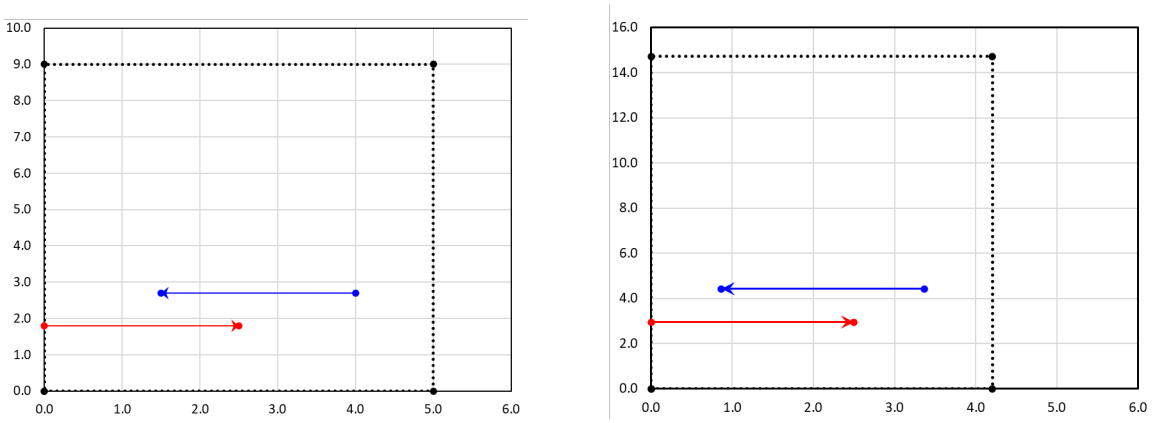
E.2.4.5 Non-expansions

The final class of non-idealities that must be considered are non-expansions. These encompasses cases where the solution to the equations indicate inversions of the cell ($\varepsilon_r < 0$) or contractions of the cell ($0 < \varepsilon_r < 1$). Rearranging Equation E.2.18:

$$\varepsilon_x = \frac{\frac{|\mathbf{\Phi}|}{|\phi|}(\Delta x_{1\to c} + a\Delta\hat{x}_1) - \Delta x_{1\to c}}{a\Delta\hat{x}_1}, \tag{E.2.32}$$

it is clear that as long as $\Delta\hat{x}_1 \neq 0$, there is no physical restrictions that prevent the occurrence of non-expansionary results.

An example of this is shown in Figure E.2. In this case, $\Delta x_{1\to c}$ and $\Delta\hat{x}_1$ have opposing signs (regardless of whether molecule A is defined as red or blue, and *vice versa*). Solving this with the linear equations correctly obtains the desired target distance of 2.20 units and the vector between the two clash sites is kept co-linear, but a contraction is applied in the x-direction ($\varepsilon_x = 0.841$). Intuitively, one can understand that the algorithm has decided the fastest way to separate the two sites is to 'slide' them past each other in the x-direction. Although not illustrated, the occurrence of inversion solutions also makes sense since inverting to a large enough cell could also separate the clash sites to the desired distance. Ultimately, the algorithm is working as designed, but a more sophisticated approach is required to deal with these undesirable results.



(a) (b)

Figure E.2: An example in 2-dimensions where solving the linear equations results in a contracting solution. (a) The initial clash site distance (distance between arrow-heads) is 1.35 units. (b) The final distance reaches the target of 2.20 units.

E.3 Quadratic Solutions with Deformed (Non-Cubic) Cells

As mentioned before, I seek a solution which is purely expansionary ($\varepsilon_r \geq 1$) since this guarantees that the clash sites can only get further apart. In the event that a contraction/inversion solution is obtained, the problematic dimension is ideally omitted from the solution ($\varepsilon_r = 1$) and the expansion is solved for only the remaining dimensions. Unfortunately, as shown by the co-linear first atom case, whenever an artificial value for ε_r is imposed, the relationship between the linear equations and the quadratic equation is invalidated. This also foreshadows complications down the line if/when symmetry constraints are imposed. The strategies for the two extreme cases is unaffected by these issues. Nevertheless, it would be prudent to to obtain a more general solution which explicitly upholds the distance equation. Moreover, while the cubic system was useful to gain intuition about the problem, I must generalize the problem to permit cell deformation. In a cell with lengths a, b, and c and with cell angles α, β, and γ, the relationship between the space-fixed coordinates ($\boldsymbol{r}$) and a set of fractional coordinates ($\hat{\boldsymbol{r}}$) is given by

$$\boldsymbol{r} = \boldsymbol{R}\hat{\boldsymbol{r}}, \tag{E.3.1}$$

where $\boldsymbol{R}$ is the unit cell matrix defined as:

$$\boldsymbol{R} = \begin{pmatrix} a & b\cos(\gamma) & c\cos(\beta) \\ 0 & b\sin(\gamma) & c\mu \\ 0 & 0 & c\omega \end{pmatrix}, \tag{E.3.2}$$

$$\mu = \frac{\cos(\alpha) - \cos(\beta)\cos(\gamma)}{\sin(\gamma)}, \tag{E.3.3}$$

$$\omega = \frac{\sqrt{W}}{\sin(\gamma)}, \tag{E.3.4}$$

$$W = 1 - \left[\cos^2(\alpha) + \cos^2(\beta) + \cos^2(\gamma)\right] + 2 \cdot \left[\cos(\alpha) \cdot \cos(\beta) \cdot \cos(\gamma)\right]. \tag{E.3.5}$$

From Equation E.2.6, it is apparent that this modification only affects the first atom position and the clash site vector before and after the expansion may be redefined as:

$$\boldsymbol{\phi}_{\text{deform}} = \begin{pmatrix} a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 \\ b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 \\ c\omega\Delta\hat{z}_1 \end{pmatrix} + \begin{pmatrix} \Delta x_{1\to c} \\ \Delta y_{1\to c} \\ \Delta z_{1\to c} \end{pmatrix} = \begin{pmatrix} p \\ q \\ r \end{pmatrix}, \tag{E.3.6}$$

and

$$\boldsymbol{\Phi}_{\text{deform}} = \begin{pmatrix} \varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 \\ \varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 \\ \varepsilon_z c\omega\Delta\hat{z}_1 \end{pmatrix} + \begin{pmatrix} \Delta x_{1\to c} \\ \Delta y_{1\to c} \\ \Delta z_{1\to c} \end{pmatrix}, \tag{E.3.7}$$

respectively. In practice, we are only interested in the total values of each element in $\boldsymbol{\phi}_{\text{deform}}$, hence for brevity I define the variables p, q, and r.

To make a square system for three variables out of four equations (1 distance equation and 3 linear equations), I first express the distance equations as

$$\begin{aligned} |\boldsymbol{\Phi}_{\text{deform}}|^2 = {} & (\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\ & + (\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\ & + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2. \end{aligned} \tag{E.3.8}$$

To solve this in conjunction with the direction information, I re-express the linear equations as ratios instead:

$$\frac{p}{q} = \frac{\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}}{\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}}, \tag{E.3.9}$$

$$\frac{p}{r} = \frac{\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}}{\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}}, \tag{E.3.10}$$

$$\frac{q}{r} = \frac{\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}}{\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}}. \tag{E.3.11}$$

Because they are taken as ratios of each other, only two out of three are linearly independent and we must select any two out of the three to solve in conjunction with Equation E.3.8. This will form the basis of all solutions going forward.

Before proceeding further to derive the analytical solutions of these equations, let us first consider if the rules established earlier still hold true.

- **Rule 1:** The working principle of the expansions has not changed and if all the $\Delta\hat{r}_1 = 0$, then can be no valid expansion isotropic/anisotropic that can lead to separation. As such, it is still true that if $\Delta\hat{\boldsymbol{r}}_1 = \mathbf{0}$, the algorithm can be terminated.

- **Rule 2:** If the clash sites are coincidental, this implies $p = q = r = 0$. As before, no directional information can be derived from Equations E.3.9–E.3.11. Yet, the expansion factors are not eliminated from the distance equation (Equation E.3.8), hence with only one equation available to solve, we can only execute an isotropic expansion.

- **Rule 3:** As per Rule 1, the effects of any ε_r are eliminated from Equation E.3.8 if their corresponding $\Delta\hat{r}_1 = 0$, hence it is still true that we should constrain $\varepsilon_r = 1$ in this case.

- **Rule 4:** This rule directly highlights the limitations of the linear-equation solution and directly motivates a quadratic solution that explicitly uses the distance equation.

- **Rule 5:** The issue that arises with this rule stems from the introduction of deformation into the unit cell. In the cubic cell, it follows exactly that if any of the clash site directional vector elements are 0 (i.e., p, q, or $r = 0$), when adhering to the principle of maintaining the proportionality of the original inter-clash site vector, its corresponding ε_r can only take a value of 1.

  This is no longer the case in the deformed cell, since vector elements can have dependencies on more than one ε_r (except for the z-vector element) and careful consideration of each case must be exercised. Closer analysis of the problem offers some guidance on the strategy of using the distance equation (Equation E.3.8) and the ratio/directional equations (Equations E.3.9–E.3.11):

  (a) Firstly, it must be established that for any case, the distance criteria (Equation E.3.8) must be solved. This is congruent with Rule 1, since there must be at least one available expansion factor not constrained to 1.

  (b) Of the three ratio equations, a maximum of two independent equations can be chosen. As such, only a maximum of three variables ($\varepsilon_{x,y,z}$) can be solved for (from 1 distance equation + 2 ratios). If $p, q, r \neq 0$, then this is the case.

  (c) If $p, q, r = 0$, then Rule 2 takes precedence and only one expansion factor can be solved for (mathematically, none of the ratios are defined, so only distance can be solved for).

  (d) If no more than two p, q, and/or r are zero, then their associated ratios will reduce to zero as well. Importantly, this does not reduce the number of variables we can solve for since a zero-valued ratio still give valid information about which direction *we should not expand* (i.e., the corresponding element in $\boldsymbol{\Phi}$ just remains at zero after expansion).

In short, with exception of the case where all three p, q, and r are zero, there is always the possibility to solve for three expansion factors. Extending this analysis to the cases where we have constraints on the various expansion factors:

(i) Constraints act as 'external' equations. Hence the inclusion of each constraint must come at the loss of one of the 'physical' equations.

(ii) The logical limit to this is thus having two constrained ε_r. In this case, only one equation can be utilized, which must be the distance equation (Equation E.3.8).

(iii) The most complex case is if we only have one constrained ε_r. As before, Equation E.3.8 must be used in the solution, but a choice must be made between the remaining ratio equations. Choosing either of the remaining ratios can offer physically valid solutions (with account for contractions/inversions). In this case, *two pairs of solutions* (each pair representing the quadratic solution of the distance equation applied in conjunction with one of the ratios) can be solved for.

With this, I now progress to determining the analytical solutions for this problem. Unfortunately, these solutions must be done on a case-by-case basis. From the earlier analysis, I can categorize these cases on the following factors:

1. Which of the clash site vector elements are zero (i.e., if p, q, or $r = 0$)?
2. Which of the ε_r have been constrained to 1 ?
3. Which ε_r have inter-dependencies due to space group symmetry?

A change to any of these factors will lead to a slightly different problem formulation, largely due to how the ratio equations need to defined. For now, I ignore the last factor. In relation to the first factor, these elements are determined at the start, based on the un-expanded cell. With regards to the second factor, as per Rule 3, ε_r may be set to 1 if $\Delta\hat{r}_1 = 0$. However, this is not the only occasion where this might occur. As alluded to earlier, when met with contraction/inversion solutions, one would prefer to manually constrain the corresponding dimension (e.g., if $\varepsilon_x = 0.8$, fix it to 1) and re-solve for the other expansion factors. With these in mind, I simply consider the general case where ε_r is constrained to 1 without any consideration of the reason for doing this ($\Delta\hat{r}_1 = 0$, or otherwise). In all cases, because we are solving a quadratic equation, a pair (or two pairs as per Rule 5-iii) of solutions are obtained. The philosophy to select between solutions will be discussed later. Here, I focus on the derivations of the analytical solutions. These are summarize in Table E.1.

E.3.1 All ε_r Constrained to 1

As per Rule 1, no further action can be taken. Exit the algorithm.

E.3.2 $p = q = r = 0$ & No ε_r Constrained to 1

As per Rule 2, none of the ratio equations can be used and only isotropic expansion can be done. Hence set $\varepsilon_x = \varepsilon_y = \varepsilon_z = \varepsilon$ and substitute into Equation E.3.8:

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= (\varepsilon a\Delta\hat{x}_1 + \varepsilon b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to\text{c}})^2 \\
&\quad + (\varepsilon b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon c\mu\Delta\hat{z}_1 + \Delta y_{1\to\text{c}})^2 \\
&\quad + (\varepsilon c\omega\Delta\hat{z}_1 + \Delta z_{1\to\text{c}})^2 \\
0 &= A\cdot\varepsilon^2 + B\cdot\varepsilon + C \\
A &= (a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1)^2 + (b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1)^2 + (c\omega\Delta\hat{z}_1)^2 \\
B &= 2\Big[(a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1)\cdot(\Delta x_{1\to\text{c}}) \\
&\quad + (b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1)\cdot(\Delta y_{1\to\text{c}}) + (c\omega\Delta\hat{z}_1)\cdot(\Delta z_{1\to\text{c}})\Big] \\
C &= (\Delta x_{1\to\text{c}})^2 + (\Delta y_{1\to\text{c}})^2 + (\Delta z_{1\to\text{c}})^2 - |\Phi_{\text{deform}}|^2 \\
\varepsilon &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned}
\tag{E.3.12}
$$

Note that this solution is unaffected by what combination of $\Delta\hat{r}_1 = 0$ as long as at least one of them is non-zero (otherwise Rule 1 takes precedence). By definition, none of the ε_r can be constrained to 1 (otherwise they must all be 1 and Rule 1 takes over).

E.3.3 At Least 1 Non-zero p, q, and r & Two ε_r Constrained to 1

Per Rule 5-(ii), when two expansion factors are constrained, I solve Equation E.3.8 for the remaining factor. Strictly speaking, these equations would still 'work' even if $p = q = r = 0$, since only the distance equation is solved. However, under these circumstances, the 'best' dimension is to expand in is unclear, hence I resort to isotropic expansion per Rule 2.

E.3.3.1 ε_x and ε_y Constrained to 1

Solving for ε_z:

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= (a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\
&\quad + (b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
0 &= A\cdot\varepsilon_z^2 + B\cdot\varepsilon_z + C \\
A &= (c\cos(\beta)\Delta\hat{z}_1)^2 + (c\mu\Delta\hat{z}_1)^2 + (c\omega\Delta\hat{z}_1)^2 \\
B &= 2\Big[(c\cos(\beta)\Delta\hat{z}_1)\cdot(a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1 + \Delta x_{1\to c}) \\
&\quad + (c\mu\Delta\hat{z}_1)\cdot(b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c}) + (c\omega\Delta\hat{z}_1)\cdot(\Delta z_{1\to c})\Big] \\
C &= (a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1 + \Delta x_{1\to c})^2 \\
&\quad + (b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c})^2 + (\Delta z_{1\to c})^2 - |\mathbf{\Phi}_{\text{deform}}|^2 \\
\varepsilon_z &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned}
\tag{E.3.13}
$$

E.3.3.2 ε_x and ε_z Constrained to 1

Solving for ε_y:

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= (a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\
&\quad + (\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + (c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2. \\
0 &= A\cdot\varepsilon_y^2 + B\cdot\varepsilon_y + C \\
A &= (b\cos(\gamma)\Delta\hat{y}_1)^2 + (b\sin(\gamma)\Delta\hat{y}_1)^2 = (b\Delta\hat{y}_1)^2 \\
B &= 2\Big[(b\cos(\gamma)\Delta\hat{y}_1)\cdot(a\Delta\hat{x}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}) \\
&\quad + (b\sin(\gamma)\Delta\hat{y}_1)\cdot(c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})\Big] \\
C &= (a\Delta\hat{x}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\
&\quad + (c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 + (c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 - |\Phi_{\text{deform}}|^2 \\
\varepsilon_y &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned}
\tag{E.3.14}
$$

E.3.3.3 ε_y and ε_z Constrained to 1

Solving for ε_x:

$$\begin{aligned}|\mathbf{\Phi}_{\text{deform}}|^2 &= (\varepsilon_x a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\ &+ (b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\ &+ (c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2.\end{aligned}$$

$$\begin{aligned}0 &= A\cdot\varepsilon_x^2 + B\cdot\varepsilon_x + C \\ A &= (a\Delta\hat{x}_1)^2 \\ B &= 2\left[(a\Delta\hat{x}_1)\cdot(b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})\right] \\ C &= (b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\ &+ (b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\ &+ (c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 - |\Phi_{\text{deform}}|^2\end{aligned} \tag{E.3.15}$$

$$\varepsilon_x = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.$$

E.3.4 All ε_r Unconstrained

As per Rule 5-(a thru d), in the event which none of the expansion factors are constrained to 1, we can always apply a combination of two ratios and the distance equation to get three values of ε_r. Depending on the values of p, q, and r, the ε_r may inadvertently return to 1, however this is not considered a constraint.

E.3.4.1 Only $r = 0$

From either Equation E.3.10 or E.3.11, we can show that for $r = 0$,

$$\varepsilon_z = 1 \tag{E.3.16}$$

Equation E.3.9 provides the relationship between the other two dimensions,

$$\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = \frac{p}{q}\cdot(\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}) \tag{E.3.17}$$

then substituting these into Equation E.3.8,

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left(\frac{p}{q}\right)^2 \cdot (\varepsilon_y b \sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + (\varepsilon_y b \sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
\frac{|\mathbf{\Phi}_{\text{deform}}|^2}{\left(\frac{p}{q}\right)^2 + 1} &= (\varepsilon_y b \sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
0 &= A \cdot \varepsilon_y^2 + B \cdot \varepsilon_y + C \\
A &= (b \sin(\gamma)\Delta\hat{y}_1)^2 \\
B &= 2\left[(b \sin(\gamma)\Delta\hat{y}_1) \cdot (c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})\right] \\
C &= (c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 - \frac{|\mathbf{\Phi}_{\text{deform}}|^2}{\left(\frac{p}{q}\right)^2 + 1} \\
\varepsilon_y &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned}
\tag{E.3.18}
$$

And finally,

$$
\varepsilon_x = \frac{\frac{p}{q} \cdot (\varepsilon_y b \sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}) - (\varepsilon_y b \cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})}{a\Delta\hat{x}_1} \tag{E.3.19}
$$

E.3.4.2 Both $r = 0$ and $q = 0$, $p \neq 0$

Because $r = 0$, it can shown from Equation E.3.10 that,

$$\varepsilon_z = 1 \tag{E.3.20}$$

and then it follows from Equation E.3.9 that,

$$\varepsilon_y = 1. \tag{E.3.21}$$

With that, there is only one unknown remaining which we employ the quadratic equation to enforce the distance requirements. This results in the same solution given in Equation E.3.15.

E.3.4.3 Both $r = 0$ and $p = 0$, $q \neq 0$

Because $r = 0$, it can shown from Equation E.3.11 that,

$$\varepsilon_z = 1. \tag{E.3.22}$$

and given that $p = 0$, it follows from Equation E.3.9,

$$\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b \cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c \cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = 0. \tag{E.3.23}$$

These can substituted into Equation E.3.8,

$$\begin{aligned}|\mathbf{\Phi}_{\text{deform}}|^2 &= (\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\ &\quad + (c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2.\end{aligned}$$

$$\begin{aligned}0 &= A\cdot\varepsilon_y^2 + B\cdot\varepsilon_y + C \\ A &= (b\sin(\gamma)\Delta\hat{y}_1)^2 \\ B &= 2\left[(b\sin(\gamma)\Delta\hat{y}_1)\cdot(c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})\right] \\ C &= (c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 + (c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 - |\Phi_{\text{deform}}|^2\end{aligned} \tag{E.3.24}$$

$$\varepsilon_y = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.$$

And finally,

$$\varepsilon_x = -\frac{\varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}}{a\Delta\hat{x}_1} \tag{E.3.25}$$

E.3.4.4 Only $q = 0$

Since $q = 0$, from Equation E.3.11,

$$0 = \varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c} \tag{E.3.26}$$

and from Equation E.3.10,

$$\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = \frac{p}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}). \tag{E.3.27}$$

Substituting both of these equations into Equation E.3.8,

$$\begin{aligned}|\mathbf{\Phi}_{\text{deform}}|^2 &= \left(\frac{p}{r}\right)^2\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\ &\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2.\end{aligned}$$

$$\frac{|\mathbf{\Phi}_{\text{deform}}|^2}{\left(\frac{p}{r}\right)^2 + 1} = (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2$$

$$\begin{aligned}0 &= A\cdot\varepsilon_z^2 + B\cdot\varepsilon_z + C \\ A &= (c\omega\Delta\hat{z}_1)^2 \\ B &= 2\left[(c\omega\Delta\hat{z}_1)\cdot(\Delta z_{1\to c})\right] \\ C &= (\Delta z_{1\to c})^2 - \frac{|\mathbf{\Phi}_{\text{deform}}|^2}{\left(\frac{p}{r}\right)^2 + 1}\end{aligned} \tag{E.3.28}$$

$$\varepsilon_z = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.$$

It then follows that,

$$\varepsilon_y = -\frac{\varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}}{b\sin(\gamma)\Delta\hat{y}_1} \tag{E.3.29}$$

and,

$$\varepsilon_x = \frac{\frac{p}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}) - (\varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})}{a\Delta\hat{x}_1} \tag{E.3.30}$$

E.3.4.5 Both $q = 0$ and $p = 0$, $r \neq 0$

Given $q = 0$, from Equation E.3.11,

$$0 = \varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c} \tag{E.3.31}$$

and similarly, from $p = 0$ and Equation E.3.10,

$$0 = \varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}. \tag{E.3.32}$$

From here it is trivial to obtain ε_z based on Equation E.3.8,

$$\begin{aligned} |\mathbf{\Phi}_{\text{deform}}|^2 &= (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\ \varepsilon_z &= \frac{-\Delta z_{1\to c} \pm |\mathbf{\Phi}_{\text{deform}}|}{c\omega\Delta\hat{z}_1}. \end{aligned} \tag{E.3.33}$$

From here, we can return to the first two equations and solve for ε_y from Equation E.3.29 and ε_x from,

$$\varepsilon_x = -\frac{\varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}}{a\Delta\hat{x}_1}. \tag{E.3.34}$$

E.3.4.6 Only $p = 0$

Given that $p = 0$, it follows from Equation E.3.10 that,

$$0 = \varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}. \tag{E.3.35}$$

To solve for the remaining two expansion factors, we utilize Equation E.3.11,

$$\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c} = \frac{q}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}). \tag{E.3.36}$$

Substituting these into Equation E.3.8,

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left(\frac{q}{r}\right)^2 \cdot \left(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to\text{c}}\right)^2 \\
&\quad + \left(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to\text{c}}\right)^2
\end{aligned}
$$

$$
\frac{|\mathbf{\Phi}_{\text{deform}}|^2}{\left(\frac{q}{r}\right)^2 + 1} = \left(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to\text{c}}\right)^2
$$

$$
\begin{aligned}
0 &= A\cdot\varepsilon_z^2 + B\cdot\varepsilon_z + C \\
A &= \left(c\omega\Delta\hat{z}_1\right)^2 \\
B &= 2\left[\left(c\omega\Delta\hat{z}_1\right)\cdot\left(\Delta z_{1\to\text{c}}\right)\right] \\
C &= \left(\Delta z_{1\to\text{c}}\right)^2 - \frac{|\mathbf{\Phi}_{\text{deform}}|^2}{\left(\frac{q}{r}\right)^2 + 1}
\end{aligned}
\tag{E.3.37}
$$

$$
\varepsilon_z = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
$$

Thereafter,

$$
\varepsilon_y = \frac{\frac{q}{r}\cdot\left(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to\text{c}}\right) - \left(\varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to\text{c}}\right)}{b\sin(\gamma)\Delta\hat{y}_1} \tag{E.3.38}
$$

and ε_x is obtained from Equation E.3.34.

E.3.4.7 All $p \neq 0$, $q \neq 0$, $r \neq 0$

From Equations E.3.10 and E.3.11, we can obtain,

$$\varepsilon_x a \Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = \frac{p}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}) \tag{E.3.39}$$

and,

$$\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c} = \frac{q}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}) \tag{E.3.40}$$

Substituting these into Equation E.3.8,

$$\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left(\frac{p}{r}\right)^2\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
&\quad + \left(\frac{q}{r}\right)^2\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
&\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
\\
\frac{|\mathbf{\Phi}_{\text{deform}}|^2}{\left(\frac{p}{r}\right)^2 + \left(\frac{q}{r}\right)^2 + 1} &= (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
\\
0 &= A\cdot\varepsilon_z^2 + B\cdot\varepsilon_z + C \\
A &= (c\omega\Delta\hat{z}_1)^2 \\
B &= 2\left[(c\omega\Delta\hat{z}_1)\cdot(\Delta z_{1\to c})\right] \\
C &= (\Delta z_{1\to c})^2 - \frac{|\mathbf{\Phi}_{\text{deform}}|^2}{\left(\frac{p}{r}\right)^2 + \left(\frac{q}{r}\right)^2 + 1} \\
\\
\varepsilon_z &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned} \tag{E.3.41}$$

Following this, ε_y can be determined from Equation E.3.38 and ε_x from Equation E.3.30.

E.3.5 ε_z Constrained to 1

As per Rule 5-(iii), in the event where only one of the three expansion factors are constrained, it is is necessary to consider two solution cases representing the distance equation used in conjunction with each of the two possible ratio equations. As with the previous case of three unconstrained expansion factors, for each constrained ε_r, we must consider all possible combinations of p, q, and r value. I begin by looking at the case of $\varepsilon_z = 1$.

E.3.5.1 Only $r = 0$

This scenario is a special case due to the fact that the constraint on ε offers no 'novel' information over the ratio equations that utilize r. To expound on this, we can consider two possible solutions involving the use of the p/q relationship and the r/p (or equivalently r/q).

Solution 1

In Solution 1, $\varepsilon_z = 1$ is given and we utilize Equation E.3.9 to obtain,

$$\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = \frac{p}{q}\cdot(\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}) \quad \text{(E.3.42)}$$

One would realize that this is identical to the previous case shown in Section E.3.4.1 and as such ε_y can be solved by Equation E.3.18 and ε_x from Equation E.3.19.

Solution 2

In Solution 2, $\varepsilon_z = 1$ is given and we utilize Equation E.3.11, however in this situation this only gives the same information,

$$\varepsilon_z = 1. \quad \text{(E.3.43)}$$

That is to say, the constraint has provided no unique information over the ratio equation and thusly this is an **invalid solution**.

E.3.5.2 Both $r = 0$ and $q = 0$, $p \neq 0$

As with the previous case, the constraint on ε_z is non-informative and we only have one unique solution to consider.

Solution 1

In Solution 1, $\varepsilon_z = 1$ is given and using Equation E.3.9, it can be shown that,

$$\varepsilon_y = 1. \quad \text{(E.3.44)}$$

This is identical to Section E.3.4.2, and ε_x can be determined from Equation E.3.15.

Solution 2

In Solution 2, $\varepsilon_z = 1$ is given and we utilize Equation E.3.10, however in this situation this provides no distinct information to the constraint and thusly this is an **invalid solution**.

E.3.5.3 Both $r = 0$ and $p = 0$, $q \neq 0$

As with the previous two cases, the constraint on ε_z is non-informative and we only have one unique solution to consider.

Solution 1

In Solution 1, $\varepsilon_z = 1$ is given and using Equation E.3.9, it can be shown that,

$$\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = 0. \tag{E.3.45}$$

This is identical to Section E.3.4.3. Hence, ε_y and ε_x can be determined from Equations E.3.24 and E.3.25 respectively.

Solution 2

In Solution 2, $\varepsilon_z = 1$ is given and we utilize Equation E.3.11, however in this situation this provides no distinct information to the constraint and thusly this is an **invalid solution**.

E.3.5.4 Only $q = 0$

In this case, the constraint on ε_z provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, $\varepsilon_z = 1$ is given and using Equation E.3.11, it can be shown that,

$$\varepsilon_y = 1. \tag{E.3.46}$$

Thereafter, ε_x may be solved from Equation E.3.15.

Solution 2

In Solution 2, $\varepsilon_z = 1$ is given and using Equation E.3.10, it can be shown that,

$$\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = \frac{p}{r}\cdot(c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}) \tag{E.3.47}$$

which is the equivalent of Equation E.3.30, but rearranged and with $\varepsilon_z = 1$. Substituting this into the distance equation (Equation E.3.8),

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= (\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + \left[\left(\frac{p}{r}\right)^2 + 1\right]\cdot(c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
0 &= A\cdot\varepsilon_y^2 + B\cdot\varepsilon_y + C \\
A &= (b\sin(\gamma)\Delta\hat{y}_1)^2 \\
B &= 2\left[(b\sin(\gamma)\Delta\hat{y}_1)\cdot(c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})\right] \\
C &= (c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 + \left[\left(\frac{p}{r}\right)^2 + 1\right]\cdot(c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 - |\Phi_{\text{deform}}|^2 \\
\varepsilon_y &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned}
\tag{E.3.48}
$$

Thereafter, ε_x may be solved from Equation E.3.30.

Note that unlike the earlier three cases, both ratio equations provides unique information on one of the unconstrained expansion factors and thus two solutions may be obtained.

E.3.5.5 Both $q = 0$ and $p = 0$, $r \neq 0$

In this case, the constraint on ε_z provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, $\varepsilon_z = 1$ is given and using Equation E.3.11, it can be shown that,

$$\varepsilon_y = 1. \tag{E.3.49}$$

Thereafter, ε_x may be solved from Equation E.3.15.

Solution 2

In Solution 2, $\varepsilon_z = 1$ is given and using Equation E.3.10, it can be shown that,

$$0 = \varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}. \tag{E.3.50}$$

Applying this to Equation E.3.8,

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= (\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + (c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
0 &= A\cdot\varepsilon_y^2 + B\cdot\varepsilon_y + C \\
A &= (b\sin(\gamma)\Delta\hat{y}_1)^2 \\
B &= 2\left[(b\sin(\gamma)\Delta\hat{y}_1)\cdot(c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})\right] \\
C &= (c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 + (c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 - |\Phi_{\text{deform}}|^2 \\
\varepsilon_y &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned}
\tag{E.3.51}
$$

Thereafter, ε_x can be determined using Equation E.3.34.

E.3.5.6 Only $p = 0$

In this case, the constraint on ε_z provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, $\varepsilon_z = 1$ is given and using Equation E.3.10, it can be shown that,

$$
0 = \varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}. \tag{E.3.52}
$$

This gives the exact solution as the previous case whereby ε_y is determined from Equation E.3.51 and ε_x is determined from Equation E.3.34.

Solution 2

In Solution 2, $\varepsilon_z = 1$ is given and using Equation E.3.11, we obtain,

$$
\varepsilon_y = \frac{\frac{q}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}) - (\varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})}{b\sin(\gamma)\Delta\hat{y}_1} \tag{E.3.53}
$$

which is the same as Equation E.3.38. ε_y can be solved directly from here. ε_x is then obtained from Equation E.3.8 as per,

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= (\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\
&\quad + (\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + (c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2.
\end{aligned}
$$

$$
\begin{aligned}
0 &= A\cdot\varepsilon_x^2 + B\cdot\varepsilon_x + C \\
A &= (a\Delta\hat{x}_1)^2 \\
B &= 2\left[(a\Delta\hat{x}_1)\cdot(\varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})\right] \\
C &= (\varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\
&\quad + (\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + (c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
&\quad - |\Phi_{\text{deform}}|^2 \\
\varepsilon_x &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned}
\tag{E.3.54}
$$

Note that this is similar to Equation E.3.15 except that ε_y is not constrained to 1.

E.3.5.7 All $p \neq 0$, $q \neq 0$, $r \neq 0$

In this case, the constraint on ε_z provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, $\varepsilon_z = 1$ is given and using Equation E.3.11, it can be shown that,

$$
\varepsilon_y = \frac{\frac{q}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}) - (\varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})}{b\sin(\gamma)\Delta\hat{y}_1}.
\tag{E.3.55}
$$

This is identical to the previous case: ε_y is determined via Equation E.3.38 and ε_x is determined from Equation E.3.54.

Solution 2

In Solution 2, $\varepsilon_z = 1$ is given and using Equation E.3.10, we obtain,

$$
\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = \frac{p}{r}\cdot(c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}).
\tag{E.3.56}
$$

This case is identical to a previous one wherein ε_y is determined from E.3.48 and ε_x is solved from Equation E.3.30.

E.3.6 ε_y Constrained to 1

I now consider the cases where ε_y is constrained to 1, while the other two expansion factors are unconstrained.

E.3.6.1 Only $r = 0$

In this case, the constraint on ε_y provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, we first use Equation E.3.10, from which it can be shown that,

$$\varepsilon_z = 1. \tag{E.3.57}$$

Paired with the constraint $\varepsilon_y = 1$, we can obtain ε_x from Equation E.3.15.

Solution 2

In Solution 2, we instead use Equation E.3.9, from which it can be shown that,

$$\varepsilon_x a\Delta\hat{x}_1+\varepsilon_y b\cos(\gamma)\Delta\hat{y}_1+\varepsilon_z c\cos(\beta)\Delta\hat{z}_1+\Delta x_{1\to c} = \frac{p}{q}\cdot(\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1+\varepsilon_z c\mu\Delta\hat{z}_1+\Delta y_{1\to c}). \tag{E.3.58}$$

Applying this into Equation E.3.8,

$$\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left(\frac{p}{q}\right)^2\cdot(b\sin(\gamma)\Delta\hat{y}_1+\varepsilon_z c\mu\Delta\hat{z}_1+\Delta y_{1\to c})^2\\
&\quad+(b\sin(\gamma)\Delta\hat{y}_1+\varepsilon_z c\mu\Delta\hat{z}_1+\Delta y_{1\to c})^2\\
&\quad+(\varepsilon_z c\omega\Delta\hat{z}_1+\Delta z_{1\to c})^2\\
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left[\left(\frac{p}{q}\right)^2+1\right]\cdot(b\sin(\gamma)\Delta\hat{y}_1+\varepsilon_z c\mu\Delta\hat{z}_1+\Delta y_{1\to c})^2\\
&\quad+(\varepsilon_z c\omega\Delta\hat{z}_1+\Delta z_{1\to c})^2\\
0 &= A\cdot\varepsilon_z^2+B\cdot\varepsilon_z+C\\
A &= \left[\left(\frac{p}{q}\right)^2+1\right]\cdot(c\mu\Delta\hat{z}_1)^2+(c\omega\Delta\hat{z}_1)^2\\
B &= 2\left\{\left[\left(\frac{p}{q}\right)^2+1\right]\cdot(c\mu\Delta\hat{z}_1)\cdot(b\sin(\gamma)\Delta\hat{y}_1+\Delta y_{1\to c})+(c\omega\Delta\hat{z}_1)\cdot(\Delta z_{1\to c})\right\}\\
C &= \left[\left(\frac{p}{q}\right)^2+1\right](b\sin(\gamma)\Delta\hat{y}_1+\Delta y_{1\to c})^2+(\Delta z_{1\to c})^2-|\mathbf{\Phi}_{\text{deform}}|^2\\
\varepsilon_z &= \frac{-B\pm\sqrt{B^2-4AC}}{2A}.
\end{aligned} \tag{E.3.59}$$

Note that this is similar to Equation E.3.18, except we solve for ε_z instead of ε_y and ratios involving $r = 0$ are not enforced. With ε_z and ε_y defined, ε_x is obtained by rearranging the ratio equation to give,

$$\varepsilon_x = \frac{\frac{p}{q} \cdot (b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}) - (b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})}{a\Delta\hat{x}_1}. \quad \text{(E.3.60)}$$

E.3.6.2 Both $r = 0$ and $q = 0$, $p \neq 0$

In this case, the constraint on ε_y is non-informative and we only have one unique solution to consider.

Solution 1

In Solution 1, we use Equation E.3.10, from which it can be shown that,

$$\varepsilon_z = 1. \quad \text{(E.3.61)}$$

Paired with the constraint $\varepsilon_y = 1$, we can obtain ε_x from Equation E.3.15.

Solution 2

In Solution 2, we instead use Equation E.3.9, from which it can be shown that,

$$0 = \varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}. \quad \text{(E.3.62)}$$

Given the the constraint $\varepsilon_y = 1$, ε_z can be shown to be = 1. Ultimately, this is identical to the first solution.

E.3.6.3 Both $r = 0$ and $p = 0$, $q \neq 0$

In this case, the constraint on ε_y provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, we first use Equation E.3.11, from which it can be shown that,

$$\varepsilon_z = 1. \quad \text{(E.3.63)}$$

Paired with the constraint $\varepsilon_y = 1$, we can obtain ε_x from Equation E.3.15.

Solution 2

In Solution 2, we instead use Equation E.3.9, from which it can be shown that,

$$0 = \varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}. \quad \text{(E.3.64)}$$

ε_z can be determined by using this in Equation E.3.8 in conjunction with the given constraint such that:

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= (b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to\text{c}})^2 \\
&\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to\text{c}})^2.
\end{aligned}
$$

$$
\begin{aligned}
0 &= A\cdot\varepsilon_z^2 + B\cdot\varepsilon_z + C \\
A &= (c\mu\Delta\hat{z}_1)^2 + (c\omega\Delta\hat{z}_1)^2 \\
B &= 2\left[(c\mu\Delta\hat{z}_1)\cdot(b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to\text{c}}) + (c\omega\Delta\hat{z}_1)\cdot(\Delta z_{1\to\text{c}})\right] \\
C &= (b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to\text{c}})^2 + (\Delta z_{1\to\text{c}})^2 - |\Phi_{\text{deform}}|^2
\end{aligned}
\qquad \text{(E.3.65)}
$$

$$
\varepsilon_z = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
$$

And lastly, ε_x is obtained from Equation E.3.34.

E.3.6.4 Only $q = 0$

In this case, the constraint on ε_y provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, we first use Equation E.3.11, from which it can be shown that,

$$0 = \varepsilon_y b \sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}. \tag{E.3.66}$$

$$\varepsilon_z = -\frac{\varepsilon_y b \sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c}\cdot}{c\mu\Delta\hat{z}_1}, \tag{E.3.67}$$

from which ε_x can be found from,

$$\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= (\varepsilon_x a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\
&\quad + (b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2.
\end{aligned}$$

$$\begin{aligned}
0 &= A\cdot\varepsilon_x^2 + B\cdot\varepsilon_x + C \\
A &= (a\Delta\hat{x}_1)^2 \\
B &= 2\left[(a\Delta\hat{x}_1)\cdot(b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})\right] \\
C &= (b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\
&\quad + (b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
&\quad - |\Phi_{\text{deform}}|^2
\end{aligned} \tag{E.3.68}$$

$$\varepsilon_x = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.$$

Note this is similar to Equation E.3.54, except $\varepsilon_y = 1$ instead of ε_z.

Solution 2

In Solution 2, we first use Equation E.3.10, from which it can be shown that,

$$\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = \frac{p}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}). \tag{E.3.69}$$

Utilizing this in Equation E.3.8,

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= (b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + \left[\left(\frac{p}{r}\right)^2 + 1\right]\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
0 &= A\cdot\varepsilon_z^2 + B\cdot\varepsilon_z + C \\
A &= (c\mu\Delta\hat{z}_1)^2 + \left[\left(\frac{p}{r}\right)^2 + 1\right]\cdot(c\omega\Delta\hat{z}_1)^2 \\
B &= 2\left\{(c\mu\Delta\hat{z}_1)\cdot(b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c}) + \left[\left(\frac{p}{r}\right)^2 + 1\right]\cdot(c\omega\Delta\hat{z}_1)\cdot(\Delta z_{1\to c})\right\} \\
C &= (b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c})^2 + \left[\left(\frac{p}{r}\right)^2 + 1\right]\cdot(\Delta z_{1\to c})^2 - |\Phi_{\text{deform}}|^2 \\
\varepsilon_z &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned}
\tag{E.3.70}
$$

Note that this is similar to the earlier Equation E.3.48, except it is $\varepsilon_y = 1$ instead of ε_z. ε_x can then be calculated using Equation E.3.30.

E.3.6.5 Both $q = 0$ and $p = 0$, $r \neq 0$

In this case, the constraint on ε_y provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, we first use Equation E.3.11, from which it can be shown that,

$$0 = \varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}. \tag{E.3.71}$$

This is identical to an earlier case where ε_z is first determined using Equation E.3.67 and then ε_x is determined using Equation E.3.68.

Solution 2

In Solution 1, we instead use Equation E.3.10, from which it can be shown that,

$$0 = \varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}. \tag{E.3.72}$$

This leads to an identical case as before where ε_z is first determined using E.3.65 and ε_x is obtained from Equation E.3.34.

E.3.6.6 Only $p = 0$

In this case, the constraint on ε_y provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, we first use Equation E.3.10, from which it can be shown that,

$$0 = \varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}. \tag{E.3.73}$$

This leads to an identical case as before where ε_z is first determined using E.3.65 and ε_x is obtained from Equation E.3.34.

Solution 2

In Solution 2, we instead use Equation E.3.11, from which it can be shown that,

$$\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c} = \frac{q}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}). \tag{E.3.74}$$

Rearranging this, we can directly obtain a value for ε_z,

$$\varepsilon_z = \frac{(\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c}) - \frac{q}{r}\cdot(\Delta z_{1\to c})}{\frac{q}{r}\cdot(c\omega\Delta\hat{z}_1) - c\mu\Delta\hat{z}_1} \tag{E.3.75}$$

With the available values of ε_y and ε_z, the remaining ε_x can be found from Equation E.3.68.

E.3.6.7 All $p \neq 0$, $q \neq 0$, $r \neq 0$

In this case, the constraint on ε_y provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, we first use Equation E.3.11, from which it can be shown that,

$$\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c} = \frac{q}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}). \tag{E.3.76}$$

This is identical to the preceding case where ε_z is found from E.3.75 and ε_x can be found from Equation E.3.68.

Solution 2

In Solution 2, we instead use Equation E.3.10, from which it can be shown that,

$$\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = \frac{p}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}). \tag{E.3.77}$$

This is identical to an earlier case where ε_z is found from Equation E.3.70 and ε_x can then be calculated using Equation E.3.30.

E.3.7 ε_x Constrained to 1

Finally, I consider the cases where ε_x is constrained to 1, while the other two expansion factors are unconstrained.

E.3.7.1 Only $r = 0$

In this case, the constraint on ε_x provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, we first use Equation E.3.10, from which it can be shown that,

$$\varepsilon_z = 1. \tag{E.3.78}$$

Paired with the constraint $\varepsilon_x = 1$, we can obtain ε_y from Equation E.3.14.

Solution 2

In Solution 2, we instead use Equation E.3.9, from which it can be shown that,

$$\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = \frac{p}{q}\cdot(\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}). \tag{E.3.79}$$

Given that $\varepsilon_x = 1$, this can be rearranged to obtain ε_y as a function of ε_z,

$$\varepsilon_y = \frac{\frac{p}{q}\cdot(\varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}) - (\varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} + a\Delta\hat{x}_1)}{b\cos(\gamma)\Delta\hat{y}_1 - \frac{p}{q}\cdot(b\sin(\gamma)\Delta\hat{y}_1)}. \tag{E.3.80}$$

For brevity, I further group the terms as,

$$\varepsilon_y = \varepsilon_z\frac{\frac{p}{q}\cdot(c\mu\Delta\hat{z}_1) - (c\cos(\beta)\Delta\hat{z}_1)}{b\cos(\gamma)\Delta\hat{y}_1 - \frac{p}{q}\cdot(b\sin(\gamma)\Delta\hat{y}_1)} + \frac{\frac{p}{q}\cdot(\Delta y_{1\to c}) - (\Delta x_{1\to c} + a\Delta\hat{x}_1)}{b\cos(\gamma)\Delta\hat{y}_1 - \frac{p}{q}\cdot(b\sin(\gamma)\Delta\hat{y}_1)} \tag{E.3.81}$$

$$\varepsilon_y = \varepsilon_z \cdot G + I$$

These can then be substituted into Equation E.3.8,

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left(\frac{p}{q}\right)^2 \cdot (\varepsilon_y b \sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + (\varepsilon_y b \sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left[\left(\frac{p}{q}\right)^2 + 1\right] \cdot (\varepsilon_y b \sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left[\left(\frac{p}{q}\right)^2 + 1\right] \cdot [(\varepsilon_z \cdot G + I) \cdot b \sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}]^2 \\
&\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left[\left(\frac{p}{q}\right)^2 + 1\right] \cdot [\varepsilon_z \cdot G \cdot b \sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + I \cdot b \sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c}]^2 \\
&\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \qquad \text{(E.3.82)} \\
0 &= A \cdot \varepsilon_z^2 + B \cdot \varepsilon_z + C \\
A &= \left[\left(\frac{p}{q}\right)^2 + 1\right] \cdot (G \cdot b \sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1)^2 + (c\omega\Delta\hat{z}_1)^2 \\
B &= 2\left\{\left[\left(\frac{p}{q}\right)^2 + 1\right] \cdot (G \cdot b \sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1)\right. \\
&\quad \left. \cdot (I \cdot b \sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c}) + (c\omega\Delta\hat{z}_1) \cdot (\Delta z_{1\to c})\right\} \\
C &= \left[\left(\frac{p}{q}\right)^2 + 1\right](I \cdot b \sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c})^2 + (\Delta z_{1\to c})^2 - |\mathbf{\Phi}_{\text{deform}}|^2 \\
\varepsilon_z &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned}
$$

Once ε_z is determined, it is trivial to return to the previous equation and solve for ε_y.

E.3.7.2 Both $r = 0$ and $q = 0$, $p \neq 0$

In this case, the constraint on ε_x provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, we first use Equation E.3.10, from which it can be shown that,

$$\varepsilon_z = 1. \qquad \text{(E.3.83)}$$

Paired with the constraint $\varepsilon_x = 1$, we can obtain ε_y from Equation E.3.14.

Solution 2

In Solution 2, we instead use Equation E.3.9, from which it can be shown that,

$$0 = \varepsilon_y b \sin(\gamma) \Delta\hat{y}_1 + \varepsilon_z c\mu \Delta\hat{z}_1 + \Delta y_{1\to c}. \tag{E.3.84}$$

Again, we need to express ε_y as a function of ε_z. This is given by Equation E.3.29 but has been rewritten here,

$$\varepsilon_y = -\frac{\varepsilon_z c\mu \Delta\hat{z}_1 + \Delta y_{1\to c}}{b \sin(\gamma) \Delta\hat{y}_1}. \tag{E.3.85}$$

Substituting these into Equation E.3.8,

$$\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left(a\Delta\hat{x}_1 - \frac{\varepsilon_z c\mu \Delta\hat{z}_1 + \Delta y_{1\to c}}{b \sin(\gamma) \Delta\hat{y}_1} \cdot b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} \right)^2 \\
&\quad + \left(\varepsilon_z c\omega \Delta\hat{z}_1 + \Delta z_{1\to c}\right)^2 \\
0 &= A \cdot \varepsilon_z^2 + B \cdot \varepsilon_z + C \\
A &= \left(c\cos(\beta)\Delta\hat{z}_1 - \frac{c\mu\Delta\hat{z}_1 \cdot b\cos(\gamma)\Delta\hat{y}_1}{b\sin(\gamma)\Delta\hat{y}_1} \right)^2 + (c\omega\Delta\hat{z}_1)^2 \\
B &= 2\left[\left(c\cos(\beta)\Delta\hat{z}_1 - \frac{c\mu\Delta\hat{z}_1 \cdot b\cos(\gamma)\Delta\hat{y}_1}{b\sin(\gamma)\Delta\hat{y}_1} \right) \right. \\
&\quad \left. \cdot \left(a\Delta\hat{x}_1 - \frac{\Delta y_{1\to c} \cdot b\cos(\gamma)\Delta\hat{y}_1}{b\sin(\gamma)\Delta\hat{y}_1} + \Delta x_{1\to c} \right) + (c\omega\Delta\hat{z}_1) \cdot (\Delta z_{1\to c}) \right] \\
C &= \left(a\Delta\hat{x}_1 - \frac{\Delta y_{1\to c} \cdot b\cos(\gamma)\Delta\hat{y}_1}{b\sin(\gamma)\Delta\hat{y}_1} + \Delta x_{1\to c} \right)^2 \\
&\quad + (\Delta z_{1\to c})^2 \\
&\quad - |\Phi_{\text{deform}}|^2 \\
\varepsilon_z &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned} \tag{E.3.86}$$

Once ε_z is known, it is trivial to return to Equation E.3.29 and solve for ε_y,

E.3.7.3 Both $r = 0$ and $p = 0$, $q \neq 0$

In this case, the constraint on ε_x provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, we first use Equation E.3.11, from which it can be shown that,

$$\varepsilon_z = 1. \tag{E.3.87}$$

Paired with the constraint $\varepsilon_x = 1$, we can obtain ε_y from Equation E.3.14.

Solution 2

In Solution 2, we instead use Equation E.3.9, from which it can be shown that,

$$0 = \varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}. \tag{E.3.88}$$

Given $\varepsilon_x = 1$, this can be rearranged to obtain ε_y as a function of ε_z,

$$\varepsilon_y = -\frac{a\Delta\hat{x}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}.}{b\cos(\gamma)\Delta\hat{y}_1} \tag{E.3.89}$$

Substituting these into Equation E.3.8,

$$\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left(-\frac{a\Delta\hat{x}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}}{b\cos(\gamma)\Delta\hat{y}_1}\cdot b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}\right)^2 \\
&\quad + \left(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}\right)^2 \\
0 &= A\cdot\varepsilon_z^2 + B\cdot\varepsilon_z + C \\
A &= \left(c\mu\Delta\hat{z}_1 - \frac{c\cos(\beta)\Delta\hat{z}_1\cdot b\sin(\gamma)\Delta\hat{y}_1}{b\cos(\gamma)\Delta\hat{y}_1}\right)^2 + (c\omega\Delta\hat{z}_1)^2 \\
B &= 2\left[\left(c\mu\Delta\hat{z}_1 - \frac{c\cos(\beta)\Delta\hat{z}_1\cdot b\sin(\gamma)\Delta\hat{y}_1}{b\cos(\gamma)\Delta\hat{y}_1}\right)\right. \\
&\quad \left.\cdot\left(\Delta y_{1\to c} - \frac{(a\Delta\hat{x}_1 + \Delta x_{1\to c})\cdot b\sin(\gamma)\Delta\hat{y}_1}{b\cos(\gamma)\Delta\hat{y}_1}\right) + (c\omega\Delta\hat{z}_1)\cdot(\Delta z_{1\to c})\right] \\
C &= \left(\Delta y_{1\to c} - \frac{(a\Delta\hat{x}_1 + \Delta x_{1\to c})\cdot b\sin(\gamma)\Delta\hat{y}_1}{b\cos(\gamma)\Delta\hat{y}_1}\right)^2 + (\Delta z_{1\to c})^2 - |\Phi_{\text{deform}}|^2 \\
\varepsilon_z &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned} \tag{E.3.90}$$

Once ε_z is known, it is trivial to return to the previous expression and solve for ε_y.

E.3.7.4 Only $q = 0$

In this case, the constraint on ε_x provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, we first use Equation E.3.9, from which it can be shown that

$$0 = \varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}. \tag{E.3.91}$$

Paired with the constraint $\varepsilon_x = 1$, this is reminiscent of the previous case where ε_z is determined using Equation E.3.86 after which ε_y can be obtained from Equation E.3.29.

Solution 2

In Solution 2, we instead use Equation E.3.10, from which it can be shown that

$$\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = \frac{p}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}). \tag{E.3.92}$$

Given that $\varepsilon_x = 1$, this can be rearranged to obtain ε_y as a function of ε_z,

$$\varepsilon_y = \frac{\frac{p}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}) - (\varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} + a\Delta\hat{x}_1)}{b\cos(\gamma)\Delta\hat{y}_1}. \tag{E.3.93}$$

For brevity, I further group the terms as,

$$\varepsilon_y = \varepsilon_z\frac{\frac{p}{r}\cdot(c\omega\Delta\hat{z}_1) - (c\cos(\beta)\Delta\hat{z}_1)}{b\cos(\gamma)\Delta\hat{y}_1} + \frac{\frac{p}{r}\cdot(\Delta z_{1\to c}) - (\Delta x_{1\to c} + a\Delta\hat{x}_1)}{b\cos(\gamma)\Delta\hat{y}_1} \tag{E.3.94}$$

$$\varepsilon_y = \varepsilon_z\cdot G + I$$

These can then be substituted into Equation E.3.8,

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left(\frac{p}{r}\right)^2 \cdot (\varepsilon_z c\omega \Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
&\quad + (\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
|\mathbf{\Phi}_{\text{deform}}|^2 &= (\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + \left[\left(\frac{p}{r}\right)^2 + 1\right] \cdot (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
|\mathbf{\Phi}_{\text{deform}}|^2 &= [(\varepsilon_z \cdot G + I) \cdot b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}]^2 \\
&\quad + \left[\left(\frac{p}{r}\right)^2 + 1\right] \cdot (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
|\mathbf{\Phi}_{\text{deform}}|^2 &= [\varepsilon_z \cdot G \cdot b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + I \cdot b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c}]^2 \\
&\quad + \left[\left(\frac{p}{r}\right)^2 + 1\right] \cdot (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2
\end{aligned}
\tag{E.3.95}
$$

$$
\begin{aligned}
0 &= A \cdot \varepsilon_z^2 + B \cdot \varepsilon_z + C \\
A &= (G \cdot b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1)^2 + \left[\left(\frac{p}{r}\right)^2 + 1\right] \cdot (c\omega\Delta\hat{z}_1)^2 \\
B &= 2\Big\{(G \cdot b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1) \cdot (I \cdot b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c}) \\
&\quad + \left[\left(\frac{p}{r}\right)^2 + 1\right] \cdot (c\omega\Delta\hat{z}_1) \cdot (\Delta z_{1\to c})\Big\} \\
C &= (I \cdot b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c})^2 + \left[\left(\frac{p}{r}\right)^2 + 1\right] \cdot (\Delta z_{1\to c})^2 - |\mathbf{\Phi}_{\text{deform}}|^2 \\
\varepsilon_z &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned}
$$

Once ε_z is known, it is trivial to return to the previous equation to solve for ε_y.

E.3.7.5 Both $q = 0$ and $p = 0$, $r \neq 0$

In this case, the constraint on ε_x provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, we first use Equation E.3.11, from which it can be shown that,

$$
0 = \varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}. \tag{E.3.96}
$$

Paired with the constraint $\varepsilon_x = 1$, this is reminiscent of the previous case where ε_z is determined

using Equation E.3.86 after which ε_y can be obtained from Equation E.3.29.

Solution 2

In Solution 2, we instead use Equation E.3.10, from which it can be shown that,

$$0 = \varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}. \tag{E.3.97}$$

This is similar to the previous case where ε_z is first determined using Equation E.3.90 and ε_y is found from Equation E.3.89.

E.3.7.6 Only $p = 0$

In this case, the constraint on ε_x provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, we first use Equation E.3.10, from which it can be shown that,

$$0 = \varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}. \tag{E.3.98}$$

This is similar to the previous case where ε_z is first determined using Equation E.3.90 and ε_y is found from Equation E.3.89.

Solution 2

In Solution 2, we instead use Equation E.3.11, from which it can be shown that,

$$\varepsilon_y b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c} = \frac{q}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}). \tag{E.3.99}$$

This can be rearranged to express ε_y as a function of ε_z,

$$\varepsilon_y = \frac{\frac{q}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}) - (\varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})}{b\sin(\gamma)\Delta\hat{y}_1}. \tag{E.3.100}$$

For brevity, I further group the terms as,

$$\begin{gathered}\varepsilon_y = \varepsilon_z\frac{\frac{q}{r}\cdot(c\omega\Delta\hat{z}_1) - (c\mu\Delta\hat{z}_1)}{b\sin(\gamma)\Delta\hat{y}_1} + \frac{\frac{q}{r}\cdot(\Delta z_{1\to c}) - \Delta y_{1\to c}}{b\sin(\gamma)\Delta\hat{y}_1} \\ \varepsilon_y = \varepsilon_z\cdot G + I\end{gathered} \tag{E.3.101}$$

These can then be substituted into Equation E.3.8,

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= (a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\
&\quad + \left(\frac{q}{r}\right)^2 \cdot (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
&\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2
\end{aligned}
$$

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= (a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\
&\quad + \left[\left(\frac{q}{r}\right)^2 + 1\right] \cdot (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2
\end{aligned}
$$

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= [(\varepsilon_z \cdot G + I) \cdot b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} + a\Delta\hat{x}_1]^2 \\
&\quad + \left[\left(\frac{q}{r}\right)^2 + 1\right] \cdot (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2
\end{aligned}
$$

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= [\varepsilon_z \cdot G \cdot b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + I \cdot b\cos(\gamma)\Delta\hat{y}_1 + \Delta x_{1\to c} + a\Delta\hat{x}_1]^2 \\
&\quad + \left[\left(\frac{q}{r}\right)^2 + 1\right] \cdot (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2
\end{aligned}
\tag{E.3.102}
$$

$$
\begin{aligned}
0 &= A \cdot \varepsilon_z^2 + B \cdot \varepsilon_z + C \\
A &= (G \cdot b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1)^2 + \left[\left(\frac{q}{r}\right)^2 + 1\right] \cdot (c\omega\Delta\hat{z}_1)^2 \\
B &= 2\Bigg\{(G \cdot b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1) \cdot (I \cdot b\cos(\gamma)\Delta\hat{y}_1 + \Delta x_{1\to c} + a\Delta\hat{x}_1) \\
&\quad + \left[\left(\frac{q}{r}\right)^2 + 1\right] \cdot (c\omega\Delta\hat{z}_1) \cdot (\Delta z_{1\to c})\Bigg\} \\
C &= (I \cdot b\cos(\gamma)\Delta\hat{y}_1 + \Delta x_{1\to c} + a\Delta\hat{x}_1)^2 + \left[\left(\frac{q}{r}\right)^2 + 1\right] \cdot (\Delta z_{1\to c})^2 - |\mathbf{\Phi}_{\text{deform}}|^2
\end{aligned}
$$

$$
\varepsilon_z = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
$$

From here, it is trivial to return to the previous equation and solve for ε_y.

E.3.7.7 All $p \neq 0$, $q \neq 0$, $r \neq 0$

In this case, the constraint on ε_x provides unique information and we can obtain two valid solutions.

Solution 1

In Solution 1, we first use Equation E.3.10, from which it can be shown that,

$$
\varepsilon_x a\Delta\hat{x}_1 + \varepsilon_y b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = \frac{p}{r} \cdot (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}). \tag{E.3.103}
$$

Given that $\varepsilon_x = 1$, this can be rearranged to obtain ε_y as a function of ε_z. Ultimately, this is similar to a previous case where we solve for ε_z using Equation E.3.95 and ε_y is solved from

Equation E.3.94.

Solution 2

In Solution 2, we instead use Equation E.3.11, from which it can be shown that,

$$\varepsilon_y b \sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c} = \frac{q}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}). \tag{E.3.104}$$

Given that $\varepsilon_x = 1$, this can be rearranged to obtain ε_y as a function of ε_z. Ultimately, this is similar to a previous case where we solve for ε_z using Equation E.3.102 and ε_y is solved from Equation E.3.101.

E.4 Space group Symmetry Constraints

The derivations shown in the previous section assume that all three unit cell dimensions can be expanded independently. However, in practice, this is not always the case and there are certain space groups wherein two, or even three, of the cell dimensions are constrained to be equal. To account for this, the following cases must be considered:

1. "$a \neq b \neq c$": In triclinic, monoclinic, and orthorhombic space groups, all three cell dimensions are allowed to vary independently. The equations related to these space groups are those derived in the previous sections.

2. "$a = b = c$": In rhombohedral and cubic space groups, all three cell dimensions are required to be equal. Equations for these will be described in the following sections.

3. "$a = b \neq c$": In tetragonal, trigonal, and hexagonal space groups, the first two cell dimensions are required to be equal. Equations for these will be described in the following sections.

E.4.1 Rhombohedral and Cubic Space groups

Due to their space group symmetries, only the isotropic expansion is permitted for these groups. This has already been derived earlier in Equation E.3.12.

E.4.2 Tetragonal, Trigonal, and Hexagonal Space groups

Due to their space group symmetries, these space groups require that $\varepsilon_x = \varepsilon_y$. For convenience, I will use ε_x, ε_y, and ε_{xy} interchangeably. The effect of this requirement is discussed below for each of the previously discussed scenarios. As before, I derive the equations for cases where ε_r are constrained to 1 without explicitly considering why such an action might be necessary. It follows from the space group constraints that any restriction of ε_{xy} to 1 must apply to both dimensions.

E.4.2.1 All ε_r Constrained to 1

There is no change here. As per Rule 1, no further action can be taken and the algorithm must be terminated.

E.4.2.2 $p = q = r$ & No ε_r Constrained to 1

There is also no change here. An isotropic expansion is not in violation of any space group constraints, since $\varepsilon_x = \varepsilon_y = \varepsilon$. Hence in the absence of any directional information (in the form of the ratio equations), we can still return to Equation E.3.12.

E.4.2.3 Two ε_r Constrained to 1

Within this subset of scenarios, the only valid case is when $\varepsilon_x = \varepsilon_y = 1$ and we solve solely for ε_z. That is to say, only Equation E.3.13 may be used with these space groups.

E.4.2.4 ε_x or ε_y Constrained to 1

These equations were derived in Sections E.3.6 and E.3.7, however by definition they are inapplicable to these space groups due to our inability to decouple the values of ε_x and ε_y. As such, we can completely ignore all of these cases.

E.4.2.5 ε_z Constrained to 1

Section E.3.5 describes these cases wherein the distance equation is used in tandem with one of the ratio equations in order to solve for ε_x and ε_y. This led to two sets of quadratic solutions (i.e., 4 sets of ε_r). With the additional space group constraint, we can only solve for one variable (ε_{xy}) and thusly only the distance equation may be used,

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= (\varepsilon_{xy} a\Delta\hat{x}_1 + \varepsilon_{xy} b\cos(\gamma)\Delta\hat{y}_1 + c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\
&\quad + (\varepsilon_{xy} b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + (c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
\\
0 &= A\cdot\varepsilon_{xy}^2 + B\cdot\varepsilon_{xy} + C \\
A &= (a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1)^2 + (b\sin(\gamma)\Delta\hat{y}_1)^2 \\
B &= 2\big[(a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1)\cdot(c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}) \\
&\quad + (b\sin(\gamma)\Delta\hat{y}_1)\cdot(c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})\big] \\
C &= (c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 + (c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&\quad + (c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 - |\Phi_{\text{deform}}|^2 \\
\\
\varepsilon_{xy} &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned}
\tag{E.4.1}
$$

Since no ratio information is required here, the values of p, q, and r are unimportant, as long as at least one of them are non-zero (otherwise Rule 2 would have been applied earlier).

E.4.2.6 All ε_r Unconstrained

These equations are still relevant to the space group constrained systems. However, in Section E.3.4, two ratio equations were used in tandem with the distance equation in order to solve for three expansion factors. Under the constraint of $\varepsilon_x = \varepsilon_y = \varepsilon_{xy}$, we can no longer solve three equations simultaneously and instead must make a choice between either one of the ratio equations. This is similar to Sections E.3.5 thru E.3.7 where one of the ε_r were constrained 1 and two sets of quadratic solutions were found.

E.4.2.6.1 Only $r = 0$

Solution 1

In Solution 1, we first use Equation E.3.10, from which it can be shown that,

$$\varepsilon_z = 1 \tag{E.4.2}$$

In this case, we can directly solve for ε_{xy} with Equation E.4.1.

Solution 2

In Solution 2, we instead use Equation E.3.9, from which it can be shown that,

$$\varepsilon_{xy} a\Delta\hat{x}_1 + \varepsilon_{xy} b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = \frac{p}{q}\cdot(\varepsilon_{xy} b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}). \tag{E.4.3}$$

Rearranging this equation, we can obtain ε_{xy} as a function of ε_z,

$$\varepsilon_{xy} = \frac{\frac{p}{q}\cdot(\varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}) - (\varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})}{a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1 - \frac{p}{q}\cdot b\sin(\gamma)\Delta\hat{y}_1}. \tag{E.4.4}$$

For brevity, I further group the terms as,

$$\varepsilon_{xy} = \varepsilon_z \frac{\frac{p}{q}\cdot(c\mu\Delta\hat{z}_1) - c\cos(\beta)\Delta\hat{z}_1}{a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1 - \frac{p}{q}\cdot b\sin(\gamma)\Delta\hat{y}_1} + \frac{\frac{p}{q}\cdot(\Delta y_{1\to c}) - \Delta x_{1\to c}}{a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1 - \frac{p}{q}\cdot b\sin(\gamma)\Delta\hat{y}_1} \tag{E.4.5}$$

$$\varepsilon_{xy} = \varepsilon_z \cdot G + I$$

These can then be substituted into Equation E.3.8,

$$
\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left(\frac{p}{q}\right)^2 \cdot (\varepsilon_{xy} b \sin(\gamma) \Delta\hat{y}_1 + \varepsilon_z c\mu \Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&+ (\varepsilon_{xy} b \sin(\gamma) \Delta\hat{y}_1 + \varepsilon_z c\mu \Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&+ (\varepsilon_z c\omega \Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left[\left(\frac{p}{q}\right)^2 + 1\right] \cdot (\varepsilon_{xy} b \sin(\gamma) \Delta\hat{y}_1 + \varepsilon_z c\mu \Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\
&+ (\varepsilon_z c\omega \Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left[\left(\frac{p}{q}\right)^2 + 1\right] \cdot [(\varepsilon_z \cdot G + I) \cdot b \sin(\gamma) \Delta\hat{y}_1 + \varepsilon_z c\mu \Delta\hat{z}_1 + \Delta y_{1\to c}]^2 \\
&+ (\varepsilon_z c\omega \Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
|\mathbf{\Phi}_{\text{deform}}|^2 &= \left[\left(\frac{p}{q}\right)^2 + 1\right] \cdot [\varepsilon_z \cdot G \cdot b \sin(\gamma) \Delta\hat{y}_1 + \varepsilon_z c\mu \Delta\hat{z}_1 + I \cdot b \sin(\gamma) \Delta\hat{y}_1 + \Delta y_{1\to c}]^2 \\
&+ (\varepsilon_z c\omega \Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
0 &= A \cdot \varepsilon_z^2 + B \cdot \varepsilon_z + C \\
A &= \left[\left(\frac{p}{q}\right)^2 + 1\right] \cdot (G \cdot b \sin(\gamma) \Delta\hat{y}_1 + c\mu \Delta\hat{z}_1)^2 + (c\omega \Delta\hat{z}_1)^2 \\
B &= 2\left\{\left[\left(\frac{p}{q}\right)^2 + 1\right] \cdot (G \cdot b \sin(\gamma) \Delta\hat{y}_1 + c\mu \Delta\hat{z}_1)\right. \\
&\left. \cdot (I \cdot b \sin(\gamma) \Delta\hat{y}_1 + \Delta y_{1\to c}) + (c\omega \Delta\hat{z}_1) \cdot (\Delta z_{1\to c})\right\} \\
C &= \left[\left(\frac{p}{q}\right)^2 + 1\right] (I \cdot b \sin(\gamma) \Delta\hat{y}_1 + \Delta y_{1\to c})^2 + (\Delta z_{1\to c})^2 - |\mathbf{\Phi}_{\text{deform}}|^2 \\
\varepsilon_z &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned}
\tag{E.4.6}
$$

Once ε_z is determined, it is trivial to return to the previous equation to solve for ε_{xy}.

E.4.2.6.2 Both $r = 0$ and $q = 0$, $p \neq 0$

Solution 1

In Solution 1, we first use Equation E.3.10, from which it can be shown that,

$$\varepsilon_z = 1 \tag{E.4.7}$$

In this case, we can directly solve for ε_{xy} with Equation E.4.1.

Solution 2

In Solution 2, we instead use Equation E.3.9, from which it can be shown that,

$$0 = \varepsilon_{xy} b \sin(\gamma) \Delta\hat{y}_1 + \varepsilon_z c\mu \Delta\hat{z}_1 + \Delta y_{1\to c}. \tag{E.4.8}$$

Rearranging this gives,

$$\varepsilon_{xy} = -\frac{\varepsilon_z c\mu \Delta\hat{z}_1 + \Delta y_{1\to c}}{b \sin(\gamma) \Delta\hat{y}_1}. \tag{E.4.9}$$

As always, we apply this to the distance equation and obtain,

$$\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 = \Bigg(& -\frac{\varepsilon_z c\mu \Delta\hat{z}_1 + \Delta y_{1\to c}}{b \sin(\gamma) \Delta\hat{y}_1} \cdot a \Delta\hat{x}_1 \\
& - \frac{\varepsilon_z c\mu \Delta\hat{z}_1 + \Delta y_{1\to c}}{b \sin(\gamma) \Delta\hat{y}_1} \cdot b \cos(\gamma) \Delta\hat{y}_1 + \varepsilon_z c \cos(\beta) \Delta\hat{z}_1 + \Delta x_{1\to c} \Bigg)^2 \\
& + (\varepsilon_z c\omega \Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
0 = {} & A \cdot \varepsilon_z^2 + B \cdot \varepsilon_z + C \\
A = {} & \left[c \cos(\beta) \Delta\hat{z}_1 - \frac{c\mu \Delta\hat{z}_1 \cdot (b \cos(\gamma) \Delta\hat{y}_1 + a \Delta\hat{x}_1)}{b \sin(\gamma) \Delta\hat{y}_1} \right]^2 + (c\omega \Delta\hat{z}_1)^2 \\
B = {} & 2 \Bigg\{ \left[c \cos(\beta) \Delta\hat{z}_1 - \frac{c\mu \Delta\hat{z}_1 \cdot (b \cos(\gamma) \Delta\hat{y}_1 + a \Delta\hat{x}_1)}{b \sin(\gamma) \Delta\hat{y}_1} \right] \\
& \cdot \left[\Delta x_{1\to c} - \frac{\Delta y_{1\to c} \cdot (b \cos(\gamma) \Delta\hat{y}_1 + a \Delta\hat{x}_1)}{b \sin(\gamma) \Delta\hat{y}_1} \right] \\
& + (c\omega \Delta\hat{z}_1) \cdot (\Delta z_{1\to c}) \Bigg\} \\
C = {} & \left[\Delta x_{1\to c} - \frac{\Delta y_{1\to c} \cdot (b \cos(\gamma) \Delta\hat{y}_1 + a \Delta\hat{x}_1)}{b \sin(\gamma) \Delta\hat{y}_1} \right]^2 \\
& + (\Delta z_{1\to c})^2 - |\Phi_{\text{deform}}|^2 \\
\varepsilon_z = {} & \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned} \tag{E.4.10}$$

Thereafter, ε_{xy} can be determined using the previous equations.

E.4.2.6.3 Both $r = 0$ and $p = 0$, $q \neq 0$

Solution 1

In Solution 1, we first use Equation E.3.10, from which it can be shown that,

$$\varepsilon_z = 1 \tag{E.4.11}$$

In this case, we can directly solve for ε_{xy} with Equation E.4.1.

Solution 2

In Solution 2, we instead use Equation E.3.9, from which it can be shown that,

$$0 = \varepsilon_{xy} a \Delta\hat{x}_1 + \varepsilon_{xy} b \cos(\gamma) \Delta\hat{y}_1 + \varepsilon_z c \cos(\beta) \Delta\hat{z}_1 + \Delta x_{1\to c}. \tag{E.4.12}$$

Rearranging this equation, we can obtain ε_{xy} as a function of ε_z,

$$\varepsilon_{xy} = -\frac{\varepsilon_z c \cos(\beta) \Delta\hat{z}_1 + \Delta x_{1\to c}}{a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1}. \tag{E.4.13}$$

For brevity, I further group the terms as,

$$\begin{aligned} \varepsilon_{xy} &= \varepsilon_z \cdot -\frac{c\cos(\beta)\Delta\hat{z}_1}{a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1} - \frac{\Delta x_{1\to c}}{a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1} \\ \varepsilon_{xy} &= \varepsilon_z \cdot G + I \end{aligned} \tag{E.4.14}$$

These can then be substituted into Equation E.3.8,

$$\begin{aligned} |\mathbf{\Phi}_{\text{deform}}|^2 &= (\varepsilon_{xy} b \sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\ &\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\ |\mathbf{\Phi}_{\text{deform}}|^2 &= [(\varepsilon_z \cdot G + I) \cdot b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}]^2 \\ &\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\ 0 &= A\cdot\varepsilon_z^2 + B\cdot\varepsilon_z + C \\ A &= (G\cdot b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1)^2 + (c\omega\Delta\hat{z}_1)^2 \\ B &= 2[(G\cdot b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1)\cdot(I\cdot b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c}) + (c\omega\Delta\hat{z}_1)\cdot(\Delta z_{1\to c})] \\ C &= (I\cdot b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c})^2 + (\Delta z_{1\to c})^2 - |\mathbf{\Phi}_{\text{deform}}|^2 \\ \varepsilon_z &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}. \end{aligned} \tag{E.4.15}$$

With this, it is trivial to return to the previous equation and solve for ε_{xy}.

E.4.2.6.4 Only $q = 0$

Solution 1

In Solution 1, we first use Equation E.3.9, from which it can be shown that,

$$0 = \varepsilon_{xy} b \sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}. \tag{E.4.16}$$

In this case, we can reuse Equations E.4.10 and E.4.9 to determined ε_z and ε_{xy} respectively.

Solution 2

In Solution 2, we instead use Equation E.3.10, from which it can be shown that,

$$\varepsilon_{xy} a\Delta\hat{x}_1 + \varepsilon_{xy} b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c} = \frac{p}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}). \tag{E.4.17}$$

This can be rearranged to obtain ε_{xy} as a function of ε_z,

$$\varepsilon_{xy} = \frac{\frac{p}{r}\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c}) - (\varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})}{a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1}. \tag{E.4.18}$$

For brevity, I further group the terms as,

$$\begin{aligned}\varepsilon_{xy} &= \varepsilon_z \frac{\frac{p}{r}\cdot(c\omega\Delta\hat{z}_1) - (c\cos(\beta)\Delta\hat{z}_1)}{a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1} + \frac{\frac{p}{r}\cdot(\Delta z_{1\to c}) - (\Delta x_{1\to c})}{a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1} \\ \varepsilon_{xy} &= \varepsilon_z \cdot G + I\end{aligned} \tag{E.4.19}$$

These can then be substituted into Equation E.3.8,

$$\begin{aligned}|\mathbf{\Phi}_{\text{deform}}|^2 &= \left(\frac{p}{r}\right)^2\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\ &\quad + (\varepsilon_{xy} b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\ &\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\ |\mathbf{\Phi}_{\text{deform}}|^2 &= (\varepsilon_{xy} b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c})^2 \\ &\quad + \left[\left(\frac{p}{r}\right)^2 + 1\right]\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\ |\mathbf{\Phi}_{\text{deform}}|^2 &= [(\varepsilon_z \cdot G + I)\cdot b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + \Delta y_{1\to c}]^2 \\ &\quad + \left[\left(\frac{p}{r}\right)^2 + 1\right]\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\ |\mathbf{\Phi}_{\text{deform}}|^2 &= [\varepsilon_z \cdot G\cdot b\sin(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\mu\Delta\hat{z}_1 + I\cdot b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c}]^2 \\ &\quad + \left[\left(\frac{p}{r}\right)^2 + 1\right]\cdot(\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2\end{aligned} \tag{E.4.20}$$

$$\begin{aligned}0 &= A\cdot\varepsilon_z^2 + B\cdot\varepsilon_z + C \\ A &= (G\cdot b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1)^2 + \left[\left(\frac{p}{r}\right)^2 + 1\right]\cdot(c\omega\Delta\hat{z}_1)^2 \\ B &= 2\Bigg\{(G\cdot b\sin(\gamma)\Delta\hat{y}_1 + c\mu\Delta\hat{z}_1)\cdot(I\cdot b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c}) \\ &\quad + \left[\left(\frac{p}{r}\right)^2 + 1\right]\cdot(c\omega\Delta\hat{z}_1)\cdot(\Delta z_{1\to c})\Bigg\} \\ C &= (I\cdot b\sin(\gamma)\Delta\hat{y}_1 + \Delta y_{1\to c})^2 + \left[\left(\frac{p}{r}\right)^2 + 1\right]\cdot(\Delta z_{1\to c})^2 - |\mathbf{\Phi}_{\text{deform}}|^2\end{aligned}$$

$$\varepsilon_z = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.$$

Once ε_z is known, it is trivial to return to the previous equation to solve for ε_{xy}.

E.4.2.6.5 Both $q = 0$ and $p = 0$, $r \neq 0$

Solution 1

In Solution 1, we first use Equation E.3.10, from which it can be shown that,

$$0 = \varepsilon_{xy} a \Delta\hat{x}_1 + \varepsilon_{xy} b \cos(\gamma) \Delta\hat{y}_1 + \varepsilon_z c \cos(\beta) \Delta\hat{z}_1 + \Delta x_{1\to c}. \tag{E.4.21}$$

This is identical to the previous case where ε_z and ε_{xy} are determined using Equations E.4.15 and E.4.14 respectively.

Solution 2

In Solution 2, we instead use Equation E.3.11, from which it can be shown that,

$$0 = \varepsilon_{xy} b \sin(\gamma) \Delta\hat{y}_1 + \varepsilon_z c \mu \Delta\hat{z}_1 + \Delta y_{1\to c}. \tag{E.4.22}$$

This is identical to the previous case where ε_z and ε_{xy} are determined using Equations E.4.10 and E.4.9 respectively.

E.4.2.6.6 Only $p = 0$

Solution 1

In Solution 1, we first use Equation E.3.10, from which it can be shown that,

$$0 = \varepsilon_{xy} a \Delta\hat{x}_1 + \varepsilon_{xy} b \cos(\gamma) \Delta\hat{y}_1 + \varepsilon_z c \cos(\beta) \Delta\hat{z}_1 + \Delta x_{1\to c}. \tag{E.4.23}$$

This is identical to the previous case where ε_z and ε_{xy} are determined using Equations E.4.15 and E.4.14 respectively.

Solution 2

In Solution 2, we instead use Equation E.3.11, from which it can be shown that,

$$\varepsilon_{xy} b \sin(\gamma) \Delta\hat{y}_1 + \varepsilon_z c \mu \Delta\hat{z}_1 + \Delta y_{1\to c} = \frac{q}{r} \cdot (\varepsilon_z c \omega \Delta\hat{z}_1 + \Delta z_{1\to c}). \tag{E.4.24}$$

This can be rearranged to express ε_{xy} as a function of ε_z,

$$\varepsilon_{xy} = \frac{\frac{q}{r} \cdot (\varepsilon_z c \omega \Delta\hat{z}_1 + \Delta z_{1\to c}) - (\varepsilon_z c \mu \Delta\hat{z}_1 + \Delta y_{1\to c})}{b \sin(\gamma) \Delta\hat{y}_1}. \tag{E.4.25}$$

For brevity, I further group the terms as,

$$\varepsilon_{xy} = \varepsilon_z \frac{\frac{q}{r} \cdot (c\omega\Delta\hat{z}_1) - (c\mu\Delta\hat{z}_1)}{b\sin(\gamma)\Delta\hat{y}_1} + \frac{\frac{q}{r} \cdot (\Delta z_{1\to c}) - \Delta y_{1\to c}}{b\sin(\gamma)\Delta\hat{y}_1}$$

$$\varepsilon_{xy} = \varepsilon_z \cdot G + I \tag{E.4.26}$$

These can then be substituted into Equation E.3.8,

$$\begin{aligned}
|\mathbf{\Phi}_{\text{deform}}|^2 &= (\varepsilon_{xy} a\Delta\hat{x}_1 + \varepsilon_{xy} b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\
&\quad + \left(\frac{q}{r}\right)^2 \cdot (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
&\quad + (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
|\mathbf{\Phi}_{\text{deform}}|^2 &= (\varepsilon_{xy} a\Delta\hat{x}_1 + \varepsilon_{xy} b\cos(\gamma)\Delta\hat{y}_1 + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c})^2 \\
&\quad + \left[\left(\frac{q}{r}\right)^2 + 1\right] \cdot (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
|\mathbf{\Phi}_{\text{deform}}|^2 &= [(\varepsilon_z \cdot G + I) \cdot (a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1) + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 + \Delta x_{1\to c}+]^2 \\
&\quad + \left[\left(\frac{q}{r}\right)^2 + 1\right] \cdot (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
|\mathbf{\Phi}_{\text{deform}}|^2 &= \Big[\varepsilon_z \cdot G \cdot (a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1) + \varepsilon_z c\cos(\beta)\Delta\hat{z}_1 \\
&\quad + I \cdot (a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1) + \Delta x_{1\to c}\Big]^2 \\
&\quad + \left[\left(\frac{q}{r}\right)^2 + 1\right] \cdot (\varepsilon_z c\omega\Delta\hat{z}_1 + \Delta z_{1\to c})^2 \\
0 &= A \cdot \varepsilon_z^2 + B \cdot \varepsilon_z + C \\
A &= [G \cdot (a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1) + c\cos(\beta)\Delta\hat{z}_1]^2 + \left[\left(\frac{q}{r}\right)^2 + 1\right] \cdot (c\omega\Delta\hat{z}_1)^2 \\
B &= 2\Big\{[G \cdot (a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1) + c\cos(\beta)\Delta\hat{z}_1] \cdot [I \cdot (a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1) + \Delta x_{1\to c}] \\
&\quad + \left[\left(\frac{q}{r}\right)^2 + 1\right] \cdot (c\omega\Delta\hat{z}_1) \cdot (\Delta z_{1\to c})\Big\} \\
C &= [I \cdot (a\Delta\hat{x}_1 + b\cos(\gamma)\Delta\hat{y}_1) + \Delta x_{1\to c}]^2 + \left[\left(\frac{q}{r}\right)^2 + 1\right] \cdot (\Delta z_{1\to c})^2 - |\mathbf{\Phi}_{\text{deform}}|^2 \\
\varepsilon_z &= \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.
\end{aligned} \tag{E.4.27}$$

From here, it is then trivial to return to the previous equation and solve for ε_{xy}.

E.4.2.6.7 All $p \neq 0$, $q \neq 0$, $r \neq 0$

In Solution 1, we first use Equation E.3.10, from which it can be shown that

$$\varepsilon_{xy} a \Delta\hat{x}_1 + \varepsilon_{xy} b \cos(\gamma) \Delta\hat{y}_1 + \varepsilon_z c \cos(\beta) \Delta\hat{z}_1 + \Delta x_{1\to c} = \frac{p}{r} \cdot (\varepsilon_z c\omega \Delta\hat{z}_1 + \Delta z_{1\to c}). \tag{E.4.28}$$

This is identical to the previous case where ε_z and ε_{xy} are determined using Equations E.4.20 and E.4.19 respectively.

Solution 2

In Solution 2, we instead use Equation E.3.11, from which it can be shown that

$$\varepsilon_{xy} b \sin(\gamma) \Delta\hat{y}_1 + \varepsilon_z c\mu \Delta\hat{z}_1 + \Delta y_{1\to c} = \frac{q}{r} \cdot (\varepsilon_z c\omega \Delta\hat{z}_1 + \Delta z_{1\to c}). \tag{E.4.29}$$

This is identical to the previous case where ε_z and ε_{xy} are determined using Equations E.4.27 and E.4.26 respectively.

E.5 Undefined Analytical Solutions

The equations above have all been tested and verified to give valid solutions that correctly satisfy the distance equation and any ratio equations applied. However, in practice, input parameters may have less-than-compliant values, which can lead to numerical issues when trying to solve the analytical equations. The two main cases are described below.

1. **Imaginary Quadratic Solutions:** This would occur if the quadratic equations determinant $B^2 - 4AC < 0$. Numerically, I cannot formulate any shared reason (amongst all the previously discussed cases) that restricts this from happening. From a physical standpoint, I see no reason why there should ever need to be an 'imaginary' expansion to separate two points in the manner which the algorithm is executed.

2. **Divide-by-zero:** Many of the solutions rely on divisions by various parameters (or parameters subtracted from each other). There is a risk that these denominator terms equate to 0. This can happen in both the quadratic solution ($A = 0$), or other equations that explicitly use the trigonometric functions involving cell angles (e.g., $\cos\left(\frac{\pi}{2}\right) = 0$).

In any case, both of these cases should always be verified before attempting to solve any solution.

If such numerical issues are encountered, this does not necessitate a dead-end. So long as one of the expansions can be constrained to 1 (without all three $\varepsilon_r = 1$), then we can attempt to bypass the numerical issue and re-solve the problem for the constrained problem instead. By definition, this 'fix' can only be attempted if the numerical issue is encountered whilst solving for two or more ε_r. This also implies that a choice must be made between which of the two (or

Table E.1: Summary of the relevant analytical equations (given by their number) for each problem formulation. The solutions are split based on: their space group constraints (SG-Con), whether any expansion factor is constrained (Con-x, y, z), as well as if p, q, r are zero-valued. Up to two sets of solutions are possible. Within these solutions, constrained values are indicated as 'Con. 1'. This is distinguished from cases where the expansion value being 1 arises from the solution itself (indicated as '= 1')

SG-Con	Con-x	Con-y	Con-z	p	q	r	Solution 1			Solution 2		
							ε_x	ε_y	ε_z	ε_x	ε_y	ε_z
None	=1	=1	=1	Unimportant			No solution			No solution		
	Isotropic			= 0	= 0	= 0	E.3.12			Only 1 sol.		
	=1	=1	-	One Dim ≠ 0			Con. 1	Con. 1	E.3.13	Only 1 sol.		
	=1	-	=1	One Dim ≠ 0			Con. 1	E.3.14	Con. 1	Only 1 sol.		
	-	=1	=1	One Dim ≠ 0			E.3.15	Con. 1	Con. 1	Only 1 sol.		
	-	-	-	≠ 0	≠ 0	= 0	E.3.19	E.3.18	=1	Only 1 sol.		
	-	-	-	≠ 0	= 0	= 0	E.3.15	=1	=1	Only 1 sol.		
	-	-	-	= 0	≠ 0	= 0	E.3.25	E.3.24	=1	Only 1 sol.		
	-	-	-	≠ 0	= 0	≠ 0	E.3.30	E.3.29	E.3.28	Only 1 sol.		
	-	-	-	= 0	= 0	≠ 0	E.3.34	E.3.29	E.3.33	Only 1 sol.		
	-	-	-	= 0	≠ 0	≠ 0	E.3.34	E.3.38	E.3.37	Only 1 sol.		
	-	-	-	≠ 0	≠ 0	≠ 0	E.3.30	E.3.38	E.3.41	Only 1 sol.		
	-	-	=1	≠ 0	≠ 0	= 0	E.3.19	E.3.18	Con. 1	Only 1 sol.		
	-	-	=1	≠ 0	= 0	= 0	E.3.15	=1	Con. 1	Only 1 sol.		
	-	-	=1	= 0	≠ 0	= 0	E.3.25	E.3.24	Con. 1	Only 1 sol.		
	-	-	=1	≠ 0	= 0	≠ 0	E.3.15	=1	Con. 1	E.3.30	E.3.48	Con. 1
	-	-	=1	= 0	= 0	≠ 0	E.3.15	=1	Con. 1	E.3.34	E.3.51	Con. 1
	-	-	=1	= 0	≠ 0	≠ 0	E.3.34	E.3.51	Con. 1	E.3.54	E.3.38	Con. 1
	-	-	=1	≠ 0	≠ 0	≠ 0	E.3.54	E.3.38	Con. 1	E.3.30	E.3.48	Con. 1
	-	=1	-	≠ 0	≠ 0	= 0	E.3.15	Con. 1	=1	E.3.60	Con. 1	E.3.59
	-	=1	-	≠ 0	= 0	= 0	E.3.15	Con. 1	=1	Only 1 sol.		
	-	=1	-	= 0	≠ 0	= 0	E.3.15	Con. 1	=1	E.3.34	Con. 1	E.3.65

Continued on next page

Table E.1 – continued from previous page

SG-Con	Con-x	Con-y	Con-z	p	q	r	Solution 1			Solution 2		
							ε_x	ε_y	ε_z	ε_x	ε_y	ε_z
	-	=1	-	≠ 0	= 0	≠ 0	E.3.68	Con. 1	E.3.67	E.3.30	Con. 1	E.3.70
	-	=1	-	= 0	= 0	≠ 0	E.3.68	Con. 1	E.3.67	E.3.34	Con. 1	E.3.65
	-	=1	-	= 0	≠ 0	≠ 0	E.3.34	Con. 1	E.3.65	E.3.68	Con. 1	E.3.75
	-	=1	-	≠ 0	≠ 0	≠ 0	E.3.68	Con. 1	E.3.75	E.3.30	Con. 1	E.3.70
	=1	-	-	≠ 0	≠ 0	= 0	Con. 1	E.3.14	=1	Con. 1	E.3.81	E.3.82
	=1	-	-	≠ 0	= 0	= 0	Con. 1	E.3.14	=1	Con. 1	E.3.29	E.3.86
	=1	-	-	= 0	≠ 0	= 0	Con. 1	E.3.14	=1	Con. 1	E.3.89	E.3.90
	=1	-	-	≠ 0	= 0	≠ 0	Con. 1	E.3.29	E.3.86	Con. 1	E.3.94	E.3.95
	=1	-	-	= 0	= 0	≠ 0	Con. 1	E.3.29	E.3.86	Con. 1	E.3.89	E.3.90
	=1	-	-	= 0	≠ 0	≠ 0	Con. 1	E.3.89	E.3.90	Con. 1	E.3.101	E.3.102
	=1	-	-	≠ 0	≠ 0	≠ 0	Con. 1	E.3.94	E.3.95	Con. 1	E.3.101	E.3.102
$a = b = c$	Isotropic			Unimportant			E.3.12			Only 1 sol.		
$a = b \neq c$	=1	=1		Unimportant			No solution			No solution		
	Isotropic			= 0	= 0	= 0	E.3.12			Only 1 sol.		
	=1	-		One Dim ≠ 0			Con. 1	Con. 1	E.3.13	Only 1 sol.		
	Either or =1	-		Violates SG			Violates SG			Violates SG		
	Equality	=1		One Dim ≠ 0			E.4.1		Con. 1	Only 1 sol.		
	Equality	-		≠ 0	≠ 0	= 0	E.4.1		=1	E.4.5		E.4.6
	Equality	-		≠ 0	= 0	= 0	E.4.1		=1	E.4.9		E.4.10
	Equality	-		= 0	≠ 0	= 0	E.4.1		=1	E.4.14		E.4.15
	Equality	-		≠ 0	= 0	≠ 0	E.4.9		E.4.10	E.4.19		E.4.20
	Equality	-		= 0	= 0	≠ 0	E.4.14		E.4.15	E.4.9		E.4.10
	Equality	-		= 0	≠ 0	≠ 0	E.4.14		E.4.15	E.4.26		E.4.27
	Equality	-		≠ 0	≠ 0	≠ 0	E.4.19		E.4.20	E.4.26		E.4.27

three) ε_r to fix to 1. I apply two in heuristics for this choice: (1) the constrained problem should avoid using the same equations as the originally unconstrained problem, otherwise it would run into the same issues, and (2) where a p, q, or $r = 0$, it would be preferable to constrain $\varepsilon_r = 1$ in that dimension. From the established reasoning, expanding in these dimensions will typically be 'less efficient'.

E.6 Algorithm Implementation and Solution Selection

At this point, I have addressed all issues related to obtaining a set of solutions that could result in the desired separation distance. I now turn my attention to obtaining a *fully expansionary* solution without inversions (i.e., all $\varepsilon_r \geq 1$). This is best explained in the context of the full algorithmic implementation.

To begin, in Figure E.3 I show the code structure leading up to a possible use of the Shift Algorithm. Upon generation of a Sobol' vector, the crystal structure is first subject to a set of sanity checks prior to minimization. The Shift Algorithm relates to the assessment of the initial intermolecular energy $U_{\text{inter},0}$, highlighted in yellow. During this check, the Energy subroutine is called, which assesses interatomic distances and interactions, before reporting back to the main program if any clashes have been detected.

If clashes are detected, and if the user has indicated a desire to use the Shift Algorithm, the program then calls the Shift Algorithm subroutine. If a valid expansion is found, the cell is expanded accordingly. Any given call to Energy will only indicate a single clash site, hence the Shift Algorithm can only address this. After expansion, there is no guarantee that no other clashes remain in the expanded cell. As such, following any shift, it is necessary to repeat

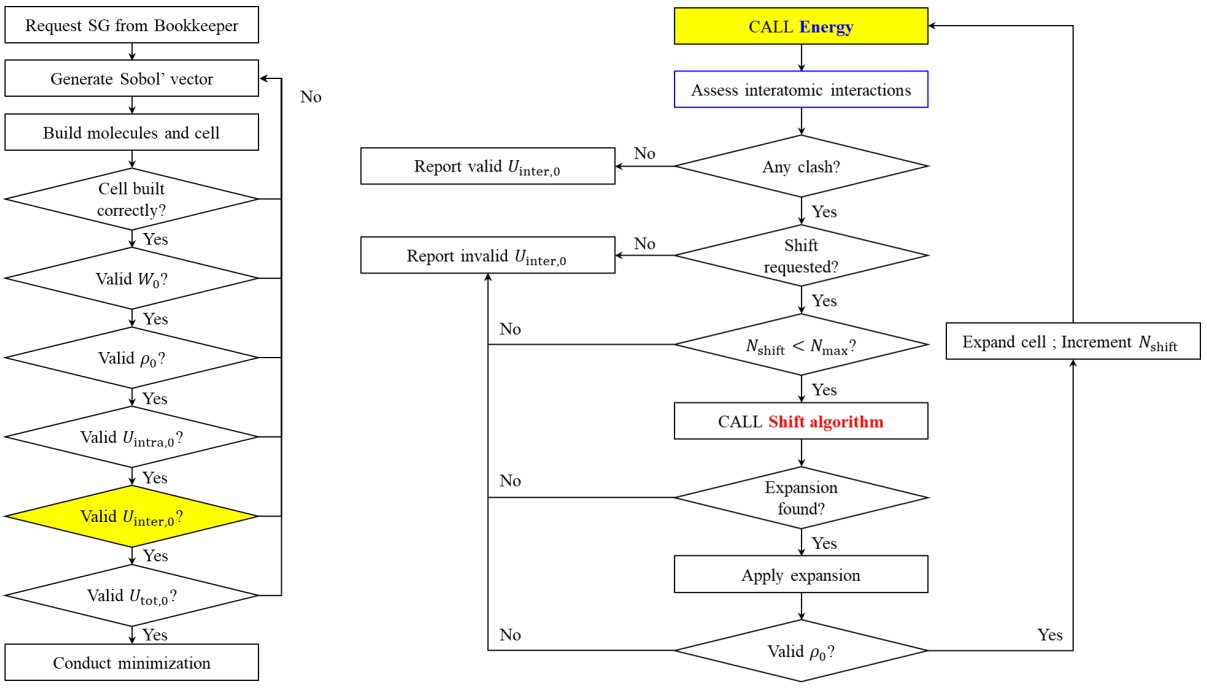


Figure E.3: The program structure leading up to a call to the Shift Algorithm subroutine.

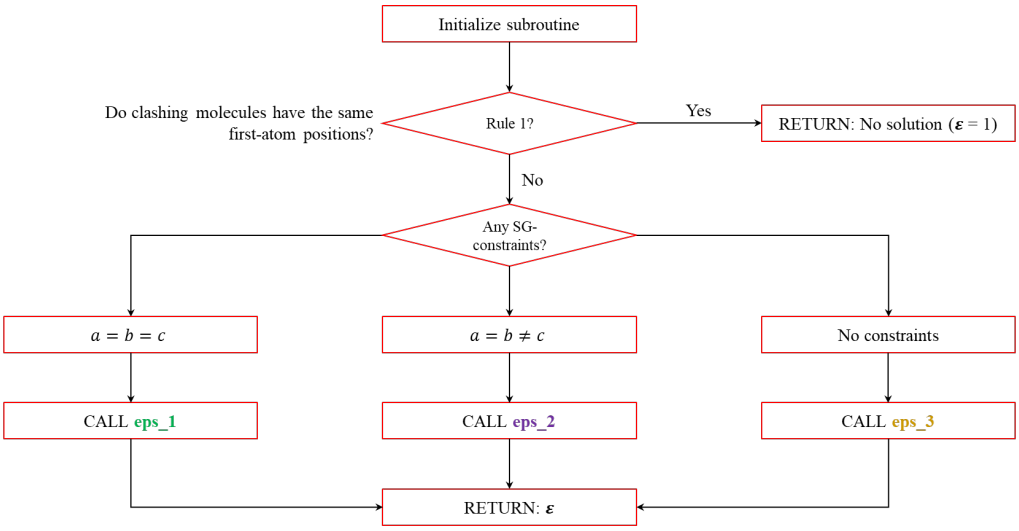


Figure E.4: The program structure within the Shift Algorithm subroutine.

this cycle. To avoid doing unnecessary shifts a cell density check is performed, since there is no point further expanding a cell that is already below the permissible density. Prior to any shift, a shift limit is also checked. A solution of the Shift Algorithm is considered to be valid only if the elements of $\boldsymbol{\varepsilon}$ are not all = 1. This a result of the subroutines within the Shift Algorithm: whenever an expansion dimension needs to be constrained (for whatever reason), its corresponding ε_r will be set to 1. The case where $\boldsymbol{\varepsilon} = 1$ naturally indicates that all expansions have ultimately been constrained to 1 and there is no valid expansion to consider. The cycle can thus be terminated to save time.

The Shift Algorithm subroutine itself (Figure E.4) does very little, in so far as to quickly check Rule 1 ("Do the clashing molecules have coincident first atoms, in which case just stop trying"), followed by calling the appropriate subroutine based on the space group constraints. Each of these subroutines are described in more detail below.

I start with the most general case where there are no space group constraints (Figure E.5). As per the earlier established rules, we can constrain $\varepsilon_r = 1$ for any $\Delta\hat{r}_1 = 0$. Due to the earlier checks done in the Shift Algorithm subroutine, there is no case where all three dimensions should be constrained in this manner. In the event where two dimensions are constrained, we can filter out to special cases where a variant of the distance equation is solved for the remaining factor. If less than two dimensions are constrained (and we solve for at least two expansion factors), this indicates that utilization of at least one ratio equation will be necessary. The validity of the ratios is verified based on whether $p = q = r = 0$, in which case an isotropic expansion is solved for. Note that when solving the isotropic expansion, we are indifferent to whether any $\Delta\hat{r}_1 = 0$ since the distance equation will inherently account for this.

If none of these 'special cases' are triggered, the appropriate set of equations to solve are selected from Table E.1. Once the relevant equations are solved, solution selection is performed by the select_sol subroutine. Details of this subroutine are given later, but in effect this subroutine will prioritize finding a fully expansionary solution. If found, this is returned within $\boldsymbol{\varepsilon}$ and

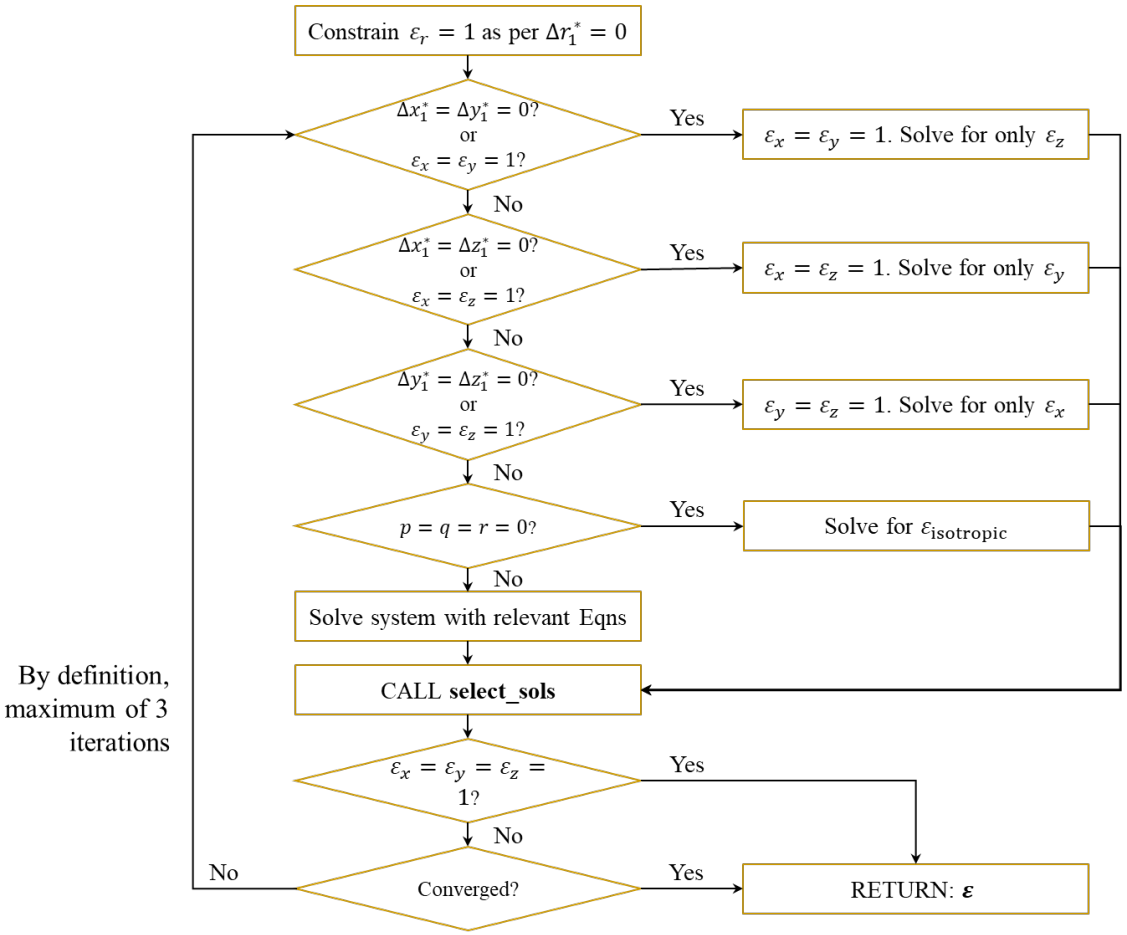


Figure E.5: The Eps_3 subroutine, which solves the expansion when there are no space group constraints.

the 'Converged' flag is set as 'True'. If this is not possible, select_sol will consider solutions that are contractions or inversions. Any ε_r that are < 1 will be set to 1 within $\boldsymbol{\varepsilon}$ and the 'Converged' flag is set as 'False'. Importantly, this will not alter any ε_r that were already constrained to 1 at the start of select_sol.

Upon returning from select_sol, if all $\varepsilon_r = 1$, this can only indicate that no valid expansion was found and we can immediately return to the Shift Algorithm subroutine. If not, the subroutine checks if the problem is converged (that is a solution with all $\varepsilon_r \geq 1$ has been found), in which case it also returns. In the final case, the problem is not converged and there remains a possible constrained formulation of the problem which can be re-attempted. The subroutine loops back to the start, where it repeats this process. Logically, a maximum of only three such attempts are ever possible. This is the case if the initial loop was completely unconstrained and one ε_r is constrained per loop (there is no restriction that only one factor can be constrained per loop). By the start of the third loop, two $\varepsilon_r = 1$ and any further constraint of the last ε_r will force termination.

The case where $a = b \neq c$ (Figure E.6), follows a very similar path to the previous case. As before, the subroutine starts by constraining $\varepsilon_r = 1$ based on $\Delta\hat{r}_1 = 0$. Unlike before, there is the

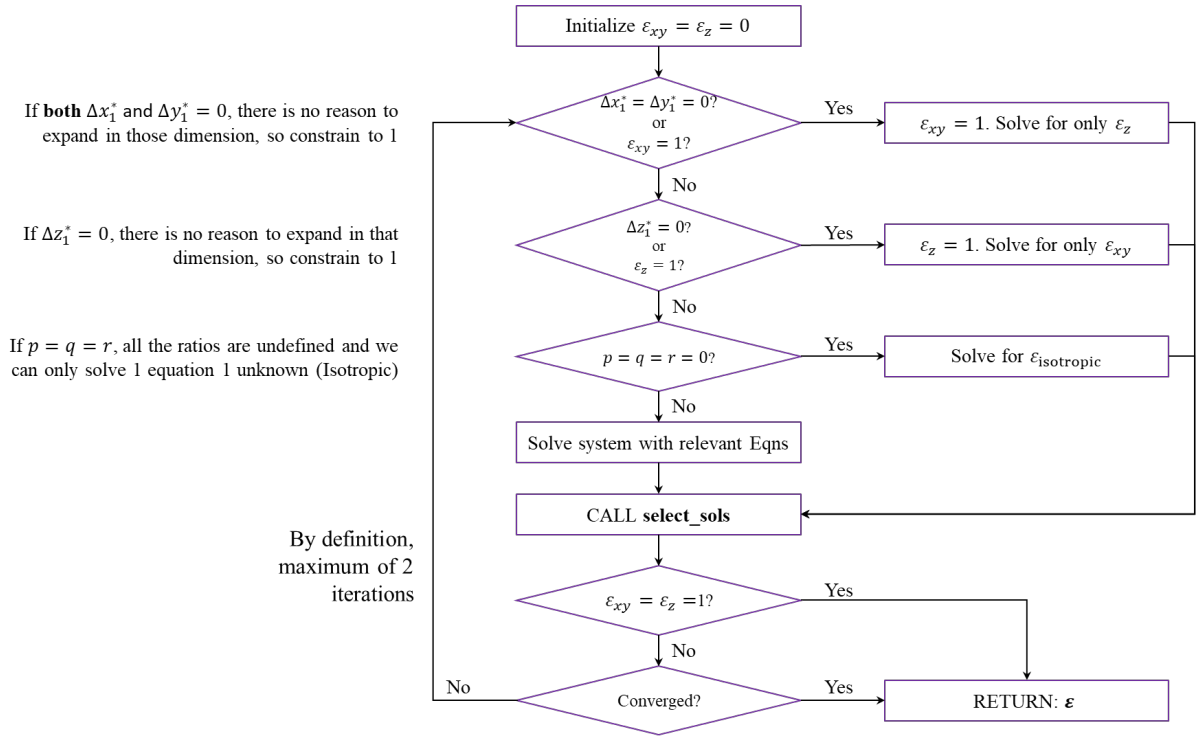


Figure E.6: The Eps_2 subroutine, which solves the expansion when $a = b \neq c$.

added complexity of the constraint on $\varepsilon_x = \varepsilon_y = \varepsilon_{xy}$. If $\Delta\hat{x}_1 = 0$ but $\Delta\hat{y}_1 \neq 0$ (and vice versa), it is still possible for ε_{xy} to contribute to separation through the non-zero dimension. Based on this, I only constrain $\varepsilon_x = \varepsilon_y = 1$ if both $\Delta\hat{x}_1$ and $\Delta\hat{y}_1 = 0$. For $\Delta\hat{z}_1$ there is no such complication. Other than this, the same algorithmic logic is followed, per the previous subroutine. Since there is a maximum of two independent expansion parameters to solve for, the procedure can only repeat a maximum of twice.

Lastly, in the case where $a = b = c$ (Figure E.7), we know that there can only be an isotropic expansion. Unlike before, there is no need to check if any $\Delta\hat{r}_1 = 0$, since if we apply the same logic as the prior cases, it would be necessary to constrain $\varepsilon_x = \varepsilon_y = \varepsilon_z = 1$ only if all three

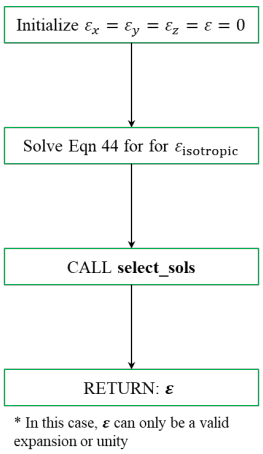


Figure E.7: The Eps_1 subroutine, which solves the expansion when $a = b = c$.

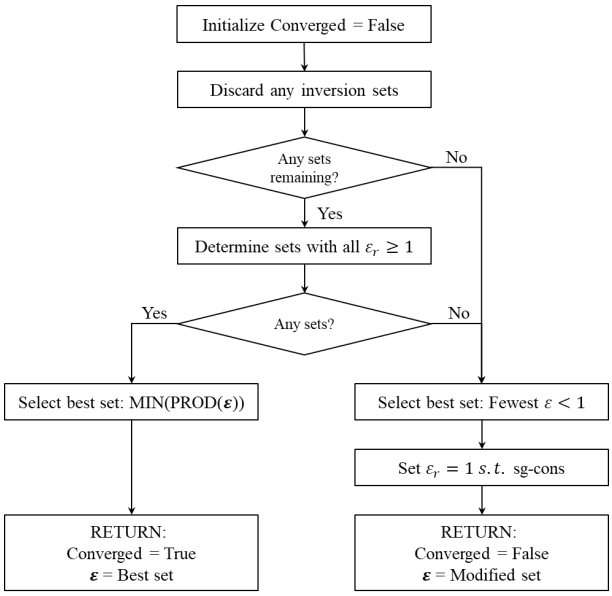


Figure E.8: The select_sol subroutine, which carries out the solution selection procedure.

$\Delta\hat{r}_1 = 0$. This, however, is already excluded by the Rule 1 check done in the Shift Algorithm subroutine. Moreover, the $p = q = r = 0$ check is irrelevant to the isotropic expansion since we do not utilize any of the ratio equations. In short, the distance equation can be directly solved and solution selection is executed before returning to the Shift Algorithm subroutine.

Finally, I explain the selection of solutions given a collection of solutions from the quadratic equations. For every quadratic equation solved, there will be two sets of candidates $\boldsymbol{\varepsilon}_{\text{candidate}}$. Given that some of the cases in Table E.1 have two quadratic solutions, this results in a maximum of four $\boldsymbol{\varepsilon}_{\text{candidate}}$. To recap, each set of solutions are classified according to the hierarchy of:

1. Fully expansionary, if *all* $\varepsilon_r \geq 1$
2. Contain contractions, if *any* $0 < \varepsilon_r < 1$, but no $\varepsilon_r < 0$
3. Contain inversions, if *any* $\varepsilon_r < 0$

Within the select_sol subroutine (Figure E.8), inversion solutions are initially disregarded as they are the most 'physically problematic'. If there are any sets remaining, the subroutine first tries to find a fully expansionary solution. If more than one $\boldsymbol{\varepsilon}_{\text{candidate}}$ is fully expansionary, the subroutine opts for the candidate that has the smallest product of expansion factors (i.e., taking the solution with the smallest value of $\varepsilon_x \times \varepsilon_y \times \varepsilon_z$). This is done in the spirit of minimizing the increase to the cell volume, otherwise there is greater risk of failing the density check post-Shift Algorithm, or adding unnecessary time to the subsequent energy minimization. Assessing the product of expansion factors is not an exact indicator of the change in cell volume, but this is a convenient proxy for rigorously determining each expanded cell volume.

In the event that no fully expansionary $\boldsymbol{\varepsilon}_{\text{candidate}}$ can be found, the subroutine will instead attempt to modify a solution which contains contractions (but not inversions). This simply

involves replacing any element in $\boldsymbol{\varepsilon}_{\text{candidate}}$ that was originally < 1 with 1. Note that space group constraints are naturally enforced here since the solution sets themselves are obtained from problem formulations that incorporate the space group constraints (e.g., if $a = b$ is a space group constraint, it is not possible for the candidate solution to have $\varepsilon_x > 1$ and $\varepsilon_y < 1$). The nature of this modification is that factors which had valid expansions are kept >1, while expansion factors which were already = 1 are left unchanged. Only elements with $\varepsilon < 1$ are modified. At the end, the modified $\boldsymbol{\varepsilon}$ will only contain elements ≥1. In this manner, if the solution generation loop is reran, elements in $\boldsymbol{\varepsilon}$ that are = 1 reflect the increasingly constrained problem. It should be stressed that select_sol will choose to modify the $\boldsymbol{\varepsilon}_{\text{candidate}}$ which has the fewest number of $\varepsilon_r < 1$. This is done on the principle of constraining as few factors as possible in order to maximize the possibility of finding a good expansion solution. If all $\boldsymbol{\varepsilon}_{\text{candidate}}$ contain an inversion, then the same modification procedure is ran, per the case with only solutions with contractions.

Appendix F

CE–755 Validation Against Experimental Sublimation Energies

Table F.1: Comparison of experimentally-derived lattice energies ($U_{\text{latt}}^{\text{exp}} \approx -\Delta H_{\text{sub}}^{\text{exp}} - 2RT$) with the DFT-D computed lattice energies ($U_{\text{latt}}^{\text{DFT}}$). The errors are computed as ($\Delta U_{\text{latt}} = U_{\text{latt}}^{\text{exp}} - U_{\text{latt}}^{\text{DFT}}$). In some cases, the same structure is referenced using different Refcodes in the CE–755 database and in the experimental data.

CE–755 Refcode	$U_{\text{latt}}^{\text{DFT}}$ (kJ/mol)	Exp. Refcode	Source	$\Delta H_{\text{sub}}^{\text{exp}}$ (kJ/mol)	T (K)	$U_{\text{latt}}^{\text{exp}}$ (kJ/mol)	ΔU_{latt} (kJ/mol)
ACETAC07	−73.15	ACETAC07	[316]	67.00	221.50	−70.68	2.46
ADAMAN08	−69.28	ADAMAN08	[264]	59.00	298.15	−63.96	5.32
AMBNAC12	−125.91	AMBNAC04	[317]	116.10	298.15	−121.06	4.85
AMBNAC12	−125.91	AMBNAC04	[318]	116.10	298.15	−121.06	4.85
AMPYRE	−104.43	AMPYRE	[319]	87.10	298.15	−92.06	12.37
AMPYRE	−104.43	AMPYRE01	[320]	88.10	298.15	−93.06	11.37
AMPYRE	−104.43	AMPYRE01	[319]	87.10	298.15	−92.06	12.37
AMPYRM11	−96.08	AMPYRM11	[319]	76.50	298.15	−81.46	14.62
ANTCEN16	−113.00	ANTCEN16	[269]	97.90	298.15	−102.86	10.14
BAGFIY01	−102.83	BAGFIY01	[321]	85.50	298.15	−90.46	12.37
BENZAC12	−102.67	BENZAC12	[322]	90.00	298.15	−94.96	7.71
BENZEN06	−55.02	BENZEN06	[270]	44.40	298.15	−49.36	5.66
BESKAL04	−127.39	BESKAL04	[323]	130.40	298.15	−135.36	−7.96
BIPYRL04	−96.88	BIPYRL04	[271]	81.80	298.15	−86.76	10.13
BIPYRL04	−96.88	BIPYRL04	[272]	81.90	298.15	−86.86	10.03
BIPYRL04	−96.88	BIPYRL04	[273]	75.00	298.15	−79.96	16.93
BNZQUI03	−76.24	BNZQUI03	[274]	66.70	298.15	−71.66	4.58
BNZQUI03	−76.24	BNZQUI03	[274]	66.70	298.15	−71.66	4.58
BNZQUI03	−76.24	BNZQUI03	[275]	62.80	298.15	−67.76	8.48
BNZQUI03	−76.24	BNZQUI03	[276]	62.76	298.15	−67.72	8.52
BZDMAZ02	−111.69	BZDMAZ02	[324]	101.20	298.15	−106.16	5.53

Continued on next page

Table F.1 – continued from previous page

CE–755 Refcode	$U_{\text{latt}}^{\text{DFT}}$ (kJ/mol)	Exp. Refcode	Source	$\Delta H_{\text{sub}}^{\text{exp}}$ (kJ/mol)	T (K)	$U_{\text{latt}}^{\text{exp}}$ (kJ/mol)	ΔU_{latt} (kJ/mol)
BZDMAZ02	−111.69	BZDMAZ02	[325]	107.00	298.15	−111.96	−0.27
BZDMAZ02	−111.69	BZDMAZ02	[324]	102.20	298.15	−107.16	4.53
BZDMAZ02	−111.69	BZDMAZ02	[285]	98.90	298.15	−103.86	7.83
BZDMAZ02	−111.69	BZDMAZ02	[326]	94.30	298.15	−99.26	12.43
BZPHAN01	−124.38	BZPHAN01	[277]	106.00	298.15	−110.96	13.42
CLBZAC01	−116.31	CLBZAC01	[327]	106.30	298.15	−111.26	5.05
CLBZAC01	−116.31	CLBZAC01	[328]	100.90	298.15	−105.86	10.45
CLMETH03	−33.15	CLMETH03	[278]	31.60	151.00	−34.11	−0.96
CRYSEN	−138.77	CRYSEN	[279]	123.40	298.15	−128.36	10.41
CUBANE	−56.48	CUBANE	[280]	55.20	298.15	−60.16	−3.68
CUKCIU03	−54.38	CUKCIU03	[281]	35.60	272.00	−40.12	14.26
CYHEXO	−86.79	CYHEXO	[282]	75.00	298.15	−79.96	6.83
DCLANT10	−123.74	DCLANT10	[283]	113.90	346.00	−119.65	4.09
DCLBQN	−82.73	DCLBQN	[276]	69.90	298.15	−74.86	7.88
DITHAN02	−84.60	DITHAN02	[284]	63.00	298.15	−67.96	16.65
DITHAN02	−84.60	DITHAN01	[285]	68.90	298.15	−73.86	10.75
DITHAN02	−84.60	DITHAN01	[284]	63.00	298.15	−67.96	16.65
DMSULO04	−83.79	DMSULO02	[286]	77.00	298.15	−81.96	1.83
DNBENZ11	−85.32	DNBENZ11	[287]	87.00	298.15	−91.96	−6.63
DTENYL02	−96.43	DTENYL02	[288]	85.20	298.15	−90.16	6.27
FABPON11	−138.61	FABPON11	[289]	123.30	298.15	−128.26	10.35
FORMAC02	−67.58	FORMAC02	[264]	60.50	274.50	−65.06	2.51
FULPIM	−96.27	FULPIM	[329]	88.40	298.15	−93.36	2.91
HCLBNZ13	−100.14	HCLBNZ13	[283]	90.50	298.15	−95.46	4.69
HEXANE01	−51.67	HEXANE01	[290]	50.80	178.00	−53.76	−2.09
HXMTAM10	−85.17	HXMTAM10	[291]	79.60	298.15	−84.56	0.61
HXMTAM10	−85.17	HXMTAM08	[292]	74.90	298.15	−79.86	5.31
HXQUIN14	−91.43	HXQUIN11	[267]	109.00	298.15	−113.96	−22.52
HXQUIN14	−91.43	HXQUIN11	[268]	108.80	298.15	−113.76	−22.32
HXQUIN14	−91.43	HXQUIN11	[266]	89.50	298.15	−94.46	−3.02
HXQUIN14	−91.43	HXQUIN11	[266]	89.00	298.15	−93.96	−2.52
HXQUIN14	−91.43	HXQUIN11	[267]	87.90	298.15	−92.86	−1.42
IMAZOL06	−91.12	IMAZOL06	[324]	81.40	298.15	−86.36	4.76
MALIAC12	−117.02	MALIAC12	[330]	105.40	298.15	−110.36	6.66
MELAMI06	−150.94	MELAMI06	[264]	123.30	432.00	−130.48	20.46
MELAMI06	−150.94	MELAMI05	[265]	120.00	298.15	−124.96	25.98
NAPHTA31	−83.65	NAPHTA31	[293]	72.70	298.15	−77.66	5.99
NAPHTA31	−83.65	NAPHTA15	[294]	73.20	298.15	−78.16	5.49

Continued on next page

Table F.1 – continued from previous page

CE–755 Refcode	$U_{\mathrm{latt}}^{\mathrm{DFT}}$ (kJ/mol)	Exp. Refcode	Source	$\Delta H_{\mathrm{sub}}^{\mathrm{exp}}$ (kJ/mol)	T (K)	$U_{\mathrm{latt}}^{\mathrm{exp}}$ (kJ/mol)	ΔU_{latt} (kJ/mol)
NAPHTA31	−83.65	NAPHTA15	[295]	73.00	298.15	−77.96	5.69
NAPHTA31	−83.65	NAPHTA15	[296]	72.90	298.15	−77.86	5.79
NAPHTA31	−83.65	NAPHTA15	[297]	72.70	298.15	−77.66	5.99
NAPHTA31	−83.65	NAPHTA15	[279]	72.60	298.15	−77.56	6.09
NAPHTA31	−83.65	NAPHTA15	[298]	72.60	298.15	−77.56	6.09
NAPHTA31	−83.65	NAPHTA15	[299]	72.50	298.15	−77.46	6.19
NAPHTA31	−83.65	NAPHTA15	[300]	72.40	298.15	−77.36	6.29
NAPHTA31	−83.65	NAPHTA15	[301]	72.30	298.15	−77.26	6.39
NAPHTA31	−83.65	NAPHTA15	[302]	72.10	298.15	−77.06	6.59
NAPHTA31	−83.65	NAPHTA15	[303]	70.40	298.15	−75.36	8.29
NAPHTA31	−83.65	NAPHTA15	[304]	66.50	298.15	−71.46	12.19
NAPHTA31	−83.65	NAPHTA15	[285]	64.00	298.15	−68.96	14.69
NAPOAC01	−127.98	NAPOAC01	[317]	113.60	298.15	−118.56	9.42
OCTANE10	−67.34	OCTANE10	[290]	68.10	216.00	−71.69	−4.35
PENTAN01	−43.33	PENTAN01	[290]	42.00	143.00	−44.38	−1.05
PERLEN06	−148.57	PERLEN06	[305]	129.60	415.00	−136.50	12.07
PFBZAC	−92.12	PFBZAC01	[332]	91.50	298.15	−96.46	−4.34
PHENAN08	−106.09	PHENAN08	[301]	90.90	298.15	−95.86	10.23
PHENAZ04	−106.40	PHENAZ04	[306]	94.30	354.00	−100.19	6.22
PNCLBZ	−93.14	PNCLBZ	[158]	87.03	298.00	−91.98	1.16
PRMDIN01	−65.09	PRMDIN02	[158]	58.30	298.15	−63.26	1.83
PTOXEC	−76.73	PTOXEC	[307]	87.90	298.15	−92.86	−16.12
PTOXEC	−76.73	PTOXEC	[158]	87.86	298.00	−92.82	−16.09
PYRAZI01	−65.26	PYRAZI01	[308]	56.20	298.15	−61.16	4.10
PYRENE03	−113.74	PYRENE03	[279]	100.30	298.15	−105.26	8.49
SALMID04	−115.45	SALMID02	[333]	101.90	298.15	−106.86	8.59
SALMID04	−115.45	SALMID02	[334]	99.30	298.15	−104.26	11.19
SUCACB03	−132.44	SUCACB03	[335]	123.10	298.15	−128.06	4.38
SUCANH12	−78.54	SUCANH12	[309]	80.70	298.15	−85.66	−7.12
SUCANH12	−78.54	SUCANH12	[309]	80.70	298.15	−85.66	−7.12
TCYETY01	−85.42	TCYETY01	[310]	81.20	298.15	−86.16	−0.73
TETDAM03	−70.45	TETDAM03	[311]	62.20	298.15	−67.16	3.29
TOXOCN	−72.01	TOXOCN	[312]	79.60	298.15	−84.56	−12.55
TOXOCN	−72.01	TOXOCN	[158]	79.50	298.00	−84.45	−12.44
TRAZOL03	−91.28	TRAZOL03	[336]	83.00	298.15	−87.96	3.32
TRAZOL03	−91.28	TRAZOL03	[158]	84.10	298.00	−89.05	2.22
TRIPHE13	−133.93	TRIPHE13	[279]	120.10	298.15	−125.06	8.88
TRITAN03	−109.71	TRITAN03	[313]	93.20	298.15	−98.16	11.56

Continued on next page

Table F.1 – continued from previous page

CE–755 Refcode	$U_{\mathrm{latt}}^{\mathrm{DFT}}$ (kJ/mol)	Exp. Refcode	Source	$\Delta H_{\mathrm{sub}}^{\mathrm{exp}}$ (kJ/mol)	T (K)	$U_{\mathrm{latt}}^{\mathrm{exp}}$ (kJ/mol)	ΔU_{latt} (kJ/mol)
TRIZIN01	−59.88	TRIZIN01	[314]	54.20	298.15	−59.16	0.72
TRIZIN01	−59.88	TRIZIN01	[158]	58.16	222.00	−61.85	−1.97
TRIZIN01	−59.88	TRIZIN01	[285]	56.50	298.15	−61.46	−1.58
TRIZIN01	−59.88	TRIZIN01	[314]	54.18	298.15	−59.14	0.74
TROXAN	−56.98	TROXAN	[315]	56.60	298.15	−61.56	−4.58
XAGWIM	−112.43	XAGWIM	[337]	100.80	298.15	−105.76	6.67
ZZZKPE03	−103.02	ZZZKPE03	[338]	85.90	343.00	−91.60	11.42

Appendix G

Training/Test Set Partitioning of the CE–755

Table G.1: Training set structures partitioned from the CE–755. This comprises 604 crystal structures: 356 non-hydrogen bonding (NHB) and 248 hydrogen bonding (HB) structures. No multicomponent crystals are included in the dataset, hence any structure with $Z' > 1$ has repeated units of the same molecule.

Refcode	HB?	Z'	nAtms	Chemical Formula
PERLEN06	NHB	1	32	$C_{20}H_{12}$
HEXANE01	NHB	1	20	C_6H_{14}
CUBANE	NHB	1	16	C_8H_8
PYRENE03	NHB	1	26	$C_{16}H_{10}$
PHENAN08	NHB	1	24	$C_{14}H_{10}$
PENTAN01	NHB	1	17	C_5H_{12}
NAPHTA31	NHB	1	18	$C_{10}H_8$
BAPOCM10	NHB	1	48	$C_{18}H_{30}$
CRYSEN	NHB	1	30	$C_{18}H_{12}$
MEYCEY	NHB	1	21	$C_{11}H_{10}$
BCPROP02	NHB	1	16	C_6H_{10}
BENZEN06	NHB	1	12	C_6H_6
ADAMAN08	NHB	1	26	$C_{10}H_{16}$
OCTANE10	NHB	1	26	C_8H_{18}
PRMDIN01	NHB	1	10	$C_4H_4N_2$
NAPTYR11	NHB	1	16	$C_8H_6N_2$
BAYJUG	NHB	1	30	$C_{10}H_{18}N_2$
BITVUV	NHB	1	12	C_4H_7N
PHENAZ04	NHB	1	22	$C_{12}H_8N_2$
XAKPAC01	NHB	1	20	$C_{10}H_5N_5$
BACJUN	NHB	2	10/10	$C_4H_4N_2$
TCYETY01	NHB	1	10	C_6N_4
TRIZIN01	NHB	1	9	$C_3H_3N_3$
TETDAM03	NHB	1	20	$C_6H_{12}N_2$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
BOXRUE	NHB	1	25	$C_7H_{11}N_7$
CUVWEV	NHB	1	19	C_9H_9N
HXMTAM10	NHB	1	22	$C_6H_{12}N_4$
TELPIJ	NHB	1	11	$C_3H_5N_3$
CUXXAW	NHB	1	26	$C_{14}H_9N_3$
PYRAZI01	NHB	1	10	$C_4H_4N_2$
DMDZCD	NHB	1	32	$C_{10}H_{20}N_2$
PYRDNA07	NHB	1	11	C_5H_5N
CYCYPR	NHB	1	12	$C_6H_3N_3$
TCYETY	NHB	1	10	C_6N_4
CTCDEC	NHB	1	25	$C_{11}H_{13}N$
QORPAR	NHB	1	18	$C_8H_5N_5$
DAZNAP	NHB	1	16	$C_8H_6N_2$
BEBXOV	NHB	1	19	$C_9H_6N_4$
BURXAN	NHB	1	24	$C_{10}H_{10}O_4$
BEZWIM	NHB	1	21	$C_8H_{10}O_3$
SUCANH12	NHB	1	11	$C_4H_4O_3$
TROXAN	NHB	1	12	$C_3H_6O_3$
AVIJEW	NHB	1	23	$C_8H_{12}O_3$
BONNEA	NHB	1	17	$C_7H_8O_2$
COUMAR14	NHB	2	17/17	$C_9H_6O_2$
CEGPOV	NHB	1	20	$C_9H_8O_3$
FURANE10	NHB	1	9	C_4H_4O
BNZQUI03	NHB	1	12	$C_6H_4O_2$
AXOSOW	NHB	1	8	C_3H_4O
DEXBEN01	NHB	1	26	$C_{10}H_{14}O_2$
AYOHEB	NHB	1	18	C_9H_8O
CUKCIU03	NHB	1	14	$C_4H_8O_2$
COUMAR01	NHB	1	17	$C_9H_6O_2$
BCYOAN10	NHB	1	23	$C_{10}H_{10}O_3$
BICVIS	NHB	1	18	$C_6H_8O_4$
PTOXEC	NHB	1	20	$C_5H_{10}O_5$
CYHEXO	NHB	1	16	$C_6H_8O_2$
CAMPQU03	NHB	1	26	$C_{10}H_{14}O_2$
BUTFUT	NHB	1	25	$C_9H_{12}O_4$
COUMAR13	NHB	1	17	$C_9H_6O_2$
ASUKAB	NHB	1	23	$C_9H_{12}O_2$
AMITEW	NHB	1	19	$C_{10}H_6O_3$
BAKQUA	NHB	1	26	$C_{10}H_{10}N_2O_4$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
BERPIX	NHB	1	31	$C_{13}H_{13}N_3O_2$
BAPBUQ03	NHB	1	27	$C_{14}H_{10}N_2O$
BARKAH	NHB	1	21	$C_9H_9NO_2$
AMUVEL	NHB	1	22	$C_{10}H_9NO_2$
AVABII	NHB	1	25	$C_{10}H_{13}NO$
ABORER01	NHB	1	27	$C_{14}H_{10}N_2O$
AKEFIH	NHB	1	23	$C_9H_9N_3O_2$
ACSKAT01	NHB	1	24	$C_{11}H_{11}NO$
APUHOJ	NHB	1	28	$C_{10}H_{14}N_2O_2$
BAZPUO	NHB	1	26	$C_9H_{12}N_2O_3$
BEKLIP	NHB	1	25	$C_{12}H_8N_2O_3$
ADIVUF	NHB	1	30	$C_{14}H_8N_6O_2$
BAPBUQ	NHB	1	27	$C_{14}H_{10}N_2O$
XAPCAS	NHB	1	9	C_3F_6
YICFEX01	NHB	1	15	C_7H_7F
VEHCET	NHB	1	18	$C_6H_8F_4$
FACFOE	NHB	1	12	$C_6H_4F_2$
ADAREE	NHB	1	18	$C_6H_6F_6$
BAZSEA	NHB	1	16	C_8F_8
AXUBUR01	NHB	1	14	C_8H_5F
XOGJEK	NHB	1	15	$C_7H_6F_2$
XOYDIB	NHB	1	22	$C_{10}H_7F_3N_2$
BOQSIJ	NHB	1	34	$C_{18}H_{14}FN$
DATLER	NHB	1	11	$C_5H_2F_3N$
DATLAN	NHB	1	11	$C_5H_3F_2N$
GOZXIF	NHB	1	18	$C_8H_6FN_3$
DAHXAO	NHB	1	28	$C_{12}H_{12}F_3N$
LOTFAE	NHB	1	25	$C_{13}H_9FN_2$
BOXQEK	NHB	1	31	$C_{17}H_{12}FN$
AQOZUB01	NHB	1	18	$C_{10}H_5FN_2$
CAXPAT	NHB	1	17	$C_9H_4F_3N$
AKOYAA	NHB	1	25	$C_{12}H_6F_6O$
FOQJIH	NHB	1	27	$C_{12}H_9F_3O_3$
DFNAPQ	NHB	1	18	$C_{10}H_4F_2O_2$
CAXPUN	NHB	1	17	$C_9H_5FO_2$
MOWDIO	NHB	1	22	$C_{12}H_4F_4O_2$
VUCJAH	NHB	1	23	$C_{11}H_6F_4O_2$
BIKCII	NHB	1	24	$C_{13}H_4F_6O$
AYUZAW	NHB	1	32	$C_{13}H_{15}F_2NO$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
EYAVEF	NHB	1	26	$C_8H_7F_6N_3O_2$
MOWDAG	NHB	1	24	$C_{14}H_4F_2N_2O_2$
SUGRUK	NHB	1	27	$C_{12}H_{10}F_3NO$
ADIGEA	NHB	1	25	$C_{10}H_9F_2N_3O$
AKUNUQ	NHB	1	22	$C_8H_7F_3N_2O_2$
GOSJEG	NHB	1	22	$C_{10}H_7F_2NO_2$
BONWAE	NHB	1	21	$C_7H_7F_3N_2O_2$
ELULIH	NHB	1	22	$C_{10}H_5F_3N_2O_2$
BETMET	NHB	1	15	$C_8H_3F_2NO$
DECPYR10	NHB	1	26	$C_{16}Cl_{10}$
ZZZBLP03	NHB	1	24	$C_{14}H_8Cl_2$
PNCLBZ	NHB	1	12	C_6HCl_5
DCLETH02	NHB	1	8	$C_2H_4Cl_2$
CLFULV01	NHB	2	18/18	$C_{10}Cl_8$
XAXCOQ03	NHB	1	8	C_2H_5Cl
AHCHEX01	NHB	1	18	$C_6H_6Cl_6$
ABUMIT	NHB	1	12	$C_6H_4Cl_2$
BAHCUI	NHB	1	23	$C_9H_6Cl_8$
HCCYHD	NHB	1	18	$C_6H_6Cl_6$
RIVDOQ	NHB	1	8	$C_2H_4Cl_2$
PIRPEL	NHB	1	18	$C_6H_6Cl_6$
MECHLF12	NHB	1	8	$C_2H_3Cl_3$
MCBENZ	NHB	2	12/12	C_6H_5Cl
YAFHUI	NHB	1	18	$C_{10}H_2Cl_6$
CLFORM03	NHB	1	5	$CHCl_3$
FEMKUD	NHB	1	19	$C_7H_8Cl_4$
HCCYHG02	NHB	1	18	$C_6H_6Cl_6$
TCHLBZ	NHB	1	12	$C_6H_3Cl_3$
DCLANT10	NHB	1	24	$C_{14}H_8Cl_2$
BOTYIT	NHB	1	10	$C_4H_2Cl_2N_2$
GUJLAZ	NHB	1	15	$C_7H_4ClN_3$
EWUZAZ	NHB	1	13	$C_7H_3Cl_2N$
DCPYAZ	NHB	1	10	$C_4H_2Cl_2N_2$
IBUZIP	NHB	1	22	$C_{12}H_6Cl_2N_2$
BAWNIW02	NHB	3	14/14/14	$C_8Cl_4N_2$
CLAZBZ02	NHB	1	24	$C_{12}H_8Cl_2N_2$
BAWNIW	NHB	1	14	$C_8Cl_4N_2$
CLBECN03	NHB	1	13	$C_7H_2Cl_3N$
DOJFAL	NHB	1	20	$C_8H_6Cl_3N_3$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
BOPHUJ	NHB	2	18/18	$C_8H_7ClN_2$
GOLYAI	NHB	1	16	$C_8H_5ClN_2$
HCZCHP	NHB	1	16	$C_6H_3Cl_6N$
DAYFIV	NHB	1	13	$C_7H_3Cl_2N$
XOZFOK	NHB	1	26	$C_{12}H_{11}ClO_2$
CLBENQ	NHB	1	12	$C_6H_3ClO_2$
LEJDEL	NHB	1	22	$C_{12}H_7ClO_2$
WERNUF01	NHB	1	17	$C_8H_6Cl_2O$
AHUGUH	NHB	1	19	$C_7H_7Cl_3O_2$
DCLBQN	NHB	1	12	$C_6H_2Cl_2O_2$
LEJDAH	NHB	1	22	$C_{12}H_7ClO_2$
HUGQOQ	NHB	1	14	$C_7H_4Cl_2O$
BAGYIR	NHB	1	14	$C_4H_5Cl_3O_2$
GISXEM	NHB	1	16	$C_7H_5Cl_3O$
BEGPIN	NHB	2	24/24	$C_7H_9ClN_4O_3$
COHSAV	NHB	1	24	$C_{10}H_9Cl_2NO_2$
BEGCOF	NHB	1	30	$C_{12}H_{13}ClN_2O_2$
CETLAQ	NHB	1	16	$C_8H_4ClNO_2$
AKETAN	NHB	1	29	$C_{13}H_{10}ClN_3O_2$
GOQJUU	NHB	1	20	$C_9H_6Cl_2N_2O$
BAFLEZ	NHB	1	19	$C_7H_5ClN_2O_4$
BEGPEJ	NHB	1	24	$C_7H_9ClN_4O_3$
NAFDAC	NHB	1	12	C_6H_4FCl
NAFCUV	NHB	1	12	C_6H_4FCl
XEZHAM	NHB	1	22	$C_{12}H_8FCl$
XEZHEQ	NHB	2	22/22	$C_{12}H_8FCl$
FAVTAX	NHB	1	22	$C_{12}H_4F_4Cl_2$
AQUSEM	NHB	2	21/21	$C_9H_6F_3ClN_2$
AWACAE	NHB	2	16/16	$C_8H_4FClN_2$
RAGPEX	NHB	2	17/17	$C_9H_4F_2ClN$
JEHYIF	NHB	1	12	$C_6HFCl_2N_2$
RAGPEX01	NHB	1	17	$C_9H_4F_2ClN$
LAPLAR01	NHB	1	16	$C_8H_3FCl_2N_2$
AJOYEF	NHB	1	23	$C_{11}H_7F_2ClO_2$
ILAMUF	NHB	1	14	$C_7H_3F_2ClO$
ILAMOZ	NHB	1	14	C_7H_4FClO
BAYXAD	NHB	2	26/26	$C_{13}H_6F_5ClO$
KUSQOG	NHB	1	20	$C_9H_6F_3ClO$
JIHNOG	NHB	1	17	$C_7H_2F_3Cl_2NO_2$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
IYIXUJ	NHB	1	23	$C_{11}H_7FClNO_2$
IYIYAQ	NHB	1	20	$C_{10}H_4F_2ClNO_2$
QUZTAH	NHB	1	22	$C_{10}H_2F_3Cl_4NO_2$
NPDTOL02	NHB	2	30/30	$C_{18}H_8S_4$
AKOZOR	NHB	1	30	$C_{18}H_8S_4$
CUGCUF	NHB	1	30	$C_{16}H_{12}S_2$
ACUWAZ	NHB	1	30	$C_{18}H_{10}S_2$
LIQWEN	NHB	1	12	C_3H_8S
AYOJAZ	NHB	1	18	C_9H_8S
TRITAN03	NHB	1	12	$C_3H_6S_3$
ASOPAB	NHB	1	29	$C_{14}H_{10}S_5$
DITHAN02	NHB	1	14	$C_4H_8S_2$
DTENYL02	NHB	1	16	$C_8H_6S_2$
ZAQFAB	NHB	1	18	$C_8H_4S_6$
KOZXIJ	NHB	1	34	$C_{18}H_{14}S_2$
PADQAN	NHB	1	9	C_2H_6S
YOTTOT	NHB	1	26	$C_{15}H_9NS$
DIZVOZ	NHB	1	18	$C_7H_6N_4S$
AYAVOM	NHB	1	21	$C_{10}H_8N_2S$
DEWYOV	NHB	1	28	$C_{11}H_{14}N_2S$
AXORIQ	NHB	1	15	$C_4H_7N_3S$
FANJOT01	NHB	1	16	$C_6H_6N_2S_2$
POZYUB02	NHB	1	19	$C_{10}H_6N_2S$
QUCGAZ	NHB	1	24	$C_{13}H_8N_2S$
POZYUB04	NHB	1	19	$C_{10}H_6N_2S$
ZUCMUI	NHB	1	27	$C_{12}H_{12}N_2S$
AFEWAK	NHB	1	18	$C_9H_5N_3S$
ZUJSIJ	NHB	1	19	$C_{10}H_7NS$
DMITDZ	NHB	1	17	$C_6H_7N_3S$
TOVDIU	NHB	1	29	$C_{14}H_{12}O_2S$
DMETSO	NHB	1	10	C_2H_6OS
DMETSO05	NHB	1	10	C_2H_6OS
AKOXUV	NHB	1	19	$C_8H_8OS_2$
GUFJAS	NHB	1	13	C_5H_6OS
CANRAM	NHB	1	30	$C_{15}H_{12}O_2S$
GUFJIA	NHB	1	18	$C_7H_8O_2S$
ASIBOT	NHB	1	27	$C_{10}H_{14}OS_2$
CABYAG	NHB	1	17	C_9H_6OS
BOZFEC	NHB	1	26	$C_{12}H_{10}O_2S_2$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
CHPDTP	NHB	1	20	$C_{11}H_6OS_2$
GOWPUG	NHB	1	26	$C_{11}H_{10}O_3S_2$
ATOMAZ	NHB	2	30/30	$C_{14}H_8O_6S_2$
KUGKIJ	NHB	1	36	$C_{16}H_{15}NO_3S$
XUMCIU	NHB	1	30	$C_{13}H_{13}NO_2S$
SUFKUC	NHB	1	24	$C_{12}H_9NOS$
PUHBIG	NHB	1	40	$C_{19}H_{17}NO_2S$
ZUJWEJ	NHB	2	33/33	$C_{17}H_{13}NOS$
XOZNIM	NHB	1	43	$C_{20}H_{17}N_3O_2S$
ZUJRII	NHB	1	23	$C_{11}H_9NOS$
COXKAE	NHB	1	34	$C_{12}H_{16}N_2O_3S$
XOYJED	NHB	1	35	$C_{17}H_{15}NOS$
XUMBIT	NHB	1	33	$C_{14}H_{15}NO_2S$
NOTYAZ	NHB	1	23	$C_9H_9NO_3S$
LINRUV	NHB	1	24	$C_{10}H_{10}F_2S_2$
VAFFUF	NHB	1	26	$C_{10}H_8F_6S_2$
IJIYAD	NHB	1	27	$C_{13}H_4F_8S_2$
FEBROU	NHB	1	27	$C_{13}H_4F_8S_2$
GOCPEW	NHB	1	24	$C_{12}H_6F_2N_2S_2$
OPAJIA	NHB	1	24	$C_{13}H_4F_5NS$
NOSNOY	NHB	1	25	$C_{10}H_9F_3N_2S$
GAXKAU	NHB	1	26	$C_{11}H_{11}F_2NS$
LIQJUS	NHB	1	15	$C_6H_5F_2NS$
COZXAT	NHB	1	28	$C_{13}H_7F_6NS$
DFPSLO	NHB	1	25	$C_{12}H_8F_2O_2S$
LUZXUA	NHB	1	28	$C_{14}H_9F_3OS$
NUQGUD	NHB	1	19	$C_8H_7FO_2S$
XUPPEE	NHB	1	19	$C_7H_5FO_4S_2$
YELLIN	NHB	1	19	$C_{10}H_2F_4O_2S$
COWTIT	NHB	1	21	$C_{10}H_5F_3O_2S$
UKUNAR	NHB	1	28	$C_{14}H_9F_3OS$
HABVIQ	NHB	1	27	$C_{13}H_{12}N_2$
JOSLOU	NHB	1	25	$C_{13}H_{11}N$
IDAHUR	NHB	1	14	C_6H_6FN
FIRLAU	NHB	1	10	$C_4H_3F_2N$
CEBZIU	NHB	1	30	$C_{10}H_{11}F_3N_2O_4$
QOFGAV	NHB	1	27	$C_{11}H_{12}O_2S_2$
ACENAP03	NHB	2	22/22	$C_{12}H_{10}$
BZPHAN01	NHB	1	30	$C_{18}H_{12}$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
CIMHUB	NHB	1	38	$C_{14}H_{24}$
HESDUE	NHB	2	20/20	$C_{12}H_8$
AYUXAU	NHB	1	32	$C_{16}H_{16}$
DIMNAP	NHB	1	24	$C_{12}H_{12}$
BCODYN01	NHB	1	24	$C_{16}H_8$
CIKBOO	NHB	1	34	$C_{18}H_{16}$
DACXIQ	NHB	1	30	$C_{16}H_{14}$
BASCON01	NHB	1	28	$C_{16}H_{12}$
BNORBN11	NHB	1	36	$C_{14}H_{22}$
AYOHOL01	NHB	1	16	C_9H_6O
ACNAQU01	NHB	1	20	$C_{12}H_6O_2$
BODCOM	NHB	1	18	$C_{10}H_4O_4$
ACETPH	NHB	1	17	C_8H_8O
BETBEH10	NHB	1	20	$C_{10}H_8O_2$
AWODIB	NHB	1	25	$C_{15}H_7N_3$
BUPQAE01	NHB	1	16	C_8H_7N
CEKGUU01	NHB	1	13	C_3H_9N
ADIVIT	NHB	1	22	$C_{12}H_4N_6$
BEKFEF	NHB	1	16	$C_6H_8N_2$
AMIMZE10	NHB	1	22	$C_8H_{10}N_4$
AZPACN	NHB	1	22	$C_{12}H_8N_2$
ABOVIY03	NHB	1	32	$C_{18}H_{12}N_2$
CECMOM	NHB	1	16	$C_8H_5NO_2$
ADADUG	NHB	1	44	$C_{22}H_{18}N_2O_2$
ADOXID	NHB	1	17	$C_5H_7N_3O_2$
AHIVEU	NHB	1	28	$C_{12}H_{11}NO_4$
VORKUJ	NHB	1	18	$C_7H_6S_5$
DMSULO04	NHB	1	11	$C_2H_6O_2S$
SAZZUO	NHB	1	12	$C_4H_4S_4$
BUTSUL10	NHB	1	15	$C_4H_8O_2S$
YINTUK	NHB	1	52	$C_{20}H_{30}S_2$
ZOBKAC	NHB	1	21	$C_9H_{10}S_2$
XAFVEH	NHB	1	16	C_7H_8S
THBZPS10	NHB	1	24	$C_{13}H_{10}S$
MAJJOX	NHB	1	17	$C_9H_6S_2$
THIEPO10	NHB	1	15	$C_6H_6O_2S$
TANTOX03	NHB	1	24	$C_{12}H_8O_2S_2$
BZTIEP	NHB	1	21	$C_{10}H_8O_2S$
NEFMIW	NHB	1	21	$C_9H_8O_3S$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
DBTHPS04	NHB	1	23	$C_{12}H_8O_2S$
CICVEQ	NHB	1	21	$C_8H_{10}O_2S$
HTIEPO	NHB	1	17	$C_6H_8O_2S$
THPYDO	NHB	1	14	$C_5H_6O_2S$
DOXJAD	NHB	1	20	$C_9H_8O_2S$
TOXANO	NHB	1	24	$C_{13}H_8O_2S$
DOXDIS10	NHB	1	8	$C_3O_2S_3$
SOBPAB	NHB	1	24	$C_{11}H_8O_4S$
AQIGOW	NHB	1	22	$C_9H_{10}OS_2$
POTLEP	NHB	1	16	$C_7H_4O_3S_2$
NOCMEX	NHB	1	21	$C_8H_6O_2S_5$
GUNYIY	NHB	1	25	$C_9H_9NO_4S_2$
IFEVAQ	NHB	1	23	$C_{10}H_8O_2S_3$
KADCIC	NHB	1	23	$C_{10}H_8O_2S_3$
CIGYIA	NHB	1	10	$C_2H_4O_2S_2$
JARDIQ	NHB	1	11	$C_2H_4O_2S_3$
DAQNEP	NHB	1	20	$C_{10}H_6O_2S_2$
DIMYON10	NHB	1	13	$C_4HClN_2O_4S$
GUJZER	NHB	1	17	$C_9H_4Cl_2OS$
GUZKAQ	NHB	1	20	$C_{11}H_6ClNS$
CIWBAL10	NHB	1	15	$C_8H_4Cl_2S$
HIGKOY	NHB	1	15	$C_7H_3Cl_2NOS$
ACIQUB	NHB	1	13	$C_6Cl_4N_2S$
BEMKAI	NHB	1	16	$C_6H_2ClNO_3S_3$
HIMRIH	NHB	1	10	$C_4Cl_2N_2S_2$
CBITAZ	NHB	1	14	C_7H_4ClNS
CTHDZO	NHB	1	9	$C_3Cl_2N_2OS$
IPUVIA	NHB	1	15	$C_4H_5N_3O_2S$
BIRYIM	NHB	1	24	$C_{10}H_9NO_3S$
KABJED	NHB	1	19	$C_6H_6N_4O_2S$
EQIREE	NHB	1	23	$C_{10}H_8ClNO_2S$
FIQTIJ	NHB	1	19	$C_8H_7NO_2S$
BAFLUR	NHB	1	22	$C_8H_8N_4OS$
COLGAO	NHB	1	25	$C_9H_{11}NO_3S$
DUJHEV	NHB	1	18	$C_6H_7NO_3S$
HESTUV	NHB	1	21	$C_8H_7N_3O_2S$
THTOXB	NHB	1	12	$C_2F_4O_4S_2$
TOPTEZ	NHB	1	23	$C_{10}H_5Cl_3O_4S$
CLPSUL02	NHB	1	25	$C_{12}H_8Cl_2O_2S$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
KONKAZ	NHB	1	17	$C_7H_6F_2OS$
FBUOFB	NHB	1	10	$C_2F_4O_2S_2$
FOHXUV	NHB	1	10	$C_2F_2Cl_2O_2S_2$
ZZZDQA01	NHB	1	18	$C_7H_7ClO_2S$
CTHETO	NHB	1	12	$C_3H_5ClO_2S$
QIBRAU	NHB	1	25	$C_{12}H_9ClO_2S$
KONJUS	NHB	1	17	$C_7H_6F_2OS$
WULPID	NHB	1	22	$C_{10}H_6Cl_2O_3S$
DUJQEF	NHB	1	11	$C_4H_2Cl_2O_2S$
VAZGIM	NHB	1	14	$C_5H_4Cl_2O_2S$
KIMNAV	NHB	1	11	$C_2F_4O_3S_2$
CORDUI	NHB	1	11	$C_2H_5ClO_2S$
JEWBAP	NHB	1	23	$C_{10}H_9ClO_2S$
YIQBIK	NHB	1	19	$C_7H_7ClO_3S$
VALVAG	NHB	1	20	$C_7H_6F_4O_2S$
IGAJIK	NHB	1	25	$C_{12}H_8F_2O_2S$
VINSAN	NHB	1	16	$C_8H_3F_4N$
DCDHNQ01	NHB	1	20	$C_{10}H_4Cl_2O_4$
LIWFUT	NHB	1	19	$C_9H_7NO_2$
MEYNAF	NHB	1	16	$C_6H_7NO_2$
BIGUAN01	HB	1	14	$C_2H_7N_5$
FITZEO01	HB	1	26	$C_{14}H_9N_3$
MELAMI06	HB	1	15	$C_3H_6N_6$
PIWXEY	HB	1	22	$C_6H_{14}N_2$
TRAZOL03	HB	1	8	$C_2H_3N_3$
PURINE	HB	1	13	$C_5H_4N_4$
BZDMAZ02	HB	1	15	$C_7H_6N_2$
OCAPUE	HB	1	11	$C_5H_4N_2$
AMPYRE	HB	1	13	$C_5H_6N_2$
AMMEPY01	HB	1	16	$C_6H_8N_2$
IMAZOL06	HB	1	9	$C_3H_4N_2$
ITOBAU	HB	1	17	$C_5H_{11}N$
METHOL04	HB	1	6	CH_4O
ACETAC07	HB	1	8	$C_2H_4O_2$
BENZAC12	HB	1	15	$C_7H_6O_2$
MALIAC12	HB	1	12	$C_4H_4O_4$
MALIAC13	HB	1	12	$C_4H_4O_4$
TELXAJ	HB	1	12	$C_4H_6O_2$
NAPOAC01	HB	1	21	$C_{11}H_8O_2$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
SUCACB03	HB	1	14	$C_4H_6O_4$
JUBBIR	HB	1	24	$C_{12}H_8O_4$
BESKAL04	HB	1	17	$C_7H_6O_4$
BESGEL	HB	1	23	$C_{13}H_8O_2$
ABINOS02	HB	1	20	$C_5H_{10}O_5$
SALMID07	HB	1	17	$C_7H_7NO_2$
FUMRAM01	HB	1	13	$C_4H_5NO_3$
SALMID04	HB	1	17	$C_7H_7NO_2$
KUXPUP	HB	1	15	$C_6H_5NO_3$
KUXPUP01	HB	1	15	$C_6H_5NO_3$
AZUDIJ	HB	1	24	$C_8H_{11}N_3O_2$
SEPNUX	HB	1	15	$C_5H_4N_2O_4$
REZDUY	HB	1	15	$C_5H_5N_3O_2$
OVUKAS	HB	1	27	$C_{10}H_{10}N_4O_3$
MALEHY14	HB	1	12	$C_4H_4N_2O_2$
MALEHY11	HB	1	12	$C_4H_4N_2O_2$
XAGWIM	HB	1	13	$C_5H_5NO_2$
KUFVAI03	HB	1	20	$C_9H_8N_2O$
ZAPGII	HB	1	15	$C_3H_5N_5O_2$
EGIRUJ	HB	1	22	$C_{10}H_8N_2O_2$
EDOMUI	HB	1	21	$C_8H_9NO_3$
QICVAZ01	HB	2	11/11	$C_3H_3N_3O_2$
HOJZUB	HB	1	14	$C_4H_4F_3NO_2$
JEGTUN	HB	1	18	$C_7H_5F_3N_2O$
GEHSIW	HB	1	19	$C_9H_4F_3NO_2$
RAGQIC	HB	1	19	$C_9H_5F_3N_2$
MUMSET	HB	1	11	$C_4HF_3N_2O$
CAFVAG	HB	1	16	$C_6H_4F_4N_2$
GEMRIB	HB	1	22	$C_9H_9F_2NO$
APUREK	HB	1	16	$C_6H_3F_4N_3$
GANBEF	HB	1	23	$C_9H_7F_4NO_2$
BENAFP	HB	1	16	C_7H_6FNO
IKESIC	HB	1	16	$C_7H_2F_4O_3$
UDUXEY	HB	1	12	C_5HF_4NO
FACETC01	HB	1	8	$C_2H_3FO_2$
COVJIG	HB	1	15	$C_7H_5FO_2$
DUTREQ	HB	1	15	$C_7H_4F_2O_2$
WESBEB	HB	1	19	$C_8H_7FO_3$
ODUSEM	HB	1	18	$C_8H_2F_4O_4$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
PFBZAC	HB	1	15	$C_7HF_5O_2$
GISZIS	HB	1	15	$C_7H_4F_2O_2$
AZCFOM01	HB	1	14	$C_2H_4Cl_2N_6$
AMCLPY	HB	1	13	$C_5H_5ClN_2$
CLTRZL	HB	1	8	$C_2H_2ClN_3$
AGIDOL	HB	1	23	$C_{11}H_8ClNO_2$
ABEROP	HB	1	18	$C_8H_6ClN_3$
ASATET	HB	1	16	C_7H_6ClNO
CANAPQ01	HB	2	20/20	$C_{10}H_6ClNO_2$
ASUWAN	HB	1	16	$C_6H_6ClNO_2$
GOTJUW	HB	1	15	$C_7H_5ClN_2$
BOMSED	HB	1	21	$C_7H_9ClN_2O_2$
BEXHOE	HB	1	17	$C_8H_4ClNO_3$
AFIGUS	HB	1	19	$C_8H_7Cl_2NO$
AZERUU	HB	2	17/17	$C_8H_5ClO_3$
CHPYRD	HB	1	12	C_5H_4ClNO
GADHAU	HB	1	9	$C_3HCl_3O_2$
DCDHNQ	HB	1	20	$C_{10}H_4Cl_2O_4$
CHXIQL	HB	1	18	C_9H_6ClNO
HOFHEQ	HB	1	14	$C_6H_4Cl_2O_2$
CLBZAC01	HB	2	15/15	$C_7H_5ClO_2$
COTFEX	HB	1	26	$C_{12}H_8Cl_2O_4$
BETXOQ	HB	1	18	$C_8H_4Cl_2O_4$
DECZEQ	HB	1	21	$C_9H_8ClNO_2$
MOVJAL	HB	2	10/10	$C_3H_4N_2S$
AKOVOL02	HB	2	12/12	C_5H_5NS
PHESAZ08	HB	1	23	$C_{12}H_9NS$
BONVOS	HB	1	20	$C_9H_7N_3S$
AGAKUP01	HB	1	27	$C_{14}H_9N_3S$
AGAKUP	HB	2	27/27	$C_{14}H_9N_3S$
AKOVOL	HB	1	12	C_5H_5NS
DERBEH	HB	2	16/16	$C_4H_6O_4S_2$
DITFIX	HB	1	38	$C_{14}H_{18}O_2S_4$
AQIGUC	HB	1	30	$C_{14}H_{12}O_2S_2$
CAPMSU	HB	1	20	$C_8H_8O_3S$
CAFKEZ	HB	1	26	$C_{10}H_{12}O_3S$
APOXOS01	HB	1	26	$C_{10}H_{10}N_4OS$
HTDZDX10	HB	1	15	$C_3H_3N_3O_5S$
HEVZAL	HB	1	25	$C_{10}H_7N_3O_4S$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
LELVIK	HB	2	18/18	$C_4H_9NO_3S$
BEJWAR	HB	1	19	$C_5H_{10}O_4$
AHEJAZ07	HB	1	20	$C_8H_{10}O_2$
FURDCB01	HB	1	15	$C_6H_4O_5$
BAGFIY	HB	1	16	$C_6H_8N_2$
ATCPEN01	HB	1	14	$C_6H_4N_4$
ACYMPZ	HB	1	16	$C_6H_5N_5$
BAZOCT	HB	1	24	$C_{10}H_{12}N_2$
ATCPEN	HB	1	14	$C_6H_4N_4$
BAGFIY01	HB	1	16	$C_6H_8N_2$
AMNTPY	HB	1	15	$C_5H_5N_3O_2$
AMNPHA	HB	1	17	$C_6H_6N_2O_3$
BOTZIU	HB	1	19	C_8H_9NS
AKEFUT	HB	1	34	$C_{11}H_{19}N_3S$
QEKXUZ	HB	1	13	$C_3H_8S_2$
AFEPAD	HB	1	24	$C_9H_{12}N_2S$
QEKXOT	HB	1	10	$C_2H_6S_2$
ATTDAZ02	HB	1	10	$C_2H_3N_3S_2$
QEKYAG	HB	1	16	$C_4H_{10}S_2$
MINSUZ01	HB	1	14	$C_6H_6S_2$
DASWOM	HB	1	23	$C_9H_{10}N_2OS$
CABPYR11	HB	1	15	$C_6H_6N_2S$
CUZXOK	HB	1	20	$C_5H_{11}N_3S$
CORZIT	HB	1	22	$C_{10}H_8N_2OS$
BZOXZT	HB	1	15	C_7H_5NOS
CASHOT03	HB	1	11	$C_3H_5NS_2$
BIPTID01	HB	1	19	$C_7H_8N_2OS$
CEHQEM	HB	1	12	$C_3H_3N_3S_3$
CEBDES	HB	1	13	$C_4H_5NOS_2$
CONFAM	HB	1	10	C_2H_5NOS
YIFGAW	HB	1	14	$C_3H_6N_2O_2S$
TAVMAH	HB	1	23	$C_9H_{10}N_2OS$
ZUNMOM	HB	1	24	$C_9H_8N_2O_3S_2$
AZANIB	HB	1	32	$C_{12}H_{15}N_3OS$
UMIQEO	HB	1	16	$C_5H_5NO_3S_2$
PAHPEV	HB	1	24	$C_{10}H_8N_2O_3S$
OLUCUT	HB	1	15	$C_6H_6O_2S$
MPYDCY	HB	1	15	$C_6H_5NO_2S$
CAGJEC01	HB	1	14	$C_3H_6N_4S$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
FEDKIJ	HB	1	19	$C_8H_6N_2S_3$
ANLINT20	HB	1	16	$C_7H_6N_2S$
DOHGOW	HB	1	12	$C_3H_4N_2S_3$
CAJWUG02	HB	1	16	$C_7H_6N_2S$
JIWWUK	HB	1	15	$C_6H_5N_3S$
IKIGEO	HB	1	15	$C_6H_6N_2S$
BOTZOA	HB	1	18	$C_7H_6N_4S$
ANTZOL10	HB	1	22	$C_9H_{10}N_2S$
BEZSAC	HB	1	17	$C_8H_5N_3S$
DSCARA01	HB	1	14	$C_4H_6O_2S_2$
TPENAC02	HB	1	12	$C_5H_4O_2S$
RONVAR01	HB	1	16	$C_7H_6O_2S$
EVOQEL	HB	1	15	$C_7H_4O_3S$
RONVEV	HB	1	16	$C_7H_6OS_2$
VINWUM	HB	1	16	$C_4H_8O_2S_2$
AZAZEI	HB	1	15	$C_6H_4O_4S$
TADCOU	HB	1	19	$C_8H_8O_2S$
WOPCEM	HB	1	18	$C_6H_6O_3S_3$
VINXAT	HB	1	19	$C_5H_8O_4S_2$
VINXEX	HB	1	19	$C_5H_8O_4S_2$
LOMWOZ	HB	1	19	$C_6H_7NO_3S_2$
HUCDIU	HB	1	14	$C_6H_3ClN_2OS$
CANVAP	HB	1	25	$C_{10}H_{10}N_2O_2S$
IGINUI	HB	1	19	$C_5H_7N_3O_3S$
ATDZDX	HB	1	14	$C_3H_5N_3O_2S$
BIXJAU	HB	1	16	$C_3H_8N_2O_2S$
BETBOR	HB	1	18	$C_5H_8N_2O_2S$
DEGWAQ	HB	1	22	$C_7H_{10}N_2O_2S$
FUDTUV	HB	1	21	$C_{10}H_7NO_2S$
CEFMUX	HB	1	22	$C_9H_8N_2O_2S$
BUDYEF	HB	1	22	$C_7H_{10}N_2O_2S$
MIFSID	HB	1	22	$C_8H_{10}O_3S$
HYTHIO	HB	1	13	$C_3H_6O_3S$
ELOBOW	HB	1	28	$C_9H_{13}NO_4S$
HAKPIW	HB	1	25	$C_7H_{13}NO_3S$
MUCHUP	HB	1	13	$CH_4O_6S_2$
LOJCOD	HB	1	24	$C_9H_{10}O_4S$
DIRMIA	HB	1	11	CH_5NO_3S
GULDIB	HB	1	30	$C_9H_{16}O_4S$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
PAWZAQ	HB	1	24	$C_9H_7F_3O_4S$
XUCKEN	HB	1	20	$C_{10}H_2F_6N_2$
GUFMAV	HB	1	14	$C_6H_2F_4O_2$
MEWZIZ	HB	1	12	$C_3H_3F_3O_3$
BIJVEY	HB	1	12	$C_4H_3F_3O_2$
LIDGEL	HB	1	18	$C_8HF_7N_2$
REZBIJ	HB	1	18	$C_8H_6F_2N_2$
IJESIC	HB	1	12	$C_2H_4FN_5$
WIXKOF	HB	1	15	$C_5H_4F_3N_3$
FLUANA	HB	1	14	$C_6H_2F_2O_4$
EWAZEK	HB	1	17	$C_7H_5F_2N_3$
IPUNEN	HB	1	12	$C_4H_5FO_2$
NIFKIW	HB	1	13	$C_3H_3F_3N_4$
CIYNED01	HB	1	20	$C_9H_5F_3O_3$
HOLTOS	HB	1	15	$C_7H_3ClN_4$
CANNUA	HB	1	17	$C_7H_7ClO_2$
CIPZIL	HB	1	18	$C_8H_2Cl_4O_4$
JEVVEM	HB	1	16	$C_6H_7ClN_2$
FELPIV	HB	1	10	$C_2HCl_3N_4$
IPAPEV	HB	1	18	$C_8H_4Cl_2N_2O_2$
COMFRN	HB	1	17	$C_7H_5ClO_4$
CONVIL01	HB	1	15	$C_6H_5ClO_3$
DUCSUQ	HB	1	11	$C_3H_2Cl_3N_3$
AMCPHO	HB	1	15	C_6H_6ClNO
CLANAC11	HB	1	14	$C_6H_2Cl_2O_4$
HIHNIV01	HB	1	15	$C_5H_5N_3O_2$
AMBACO08	HB	1	17	$C_7H_7NO_2$
AFIGOO	HB	1	9	$C_2H_3NO_3$
BEKBOI	HB	1	16	$C_7H_5NO_3$
DMEGLY01	HB	1	16	$C_4H_8N_2O_2$
COXGAX	HB	1	20	$C_9H_8N_2O$
IFUWIR	HB	1	10	$C_3H_4N_2O$
APMCOX	HB	1	14	$C_5H_5N_3O$
IHULUS	HB	1	19	$C_{10}H_6N_2O$
CEJMAF	HB	1	16	$C_7H_5N_3O$
APYRCX	HB	1	15	$C_6H_6N_2O$
CYACAC02	HB	1	9	$C_3H_3NO_2$
ABEGEU	HB	1	14	C_7H_5NO
IFESUH	HB	1	16	$C_6H_6N_2O_2$

Continued on next page

Table G.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
GOMCIU	HB	1	19	$C_8H_7N_3O$
AMBNAC18	HB	1	17	$C_7H_7NO_2$
EVIGIZ	HB	1	19	$C_9H_7NO_2$
FOMWAF	HB	1	13	$C_5H_5NO_2$
BESPOF	HB	1	18	$C_6H_8N_2O_2$
GOLCOB	HB	1	19	C_8H_9NO
GOFQEY	HB	1	12	$C_4H_4N_2O_2$
CORYOZ	HB	1	16	$C_5H_6N_4O$
IFETES	HB	1	16	$C_6H_6N_2O_2$
BELBUP03	HB	1	14	$C_5H_4N_4O$
AMBNAC12	HB	1	17	$C_7H_7NO_2$
ACAMOX	HB	1	17	$C_4H_{10}N_2O$
AJAQAE	HB	1	16	$C_3H_7N_5O$
FIWFOH01	HB	1	19	$C_6H_{10}N_2O$
DOKVIJ	HB	1	20	$C_8H_6N_4O_2$
ACDHAL01	HB	1	16	$C_5H_7NO_3$
AMPHOL01	HB	1	15	C_6H_7NO
BAQLUD	HB	1	19	$C_7H_4N_2O_6$
EXIQOS	HB	1	14	$C_5H_6N_2O$
GIJBUX	HB	1	19	$C_7H_9NO_2$
BOZGED	HB	1	17	$C_5H_5N_3O_4$
BIBREL	HB	1	16	$C_5H_8N_2O$
BZTZAD	HB	1	20	$C_8H_7N_3O_2$
DUJGIZ	HB	1	14	$C_5H_5NO_3$
HXQUIN14	HB	1	18	C_9H_7NO
BUPXOC	HB	1	15	$C_7H_5NO_2$
GIKSUQ	HB	1	20	$C_9H_8N_2O$
GEYSEI	HB	1	18	$C_6H_8N_2O_2$
EPAHEJ	HB	1	14	$C_5H_5NO_3$
FERFOZ	HB	1	14	$C_2H_2N_8O_2$
HACMUT	HB	1	14	$C_4H_5N_3O_2$
BELBUP04	HB	1	14	$C_5H_4N_4O$
PECBII	HB	1	18	$C_{10}H_2N_4O_2$
BELBUP02	HB	1	14	$C_5H_4N_4O$

Table G.2: Test set structures partitioned from the CE–755. This comprises 151 crystal structures: 85 non-hydrogen bonding (NHB) and 66 hydrogen bonding (HB) structures. No multi-component crystals are included in the dataset, hence any structure with $Z' > 1$ has repeated units of the same molecule.

Refcode	HB?	Z'	nAtms	Chemical Formula
BADAMN10	NHB	1	50	$C_{20}H_{30}$
TRIPHE13	NHB	1	30	$C_{18}H_{12}$
ANTCEN16	NHB	1	24	$C_{14}H_{10}$
MEYCIC	NHB	1	11	C_5H_6
DAZNAQ	NHB	1	16	$C_8H_6N_2$
DAZNAR	NHB	1	16	$C_8H_6N_2$
BIPYRL04	NHB	1	20	$C_{10}H_8N_2$
TCYCPR01	NHB	1	13	$C_7H_2N_4$
CYTCOD10	NHB	1	17	C_9H_7N
COSNAZ	NHB	1	22	$C_6H_{12}N_4$
TOXOCN	NHB	1	16	$C_4H_8O_4$
AXUDED	NHB	1	18	$C_8H_8O_2$
QOTDEL	NHB	1	23	$C_8H_{10}O_5$
AXOREM	NHB	1	18	$C_9H_6O_3$
BEYDEO01	NHB	1	18	$C_6H_8O_4$
ABEWAG	NHB	1	9	C_2H_6O
APINIY	NHB	1	29	$C_{14}H_{11}NO_3$
BAGRAC	NHB	1	26	$C_{11}H_{10}N_2O_3$
ARIDUB	NHB	1	20	$C_8H_8N_2O_2$
AMIXPO	NHB	1	25	$C_{10}H_{10}N_2O_3$
YICFEX	NHB	1	15	C_7H_7F
CAXNUL01	NHB	1	18	$C_{10}H_4F_4$
BUCNEU	NHB	1	25	$C_{13}H_7F_4N$
CATXEC	NHB	1	23	$C_{13}H_5F_4N$
AYOHUR	NHB	1	18	$C_9H_4F_4O$
DUCTAZ	NHB	1	23	$C_{11}H_8F_2O_2$
SUDYOI	NHB	1	22	$C_{10}H_5F_3N_2O_2$
FADBEU	NHB	1	25	$C_{10}H_8F_4N_2O$
DUBPUO	NHB	1	32	$C_{13}H_{13}F_3N_2O$
ZZZBLP02	NHB	1	24	$C_{14}H_8Cl_2$
TCCYHX10	NHB	2	18/18	$C_6H_9Cl_3$
CLMETH03	NHB	2	5/5	CH_3Cl
HCLTRL	NHB	1	22	$C_{10}H_6Cl_6$
HCLBNZ13	NHB	1	12	C_6Cl_6
CACTAD	NHB	1	29	$C_{11}H_{12}Cl_3N_3$

Continued on next page

Table G.2 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
CIPWOO	NHB	1	19	$C_9H_7ClN_2$
GETVEH	NHB	1	11	$C_5H_2Cl_3N$
AZIBER	NHB	1	14	$C_7H_4Cl_2O$
AHUHAO	NHB	1	29	$C_{11}H_{13}Cl_3O_2$
CDCDOX	NHB	1	14	$C_4H_6Cl_2O_2$
FICFON	NHB	1	24	$C_{10}H_9Cl_2NO_2$
FOXWEW	NHB	1	23	$C_{11}H_8ClNO_2$
IHERES	NHB	1	5	CH_2FCl
AWOBIZ	NHB	1	13	C_7H_3FClN
ILANAM	NHB	1	14	$C_7H_3F_2ClO$
BIRQOL	NHB	1	34	$C_{18}H_{10}FCl_2NO_2$
FABPON11	NHB	1	18	$C_9H_6S_3$
THIANT05	NHB	1	22	$C_{12}H_8S_2$
LEVMUU	NHB	1	15	C_8H_6S
FANJOT	NHB	1	16	$C_6H_6N_2S_2$
ZUCNAP	NHB	1	30	$C_{13}H_{14}N_2S$
DOSZUG	NHB	2	15/15	$C_8H_4N_2S$
AJEVOB	NHB	1	20	$C_8H_8O_3S$
RIQKAE	NHB	1	22	$C_8H_{10}O_2S_2$
FORROU	NHB	1	14	$C_5H_6OS_2$
NOVSUP	NHB	1	33	$C_{17}H_{13}NOS$
XORSIJ	NHB	1	33	$C_{13}H_{15}NO_2S_2$
CUBTEB	NHB	1	29	$C_{16}H_9NOS_2$
MAKQIZ	NHB	1	24	$C_{14}H_4F_4S_2$
IJEPEU	NHB	1	29	$C_{14}H_5F_8NS$
APATER	NHB	1	20	$C_{11}H_4F_3NS$
CEXVAF	NHB	1	22	$C_{11}H_5F_3O_2S$
FOCJOZ	NHB	1	25	$C_{12}H_9FOS_2$
CIKBUU	NHB	2	36/36	$C_{20}H_{16}$
ETIFAP	NHB	2	26/26	$C_{16}H_{10}$
BIZQEJ	NHB	1	31	$C_{19}H_{12}$
AMIMZD10	NHB	1	25	$C_9H_{12}N_4$
DNBENZ11	NHB	1	16	$C_6H_4N_2O_4$
QEKYEK	NHB	1	19	$C_5H_{12}S_2$
CELDUS	NHB	1	30	$C_{17}H_{12}S$
SAFGIP	NHB	1	14	C_7H_6S
IPRNSO01	NHB	1	16	$C_5H_8O_2S$
TOXDTH12	NHB	1	12	$C_2H_4O_4S_2$
MOCVEF	NHB	1	20	$C_6H_8N_2O_2S_2$

Continued on next page

Table G.2 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
FUNFEA	NHB	1	17	$C_8H_4O_2S_3$
CEBBIX	NHB	2	9/9	$C_3Cl_2N_2S_2$
FEQYAC	NHB	1	19	$C_{10}H_4Cl_2N_2S$
AQOTUY	NHB	1	16	$C_6H_5ClO_3S$
BETBUX	NHB	1	18	$C_5H_8N_2O_2S$
JIYYAT	NHB	1	21	$C_8H_8N_2O_2S$
THTOXC	NHB	1	12	$C_2Cl_4O_4S_2$
TECFIQ	NHB	1	18	$C_7H_7ClO_2S$
IFOVOR	NHB	1	20	$C_8H_8F_2OS$
KONJIG	NHB	1	17	$C_7H_6F_2OS$
WULPEZ	NHB	1	25	$C_{11}H_9ClO_3S$
AMPYRM11	HB	1	12	$C_4H_5N_3$
HIFCIJ	HB	1	20	$C_8H_{10}N_2$
DAMTRZ11	HB	1	12	$C_2H_5N_5$
FULPIM	HB	1	12	$C_4H_6N_2$
FORMAC02	HB	1	5	CH_2O_2
ZZZKPE03	HB	1	20	$C_6H_{12}O_2$
ZEFPUX	HB	1	15	$C_4H_{10}O$
HIHNIV	HB	1	15	$C_5H_5N_3O_2$
NMALAM01	HB	1	15	$C_3H_5N_3O_4$
MALEHY01	HB	1	12	$C_4H_4N_2O_2$
JEFTEU	HB	1	18	$C_6H_7N_3O_2$
UTODOX	HB	1	20	$C_8H_5F_4NO_2$
KUSJAL	HB	1	14	$C_6H_5F_2N$
VEPXOG	HB	1	13	$C_5H_2F_4N_2$
ABENEB	HB	1	16	$C_7H_5FO_3$
WESBEB01	HB	1	19	$C_8H_7FO_3$
IKAYEZ	HB	1	15	$C_7H_3F_3O_2$
FARMIW	HB	1	15	$C_7H_5ClN_2$
GAPPIY	HB	1	12	$C_4H_4ClN_3$
BAWVAW	HB	1	21	$C_{10}H_8ClNO$
CLACET02	HB	1	8	$C_2H_3ClO_2$
DOQQEH	HB	2	14/14	$C_6H_3Cl_2NO_2$
AJIVIZ	HB	1	13	$C_5H_4ClNO_2$
ABALAS	HB	1	21	$C_{10}H_8N_2S$
ADAHET	HB	1	20	$C_8H_7N_3S_2$
XOPSED	HB	1	18	$C_7H_6O_4S$
GUMFUS	HB	2	18/18	$C_5H_{10}O_2S$
AYUBUS	HB	1	23	$C_6H_{12}N_2O_2S$

Continued on next page

Table G.2 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula
CACDOA	HB	1	20	$C_8H_{10}O_2$
AHEJAZ08	HB	1	20	$C_8H_{10}O_2$
ACMCPY	HB	1	21	$C_9H_6N_6$
BILHOV	HB	1	19	$C_5H_{12}N_2$
AMNIFL	HB	1	27	$C_{13}H_{10}N_2O_2$
JUJPEL	HB	1	13	C_6H_6S
DASKAK01	HB	1	25	$C_{12}H_9N_3S$
CIRXAF	HB	1	24	$C_{12}H_9NOS$
AQEZUR	HB	1	24	$C_{10}H_{10}O_2S_2$
CIJNEO	HB	1	31	$C_{13}H_{13}N_3OS$
DUTZEY	HB	1	13	$C_4H_5N_3S$
AMTHTZ	HB	1	8	CH_2N_4S
DMCTDZ	HB	1	9	$C_2H_2N_2S_3$
VUTWAL	HB	1	21	$C_9H_8O_3S$
ZZZLWW03	HB	1	16	$C_7H_6O_2S$
USUMOL	HB	1	17	$C_6H_6O_3S_2$
CHTDZO	HB	1	14	$C_4H_3N_3O_3S$
AFIMEJ	HB	1	22	$C_9H_8N_2O_2S$
FUZNAQ	HB	1	17	$C_6H_7NO_2S$
COBFUU	HB	1	21	$C_8H_8O_4S$
HOMSAN	HB	1	21	$C_7H_9NO_3S$
LIDGAH	HB	1	18	$C_8H_5F_3N_2$
MONPEM	HB	1	14	$C_6HF_3N_4$
MEKMOF	HB	1	20	$C_{10}H_5F_3O_2$
CLCROT10	HB	1	12	$C_4H_5ClO_2$
DOKWUW	HB	1	16	$C_6H_6Cl_2N_2$
ACIRAI	HB	1	16	$C_6H_4Cl_4N_2$
INIGEU	HB	1	16	$C_7H_6N_2O$
AFAPUT	HB	1	20	$C_8H_9NO_2$
EREFEP	HB	1	12	$C_3H_2N_6O$
AMBACO05	HB	1	17	$C_7H_7NO_2$
AJOQUL03	HB	1	18	$C_7H_7N_3O$
AYUMUD	HB	1	17	$C_7H_7NO_2$
ISNICA02	HB	1	14	$C_6H_5NO_2$
FUZZAB	HB	1	16	$C_6H_6N_2O_2$
IFETAO	HB	1	16	$C_6H_6N_2O_2$
FORAMO01	HB	1	8	CH_4N_2O
IGAMEK	HB	1	13	$C_5H_4N_2O_2$

Appendix H

Gradients Involving Tang–Toennies Damping

To use the damped Buckingham potential (dBHP) in the various CSP (and parameter estimation) computations, analytical gradients have been derived for the dBHP, analogous to those derived for the BHP and MIP in Sections 3.2.1.3 and 4.3.2. For convenience, I first rewrite the dBHP pair potential function from Section 6.1.2,

$$U^{\mathrm{FF,dBHP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}(r_{\mathbf{0}Ii,\boldsymbol{l}Jj}) = A^{t,t'} \exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right) - f_6 \frac{C^{t,t'}}{r^6_{\mathbf{0}Ii,\boldsymbol{l}Jj}}, \qquad t = t(Ii), \quad t' = t'(Jj), \tag{H.0.1}$$

and a generalized form of the Tang–Toennies damping function,

$$f_n(B^{t,t'}, r_{\mathbf{0}Ii,\boldsymbol{l}Jj}) = 1 - \exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right) \sum_{k=0}^{n} \left[\frac{\left(\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right)^k}{k!} \right], \qquad t = t(Ii), \quad t' = t'(Jj), \tag{H.0.2}$$

where the Tang–Toennies damping factor has been replaced using $\beta^{t,t'} = (B^{t,t'})^{-1}$.

Derivatives with respect to Euclidean Distance

Using chain-rule, I derive the first and second derivatives of the dBHP with respect to $r_{\mathbf{0}Ii,\boldsymbol{l}Jj}$, which are given by

$$\begin{aligned} \frac{\partial U^{\mathrm{FF,dBHP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} = &- \frac{A^{t,t'}}{B^{t,t'}} \exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right) + f_6 \frac{6C^{t,t'}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^7} \\ &- \frac{\partial f_6}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \frac{C^{t,t'}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^6}, \qquad t = t(Ii), \quad t' = t'(Jj), \end{aligned} \tag{H.0.3}$$

and

$$\begin{aligned} \frac{\partial^2 U^{\mathrm{FF,dBHP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial (r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2} = &\frac{A^{t,t'}}{(B^{t,t'})^2} \exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right) - f_6 \frac{42C^{t,t'}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^8} + \frac{\partial f_6}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} \frac{12C^{t,t'}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^7} \\ &- \frac{\partial^2 f_6}{\partial (r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^2} \frac{C^{t,t'}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^6}, \qquad t = t(Ii), \quad t' = t'(Jj), \end{aligned} \tag{H.0.4}$$

respectively. These are analogues to Equations 3.2.45 and 3.2.66 for the native BHP. The corresponding first and second derivatives of the (generalized) damping function with respect to $r_{\mathbf{0}Ii,\boldsymbol{l}Jj}$ are given by

$$
\begin{aligned}
\frac{\partial f_n}{\partial r_{\mathbf{0}Ii,\mathbf{l}Jj}} =& \frac{\exp\left(-\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right)}{B^{t,t'}} \sum_{k=0}^{n} \left[\frac{\left(\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right)^k}{k!} \right] \\
& - \exp\left(-\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right) \sum_{k=0}^{n} \left[\frac{k\left(B^{t,t'}\right)^{-k} \left(r_{\mathbf{0}Ii,\mathbf{l}Jj}\right)^{k-1}}{k!} \right] \\
=& \exp\left(-\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right) \left\{ \frac{1}{B^{t,t'}} \sum_{k=0}^{n} \left[\frac{\left(\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right)^k}{k!} \right] - \sum_{k=1}^{n} \left[\frac{\left(B^{t,t'}\right)^{-k} \left(r_{\mathbf{0}Ii,\mathbf{l}Jj}\right)^{k-1}}{(k-1)!} \right] \right\}, \\
& t = t(Ii), \quad t' = t'(Jj),
\end{aligned}
\tag{H.0.5}
$$

and

$$
\begin{aligned}
\frac{\partial^2 f_n}{\partial (r_{\mathbf{0}Ii,\mathbf{l}Jj})^2} =& -\frac{\exp\left(-\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right)}{B^{t,t'}} \left\{ \frac{1}{B^{t,t'}} \sum_{k=0}^{n} \left[\frac{\left(\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right)^k}{k!} \right] - \sum_{k=1}^{n} \left[\frac{\left(B^{t,t'}\right)^{-k} \left(r_{\mathbf{0}Ii,\mathbf{l}Jj}\right)^{k-1}}{(k-1)!} \right] \right\} \\
& + \exp\left(-\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right) \left\{ \frac{1}{B^{t,t'}} \sum_{k=1}^{n} \left[\frac{\left(B^{t,t'}\right)^{-k} \left(r_{\mathbf{0}Ii,\mathbf{l}Jj}\right)^{k-1}}{(k-1)!} \right] \right. \\
& \left. - \sum_{k=2}^{n} \left[\frac{\left(B^{t,t'}\right)^{-k} \left(r_{\mathbf{0}Ii,\mathbf{l}Jj}\right)^{k-2}}{(k-2)!} \right] \right\} \\
=& \exp\left(-\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right) \left\{ -\frac{1}{(B^{t,t'})^2} \sum_{k=0}^{n} \left[\frac{\left(\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right)^k}{k!} \right] \right. \\
& \left. + \frac{2}{B^{t,t'}} \sum_{k=1}^{n} \left[\frac{\left(B^{t,t'}\right)^{-k} \left(r_{\mathbf{0}Ii,\mathbf{l}Jj}\right)^{k-1}}{(k-1)!} \right] - \sum_{k=2}^{n} \left[\frac{\left(B^{t,t'}\right)^{-k} \left(r_{\mathbf{0}Ii,\mathbf{l}Jj}\right)^{k-2}}{(k-2)!} \right] \right\}, \\
& t = t(Ii), \quad t' = t'(Jj),
\end{aligned}
\tag{H.0.6}
$$

respectively. These gradients have been implemented into the relevant CSO–RM subroutines for use in the various CSP calculations.

Derivatives for Solving the Parameter Estimation

In service of the parameter estimation, derivatives with respect to the dBHP parameters are also necessary. The first-derivatives with respect to the FF parameters are given by:

$$
\left.\frac{\partial U^{\mathrm{FF,dBHP}}_{\mathbf{0}Ii,\mathbf{l}Jj}}{\partial A^{t,t'}}\right|_{r_{\mathbf{0}Ii,\mathbf{l}Jj}} = \exp\left(-\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right), \qquad t = t(Ii), \quad t' = t'(Jj),
\tag{H.0.7}
$$

$$
\left.\frac{\partial U^{\mathrm{FF,dBHP}}_{\mathbf{0}Ii,\mathbf{l}Jj}}{\partial B^{t,t'}}\right|_{r_{\mathbf{0}Ii,\mathbf{l}Jj}} = \frac{A^{t,t'} r_{\mathbf{0}Ii,\mathbf{l}Jj}}{(B^{t,t'})^2} \exp\left(-\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right) - \frac{\partial f_6}{\partial B^{t,t'}} \frac{C^{t,t'}}{(r_{\mathbf{0}Ii,\mathbf{l}Jj})^6}, \qquad t = t(Ii), \quad t' = t'(Jj),
\tag{H.0.8}
$$

$$\left.\frac{\partial U^{\mathrm{FF,dBHP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial C^{t,t'}}\right|_{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}} = -\frac{f_6}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^6}, \qquad t = t(Ii), \quad t' = t'(Jj), \tag{H.0.9}$$

which are analogues to Equations 4.3.21–4.3.23. The cross-derivatives with respect to the FF parameter and Euclidean distances are given by:

$$\frac{\partial^2 U^{\mathrm{FF,dBHP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\partial A^{t,t'}} = -\frac{1}{B^{t,t'}}\exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right), \qquad t = t(Ii), \quad t' = t'(Jj), \tag{H.0.10}$$

$$\begin{aligned}\frac{\partial^2 U^{\mathrm{FF,dBHP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\partial B^{t,t'}} =& \frac{A^{t,t'}}{(B^{t,t'})^2}\left(1-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right)\exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right)\\ &-\frac{\partial^2 f_6}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\partial B^{t,t'}}\frac{C^{t,t'}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^6}+\frac{\partial f_6}{\partial B^{t,t'}}\frac{6C^{t,t'}}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^7},\\ & t = t(Ii), \quad t' = t'(Jj),\end{aligned} \tag{H.0.11}$$

$$\begin{aligned}\frac{\partial^2 U^{\mathrm{FF,dBHP}}_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\partial C^{t,t'}} =& -\frac{\frac{\partial f_6}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^6 - 6f_6(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^5}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^{12}}\\ =& \frac{6f_6}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^7}-\frac{\partial f_6}{\partial r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}\frac{1}{(r_{\mathbf{0}Ii,\boldsymbol{l}Jj})^6}, \qquad t = t(Ii), \quad t' = t'(Jj),\end{aligned} \tag{H.0.12}$$

which are analogues to Equations 4.3.36–4.3.38. To evaluate these, the remaining quantities required are the derivatives of the damping function with respect to the $B^{t,t'}$ parameter, and its cross-derivative with the Euclidean distance. These are given by

$$\begin{aligned}\frac{\partial f_n}{\partial B^{t,t'}} =& -\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{(B^{t,t'})^2}\exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right)\sum_{k=0}^{n}\left[\frac{\left(\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right)^k}{k!}\right]\\ &+\exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right)\sum_{k=0}^{n}\left[\frac{k\left(r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\right)^k(B^{t,t'})^{-k-1}}{k!}\right]\\ =&\exp\left(-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right)\left\{\sum_{k=1}^{n}\left[\frac{\left(r_{\mathbf{0}Ii,\boldsymbol{l}Jj}\right)^k(B^{t,t'})^{-k-1}}{(k-1)!}\right]-\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{(B^{t,t'})^2}\sum_{k=0}^{n}\left[\frac{\left(\frac{r_{\mathbf{0}Ii,\boldsymbol{l}Jj}}{B^{t,t'}}\right)^k}{k!}\right]\right\},\\ & t = t(Ii), \quad t' = t'(Jj),\end{aligned} \tag{H.0.13}$$

and

$$
\begin{aligned}
\frac{\partial^2 f_n}{\partial r_{\mathbf{0}Ii,\mathbf{l}Jj}\partial B^{t,t'}} =& \frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{(B^{t,t'})^2}\exp\left(-\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right)\left\{\frac{1}{B^{t,t'}}\sum_{k=0}^{n}\left[\frac{\left(\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right)^k}{k!}\right]\right. \\
&\left. -\sum_{k=1}^{n}\left[\frac{\left(B^{t,t'}\right)^{-k}\left(r_{\mathbf{0}Ii,\mathbf{l}Jj}\right)^{k-1}}{(k-1)!}\right]\right\} \\
&+\exp\left(-\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right)\left\{-\frac{1}{(B^{t,t'})^2}\sum_{k=0}^{n}\left[\frac{\left(\frac{r_{\mathbf{0}Ii,\mathbf{l}Jj}}{B^{t,t'}}\right)^k}{k!}\right]\right. \\
&\left. -\frac{1}{B^{t,t'}}\sum_{k=1}^{n}\left[\frac{\left(r_{\mathbf{0}Ii,\mathbf{l}Jj}\right)^k\left(B^{t,t'}\right)^{-k-1}}{(k-1)!}\right]\right. \\
&\left. +\sum_{k=1}^{n}\left[\frac{k\left(B^{t,t'}\right)^{-k-1}\left(r_{\mathbf{0}Ii,\mathbf{l}Jj}\right)^{k-1}}{(k-1)!}\right]\right\}, \\
& t = t(Ii), \quad t' = t'(Jj),
\end{aligned}
\tag{H.0.14}
$$

respectively. These gradients have been implemented into the relevant CSO–RM subroutines which interface with CrystalEstimator during the parameter estimation.

With this, all necessary dBHP gradient terms required for the current work have been derived. These gradients can be used in conjunction with the other expressions in Sections 3.2.1.3 and 4.3.2 to 'interconvert' between expressions in terms of the rigid-body variables, $\mathbf{\Omega}$. The validity of these gradient expressions for rigid-body minimization and parameter estimation have been verified using the gradient check utility implemented in the NAG subroutines called by CSO–RM and CrystalEstimator.

Appendix I

Parameter Bounds and Reference Potentials for Parameter Estimation

I.1 Reference Potentials for CrystalEstimator

To conduct parameter estimation of HAIEFF FFs with CrystalEstimator, it is necessary to provide the program with a set of reference parameters. In the case of the BHP and dBHP, the FIT FF [4] was selected for this purpose. The parameters associated with this FF are tabulated in Table I.1.

Table I.1: FIT pair potential [4] parameters used in this work. Note that three hydrogen types are defined: a 'general' hydrogen (H), a nitrogen-bonded hydrogen (H_N), and an oxygen-bonded hydrogen (H_O). If only polarized hydrogens (H_p) are used, these assume the values of H_N.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å^{-1})	$C^{t,t'}$ (eV Å^{-6})
C	C	3832.1293465306	0.2777777778	25.2868321501
C	H	689.5335380168	0.2724795640	5.9789441613
C	O	3022.8361459857	0.2645502646	17.1601565267
C	N	3179.4998934425	0.2710027100	19.0066223013
C	F	3800.8152104990	0.2577319588	14.8726560911
C	Cl	6060.1690226192	0.2812939522	45.0402983577
C	S	3990.9899293835	0.2898550725	38.9566133819
C	H_N	446.9514563216	0.2421307506	2.3737585513
H	H	124.0710991346	0.2673796791	1.4136912473
H	O	543.9135045046	0.2597402597	4.0574326220
H	N	572.1027690869	0.2659574468	4.4940201588
H	F	683.8990343100	0.2531645570	3.5165646598
H	Cl	1090.4354757570	0.2758620690	10.6495518019
H	S	718.1180898000	0.2840909091	9.2110951162
H	H_N	80.4221337879	0.2380952381	0.5612632593
O	O	2384.4545784319	0.2525252525	11.6452298285
O	N	2508.0330894251	0.2583979328	12.8982789066
O	F	2998.1351263393	0.2463054187	10.0928856955
O	Cl	4780.3443766693	0.2677376171	30.5652588367

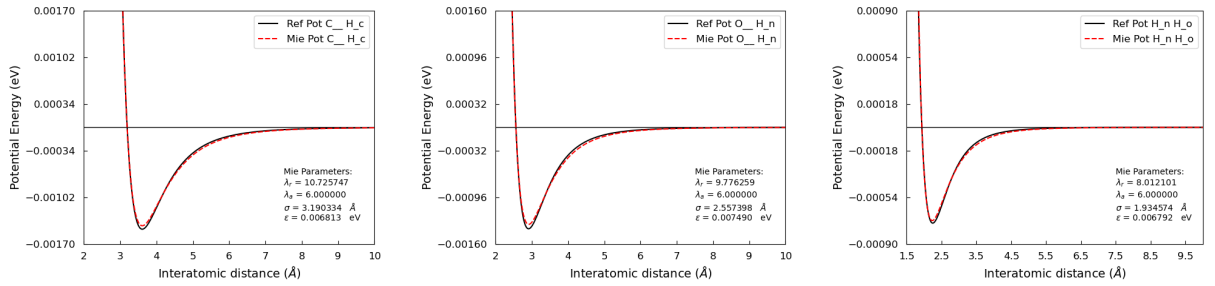
Continued on next page

Table I.1 – continued from previous page

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å^{-1})	$C^{t,t'}$ (eV Å^{-6})
O	S	3148.1475508461	0.2754820937	26.4367469763
O	H_N	352.5614339957	0.2320185615	1.6108806376
N	N	2638.0162719594	0.2645502646	14.2861584702
N	F	3153.5187004366	0.2518891688	11.1788995657
N	Cl	5028.0940488363	0.2743484225	33.8541393459
N	S	3311.3057800863	0.2824858757	29.2813916862
N	H_N	370.8335442892	0.2369668246	1.7842144857
F	F	3769.7569570400	0.2403846154	8.7474737006
F	Cl	6010.6485237038	0.2607561930	26.4908179774
F	S	3958.3776686146	0.2680965147	22.9126491552
F	H_N	443.2992052003	0.2267573696	1.3961454075
Cl	Cl	9583.6140332694	0.2849002849	80.2246981396
Cl	S	6311.3927930257	0.2936857562	69.3885845134
Cl	H_N	706.8136603132	0.2447980416	4.2280817441
S	S	4156.4360636368	0.3030303030	60.0161268591
S	H_N	465.4797893807	0.2512562814	3.6569861181
H_N	H_N	52.1291392444	0.2145922747	0.2228325646
H_O	C	299.8204425427	0.2421307506	2.3737585513
H_O	H	53.9481400082	0.2380952381	0.5612632593
H_O	O	236.5024739690	0.2320185615	1.6108806376
H_O	N	248.7596265454	0.2369668246	1.7842144857
H_O	F	297.3704683185	0.2267573696	1.3961454075
H_O	Cl	474.1391518766	0.2447980416	4.2280817441
H_O	S	312.2494724492	0.2512562814	3.6569861181
H_O	H_N	34.9688570796	0.2145922747	0.2228325646
H_O	H_O	23.4575322589	0.2145922747	0.2228325646

For the parametrization of Mie potentials with CrystalEstimator, a set of reference Mie FF parameters were estimated so as to reproduce the FIT FF [4]. A Python script was written to conduct the following steps for each atom type pairing (t, t'):

1. $\lambda_a^{t,t'}$ is fixed to 6.0, in accordance with the leading term of the long-range pairwise London dispersion expansion.
2. $\sigma^{t,t'}$ is set to the x-intercept of the corresponding FIT pair potential. Importantly, the intercept corresponding to the unphysical short-range attractive divergence of the BHP is disregarded.
3. Finally, $\lambda_r^{t,t'}$ and $\varepsilon^{t,t'}$ are estimated by least-squares optimization, using functions available from the SciPy.optimize package. For reference data in this optimization, 10,000 equally spaced data points are sampled from the FIT pair potential. This data begins from 0.10–0.20 Å below the $\sigma^{t,t'}$ value, up to 10.0 Å.



(a) (b) (c)

Figure I.1: Exemplars of the FIT-analogue determined using a Mie potential for (a) the C/H interaction, (b) the O/H_N interaction, and (c) the H_N/H_O interaction. The FIT reference (—) and the derived Mie potential (dashed —) are overlaid.

In principle, $\varepsilon^{t,t'}$ could also have been fixed to the energy well minimum of the FIT pair potential. However, I choose not to do this so as to avoid over-constraining the parameter fitting and enable better agreement between the FIT FF and the Mie-analogue over the *entire* fitting region. In any case, even without constraining $\varepsilon^{t,t'}$, its optimized value is often close to the expected value of the FIT pair potential energy minimum. Across all of the pair potentials, the absolute deviation between the optimized $\varepsilon^{t,t'}$ (after accounting for $C^{t,t'}_{\text{Mie}}$) and the FIT energy minimum does not exceed 3.6×10^{-4} eV, while the percentage deviation does not exceed 6.1%.

The fitted Mie FF is tabulated in Table I.2, with some exemplar pair potentials shown in Figure I.1. Note that during the estimation of the Mie potential parameters, to reduce non-linearity in the optimization problem, the Mie potential is defined without $C^{t,t'}_{\text{Mie}}$. For consistency with the definition of the Mie potential implemented in our CSP programs, the value of $\varepsilon^{t,t'}$ is corrected to account for $C^{t,t'}_{\text{Mie}}$ in post-processing. It is for this reason that the $\varepsilon^{t,t'}$ values presented in Figure I.1 differ from their counterparts reported in Table I.2.

Table I.2: Mie-analogue of the FIT pair potential [4] used in this work. Three hydrogen types are defined: a 'general' hydrogen (H), a nitrogen-bonded hydrogen (H_N), and an oxygen-bonded hydrogen (H_O).

t	t'	$\varepsilon^{t,t'}$ (eV)	$\sigma^{t,t'}$ (Å)	$\lambda_r^{t,t'}$ (-)	$\lambda_a^{t,t'}$ (-)
C	C	0.0040025279	3.4666951311	11.6030423830	6.0000000000
C	H	0.0014357165	3.1903335786	10.7257472739	6.0000000000
C	N	0.0036162887	3.3562893791	11.4725179492	6.0000000000
C	O	0.0040442616	3.2283628039	11.2459307242	6.0000000000
C	F	0.0034888480	3.2526425224	11.6826029884	6.0000000000
C	S	0.0052996272	3.5406303408	11.3535618260	6.0000000000
C	Cl	0.0068310261	3.4869460119	11.5231096360	6.0000000000
C	H_N	0.0013859821	2.7242405939	10.0824432890	6.0000000000
C	H_O	0.0020119835	2.5066084685	9.0255649808	6.0000000000
H	H	0.0005314574	2.9101007503	9.7563290458	6.0000000000

Continued on next page

Table I.2 – continued from previous page

t	t'	$\varepsilon^{t,t'}$ (eV)	$\sigma^{t,t'}$ (Å)	$\lambda_{\mathrm{r}}^{t,t'}$ (-)	$\lambda_{\mathrm{a}}^{t,t'}$ (-)
H	N	0.0012957173	3.0884582889	10.5882777845	6.0000000000
H	O	0.0014531181	2.9673270844	10.3498950465	6.0000000000
H	F	0.0012258331	3.0061471216	10.8346187984	6.0000000000
H	S	0.0019409973	3.2429151067	10.4344098769	6.0000000000
H	Cl	0.0024657500	3.2044078748	10.6292450071	6.0000000000
H	$\mathrm{H_N}$	0.0005126961	2.4791046054	9.0831779854	6.0000000000
H	$\mathrm{H_O}$	0.0007893279	2.2376603244	7.8185390849	6.0000000000
N	N	0.0032483522	3.2528022516	11.3488447878	6.0000000000
N	O	0.0036139258	3.1316678310	11.1291849002	6.0000000000
N	F	0.0030932415	3.1602309087	11.5782628787	6.0000000000
N	S	0.0048484098	3.4195457703	11.2041493124	6.0000000000
N	Cl	0.0061941995	3.3735172438	11.3866597581	6.0000000000
N	$\mathrm{H_N}$	0.0012168397	2.6505791523	10.0054624585	6.0000000000
N	$\mathrm{H_O}$	0.0017731963	2.4358479783	8.9152368036	6.0000000000
O	O	0.0040037998	3.0170440370	10.9141049522	6.0000000000
O	F	0.0033942209	3.0508430263	11.3797902535	6.0000000000
O	S	0.0054956779	3.2802637945	10.9563498382	6.0000000000
O	Cl	0.0069546453	3.2423589350	11.1549621781	6.0000000000
O	$\mathrm{H_N}$	0.0013319211	2.5573982604	9.7762592941	6.0000000000
O	$\mathrm{H_O}$	0.0019608792	2.3426801709	8.6719755030	6.0000000000
F	F	0.0028863417	3.0812558867	11.8336406165	6.0000000000
F	S	0.0047638461	3.3033643944	11.3910440930	6.0000000000
F	Cl	0.0060071730	3.2663882468	11.5905666524	6.0000000000
F	$\mathrm{H_N}$	0.0010798221	2.6130916386	10.3329613575	6.0000000000
F	$\mathrm{H_O}$	0.0015419760	2.4156798420	9.3207495472	6.0000000000
S	S	0.0069025790	3.6264744378	11.1244979039	6.0000000000
S	Cl	0.0090027151	3.5641570397	11.2792095802	6.0000000000
S	$\mathrm{H_N}$	0.0019939050	2.7338340743	9.6882527900	6.0000000000
S	$\mathrm{H_O}$	0.0029662527	2.4967737385	8.5569860633	6.0000000000
Cl	Cl	0.0116429733	3.5080972803	11.4446166585	6.0000000000
Cl	$\mathrm{H_N}$	0.0024234097	2.7261702003	9.9605741494	6.0000000000
Cl	$\mathrm{H_O}$	0.0035434445	2.5029629828	8.8816643160	6.0000000000
$\mathrm{H_N}$	$\mathrm{H_N}$	0.0004329687	2.1653604686	8.5902283308	6.0000000000
$\mathrm{H_N}$	$\mathrm{H_O}$	0.0007200355	1.9345743418	8.0121011281	6.0000000000
$\mathrm{H_O}$	$\mathrm{H_O}$	0.0013388225	1.6219646181	6.0754226046	6.0000000000

I.2 Lower Bounds for Mie σ Parameter

By the formulation of the Mie potential used in this work, the $\sigma^{t,t'}$ parameter approximates the position of the repulsive wall of the Mie potential. Recognizing that in the vicinity of this repulsive wall, $\frac{\partial U^{\mathrm{FF}}}{\partial r} < 0$, implying that the thermodynamic driving force favors separation of

the participating atoms. As such, the value of $\sigma^{t,t'}$ should also roughly correspond to the lower bound of observed interatomic distance between atom types t and t'. Based on this logic, I estimate appropriate lower bound values for $\sigma^{t,t'}$. First, the entire CE–755 dataset is surveyed using the CCDC Python API and the shortest contact distance for each pair of atom types, tt', is identified. $\sigma^{\mathrm{FIT},t,t'}$ values have already been determined per the parameter estimation procedure used to define the Mie-analogue. Since the potential uses $\sigma^{\mathrm{FIT},t,t'}$ as a reference, $\sigma^{t,t'}_{\mathrm{lbd}}$ must also encompass this reference value to avoid violating the lower bound at the initial parameter vector. Altogether, $\sigma^{t,t'}_{\mathrm{lbd}}$ is determined according to,

$$\sigma^{t,t'}_{\mathrm{lbd}} = \min(\sigma^{\mathrm{FIT},t,t'}, \sigma^{\mathrm{CE755},t,t'}) - 0.1, \tag{I.2.1}$$

where $\sigma^{\mathrm{CE755},t,t'}$ is the shortest contact distance found for tt' in the CE–755. A slight tolerance is permitted with the deduction of 0.1 Å. The results of this are tabulated in Table I.3.

Table I.3: Values of $\sigma^{t,t'}_{\mathrm{lbd}}$ determined from $\sigma^{\mathrm{FIT},t,t'}$ and $\sigma^{\mathrm{CE755},t,t'}$.

t	t'	$\sigma^{\mathrm{FIT},t,t'}$ (Å)	$\sigma^{\mathrm{CE755},t,t'}$ (Å)	$\sigma^{t,t'}_{\mathrm{lbd}}$ (Å)
C	C	3.4667	2.6808	2.5808
C	H	3.1903	2.3919	2.2919
C	N	3.3563	2.2058	2.1058
C	O	3.2284	2.7815	2.6815
C	F	3.2526	2.3195	2.2195
C	S	3.5406	3.2642	3.1642
C	Cl	3.4869	3.2858	3.1858
C	$\mathrm{H_N}$	2.7242	2.3741	2.2741
C	$\mathrm{H_O}$	2.5066	2.2984	2.1984
H	H	2.9101	1.9073	1.8073
H	N	3.0885	1.9740	1.8740
H	O	2.9673	1.8831	1.7831
H	F	3.0061	2.2874	2.1874
H	S	3.2429	2.4870	2.3870
H	Cl	3.2044	2.5184	2.4184
H	$\mathrm{H_N}$	2.4791	2.1305	2.0305
H	$\mathrm{H_O}$	2.2377	2.1469	2.0469
N	N	3.2528	2.2519	2.1519
N	O	3.1317	2.5787	2.4787
N	F	3.1602	2.9673	2.8673
N	S	3.4195	2.9017	2.8017
N	Cl	3.3735	2.8572	2.7572
N	$\mathrm{H_N}$	2.6506	1.6950	1.5950
N	$\mathrm{H_O}$	2.4358	1.5188	1.4188
O	O	3.0170	2.5148	2.4148
O	F	3.0508	2.8608	2.7608

Continued on next page

Table I.3 – continued from previous page

t	t'	$\sigma^{\mathrm{FIT},t,t'}$ (Å)	$\sigma^{\mathrm{CE755},t,t'}$ (Å)	$\sigma_{\mathrm{lbd}}^{t,t'}$ (Å)
O	S	3.2803	2.7606	2.6606
O	Cl	3.2424	2.8918	2.7918
O	H_N	2.5574	1.6179	1.5179
O	H_O	2.3427	1.4894	1.3894
F	F	3.0813	2.3486	2.2486
F	S	3.3034	3.0108	2.9108
F	Cl	3.2664	3.0762	2.9762
F	H_N	2.6131	2.0197	1.9197
F	H_O	2.4157	2.4366	2.3157
S	S	3.6265	3.0426	2.9426
S	Cl	3.5642	3.3548	3.2548
S	H_N	2.7338	2.1115	2.0115
S	H_O	2.4968	2.1006	2.0006
Cl	Cl	3.5081	3.2427	3.1427
Cl	H_N	2.7262	2.6505	2.5505
Cl	H_O	2.5030	2.6007	2.4030
H_N	H_N	2.1654	2.1452	2.0452
H_N	H_O	1.9346	2.1299	1.8346
H_O	H_O	1.6220	2.0659	1.5220

Appendix J

Optimized Force-field Parameters for Point-charge and Multipole Models with/without PCM

Below, I tabulate the optimized FF parameters obtained by the fitting schemes described in Section 6.1.6. Ground-state electron densities are determined at the PBE0/6-311G(d,p) level of theory. Where PCM is used, $\varepsilon_{\mathrm{PCM}}$ = 11.0 is applied. Point-charges are fitted using G09's HLYGAt method, while rank 4 multipoles are obtained using GDMAv2.2 (RADIUS H 0.65).

J.1 Point-charge Electrostatic Model, without PCM

Table J.1: Optimized FF for point-charge electrostatics without PCM, by Scheme 1.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3210.9781901152	0.2777777778	24.5420470151
C	H	446.2344606496	0.2724795640	4.8502671164
C	O	2657.9998158422	0.2645502646	15.6990146374
C	N	3391.5876194700	0.2710027100	22.1703391498
C	F	2524.8546758312	0.2577319588	8.0384181156
C	Cl	6367.9383684838	0.2812939522	43.8089008361
C	S	3663.4086204347	0.2898550725	40.3903861941
C	H_p	204.2636360888	0.2421307506	0.0000000000
H	H	108.5801567767	0.2673796791	1.5358412749
H	O	258.7166702015	0.2597402597	2.0179381126
H	N	227.1857318972	0.2659574468	1.9911516619
H	F	412.3984542151	0.2531645570	2.8077495807
H	Cl	540.2958323554	0.2758620690	7.5399649635
H	S	818.8756320097	0.2840909091	18.0212342526
H	H_p	11.5518967256	0.2380952381	0.0000000000
O	O	5904.9292235596	0.2525252525	16.5416633657
O	N	5142.3923302677	0.2583979328	14.7734122145
O	F	2480.6938762071	0.2463054187	4.1933868351
O	Cl	5153.7929805017	0.2677376171	29.7231992544
O	S	2437.4829228706	0.2754820937	19.7834982626

Continued on next page

Table J.1 – continued from previous page

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
O	H_p	0.0912966062	0.2320185615	0.0000000000
N	N	5825.6113155649	0.2645502646	22.8904965019
N	F	3371.6547969124	0.2518891688	16.3885849475
N	Cl	7702.4851518948	0.2743484225	49.7035117375
N	S	4773.0531379841	0.2824858757	38.6758399295
N	H_p	0.0000000000	0.2369668246	0.0000000000
F	F	3616.8770961878	0.2403846154	7.8285954600
F	Cl	14168.8033147987	0.2607561930	36.1928079577
F	S	4349.9795378503	0.2680965147	22.1879449461
F	H_p	0.6917949978	0.2267573696	0.0000000000
Cl	Cl	11359.1421050943	0.2849002849	89.5399407546
Cl	S	6926.4158525769	0.2936857562	97.4814914458
Cl	H_p	133.6578989676	0.2447980416	0.0000000000
S	S	4217.9364215146	0.3030303030	126.0332401314
S	H_p	60.2247165535	0.2512562814	0.0000000000
H_p	H_p	0.0000000000	0.2145922747	0.0000000000

Table J.2: Optimized FF for point-charge electrostatics without PCM, by Scheme 2.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	2958.4324843196	0.2777777778	22.9801217214
C	H	439.4717139193	0.2724795640	5.0786043877
C	O	3425.6028328511	0.2645502646	20.3574097440
C	N	3939.4425648166	0.2710027100	25.2268529401
C	F	2524.3361783412	0.2577319588	9.7943722982
C	Cl	5978.6988980900	0.2812939522	42.4685533609
C	S	2978.4960834521	0.2898550725	39.3790807883
C	H_p	0.0000000000	0.2421307506	0.0000000000
H	H	90.8908353086	0.2673796791	1.2566668550
H	O	204.5811840105	0.2597402597	1.0059501524
H	N	210.7027777181	0.2659574468	1.5873268131
H	F	278.4769288760	0.2531645570	0.9754974731
H	Cl	475.5593471809	0.2758620690	6.9660491013
H	S	683.1690384175	0.2840909091	15.9130963193
H	H_p	0.0000000000	0.2380952381	0.0000000000
O	O	1889.4061471280	0.2525252525	7.2670889666
O	N	3482.9096623026	0.2583979328	23.5572704668
O	F	2463.1344545319	0.2463054187	5.6921216080
O	Cl	3968.1418257254	0.2677376171	24.3433993385
O	S	2289.7447646655	0.2754820937	16.6293636462
O	H_p	19.9691786961	0.2320185615	0.0000000000

Continued on next page

Table J.2 – continued from previous page

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
N	N	3784.7227858833	0.2645502646	21.8343345854
N	F	4335.1416455429	0.2518891688	14.9391562225
N	Cl	4794.0374151653	0.2743484225	42.2895692767
N	S	1618.2180660447	0.2824858757	33.5089288965
N	H_p	0.0000000000	0.2369668246	0.0000000000
F	F	2242.5252990834	0.2403846154	4.9352842614
F	Cl	15723.1590906737	0.2607561930	37.3796987067
F	S	3106.5334882700	0.2680965147	14.2968532612
F	H_p	0.0000000000	0.2267573696	0.0000000000
Cl	Cl	11125.9653547009	0.2849002849	89.7040667907
Cl	S	4079.3804050838	0.2936857562	67.7668636978
Cl	H_p	402.6489212606	0.2447980416	0.0000000000
S	S	4150.1225717844	0.3030303030	124.8156210031
S	H_p	0.0000000000	0.2512562814	0.0000000000
H_p	H_p	0.0000000000	0.2145922747	0.0000000000

No solution was found for Scheme 3 using point-charge electrostatics without PCM.

Table J.3: Optimized FF for point-charge electrostatics without PCM, by Scheme 4.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3013.7596713142	0.2777777778	23.9122690950
C	H	399.7379623063	0.2724795640	4.8239340845
C	O	2833.9830894924	0.2645502646	18.0417682982
C	N	4084.2671447098	0.2710027100	26.2460936451
C	F	2889.4032058304	0.2577319588	10.0082066357
C	Cl	5630.1384147926	0.2812939522	42.1835842360
C	S	3451.6888723279	0.2898550725	41.1678139241
C	H_p	252.5859227101	0.2421307506	1.4550819798
H	H	96.0846439382	0.2673796791	1.5647498774
H	O	192.3254163556	0.2597402597	0.9124532198
H	N	156.6119947451	0.2659574468	0.4719776723
H	F	379.1364563530	0.2531645570	2.1974847141
H	Cl	386.8369700006	0.2758620690	6.3484198189
H	S	631.9023067772	0.2840909091	16.0336568160
H	H_p	378.9449907594	0.2380952381	2.8193096733
O	O	4819.9978895950	0.2525252525	13.7953210728
O	N	6973.9620204759	0.2583979328	23.1280459132
O	F	2162.7741949688	0.2463054187	5.2973798653
O	Cl	3389.3740247116	0.2677376171	24.6067319272

Continued on next page

Table J.3 – continued from previous page

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
O	S	2767.2308249246	0.2754820937	16.9027161273
O	H_p	0.0000000000	0.2320185615	3.0629036997
N	N	10142.4174214313	0.2645502646	28.3481786305
N	F	2735.5970163065	0.2518891688	11.4680164274
N	Cl	7816.5826548198	0.2743484225	53.7868371386
N	S	3264.7544988559	0.2824858757	41.1397561234
N	H_p	0.0000000000	0.2369668246	7.4589073142
F	F	3442.3348085353	0.2403846154	6.8698754134
F	Cl	13361.2110256044	0.2607561930	32.5390889134
F	S	3942.7616381758	0.2680965147	19.3739710723
F	H_p	201.8430296988	0.2267573696	0.6084098839
Cl	Cl	11005.5287411681	0.2849002849	91.2487963475
Cl	S	8068.4107512298	0.2936857562	94.4928157815
Cl	H_p	1173.5526156966	0.2447980416	0.0000000000
S	S	3833.3427039690	0.3030303030	132.2034044036
S	H_p	4.1099021059	0.2512562814	5.1569515122
H_p	H_p	975.9433860281	0.2145922747	0.0000000000

Table J.4: Optimized FF for point-charge electrostatics without PCM, by Scheme 5.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	2917.2296252723	0.2777777778	22.9825158683
C	H	438.7398009913	0.2724795640	5.5376320229
C	O	2828.7370785366	0.2645502646	18.2948432649
C	N	3628.3533370411	0.2710027100	23.2835590857
C	F	2560.7239836503	0.2577319588	9.3780064620
C	Cl	6590.0681724417	0.2812939522	45.5643757589
C	S	3384.4574395405	0.2898550725	40.6411426264
C	H_N	1364.2085188676	0.2421307506	7.6217625113
C	H_O	0.0000000000	0.2421307506	0.0000000000
H	H	86.7378240940	0.2673796791	1.1937351986
H	O	193.3267384057	0.2597402597	0.6551255721
H	N	208.6611904442	0.2659574468	1.9035653890
H	F	437.6045581912	0.2531645570	2.7941044003
H	Cl	420.8684967081	0.2758620690	6.7627671509
H	S	665.4699561055	0.2840909091	16.9194245851
H	H_N	142.8785551995	0.2380952381	0.0000000000
H	H_O	299.9878214843	0.2380952381	1.5529959408
O	O	9296.1558304218	0.2525252525	19.1634025764
O	N	8687.3037965579	0.2583979328	20.8259096144
O	F	2458.7289295811	0.2463054187	6.3495391334

Continued on next page

Table J.4 – continued from previous page

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
O	Cl	3174.8257907393	0.2677376171	25.6676475252
O	S	2234.5463298820	0.2754820937	18.2781887189
O	H_N	0.0000000000	0.2320185615	3.8066638238
O	H_O	0.0000000000	0.2320185615	8.2061418940
N	N	5177.8650707041	0.2645502646	23.5039477550
N	F	3974.4505187969	0.2518891688	13.9933972397
N	Cl	6474.0633322589	0.2743484225	50.0502119917
N	S	5117.8875354187	0.2824858757	50.6657435204
N	H_N	0.0000000000	0.2369668246	2.1516315745
N	H_O	74.9662127203	0.2369668246	13.7742366311
F	F	3293.4914835232	0.2403846154	7.0290181008
F	Cl	11497.4862074938	0.2607561930	27.2900442539
F	S	3440.8316799695	0.2680965147	16.0199811281
F	H_N	179.0304418063	0.2267573696	0.0000000000
F	H_O	3056.3699637294	0.2267573696	0.0000000000
Cl	Cl	11004.2367445141	0.2849002849	88.7970421102
Cl	S	5258.7423050695	0.2936857562	80.5662938947
Cl	H_N	15507.4519421121	0.2447980416	30.9429504484
Cl	H_O	3439.0984274601	0.2447980416	0.0000000000
S	S	3913.6043643707	0.3030303030	129.0178603608
S	H_N	0.0000000000	0.2512562814	9.5178169526
S	H_O	8.4182559701	0.2512562814	2.1619302496
H_N	H_N	1034.2020873000	0.2145922747	0.0000000000
H_N	H_O	1026.2470260097	0.2145922747	0.0000000000
H_O	H_O	682.9858451834	0.2145922747	0.0000000000

Table J.5: Optimized FF for point-charge electrostatics without PCM, by Scheme 6.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	100123.6709884793	0.2144619749	20.2245128376
C	H	598.9537281326	0.2658314497	5.9041322056
C	O	2627.9966210773	0.2660815845	18.0938634409
C	N	7328.4637005735	0.2571220954	23.3403582127
C	F	140361.0260675988	0.1914195355	7.7763985745
C	Cl	375492.2179891188	0.2106645102	38.8693574196
C	S	410502.5115864910	0.2053325342	34.9819649694
C	H_N	3060.6072827701	0.2191598734	6.1870099569
C	H_O	830738.4226767645	0.1247236685	1.1258587341
H	H	37.3253588777	0.3015558976	1.5468856933
H	O	117.5651016532	0.2819924156	1.6716196824
H	N	211.8970768778	0.2680652477	2.1498982170

Continued on next page

Table J.5 – continued from previous page

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
H	F	97.2864051941	0.3063358243	4.4530782296
H	Cl	640.1949950682	0.2726147347	7.9907899741
H	S	497.8082763950	0.2949397002	18.2718671302
H	H_N	14.8722596762	0.3350707193	0.8114802867
H	H_O	474.3279771802	0.2201707538	0.9067765367
O	O	243151.5827176615	0.1973667338	13.7785426553
O	N	281401.4703151011	0.1959457957	13.4418793723
O	F	247829.0305763674	0.1762428756	3.8384247266
O	Cl	2919.3463392305	0.2737547096	29.0580906241
O	S	878339.0410340804	0.1764818652	13.5791662711
O	H_N	1201.9535879202	0.0892297993	3.3131552714
O	H_O	524.3160289698	0.0845445810	7.2289708296
N	N	32850.9010002042	0.2230771786	18.2365102954
N	F	1374.9652128227	0.2791680192	15.3589496830
N	Cl	82489.1919840513	0.2256516512	45.8609220782
N	S	82118.3631565357	0.2274508961	44.3532241433
N	H_N	188842.5508865030	0.1101075417	2.1136961999
N	H_O	2918.5174282638	0.1434007819	9.0951389862
F	F	982.0642961324	0.2743667982	9.7388501789
F	Cl	20730.4191768807	0.2516512966	34.3924884476
F	S	1783836.5012843704	0.1762396646	15.6926560506
F	H_N	479.1079433403	0.2008491273	0.5464810530
F	H_O	102398.2659196005	0.1814067226	2.9479343271
Cl	Cl	24408.4581326576	0.2669833343	86.8471614286
Cl	S	1245.1911738067	0.3337324431	83.2108058162
Cl	H_N	301.8462000990	0.3671376560	53.7312006455
Cl	H_O	1026.4704213718	0.2828088484	3.9253131041
S	S	23198.6273225503	0.2612814752	120.3081967712
S	H_N	6310.3594020364	0.1322292697	9.5935900131
S	H_O	690.4850982997	0.1320154267	3.1215737611
H_N	H_N	25.3586689555	0.3395397433	0.3309531076
H_N	H_O	232.4363829858	0.2607178184	1.4135072045
H_O	H_O	4.5151901757	0.4910103061	0.8739087556

Table J.6: Optimized FF for point-charge electrostatics without PCM, by Scheme 7.

t	t'	$\varepsilon^{t,t'}$ (eV)	$\sigma^{t,t'}$ (Å)	$\lambda_r^{t,t'}$ (-)	$\lambda_a^{t,t'}$ (-)
C	C	0.0049874500	3.3912945495	18.8888326897	6.0000000000
C	H	0.0019584842	2.9772336897	10.4314616550	6.0000000000
C	N	0.0051225297	3.3130157338	13.7334838782	6.0000000000
C	O	0.0052516273	3.1521754369	13.0266078754	6.0000000000

Continued on next page

Table J.6 – continued from previous page

t	t'	$\varepsilon^{t,t'}$ (eV)	$\sigma^{t,t'}$ (Å)	$\lambda_{\mathrm{r}}^{t,t'}$ (-)	$\lambda_{\mathrm{a}}^{t,t'}$ (-)
C	F	0.0031998867	3.1805222978	22.4503911000	6.0000000000
C	S	0.0108101866	3.4459198967	38.1024259037	6.0000000000
C	Cl	0.0082341735	3.5172870576	19.4481990411	6.0000000000
C	$\mathrm{H_N}$	0.0050281704	2.6459475490	12.6043472809	6.0000000000
C	$\mathrm{H_O}$	0.0000146189	2.5867519418	59.5451036840	6.0000000000
H	H	0.0002809004	3.2420110837	6.1003847133	6.0000000000
H	N	0.0004900602	3.2005848153	7.3550837819	6.0000000000
H	O	0.0001860832	3.5890604602	6.8868084846	6.0000000000
H	F	0.0006816219	3.2868136513	6.1052895845	6.0000000000
H	S	0.0047410789	2.9657739030	7.6099558312	6.0000000000
H	Cl	0.0013604491	3.3319599673	6.0722183677	6.0000000000
H	$\mathrm{H_N}$	0.0000016784	6.0952113055	6.0775030790	6.0000000000
H	$\mathrm{H_O}$	0.0004517127	2.5856982664	21.0216194765	6.0000000000
N	N	0.0043027859	3.2916960785	15.6861707409	6.0000000000
N	O	0.0027454154	3.3688855921	16.7676348795	6.0000000000
N	F	0.0020795200	3.3836805057	8.1277993733	6.0000000000
N	S	0.0092986471	3.3612465982	18.5906138500	6.0000000000
N	Cl	0.0143531696	3.2817221326	22.7503628529	6.0000000000
N	$\mathrm{H_N}$	0.1376602090	1.5956623021	55.9485708456	6.0000000000
N	$\mathrm{H_O}$	0.7704899557	1.4188000000	59.0998613734	6.0000000000
O	O	0.0030920730	3.4041945898	13.6009813200	6.0000000000
O	F	0.0014746133	3.0602982075	45.7456229106	6.0000000000
O	S	0.0052960797	3.1267772175	15.7368225691	6.0000000000
O	Cl	0.0067743938	3.1876565782	9.5088097685	6.0000000000
O	$\mathrm{H_N}$	0.1960844207	1.5179000000	58.2537921916	6.0000000000
O	$\mathrm{H_O}$	0.6126408085	1.3894000000	59.7144431455	6.0000000000
F	F	0.0021128402	3.2302463430	9.7122510784	6.0000000000
F	S	0.0068027802	3.3225252491	55.8587037704	6.0000000000
F	Cl	0.0050842549	3.4349346487	23.4883189885	6.0000000000
F	$\mathrm{H_N}$	0.0000669290	2.2475968632	57.9901406487	6.0000000000
F	$\mathrm{H_O}$	0.0000286782	3.4790340720	23.8541153845	6.0000000000
S	S	0.0324020138	3.2422499270	17.3940149181	6.0000000000
S	Cl	0.0072631883	3.3757522959	6.0771754294	6.0000000000
S	$\mathrm{H_N}$	0.1376876752	2.0115000000	59.6982327710	6.0000000000
S	$\mathrm{H_O}$	0.0000058051	2.0009269231	6.0563221709	6.0000000000
Cl	Cl	0.0123725250	3.5586151108	12.0071606592	6.0000000000
Cl	$\mathrm{H_N}$	0.0021702816	3.7079375976	6.0989280952	6.0000000000
Cl	$\mathrm{H_O}$	0.0000003112	10.8288779538	6.0640949774	6.0000000000
$\mathrm{H_N}$	$\mathrm{H_N}$	0.0000002260	9.7638301963	6.0611070400	6.0000000000
$\mathrm{H_N}$	$\mathrm{H_O}$	0.0010321820	3.1241596746	6.1467426712	6.0000000000
$\mathrm{H_O}$	$\mathrm{H_O}$	0.0000090963	5.4275355912	6.0825720228	6.0000000000

J.2 Point-charge Electrostatic Model, with PCM

Table J.7: Optimized FF for point-charge electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 1.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3054.0039690254	0.2777777778	23.8841970184
C	H	451.6381330322	0.2724795640	4.6574578919
C	O	2620.3377558094	0.2645502646	14.2709531817
C	N	3366.7430494099	0.2710027100	20.3220849662
C	F	3682.8164116398	0.2577319588	10.9686291435
C	Cl	5751.1051739138	0.2812939522	43.3946077779
C	S	3655.6225371480	0.2898550725	43.9766460801
C	H_p	249.5689696033	0.2421307506	0.0000000000
H	H	93.8407475324	0.2673796791	1.5163900891
H	O	232.3910109338	0.2597402597	0.0088756494
H	N	293.6494787366	0.2659574468	1.1931831396
H	F	326.2971257420	0.2531645570	0.8617462967
H	Cl	483.3726540934	0.2758620690	5.5702173365
H	S	683.7708016100	0.2840909091	11.7957664800
H	H_p	0.0000000000	0.2380952381	0.0000000000
O	O	6334.9952743819	0.2525252525	20.3516780405
O	N	6120.1042451840	0.2583979328	20.0924312509
O	F	2790.3530259640	0.2463054187	4.2620818660
O	Cl	5351.5034081876	0.2677376171	24.6995004318
O	S	2664.3694848971	0.2754820937	15.5476421089
O	H_p	14.8635962053	0.2320185615	0.0000000000
N	N	7291.2308335311	0.2645502646	25.1491201536
N	F	5197.5767988627	0.2518891688	18.0209866543
N	Cl	7265.2087117977	0.2743484225	46.9612792850
N	S	3188.6922071026	0.2824858757	25.8247931545
N	H_p	0.0000000000	0.2369668246	0.0000000000
F	F	2551.8322102111	0.2403846154	5.1417716703
F	Cl	13920.7890562938	0.2607561930	33.1038830409
F	S	3223.6230653145	0.2680965147	12.7654934144
F	H_p	250.0073604080	0.2267573696	0.0000000000
Cl	Cl	9905.7624448900	0.2849002849	82.6563040021
Cl	S	12583.2797526586	0.2936857562	113.8339201844
Cl	H_p	0.0000000000	0.2447980416	0.0000000000
S	S	4512.7882825471	0.3030303030	128.6661799627
S	H_p	172.5408092286	0.2512562814	0.0000000000
H_p	H_p	0.0000000000	0.2145922747	0.0000000000

Table J.8: Optimized FF for point-charge electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 2.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	2968.4960237325	0.2777777778	23.6832149760
C	H	423.4775088593	0.2724795640	4.3029804319
C	O	3293.5616883163	0.2645502646	17.6950043131
C	N	3897.3673142310	0.2710027100	22.3760735755
C	F	3289.0340259408	0.2577319588	11.2150613892
C	Cl	6314.2080810067	0.2812939522	44.2560086155
C	S	3395.5885736915	0.2898550725	42.2724417272
C	H_p	0.0000000000	0.2421307506	0.0000000000
H	H	80.2248258729	0.2673796791	1.4580419801
H	O	223.2034484773	0.2597402597	0.0000000000
H	N	275.5366387318	0.2659574468	0.9997322598
H	F	286.8686018481	0.2531645570	0.2194083953
H	Cl	391.2055402292	0.2758620690	4.9081910743
H	S	613.3426019430	0.2840909091	10.6617830040
H	H_p	0.0000000000	0.2380952381	0.0000000000
O	O	1942.7998537815	0.2525252525	8.3687271615
O	N	3820.1281450847	0.2583979328	18.9777621073
O	F	2336.9938675752	0.2463054187	4.6652804068
O	Cl	2651.5331353795	0.2677376171	17.3592212339
O	S	2775.1131385498	0.2754820937	10.6560848623
O	H_p	49.8986916311	0.2320185615	0.0000000000
N	N	5971.1976132792	0.2645502646	25.5736723968
N	F	3856.1216526644	0.2518891688	11.7883147862
N	Cl	6825.8359758746	0.2743484225	44.3707366569
N	S	2569.1323502650	0.2824858757	31.7524508778
N	H_p	0.0000000000	0.2369668246	0.0000000000
F	F	2300.1947412079	0.2403846154	4.3990751085
F	Cl	11896.1116549046	0.2607561930	28.5929263582
F	S	3567.0025152576	0.2680965147	13.6186264366
F	H_p	174.1138737897	0.2267573696	0.0000000000
Cl	Cl	10465.2832930892	0.2849002849	85.0323864709
Cl	S	6647.9488264484	0.2936857562	81.5878067572
Cl	H_p	840.5178829697	0.2447980416	0.0000000000
S	S	4822.0380884310	0.3030303030	143.0141693015
S	H_p	145.3887884635	0.2512562814	0.0000000000
H_p	H_p	0.0000000000	0.2145922747	0.0000000000

Table J.9: Optimized FF for point-charge electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 3.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3135.3753544644	0.2777777778	24.2342721711
C	H	435.3851826490	0.2724795640	4.3730044115
C	O	2956.5545489224	0.2645502646	16.0508666052
C	N	3771.4164926557	0.2710027100	21.8324740446
C	F	3423.5008971243	0.2577319588	11.3262092799
C	Cl	6275.3576346842	0.2812939522	44.6273782618
C	S	3581.5222072885	0.2898550725	42.7519049679
C	H_p	460.8287214221	0.2421307506	2.7699245201
H	H	98.9611461562	0.2673796791	1.6963040553
H	O	240.1195173888	0.2597402597	0.0000000000
H	N	250.6359438220	0.2659574468	0.0351924756
H	F	320.2130895557	0.2531645570	0.5054545410
H	Cl	422.2951292994	0.2758620690	5.0878192322
H	S	644.2599977237	0.2840909091	11.2091967141
H	H_p	310.5278721065	0.2380952381	2.6669665260
O	O	4489.9355834468	0.2525252525	13.5497645230
O	N	7991.6048228958	0.2583979328	27.3362562211
O	F	2554.2148984680	0.2463054187	5.9244438987
O	Cl	2912.5324566801	0.2677376171	20.6011409753
O	S	2860.1118256978	0.2754820937	12.5228659913
O	H_p	0.0000000000	0.2320185615	0.2217221101
N	N	11201.9076626023	0.2645502646	33.0818320033
N	F	4770.7670648866	0.2518891688	13.1957917030
N	Cl	8321.3638088127	0.2743484225	49.6122263477
N	S	3704.4023839561	0.2824858757	35.6754146104
N	H_p	0.0000000000	0.2369668246	2.5781352126
F	F	2441.5952428242	0.2403846154	4.6156460058
F	Cl	11111.7741424328	0.2607561930	25.0266869397
F	S	3218.2075568362	0.2680965147	11.0484232396
F	H_p	259.3876871431	0.2267573696	0.0000000000
Cl	Cl	10621.4334845657	0.2849002849	85.1694039171
Cl	S	4834.4400763508	0.2936857562	71.7350076639
Cl	H_p	1240.1334680693	0.2447980416	0.0000000000
S	S	4783.6183636571	0.3030303030	141.6949728086
S	H_p	406.9051006881	0.2512562814	4.7227408323
H_p	H_p	0.0000000000	0.2145922747	0.3044564649

Table J.10: Optimized FF for point-charge electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 4.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	2927.3060586749	0.2777777778	23.8662498051
C	H	418.6554030735	0.2724795640	4.5439703523
C	O	2773.3425311509	0.2645502646	16.4343742688
C	N	3812.9412638382	0.2710027100	23.4300521632
C	F	3329.8553976916	0.2577319588	11.3740681227
C	Cl	6004.1748781962	0.2812939522	44.3685141360
C	S	3231.8716555812	0.2898550725	42.0507637142
C	H_p	160.8446533793	0.2421307506	0.6726187455
H	H	85.1307417563	0.2673796791	1.6315709335
H	O	254.4450219482	0.2597402597	0.0000000000
H	N	246.4951898938	0.2659574468	0.0000000000
H	F	359.9087337735	0.2531645570	0.7936446829
H	Cl	365.2247308649	0.2758620690	4.7279952902
H	S	605.4920224121	0.2840909091	11.7050599494
H	H_p	203.8173574556	0.2380952381	2.4903699847
O	O	4480.7942860704	0.2525252525	14.3715985462
O	N	7196.6797556415	0.2583979328	22.5125353386
O	F	2714.4200855409	0.2463054187	5.8586528455
O	Cl	2688.4004813428	0.2677376171	20.2476949452
O	S	2171.8791449749	0.2754820937	9.7514046512
O	H_p	0.0000000000	0.2320185615	0.6968399567
N	N	12571.2600065186	0.2645502646	31.6064370848
N	F	7034.4244578159	0.2518891688	14.4336287420
N	Cl	7964.2720373865	0.2743484225	48.6802939059
N	S	3269.6549710753	0.2824858757	31.4258402177
N	H_p	0.0000000000	0.2369668246	5.7468618329
F	F	2628.2614532872	0.2403846154	4.9565238651
F	Cl	9943.0766357422	0.2607561930	22.5322017294
F	S	3311.6362785221	0.2680965147	12.6306662060
F	H_p	296.9967534373	0.2267573696	0.7837471813
Cl	Cl	10229.3670041553	0.2849002849	85.5019883239
Cl	S	4766.5163653024	0.2936857562	74.7782062816
Cl	H_p	1165.1227637834	0.2447980416	0.0000000000
S	S	3895.2147190012	0.3030303030	135.7214252390
S	H_p	398.6498838408	0.2512562814	6.0261922625
H_p	H_p	0.0000000000	0.2145922747	0.0000000000

Table J.11: Optimized FF for point-charge electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 5.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	2818.0867238271	0.2777777778	23.2171024282
C	H	438.8345800157	0.2724795640	4.9494510083
C	O	2892.9261712855	0.2645502646	16.0668328219
C	N	3805.1719246381	0.2710027100	21.2850945244
C	F	2947.1735979504	0.2577319588	10.4482845167
C	Cl	6291.3370785613	0.2812939522	45.4371191143
C	S	3488.4968553824	0.2898550725	42.7927415237
C	$\mathrm{H_N}$	1023.1554446737	0.2421307506	5.3159642940
C	$\mathrm{H_O}$	0.0000000000	0.2421307506	2.8907815812
H	H	76.4551261425	0.2673796791	1.3665065168
H	O	246.4041471325	0.2597402597	0.0000000000
H	N	269.9975236504	0.2659574468	1.0773623772
H	F	348.3264657392	0.2531645570	0.9801498420
H	Cl	455.3722130529	0.2758620690	5.9375535797
H	S	597.3032638371	0.2840909091	11.6739816053
H	$\mathrm{H_N}$	66.3880221449	0.2380952381	0.6407035494
H	$\mathrm{H_O}$	258.0689999558	0.2380952381	1.9165343998
O	O	7540.0497996098	0.2525252525	16.7402697123
O	N	8444.2516271199	0.2583979328	22.4111396604
O	F	2175.2197271982	0.2463054187	5.6209173189
O	Cl	2943.6851587266	0.2677376171	22.4900465981
O	S	2870.4413063354	0.2754820937	16.0019469288
O	$\mathrm{H_N}$	0.0000000000	0.2320185615	0.5337526940
O	$\mathrm{H_O}$	0.0000000000	0.2320185615	4.0547835651
N	N	6141.0506073041	0.2645502646	25.6028686972
N	F	4784.6061296438	0.2518891688	14.0526762796
N	Cl	7947.5073621339	0.2743484225	51.8732428573
N	S	3003.0630244616	0.2824858757	36.3516278669
N	$\mathrm{H_N}$	16.6899455025	0.2369668246	0.0451168231
N	$\mathrm{H_O}$	62.8644037465	0.2369668246	9.9326126995
F	F	2636.0631038647	0.2403846154	5.3865520620
F	Cl	12660.4106382197	0.2607561930	28.9242976925
F	S	3653.5952446297	0.2680965147	11.5401817204
F	$\mathrm{H_N}$	284.9642011349	0.2267573696	0.0000000000
F	$\mathrm{H_O}$	5943.9498862573	0.2267573696	2.4260862000
Cl	Cl	10379.3933174934	0.2849002849	84.8862564437
Cl	S	6344.6631599039	0.2936857562	76.0473118476
Cl	$\mathrm{H_N}$	1335.1383599892	0.2447980416	0.0000000000
Cl	$\mathrm{H_O}$	2097.8029626848	0.2447980416	0.0000000000
S	S	3979.9720020738	0.3030303030	136.6116898372

Continued on next page

Table J.11 – continued from previous page

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
S	H_N	447.1229570903	0.2512562814	5.7222820403
S	H_O	122.6898130289	0.2512562814	0.1374598850
H_N	H_N	0.0000000000	0.2145922747	0.2486147106
H_N	H_O	965.6437131296	0.2145922747	2.9282508692
H_O	H_O	0.0000000000	0.2145922747	0.0000000000

Table J.12: Optimized FF for point-charge electrostatics with PCM (ε_{PCM} = 11), by Scheme 6.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	195715.7402245403	0.2045009335	20.4854313163
C	H	997.7645612514	0.2507968215	4.8886106975
C	O	2513.2919770941	0.2706430953	15.6757391869
C	N	127499.2836187431	0.2063688641	18.4046321720
C	F	23458.4789840986	0.2211204042	10.4574297797
C	Cl	1841289.4629583443	0.1906210106	35.4159527777
C	S	4013347.8410754274	0.1764555128	29.7132949567
C	H_N	36844.6905445428	0.1775598006	2.3708434539
C	H_O	10357869.4180571064	0.1054846393	1.7553508028
H	H	183.8324432568	0.2568790041	1.8328392784
H	O	221.6770403579	0.2658056140	0.9828685129
H	N	104.1650409261	0.3075967001	2.4942106013
H	F	401.3359299236	0.2545895610	1.1301127903
H	Cl	759.1456703613	0.2655742834	4.9280855458
H	S	12374.7052019556	0.2132806049	7.4543527627
H	H_N	1233.1714227483	0.1505647375	0.2733413070
H	H_O	1476.7127860638	0.2049639596	2.5040131750
O	O	583109.6097490498	0.1833101919	11.4218838466
O	N	109993.5293628790	0.2121997442	18.9207844790
O	F	402350.2706940376	0.1701179163	5.7427597058
O	Cl	8617.9536335998	0.2444524042	32.6787868896
O	S	17050999.1176321842	0.1519564294	11.1477817534
O	H_N	170.1948718743	0.0831856435	2.3332071661
O	H_O	248.4518778352	0.1393221766	4.7260531798
N	N	74708.8139749980	0.2124839472	13.5384262595
N	F	3603.0730733586	0.2626053429	15.8029075321
N	Cl	1904.9469957089	0.3121674067	58.7954803552
N	S	1142639.7078604726	0.1815309652	23.5550494103
N	H_N	257.3860048851	0.1110870746	0.4620623076
N	H_O	608.1643734186	0.1358925600	7.0466139944
F	F	5512.2810015461	0.2412602440	8.6488230695
F	Cl	9775.2729224751	0.2613475619	28.5717195127

Continued on next page

Table J.12 – continued from previous page

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
F	S	6076408.4950627992	0.1712197490	9.0772257061
F	H_N	118.3139568209	0.2927280522	1.2325741530
F	H_O	613684.5673062574	0.1630278256	2.0945607301
Cl	Cl	22985.6516429373	0.2706479865	94.9917516517
Cl	S	13247.5400069742	0.2726952557	56.2067122227
Cl	H_N	172.3980391202	0.3436239451	11.5385916363
Cl	H_O	304.4024256852	0.3271524465	2.1165029558
S	S	569588.4902126211	0.2004330965	105.6935685166
S	H_N	924688.8290038572	0.1287325293	4.2400563853
S	H_O	106.0042005982	0.2507196293	3.2096379353
H_N	H_N	9.7959190901	0.2784562337	0.1274098564
H_N	H_O	245.9125639548	0.2424214469	1.7898208494
H_O	H_O	544.9114070259	0.1102361276	0.1241825293

Table J.13: Optimized FF for point-charge electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 7.

t	t'	$\varepsilon^{t,t'}$ (eV)	$\sigma^{t,t'}$ (Å)	$\lambda_{\mathrm{r}}^{t,t'}$ (-)	$\lambda_{\mathrm{a}}^{t,t'}$ (-)
C	C	0.0048812219	3.4007779687	19.4121098726	6.0000000000
C	H	0.0019723046	2.9883740812	10.9809657953	6.0000000000
C	N	0.0048013862	3.3406606479	18.9399969169	6.0000000000
C	O	0.0037835976	3.2631635765	10.9897137460	6.0000000000
C	F	0.0034326691	3.1843317930	16.2829181342	6.0000000000
C	S	0.0103736772	3.4313555010	28.0697322731	6.0000000000
C	Cl	0.0094658467	3.5206480428	25.0181893223	6.0000000000
C	H_N	0.0030124026	2.6768244401	11.8403442910	6.0000000000
C	H_O	0.0116447968	2.3237960335	56.0793555067	6.0000000000
H	H	0.0003149849	2.9914439023	7.7288511123	6.0000000000
H	N	0.0003271339	3.5809341307	6.8562586564	6.0000000000
H	O	0.0000391151	4.4517401300	7.0764213069	6.0000000000
H	F	0.0001268993	3.8853351994	6.9657056867	6.0000000000
H	S	0.0037736720	3.0572404491	10.3002122995	6.0000000000
H	Cl	0.0009973204	3.4061113725	6.6618832171	6.0000000000
H	H_N	0.0010000421	2.2256297103	36.6207654471	6.0000000000
H	H_O	0.0023994000	2.4482928590	56.0256106797	6.0000000000
N	N	0.0051089283	3.2884726276	14.3407581426	6.0000000000
N	O	0.0041097062	3.2933354560	14.2743189982	6.0000000000
N	F	0.0015091389	3.5752874764	6.0679802025	6.0000000000
N	S	0.0120558400	3.2696311661	45.5883328769	6.0000000000
N	Cl	0.0091659649	3.1706839195	10.2039902094	6.0000000000
N	H_N	0.0017834502	1.6058189039	13.9955428682	6.0000000000
N	H_O	0.3357849192	1.4228160205	37.5526096199	6.0000000000

Continued on next page

Table J.13 – continued from previous page

t	t'	$\varepsilon^{t,t'}$ (eV)	$\sigma^{t,t'}$ (Å)	$\lambda_{\mathrm{r}}^{t,t'}$ (-)	$\lambda_{\mathrm{a}}^{t,t'}$ (-)
O	O	0.0029024758	3.3776939840	13.0397817046	6.0000000000
O	F	0.0013492242	3.0830713298	39.8415284828	6.0000000000
O	S	0.0024872821	3.2183595521	28.4422521348	6.0000000000
O	Cl	0.0058088055	3.2231448482	10.2019325712	6.0000000000
O	$\mathrm{H_N}$	0.0000540349	1.6852856977	10.0619258208	6.0000000000
O	$\mathrm{H_O}$	0.3187958039	1.3905132875	56.9917064837	6.0000000000
F	F	0.0015662908	3.2576943183	11.2906969937	6.0000000000
F	S	0.0040351075	3.3939737238	29.6279722163	6.0000000000
F	Cl	0.0027997378	3.6115163251	8.3450514893	6.0000000000
F	$\mathrm{H_N}$	0.0000144410	4.3297037409	6.8589805537	6.0000000000
F	$\mathrm{H_O}$	0.0000016611	3.6424430380	38.6369151107	6.0000000000
S	S	0.0378393292	3.2487376461	18.6251510329	6.0000000000
S	Cl	0.0090498029	3.5136433883	9.5705044426	6.0000000000
S	$\mathrm{H_N}$	0.0132851115	2.1939484612	53.8089746762	6.0000000000
S	$\mathrm{H_O}$	0.0002740725	3.0364668449	6.5590581255	6.0000000000
Cl	Cl	0.0135168667	3.5476612428	15.2887845712	6.0000000000
Cl	$\mathrm{H_N}$	0.0007365499	3.8172137759	7.5708513831	6.0000000000
Cl	$\mathrm{H_O}$	0.0000126839	5.9406617939	7.5509467551	6.0000000000
$\mathrm{H_N}$	$\mathrm{H_N}$	0.0000162121	2.0472802733	52.9404895805	6.0000000000
$\mathrm{H_N}$	$\mathrm{H_O}$	0.0028789668	2.3942583345	18.0826967888	6.0000000000
$\mathrm{H_O}$	$\mathrm{H_O}$	0.0006458403	1.5232412324	54.1871607598	6.0000000000

J.3 Multipole Electrostatic Model, without PCM

Table J.14: Optimized FF for multipole electrostatics without PCM, by Scheme 1.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3301.5730419371	0.2777777778	24.2878042115
C	H	426.4168351402	0.2724795640	4.9387729959
C	O	2533.5883805171	0.2645502646	17.0591097935
C	N	2923.0002953502	0.2710027100	22.4010218033
C	F	2431.7823302240	0.2577319588	7.8900539249
C	Cl	6386.4443299068	0.2812939522	46.1138284058
C	S	4355.8231868369	0.2898550725	47.7702019440
C	H_p	118.4767243667	0.2421307506	0.0000000000
H	H	118.2024913062	0.2673796791	1.5589623619
H	O	315.2115253587	0.2597402597	1.8274910309
H	N	433.1522579779	0.2659574468	3.6369938723
H	F	266.6198127419	0.2531645570	1.3487382932
H	Cl	741.9575888595	0.2758620690	8.6458221351
H	S	652.8198559512	0.2840909091	12.5903318657
H	H_p	0.0000000000	0.2380952381	0.0000000000
O	O	4260.7265314549	0.2525252525	11.1114303787
O	N	5066.8312614787	0.2583979328	16.9258919058
O	F	4213.7393782951	0.2463054187	6.8861806769
O	Cl	5771.7496725555	0.2677376171	23.4092066100
O	S	2806.6768368179	0.2754820937	16.1977683552
O	H_p	109.6690417733	0.2320185615	0.0000000000
N	N	4101.7103375996	0.2645502646	13.6167388505
N	F	4962.3766934392	0.2518891688	22.2268890188
N	Cl	3719.1358498598	0.2743484225	23.0278710860
N	S	2933.8541711598	0.2824858757	26.9931501206
N	H_p	206.6965933762	0.2369668246	0.0000000000
F	F	5014.9648979089	0.2403846154	9.7091321310
F	Cl	6312.3240501747	0.2607561930	18.7650981801
F	S	3897.9039729717	0.2680965147	19.5288416088
F	H_p	84.9508979762	0.2267573696	0.0000000000
Cl	Cl	10341.5798258083	0.2849002849	81.3722921364
Cl	S	6689.0712694336	0.2936857562	91.9672598528
Cl	H_p	3231.5008573126	0.2447980416	0.0000000000
S	S	5309.3937949225	0.3030303030	128.3725377434
S	H_p	475.2927328097	0.2512562814	0.0000000000
H_p	H_p	8.9222263268	0.2145922747	0.0000000000

Table J.15: Optimized FF for multipole electrostatics without PCM, by Scheme 2.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3462.3170071119	0.2777777778	24.2791531784
C	H	400.8124482178	0.2724795640	4.7637164956
C	O	3272.5745374563	0.2645502646	19.7288987454
C	N	3429.7024968772	0.2710027100	25.1034345357
C	F	1731.9931640720	0.2577319588	7.7631505093
C	Cl	5708.3708143221	0.2812939522	43.9316632046
C	S	4374.7498018220	0.2898550725	50.8891339754
C	H_p	0.0000000000	0.2421307506	0.0000000000
H	H	111.2655298766	0.2673796791	1.5388490571
H	O	266.6549378491	0.2597402597	0.8156160166
H	N	356.4624219498	0.2659574468	1.9734933433
H	F	391.1626736143	0.2531645570	2.2125798252
H	Cl	743.6497926633	0.2758620690	8.7645998419
H	S	645.8084581190	0.2840909091	11.7374400748
H	H_p	0.0000000000	0.2380952381	0.0000000000
O	O	4067.6372647326	0.2525252525	15.8117423159
O	N	5277.1040701580	0.2583979328	23.9285076385
O	F	2574.1083295458	0.2463054187	5.2709297261
O	Cl	7048.3489834318	0.2677376171	29.6652792669
O	S	2754.0038108395	0.2754820937	14.6090971218
O	H_p	39.9925983662	0.2320185615	0.0000000000
N	N	5247.5087519441	0.2645502646	21.2871859666
N	F	1885.6740127678	0.2518891688	8.5259110729
N	Cl	4118.4447917860	0.2743484225	29.3127864501
N	S	2285.6846833543	0.2824858757	32.2964711507
N	H_p	76.1279396572	0.2369668246	0.0000000000
F	F	3756.2204452811	0.2403846154	7.7432324428
F	Cl	4633.9787456365	0.2607561930	14.0991797741
F	S	2298.5084103366	0.2680965147	7.5902458919
F	H_p	43.2021662118	0.2267573696	0.0000000000
Cl	Cl	9497.9738839498	0.2849002849	79.1175934720
Cl	S	4897.2294237465	0.2936857562	61.4159523742
Cl	H_p	674.4727740418	0.2447980416	0.0000000000
S	S	5778.6961598464	0.3030303030	129.8243685675
S	H_p	201.2556428426	0.2512562814	0.0000000000
H_p	H_p	696.5487779970	0.2145922747	0.0000000000

Table J.16: Optimized FF for multipole electrostatics without PCM, by Scheme 3.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3157.0267956175	0.2777777778	23.9955651310
C	H	420.7862236580	0.2724795640	4.7985466808
C	O	2566.5296368020	0.2645502646	16.5398050960
C	N	3015.0679556519	0.2710027100	23.0749195932
C	F	2314.5800988340	0.2577319588	8.5151602103
C	Cl	6391.7853628695	0.2812939522	45.7465921868
C	S	4408.9385261757	0.2898550725	50.3489618870
C	H_p	895.4392185509	0.2421307506	4.3878582758
H	H	121.5619671628	0.2673796791	1.6896301015
H	O	296.4641905905	0.2597402597	1.4843840317
H	N	395.3503245408	0.2659574468	2.5626632827
H	F	343.2103494223	0.2531645570	1.8082588446
H	Cl	724.3972990844	0.2758620690	8.6787876719
H	S	681.2907640189	0.2840909091	12.1973325309
H	H_p	438.0288664704	0.2380952381	2.2923507535
O	O	5035.2639281498	0.2525252525	16.5356244502
O	N	7336.4464898856	0.2583979328	27.5988829778
O	F	3454.3087813095	0.2463054187	7.9142013594
O	Cl	4894.9526988283	0.2677376171	17.7072630202
O	S	2510.8332842257	0.2754820937	10.8008594650
O	H_p	26.1560753107	0.2320185615	0.4201252705
N	N	7605.2735906100	0.2645502646	23.3922321539
N	F	4539.8626230175	0.2518891688	13.8367126585
N	Cl	3658.0915555522	0.2743484225	25.0277431619
N	S	2664.0937332097	0.2824858757	22.9048569133
N	H_p	168.0432883262	0.2369668246	3.0761761910
F	F	4619.7362919728	0.2403846154	8.8093623484
F	Cl	6789.0486504299	0.2607561930	19.7658320669
F	S	2607.1815405082	0.2680965147	10.8535273655
F	H_p	355.0630918229	0.2267573696	0.7572363277
Cl	Cl	10634.8987646295	0.2849002849	82.7865060087
Cl	S	7144.6640573612	0.2936857562	96.8843389597
Cl	H_p	11141.9756363087	0.2447980416	32.0899895875
S	S	5359.9393391988	0.3030303030	128.5979357465
S	H_p	752.8001773062	0.2512562814	12.2260714988
H_p	H_p	1343.2325848277	0.2145922747	0.0000000000

Table J.17: Optimized FF for multipole electrostatics without PCM, by Scheme 4.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3067.4800178800	0.2777777778	24.0979870332
C	H	388.4387940831	0.2724795640	4.7747014388
C	O	2451.6151839088	0.2645502646	16.7148671644
C	N	2830.3835346156	0.2710027100	23.2040283197
C	F	2223.4445030148	0.2577319588	8.3328678866
C	Cl	6107.1243332478	0.2812939522	45.3741026354
C	S	4179.2242115433	0.2898550725	49.9861948831
C	H_p	635.8704479550	0.2421307506	3.3836137847
H	H	105.5405578658	0.2673796791	1.6558662406
H	O	279.1878142230	0.2597402597	1.4800945705
H	N	369.0862684518	0.2659574468	2.4777660045
H	F	322.7748527830	0.2531645570	1.8531875787
H	Cl	678.4225659665	0.2758620690	8.8049815195
H	S	592.2901130394	0.2840909091	11.9836066050
H	H_p	322.4640155917	0.2380952381	1.8613128733
O	O	4754.5773534624	0.2525252525	14.8607751237
O	N	7313.6821388938	0.2583979328	26.6377878239
O	F	3263.3759906707	0.2463054187	7.5950813556
O	Cl	4752.2254588046	0.2677376171	17.9472398021
O	S	2518.2700800351	0.2754820937	12.0592320407
O	H_p	38.8766028184	0.2320185615	1.5755680250
N	N	7935.2367413000	0.2645502646	23.3543627375
N	F	4303.0135408222	0.2518891688	13.8158828602
N	Cl	3585.4575240747	0.2743484225	25.5711866926
N	S	2378.0081696797	0.2824858757	22.7841688758
N	H_p	81.0636689674	0.2369668246	3.4892740395
F	F	4525.8302300676	0.2403846154	8.9176040147
F	Cl	6657.5403870753	0.2607561930	20.1187037538
F	S	2661.6375116068	0.2680965147	11.9416044325
F	H_p	305.2011872014	0.2267573696	0.7975284621
Cl	Cl	10149.1708890945	0.2849002849	82.4783755991
Cl	S	6828.3135557628	0.2936857562	96.9630010996
Cl	H_p	11154.8086194614	0.2447980416	31.5300916969
S	S	4561.2618209242	0.3030303030	124.3242543259
S	H_p	542.0043339494	0.2512562814	11.1919750249
H_p	H_p	1498.3923886503	0.2145922747	0.0000000000

Table J.18: Optimized FF for multipole electrostatics without PCM, by Scheme 5.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3195.4118119879	0.2777777778	24.1302247605
C	H	417.5298802115	0.2724795640	5.1538923636
C	O	2477.4097524500	0.2645502646	17.1141392439
C	N	2797.4666140366	0.2710027100	22.3239464179
C	F	1836.2477793300	0.2577319588	7.2327526411
C	Cl	6014.8112240491	0.2812939522	44.7524630115
C	S	4416.3734585395	0.2898550725	50.7775288309
C	H_N	1420.3871371688	0.2421307506	7.0231425968
C	H_O	0.0000000000	0.2421307506	0.1609497455
H	H	101.9089535975	0.2673796791	1.4603775905
H	O	275.2004477660	0.2597402597	1.1628644127
H	N	333.8588571297	0.2659574468	3.0196251922
H	F	406.2006414471	0.2531645570	2.8113296421
H	Cl	717.9200682679	0.2758620690	9.2659848882
H	S	560.6090517975	0.2840909091	11.7692349697
H	H_N	113.2778200476	0.2380952381	0.0000000000
H	H_O	510.4358514505	0.2380952381	1.9529549271
O	O	7708.0548298351	0.2525252525	19.9782486660
O	N	8098.8669738537	0.2583979328	23.8060577828
O	F	3587.2097235526	0.2463054187	7.3102746503
O	Cl	6132.2052620987	0.2677376171	26.0881972624
O	S	2140.9448253772	0.2754820937	11.6514594234
O	H_N	12.7497471380	0.2320185615	1.4280044316
O	H_O	38.6938420355	0.2320185615	4.6674411943
N	N	2612.1335111570	0.2645502646	11.4118809205
N	F	1659.6151960434	0.2518891688	10.5222461337
N	Cl	3219.0225457245	0.2743484225	24.4491072897
N	S	3175.4163048920	0.2824858757	27.7643008836
N	H_N	249.4438158847	0.2369668246	2.5369721312
N	H_O	183.9803492592	0.2369668246	9.2449977471
F	F	3631.7130149651	0.2403846154	8.3367838927
F	Cl	5297.0275806594	0.2607561930	17.6215208456
F	S	2937.7634621344	0.2680965147	11.8210104388
F	H_N	456.2624250271	0.2267573696	0.0000000000
F	H_O	0.0000000000	0.2267573696	1.2168991893
Cl	Cl	9850.2145483697	0.2849002849	81.7164004330
Cl	S	5232.7011630546	0.2936857562	82.2906290080
Cl	H_N	16104.9652419817	0.2447980416	48.2788840062
Cl	H_O	11437.3661250039	0.2447980416	22.8721349713
S	S	4994.2092517791	0.3030303030	126.8886795312

Continued on next page

Table J.18 – continued from previous page

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
S	H_N	597.3076137068	0.2512562814	13.6517066068
S	H_O	482.5878289075	0.2512562814	8.9783031100
H_N	H_N	1572.3368400376	0.2145922747	0.0000000000
H_N	H_O	1695.8184952800	0.2145922747	0.0000000000
H_O	H_O	1607.4703121235	0.2145922747	0.0000000000

Table J.19: Optimized FF for multipole electrostatics without PCM, by Scheme 6.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	18921.8387739819	0.2423331881	22.8417084345
C	H	580.8449211623	0.2657202542	5.5348953364
C	O	2516.6546231088	0.2609128464	14.9925025649
C	N	6487.0862289917	0.2540763175	22.4409324323
C	F	22460.7533976998	0.2105731323	6.8621675782
C	Cl	2847.7222945957	0.2915157529	40.6198723969
C	S	28780.5854218587	0.2506202532	48.5377876601
C	H_N	1228.2960553417	0.2366098008	5.2337238389
C	H_O	106.8523226804	0.2468618659	1.7596727944
H	H	105.5887321243	0.2714358884	1.5199126182
H	O	325.0274570193	0.2570296472	1.8472186355
H	N	310.3500217551	0.2737421811	3.8753710856
H	F	697.8350831318	0.2460327801	2.8200085838
H	Cl	810.4517154605	0.2822783328	11.8432766755
H	S	543.3209932514	0.2781688686	9.4760587776
H	H_N	51.2513693296	0.2671880662	0.5090839228
H	H_O	92.6023186053	0.2860656811	2.6290726079
O	O	29517.3472653426	0.2154194319	12.1856171955
O	N	20131.8175421081	0.2341026394	18.8020179425
O	F	6965.1858825976	0.2371406181	7.9178712788
O	Cl	6830.0080294235	0.2724691105	35.9890686456
O	S	6563.3878288881	0.2539405046	16.1202121192
O	H_N	309.9050418588	0.2026087141	2.0775570259
O	H_O	508.7478654445	0.1776954085	2.8186131572
N	N	1624.3218257942	0.2712482016	10.8016157821
N	F	11234.3847296914	0.2264308068	10.3905101550
N	Cl	10425.6781728449	0.2613003153	33.5338205538
N	S	5682.8899038301	0.2679320675	31.9377647585
N	H_N	1250.6103515655	0.1910562366	1.2311065996
N	H_O	207.9005928277	0.2203570158	5.7355442120
F	F	5972.3788991197	0.2297483536	7.3865345045
F	Cl	18247.1819721866	0.2396239898	21.3739637791

Continued on next page

Table J.19 – continued from previous page

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
F	S	39032.1515593130	0.2122011939	13.0996434897
F	H_N	7926.8943580525	0.1831646765	0.9780709715
F	H_O	475.9484348544	0.1278995673	1.9554693002
Cl	Cl	6870.6555110615	0.2956416207	90.0182042646
Cl	S	2807.6112670693	0.3136806829	74.9861335559
Cl	H_N	430.9130657880	0.3440380797	52.0430590973
Cl	H_O	1094.4936020388	0.2910691590	7.7230442970
S	S	4529.6142544956	0.3071990153	129.0057257106
S	H_N	2917.2213783889	0.1998812876	5.0956196725
S	H_O	212.3884606887	0.2754072971	8.9994676047
H_N	H_N	33.3826001285	0.3701379771	6.1330835239
H_N	H_O	27.7983375285	0.3677285042	0.7931449361
H_O	H_O	14.0055703366	0.5122602004	40.5490815978

Table J.20: Optimized FF for multipole electrostatics without PCM, by Scheme 7.

t	t'	$\varepsilon^{t,t'}$ (eV)	$\sigma^{t,t'}$ (Å)	$\lambda_r^{t,t'}$ (-)	$\lambda_a^{t,t'}$ (-)
C	C	0.0049906707	3.4063854561	18.0756417028	6.0000000000
C	H	0.0019930372	2.9524175962	11.7764203842	6.0000000000
C	N	0.0057788783	3.2369692964	13.5107191200	6.0000000000
C	O	0.0054854006	3.1201208471	13.7961632261	6.0000000000
C	F	0.0028517125	3.1354889689	21.1117895993	6.0000000000
C	S	0.0100031503	3.4659640125	18.3443848937	6.0000000000
C	Cl	0.0079345802	3.4509887018	14.1401414364	6.0000000000
C	H_N	0.0017806674	2.9631620409	7.6179516321	6.0000000000
C	H_O	0.0106009906	2.2067699531	53.8312540781	6.0000000000
H	H	0.0004201546	3.1094906311	6.4743279258	6.0000000000
H	N	0.0005164360	3.2116138431	9.3886677457	6.0000000000
H	O	0.0002277440	3.4508162436	8.1834486797	6.0000000000
H	F	0.0007852745	3.1733397013	6.7303066196	6.0000000000
H	S	0.0025821755	3.0868585245	8.5183839007	6.0000000000
H	Cl	0.0016664935	3.4483227126	6.1511350306	6.0000000000
H	H_N	0.0000899344	3.3872100232	6.7764182912	6.0000000000
H	H_O	0.0001132738	3.3127789972	11.5925513998	6.0000000000
N	N	0.0030644902	3.3274090562	15.7117476440	6.0000000000
N	O	0.0051126659	3.2316619868	14.2016525470	6.0000000000
N	F	0.0035783583	3.2246085054	13.9937327645	6.0000000000
N	S	0.0042772945	3.4376189971	9.8205596420	6.0000000000
N	Cl	0.0041634306	3.2016047449	16.2235488271	6.0000000000
N	H_N	0.0010487265	2.0202694199	6.4307190923	6.0000000000
N	H_O	0.0539685646	1.4247737685	6.3747477307	6.0000000000

Continued on next page

Table J.20 – continued from previous page

t	t'	$\varepsilon^{t,t'}$ (eV)	$\sigma^{t,t'}$ (Å)	$\lambda_{\mathrm{r}}^{t,t'}$ (-)	$\lambda_{\mathrm{a}}^{t,t'}$ (-)
O	O	0.0033209702	3.2631449485	14.5936291504	6.0000000000
O	F	0.0014000971	3.1794044801	18.2719104005	6.0000000000
O	S	0.0016728088	3.4640034953	14.7138923748	6.0000000000
O	Cl	0.0032500314	3.5218502298	9.1327409059	6.0000000000
O	$\mathrm{H_N}$	0.0000152708	1.6184510008	34.4567169604	6.0000000000
O	$\mathrm{H_O}$	0.1810652018	1.4155480057	40.9093409064	6.0000000000
F	F	0.0024432159	3.1909320453	10.2144043637	6.0000000000
F	S	0.0042589755	3.2765893588	55.1873115635	6.0000000000
F	Cl	0.0030813781	3.4540849417	15.3458015873	6.0000000000
F	$\mathrm{H_N}$	0.0000324223	2.5359935604	55.8695315497	6.0000000000
F	$\mathrm{H_O}$	0.0003669754	3.0400864747	17.9702283998	6.0000000000
S	S	0.0250644176	3.3556665230	13.7828439763	6.0000000000
S	Cl	0.0131292681	3.4192946496	10.7427455212	6.0000000000
S	$\mathrm{H_N}$	0.0140972151	2.2673196619	6.9172460431	6.0000000000
S	$\mathrm{H_O}$	0.0127704588	2.2939190098	6.2819499514	6.0000000000
Cl	Cl	0.0101803323	3.5903040875	10.2637727708	6.0000000000
Cl	$\mathrm{H_N}$	0.0106847522	3.2182788341	6.5291346130	6.0000000000
Cl	$\mathrm{H_O}$	0.0049714315	3.2222114222	18.8465930345	6.0000000000
$\mathrm{H_N}$	$\mathrm{H_N}$	0.0000925102	3.6588319640	9.1054146771	6.0000000000
$\mathrm{H_N}$	$\mathrm{H_O}$	0.0000121620	1.8365008019	56.8992091229	6.0000000000
$\mathrm{H_O}$	$\mathrm{H_O}$	0.0000253181	1.5243591930	56.8409707002	6.0000000000

J.4 Multipole Electrostatic Model, with PCM

Table J.21: Optimized FF for multipole electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 1.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3189.1079585490	0.2777777778	24.3600755367
C	H	380.8314106702	0.2724795640	4.0185556138
C	O	2482.9952444614	0.2645502646	14.7848099306
C	N	2474.7047993329	0.2710027100	18.5787072893
C	F	2762.5155187527	0.2577319588	9.5061822326
C	Cl	6303.8388435441	0.2812939522	47.2439860667
C	S	4711.6739776087	0.2898550725	54.1973872361
C	H_p	261.6398918394	0.2421307506	0.0000000000
H	H	102.0986969163	0.2673796791	1.7325513465
H	O	306.9622467594	0.2597402597	0.1511934462
H	N	451.3781603605	0.2659574468	2.7555084259
H	F	309.5620065290	0.2531645570	0.6435349476
H	Cl	664.8463195618	0.2758620690	6.4645484636
H	S	488.0629752181	0.2840909091	5.5026980193
H	H_p	0.0000000000	0.2380952381	0.0000000000
O	O	5264.7585871544	0.2525252525	17.9124896683
O	N	6653.8839633648	0.2583979328	25.4938856553
O	F	4137.9622027111	0.2463054187	7.2318856326
O	Cl	5300.7339329188	0.2677376171	18.1105841125
O	S	2161.7880318496	0.2754820937	5.7988020252
O	H_p	112.9993319264	0.2320185615	0.0000000000
N	N	4889.7128609476	0.2645502646	13.8215916829
N	F	4073.4545450594	0.2518891688	17.4717855545
N	Cl	5117.3215679542	0.2743484225	25.2662910961
N	S	1474.0539672246	0.2824858757	11.0666382672
N	H_p	201.4939822548	0.2369668246	0.0000000000
F	F	2860.4254699805	0.2403846154	5.8946911782
F	Cl	4810.6285423519	0.2607561930	12.9627841115
F	S	2685.5373125267	0.2680965147	6.6328190810
F	H_p	387.6567220810	0.2267573696	0.0000000000
Cl	Cl	9469.6061633514	0.2849002849	75.8788577510
Cl	S	4568.4750212567	0.2936857562	69.0067751758
Cl	H_p	2703.2008595713	0.2447980416	0.0000000000
S	S	5498.5873587023	0.3030303030	126.2039182909
S	H_p	548.4916492134	0.2512562814	0.0000000000
H_p	H_p	0.0000000000	0.2145922747	0.0000000000

Table J.22: Optimized FF for multipole electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 2.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3329.2131614624	0.2777777778	25.3241318983
C	H	366.9887129620	0.2724795640	3.4756015443
C	O	3155.1594365715	0.2645502646	17.5765513601
C	N	3545.1744595709	0.2710027100	22.4651092600
C	F	2001.0144533124	0.2577319588	8.2819711456
C	Cl	5230.2660771257	0.2812939522	42.8166532295
C	S	4210.5594627224	0.2898550725	51.3047490303
C	H_p	14.6042616236	0.2421307506	0.0000000000
H	H	95.4491337923	0.2673796791	1.8439557708
H	O	302.2417648962	0.2597402597	0.0000000000
H	N	397.4314068950	0.2659574468	1.3942153228
H	F	353.4050679522	0.2531645570	1.0906767524
H	Cl	552.1207564242	0.2758620690	5.5862421079
H	S	595.8887150858	0.2840909091	6.7373441376
H	H_p	0.0000000000	0.2380952381	0.0000000000
O	O	3725.4738063351	0.2525252525	13.2956511536
O	N	6539.3186278431	0.2583979328	23.7310511629
O	F	2792.2350177692	0.2463054187	5.5042280431
O	Cl	6350.5700278326	0.2677376171	25.3222776707
O	S	2789.5870796019	0.2754820937	8.0670724038
O	H_p	81.1615180545	0.2320185615	0.0000000000
N	N	7704.4640560099	0.2645502646	23.3571264150
N	F	2349.8229109168	0.2518891688	8.7290946631
N	Cl	6164.5348818822	0.2743484225	31.5030685370
N	S	2175.9430425686	0.2824858757	23.0326749278
N	H_p	86.1980406722	0.2369668246	0.0000000000
F	F	4164.3646364976	0.2403846154	7.6113805770
F	Cl	4705.5545125480	0.2607561930	13.0991064743
F	S	2405.6990522551	0.2680965147	5.9185018940
F	H_p	321.0181888665	0.2267573696	0.0000000000
Cl	Cl	9737.3599180619	0.2849002849	79.2248577585
Cl	S	3489.1196545694	0.2936857562	52.2188029066
Cl	H_p	1447.2945356443	0.2447980416	0.0000000000
S	S	5518.5347619000	0.3030303030	136.3711853373
S	H_p	685.1748611730	0.2512562814	0.0000000000
H_p	H_p	240.1522343941	0.2145922747	0.0000000000

Table J.23: Optimized FF for multipole electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 3.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3265.0723770859	0.2777777778	24.9186132081
C	H	376.3521762085	0.2724795640	3.6833412579
C	O	2906.3458207309	0.2645502646	16.7598556702
C	N	2759.1353943281	0.2710027100	19.9245093969
C	F	2339.2448297005	0.2577319588	8.9368137857
C	Cl	6278.6450436424	0.2812939522	46.5913325464
C	S	4447.8076072166	0.2898550725	52.3575914234
C	H_p	558.7193057819	0.2421307506	2.3324603079
H	H	107.7757383587	0.2673796791	1.9734391479
H	O	315.2375534612	0.2597402597	0.0619374772
H	N	391.2850372484	0.2659574468	1.2926862546
H	F	364.4902676077	0.2531645570	1.1307009175
H	Cl	680.0659953947	0.2758620690	6.7220968968
H	S	584.2829401203	0.2840909091	6.4848368942
H	H_p	278.2405739392	0.2380952381	2.3739354612
O	O	3974.0200049799	0.2525252525	13.2906544738
O	N	7699.1174871315	0.2583979328	26.4429557782
O	F	2466.9205918045	0.2463054187	5.3613908627
O	Cl	5250.8956105343	0.2677376171	17.9906824747
O	S	2376.2296695520	0.2754820937	4.8490316564
O	H_p	73.7154189538	0.2320185615	0.0000000000
N	N	9253.3322994642	0.2645502646	24.4719390068
N	F	2930.5968571717	0.2518891688	10.3846235353
N	Cl	5889.3181536460	0.2743484225	27.9657898414
N	S	2004.2143380052	0.2824858757	16.4650208040
N	H_p	160.3623144405	0.2369668246	2.0451789360
F	F	2823.8797842800	0.2403846154	5.8311775967
F	Cl	4963.6940004097	0.2607561930	13.5234787428
F	S	2714.1659661643	0.2680965147	4.6945090808
F	H_p	513.5650663135	0.2267573696	0.0000000000
Cl	Cl	10041.3782787494	0.2849002849	78.7257354202
Cl	S	4832.5833034878	0.2936857562	72.2788574078
Cl	H_p	6548.4800993580	0.2447980416	17.8715588117
S	S	5644.0092253972	0.3030303030	138.1611725832
S	H_p	1264.9839183025	0.2512562814	7.5054864075
H_p	H_p	597.4174783903	0.2145922747	0.0000000000

Table J.24: Optimized FF for multipole electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 4.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3118.4714646092	0.2777777778	24.7135809868
C	H	363.3678064189	0.2724795640	3.8789325944
C	O	2722.0981845986	0.2645502646	16.4628200940
C	N	2656.7722218774	0.2710027100	19.9889002830
C	F	1964.4983797676	0.2577319588	8.5560876903
C	Cl	6090.5638858911	0.2812939522	46.7921857578
C	S	4415.2066567315	0.2898550725	52.8894356184
C	H_p	512.6941562533	0.2421307506	2.6413387539
H	H	93.2709767674	0.2673796791	1.9253302238
H	O	305.9817801186	0.2597402597	0.0000000000
H	N	373.0856176949	0.2659574468	1.2008383106
H	F	354.8042235596	0.2531645570	1.0929404989
H	Cl	635.9707058905	0.2758620690	6.6224835128
H	S	508.5327054373	0.2840909091	5.8186683953
H	H_p	54.7453694804	0.2380952381	1.3457631256
O	O	3990.6315602123	0.2525252525	13.5550323855
O	N	7516.9109701048	0.2583979328	26.0211577313
O	F	2813.7289556110	0.2463054187	6.0891442139
O	Cl	5167.9849688511	0.2677376171	18.0509378210
O	S	2533.1257978672	0.2754820937	6.2376835367
O	H_p	68.2892707612	0.2320185615	0.0000000000
N	N	9278.0915892753	0.2645502646	24.9905110305
N	F	1937.2826523456	0.2518891688	8.2392849516
N	Cl	5664.3158916673	0.2743484225	27.5843767245
N	S	2126.7085249889	0.2824858757	17.8364647208
N	H_p	97.4630200546	0.2369668246	1.8173256556
F	F	3783.4652762315	0.2403846154	7.2529145642
F	Cl	4892.1137396321	0.2607561930	14.0081958833
F	S	2262.5938345565	0.2680965147	4.9454029560
F	H_p	550.1822231209	0.2267573696	0.0000000000
Cl	Cl	9689.5683241540	0.2849002849	78.6520283577
Cl	S	4586.9014824666	0.2936857562	72.3927442963
Cl	H_p	6290.0972739074	0.2447980416	17.6824830942
S	S	5333.0947872917	0.3030303030	139.6047198161
S	H_p	1151.2589871492	0.2512562814	7.5677801111
H_p	H_p	637.1746107280	0.2145922747	0.0000000000

Table J.25: Optimized FF for multipole electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 5.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3151.5126566640	0.2777777778	24.2944107760
C	H	369.6708130485	0.2724795640	4.2184001910
C	O	2567.9347256494	0.2645502646	15.8098912661
C	N	2933.3286024096	0.2710027100	20.6000112475
C	F	2179.2081713585	0.2577319588	8.6497913301
C	Cl	6249.6300460603	0.2812939522	47.7178677785
C	S	4194.1993785140	0.2898550725	51.5825951022
C	H_N	799.4618724753	0.2421307506	2.9175273836
C	H_O	0.0000000000	0.2421307506	2.8014324859
H	H	81.5919869331	0.2673796791	1.5889857542
H	O	315.7111324881	0.2597402597	0.0000000000
H	N	402.0235058541	0.2659574468	2.2478647879
H	F	378.5095672990	0.2531645570	1.4313313066
H	Cl	600.9281859001	0.2758620690	6.2540001112
H	S	581.6577947580	0.2840909091	7.1236140032
H	H_N	19.4000719379	0.2380952381	0.8633997997
H	H_O	528.1353521263	0.2380952381	3.6335896798
O	O	6281.9540612295	0.2525252525	18.0093103500
O	N	8180.0011095063	0.2583979328	25.2722439210
O	F	3877.0420924182	0.2463054187	7.9051175400
O	Cl	4842.8233357879	0.2677376171	19.7200710749
O	S	2307.1605842364	0.2754820937	4.8910505391
O	H_N	64.8547053198	0.2320185615	0.0000000000
O	H_O	38.5198094617	0.2320185615	0.5641982234
N	N	2734.8491355708	0.2645502646	10.8569117202
N	F	1980.1590804884	0.2518891688	8.9362910926
N	Cl	5392.1698256734	0.2743484225	26.9963821039
N	S	2227.5993096406	0.2824858757	20.9102387152
N	H_N	344.0120946588	0.2369668246	1.5374801969
N	H_O	134.0132812111	0.2369668246	3.8604210176
F	F	3218.8236080549	0.2403846154	6.7566749524
F	Cl	4139.0484917229	0.2607561930	11.5556684289
F	S	2628.0352899380	0.2680965147	5.8568656318
F	H_N	587.5121664874	0.2267573696	0.0000000000
F	H_O	0.0000000000	0.2267573696	0.9551324322
Cl	Cl	9802.1590695222	0.2849002849	78.3475014811
Cl	S	3579.3530634217	0.2936857562	58.7915234814
Cl	H_N	7782.5762071687	0.2447980416	27.1565630021
Cl	H_O	11034.6693566387	0.2447980416	21.7531557078
S	S	5191.8839422499	0.3030303030	139.2678084593

Continued on next page

Table J.25 – continued from previous page

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
S	H_N	1224.4206949138	0.2512562814	7.6846064622
S	H_O	883.3180236755	0.2512562814	8.0511087427
H_N	H_N	217.8022933821	0.2145922747	0.0000000000
H_N	H_O	739.0821736266	0.2145922747	0.0532839270
H_O	H_O	799.0229926451	0.2145922747	0.0000000000

Table J.26: Optimized FF for multipole electrostatics with PCM (ε_{PCM} = 11), by Scheme 6.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	111825.0919063579	0.2121692750	20.5217459717
C	H	893.6079164364	0.2520016550	4.0329336480
C	O	2053.9446973436	0.2704699453	17.6580764230
C	N	59294.0154591404	0.2173565468	21.8960822871
C	F	164876.8391705310	0.1894840229	8.5212472228
C	Cl	17404.7691007226	0.2601915827	45.2676935240
C	S	340052.1565796466	0.2074721451	41.8383572059
C	H_N	1183.7086706785	0.2298207108	2.4876921496
C	H_O	57.5596323919	0.1223500856	1.6503558246
H	H	35.2972170459	0.3049290911	2.4395342160
H	O	118.1045024961	0.3046374523	2.1189858873
H	N	231.0138766765	0.2968237099	3.9002027818
H	F	123.4496782646	0.2827555058	1.0398504123
H	Cl	415.8023548522	0.2923094163	9.4084917807
H	S	1768.4066659110	0.2586137904	8.0014893039
H	H_N	201217.2138414681	0.1217816588	0.7821000524
H	H_O	1501.4567774613	0.1730028835	1.1753148686
O	O	69957.0310531577	0.2057245779	11.8459289577
O	N	7307.0394505736	0.2573180584	22.3420652346
O	F	5211.9018939419	0.2347876869	6.1571825067
O	Cl	15969.9555124006	0.2410766623	17.6400101336
O	S	15059.3104182060	0.2373576215	4.7676838997
O	H_N	20.2399318141	0.3082147133	0.3630469801
O	H_O	100.7790848304	0.2064730617	1.8962199983
N	N	6838.5262511685	0.2512949334	12.4897333345
N	F	446.2226859804	0.3204468915	19.7388816224
N	Cl	3372.9378778430	0.2848743691	28.3401913702
N	S	85087.9702641500	0.2110370333	20.6641926998
N	H_N	321.3173096697	0.2407092411	1.7494854983
N	H_O	113.4641273321	0.2520541958	4.3006796944
F	F	1674.9981434415	0.2651173079	9.8735641914
F	Cl	49384.6493937995	0.2191689693	12.0569375435

Continued on next page

Table J.26 – continued from previous page

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
F	S	832360.5604101171	0.1813186215	8.0789860179
F	H_N	10833.8035254374	0.1708278599	0.6630057187
F	H_O	1827680.8135076088	0.1527079192	3.6103163698
Cl	Cl	5210.7857426352	0.2972998114	75.2356028301
Cl	S	6488.5999468747	0.2772452841	51.2224187543
Cl	H_N	447.7261513274	0.3250574893	36.6089199620
Cl	H_O	2230.3050353643	0.2731835884	14.7352626192
S	S	390051.1236657948	0.2185464738	123.2209264727
S	H_N	65678.0403897935	0.1724182417	2.1742661830
S	H_O	23.8384321739	0.4635366384	57.0316667793
H_N	H_N	11.5339154262	0.3111479805	0.9879526180
H_N	H_O	26462.8689762340	0.1514800086	0.1776449580
H_O	H_O	13.5437860881	0.3896765071	0.2989549194

Table J.27: Optimized FF for multipole electrostatics with PCM ($\varepsilon_{\mathrm{PCM}} = 11$), by Scheme 7.

t	t'	$\varepsilon^{t,t'}$ (eV)	$\sigma^{t,t'}$ (Å)	$\lambda_{\mathrm{r}}^{t,t'}$ (-)	$\lambda_{\mathrm{a}}^{t,t'}$ (-)
C	C	0.0049156859	3.3946319672	16.5497441167	6.0000000000
C	H	0.0013534472	3.0283088304	9.8270073763	6.0000000000
C	N	0.0053929306	3.2801754712	18.6432003398	6.0000000000
C	O	0.0044827027	3.1541603692	12.2533629592	6.0000000000
C	F	0.0036570992	3.1405412290	24.7912055706	6.0000000000
C	S	0.0099061742	3.4797120810	18.1490155026	6.0000000000
C	Cl	0.0085450153	3.4665674503	15.5896991737	6.0000000000
C	H_N	0.0010209889	2.8929834069	11.5499702048	6.0000000000
C	H_O	0.0115624352	2.2692871344	45.4337714345	6.0000000000
H	H	0.0004663025	2.9677382713	6.4201108492	6.0000000000
H	N	0.0004326906	3.6265806764	6.6673284589	6.0000000000
H	O	0.0002181784	3.9181968508	6.1704130374	6.0000000000
H	F	0.0002839926	3.6646991462	6.2410775253	6.0000000000
H	S	0.0010050418	3.2620285724	12.4590224948	6.0000000000
H	Cl	0.0010797768	3.5449777702	6.3503109394	6.0000000000
H	H_N	0.0068769406	2.0585366457	33.5132501852	6.0000000000
H	H_O	0.0026755588	2.4979953563	11.2921913112	6.0000000000
N	N	0.0020742646	3.3968372740	11.1683309927	6.0000000000
N	O	0.0043712309	3.3643907414	11.5714350817	6.0000000000
N	F	0.0017756535	3.4920867662	7.2650535735	6.0000000000
N	S	0.0061067701	3.2751562996	17.7259089587	6.0000000000
N	Cl	0.0041795960	3.4839764445	12.9859568656	6.0000000000
N	H_N	0.0004955486	3.0887160572	6.7981573630	6.0000000000
N	H_O	0.0060712314	1.8492558030	6.2103865145	6.0000000000

Continued on next page

Table J.27 – continued from previous page

t	t'	$\varepsilon^{t,t'}$ (eV)	$\sigma^{t,t'}$ (Å)	$\lambda_{\mathrm{r}}^{t,t'}$ (-)	$\lambda_{\mathrm{a}}^{t,t'}$ (-)
O	O	0.0026583047	3.2954501992	13.0516399855	6.0000000000
O	F	0.0012869411	3.2272600284	15.9108140313	6.0000000000
O	S	0.0007607526	3.7507331673	12.8604827730	6.0000000000
O	Cl	0.0032349825	3.4670828087	11.8578759531	6.0000000000
O	$\mathrm{H_N}$	0.0000215699	4.0535261469	6.3297034514	6.0000000000
O	$\mathrm{H_O}$	0.0000173713	1.4495303708	13.0517614752	6.0000000000
F	F	0.0022108199	3.1842484266	13.1363448877	6.0000000000
F	S	0.0017498915	3.4595518100	17.5803570157	6.0000000000
F	Cl	0.0017854059	3.4931255064	14.6682928185	6.0000000000
F	$\mathrm{H_N}$	0.0000651687	2.6094354933	34.6689594116	6.0000000000
F	$\mathrm{H_O}$	0.0016704566	3.0096869106	19.3386838488	6.0000000000
S	S	0.0309364008	3.3080859894	15.4019171804	6.0000000000
S	Cl	0.0093203728	3.3712567473	11.0507504817	6.0000000000
S	$\mathrm{H_N}$	0.0024337423	3.0483171189	8.3001876932	6.0000000000
S	$\mathrm{H_O}$	0.0054783385	2.8257909460	6.1359330645	6.0000000000
Cl	Cl	0.0112885805	3.5586042669	13.4244077378	6.0000000000
Cl	$\mathrm{H_N}$	0.0054079520	3.2779525113	6.2677005269	6.0000000000
Cl	$\mathrm{H_O}$	0.0044700105	3.2522621929	15.7396498384	6.0000000000
$\mathrm{H_N}$	$\mathrm{H_N}$	0.0000538280	3.3342030693	7.5049773569	6.0000000000
$\mathrm{H_N}$	$\mathrm{H_O}$	0.0002150362	2.4502887561	49.1005680689	6.0000000000
$\mathrm{H_O}$	$\mathrm{H_O}$	0.0002112325	3.5077195560	7.1059829267	6.0000000000

Appendix K

Distributed Point-charges for SCERP

K.1 Introduction

As pointed out by Welch & co-workers [183, 189] and Zhang [71], the representation of the crystalline electric field by atomic-charges is a rather crude approximation. The inaccuracy of an atomic-charge representation is widely recognized [166, 167, 221], and is a motivating factor for progressing to a distributed-multipole representation in the multistage CSP workflow [50]. Unfortunately, an analogous development for the SCERP methodology is currently beyond the capabilities of most quantum chemistry packages (e.g., Gaussian), which only permit charge-embedding. In view of this limitation, one avenue to attain a better representation of the molecular charge density is to employ satellite-charges in the SCERP model. Like distributed multipoles, satellite-charges can provide an anisotropic description of the molecular charge density. However, rather than have this anisotropy inherent to the electrostatic features (e.g., dipoles, quadrupoles, octupoles, etc.), satellite-charges achieve anisotropy by using multiple point-charges to describe a single atomic site. For a given atomic site, charges may be placed on the site itself (i.e., the atomic-charge), but additional off-site charges may also be added, which provide greater flexibility to describe the molecular charge density. Electrostatic models based on satellite-charges have previously been explored in CSP [93, 94], but have lost ground to more-modern distributed multipole methods. Nevertheless, to improve the model fidelity of SCERP, satellite-charges may prove useful given current methodological constraints.

K.2 The Distributed Point-charge Model

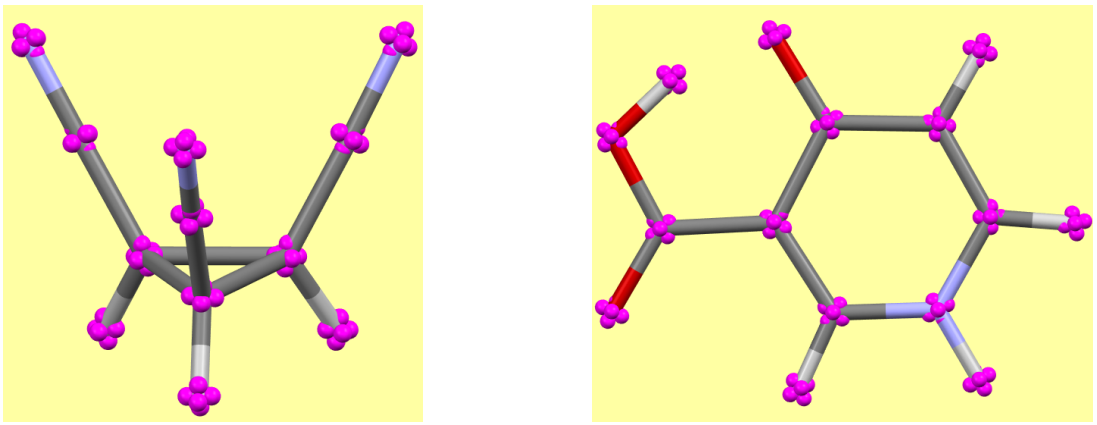
K.2.1 Background

Defining a suitable satellite-charge model involves two activities: (1) deciding the number and placement of the charges (per site), and (2) fitting appropriate values for those charges. In principle, both steps can be solved concomitantly through an appropriately formulated optimization problem. Since it involves integer decision variables (number of satellite charges), (1) tends to introduce greater numerical difficulty in the optimization. For a given placement of satellite-charges, (2) is no different to the ESP-fitting routinely done for atomic-charges.

To construct a satellite-charge model for use in SCERP, Zhang [71] had proposed adopting

the Minimum Distributed Charge Model (MDCM) [184]. As its name suggests, the MDCM method attempts to determine the *fewest* number of point-charges, their locations, and their charge-values, so as to accurately reproduce the molecular ESP for a given conformation. The emphasis on using the minimal number of charges is motivated by the desire to maximize computational efficiency when evaluating the satellite-charge model in molecular simulations (e.g., during molecular dynamics calculations). Unke *et al.* [184] have demonstrated that the MDCM method can capture *ab initio* molecular ESPs to similar quality as GDMA-derived distributed multipole models truncated between $l = 2$ and $l = 4$. On this basis, the MDCM method seems appropriate for my intentions, but from my own testing, I find that it is simply too costly to be used practically within the SCERP framework. A charge-fitting for a relatively small molecule like cis-1,2,3-tricyanocyclopropane (CYCYPR, 12 atoms) was found to be non-trivial. In practice, this cost would be further compounded by the fact that each SCERP process requires multiple rounds of charge-fitting, and for lattice energy minimization several rounds of the SCERP process may be needed.

To achieve a more tenable balance in cost and accuracy, I instead propose using the Distributed Charge Model (DCM) [398], which was developed as the progenitor to MCDM [184]. The DCM method leverages the fact that any multipole expansion truncated at $l = 2$ (or less) can be *exactly* represented by 6 (or fewer) satellite-charges. In the most general case, 6 charges are placed in an octahedral formation around an atomic site (which excludes the atomic-charge position), and analytical relationships can be used to determine their respective charge values that reproduce the $l = 2$ multipole expansion. An example of this is shown in Figure K.1. Because the charge placement is pre-determined and the charge values are analytically derived, the DCM method is extremely fast to evaluate (practically no cost in my experience). Of course, this acceleration does come with its trade-offs. Firstly, the DCM method does not attempt to minimize the number of point-charges used in the model. In my view, this is not of great



(a) (b)

Figure K.1: DCM model for the central molecule in (a) CYCYPR, and (b) KUXPUP01. DCM-determined satellite-charges are shown in magenta, while all other elements adhere to their default colors in Mercury.

importance to the SCERP workflow, since the cost of Gaussian QM calculations do not scale aggressively with the number of charges used in the embedding [189]. Secondly, unlike MDCM, DCM does not conduct fitting to an *ab initio* ESP, but instead determines its model charges analytically from the multipole expansion. As such, the accuracy of the DCM-charge model is tied to the quality of the input multipole model, which (unfortunately) is limited to $l = 2$.

Based on its cost/accuracy profile, I believe the DCM method may offer a practicable route to bolster the fidelity of the SCERP model. Even if its accuracy is limited to only an $l = 2$ multipole model, this is still a meaningful improvement from the atomic-charge model ($l = 0$) currently used in SCERP. To that end, below I investigate the viability of integrating the DCM method within the SCERP process.

K.2.2 Benchmarking the DCM Method

K.2.2.1 Methodology

To begin, I benchmark the accuracy of the DCM-derived satellite-charge model through its ability to reproduce the *ab initio* molecular ESP. Crystal structures from the CE–755 dataset are utilized as a test set here. Single-point, isolated-molecule QM calculations are conducted at the PBE0/6-311G(d,p) level of theory, using Gaussian (G09-d01), to obtain converged SCF charge densities at the crystalline conformation of each molecule.

From here, six different electrostatic models are obtained. (1) Firstly, a HLYGAt-fitted atomic-charge model is determined from Gaussian. This represents the quality of the electrostatic model currently used in the SCERP field. Following this, two distributed multipole models are derived using GDMA: (2) a distributed multipole model truncated at rank 2 (charge, dipoles, and quadrupoles), and (3) a distributed multipole model truncated at rank 4 (charge, dipoles, quadrupoles, octupoles, and hexadecapoles). The former offers the most straightforward route to use DCM, and (4) from the rank 2 multipole model a DCM-charge model is derived. The rank 4 multipole model represents the upper bound of electrostatic model fidelity routinely used in HAIEFF-model CSP. Ideally, DCM could be applied to this multipole model to obtain a corresponding satellite-charge model, but the $l = 2$ limitation of DCM prohibits this. To bypass this constraint, (5) I use the MULFIT program [177] to fit a rank 2 distributed multipole model using the rank 4 distributed multipole model as an input. Given an input rank-N multipole model, MULFIT generates a reference molecular ESP and then fits a corresponding rank-M multipole model (where $M < N$) to reproduce this reference ESP. Finally, (6) a DCM-charge model is computed from the MULFIT-derived rank 2 multipole model.

All DCM computations are performed using the DCM program script available from Ref [398]. The MULFIT program [177] is distributed as part of GDMAv2.3 [182]. For consistency with our CSP workflows, I use GDMAv2.2 for all distributed multipole calculations. In GDMAv2, the default distribution scheme relies on a grid-based quadrature [181, 182]. Here, I have chosen to disable this (i.e., $r_{\mathrm{H}}^{\mathrm{GDMA}}$ is irrelevant) and instead use the original nearest-neighbor distribution scheme ('SWITCH=0' keyword in GDMA). From testing, this signifi-

cantly accelerate the derivation of multipoles in GDMA, without leading to loss of accuracy in the molecular ESPs. Stone [181] has suggested that while the quadrature distribution scheme is preferable for deriving multipole moments that align better with chemical intuition (particularly with diffuse basis functions), the quality of the overall molecular ESP should be equally satisfactory with either distribution scheme.

Finally, to assess the quality of the molecular ESP predicted by each electrostatic model, I employ utility scripts provided in the MDCM distribution [184]. As inputs, these scripts require the *ab initio* ESP for a molecular system stored in cube file format, which can be generated using Gaussian's utility programs with the SCF-calculation's outputs. Given a set of point-charges/multipoles (e.g., from one of the six electrostatic models), the MDCM utility scripts can generate the model-predicted ESP in the equivalent cube format, and determine a root-mean-square-error (RMSE) against the *ab initio* reference cube file. The RMSE is given by

$$\begin{aligned} \mathrm{RMSE}(\boldsymbol{Q}_{\mathrm{EM}}) =& \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left[\mathrm{ESP}_j^{\mathrm{AI}} - \mathrm{ESP}_j^{\mathrm{EM}}(\boldsymbol{Q}_{\mathrm{EM}})\right]^2}, \\ & \mathrm{EM} = \mathrm{HLYGAt}, \mathrm{GDMA}_2, \mathrm{DCM}_{\mathrm{gdma2}}, \mathrm{GDMA}_4, \mathrm{MULFIT}_{\mathrm{gdma4}}, \mathrm{DCM}_{\mathrm{mulfit2}}, \end{aligned} \tag{K.2.1}$$

where $\boldsymbol{Q}_{\mathrm{EM}}$ is the input charge/multipole model for the one of the six electrostatic model (EM). $\mathrm{ESP}_j^{\mathrm{AI}}$ and $\mathrm{ESP}_j^{\mathrm{EM}}$ are, respectively, the *ab initio* and model-predicted ESP at point j of the cube file. J is the total number of reference points in the *ab initio* ESP. For each evaluation, four RMSE scores are reported: (1) a close-range RMSE for ESP-points bound by $1.2\sigma_{\mathrm{vdW}} < R < 1.66\sigma_{\mathrm{vdW}}$, (2) a medium-range RMSE for ESP-points bound by $1.66\sigma_{\mathrm{vdW}} < R < 2.20\sigma_{\mathrm{vdW}}$, (3) a long-range RMSE for ESP-points bound by $R > 2.20\sigma_{\mathrm{vdW}}$, and (4) a total RMSE over all reference points. R is the distance from the ESP reference point to an atomic centre, and σ_{vdW} is the vdW radii for the atomic centre. To quantify the improvement each electrostatic model yields over the HLYGAt reference, I also define a change in RMSE (ΔRMSE) as,

$$\begin{aligned} \Delta\mathrm{RMSE}^{\mathrm{EM}} =& \mathrm{RMSE}^{\mathrm{EM}} - \mathrm{RMSE}^{\mathrm{HLYGAt}}, \\ & \mathrm{EM} = \mathrm{GDMA}_2, \mathrm{DCM}_{\mathrm{gdma2}}, \mathrm{GDMA}_4, \mathrm{MULFIT}_{\mathrm{gdma4}}, \mathrm{DCM}_{\mathrm{mulfit2}}, \end{aligned} \tag{K.2.2}$$

such that $\Delta\mathrm{RMSE}^{\mathrm{EM}} < 0$ indicates that electrostatic model EM has a lower error in its predicted-ESP than the HLYGAt-charge model.

K.2.3 Benchmarking Results

The results of this benchmarking exercise are summarized in Figures K.2 and K.3. For the 755 crystal structures, only 753 are successfully evaluated with all six electrostatic models. When adamantane (ADAMAN08) and furan (FURANE10) are evaluated with DCM, the resulting ESP are found to be highly erroneous. Given that all other structures exhibit sensible results

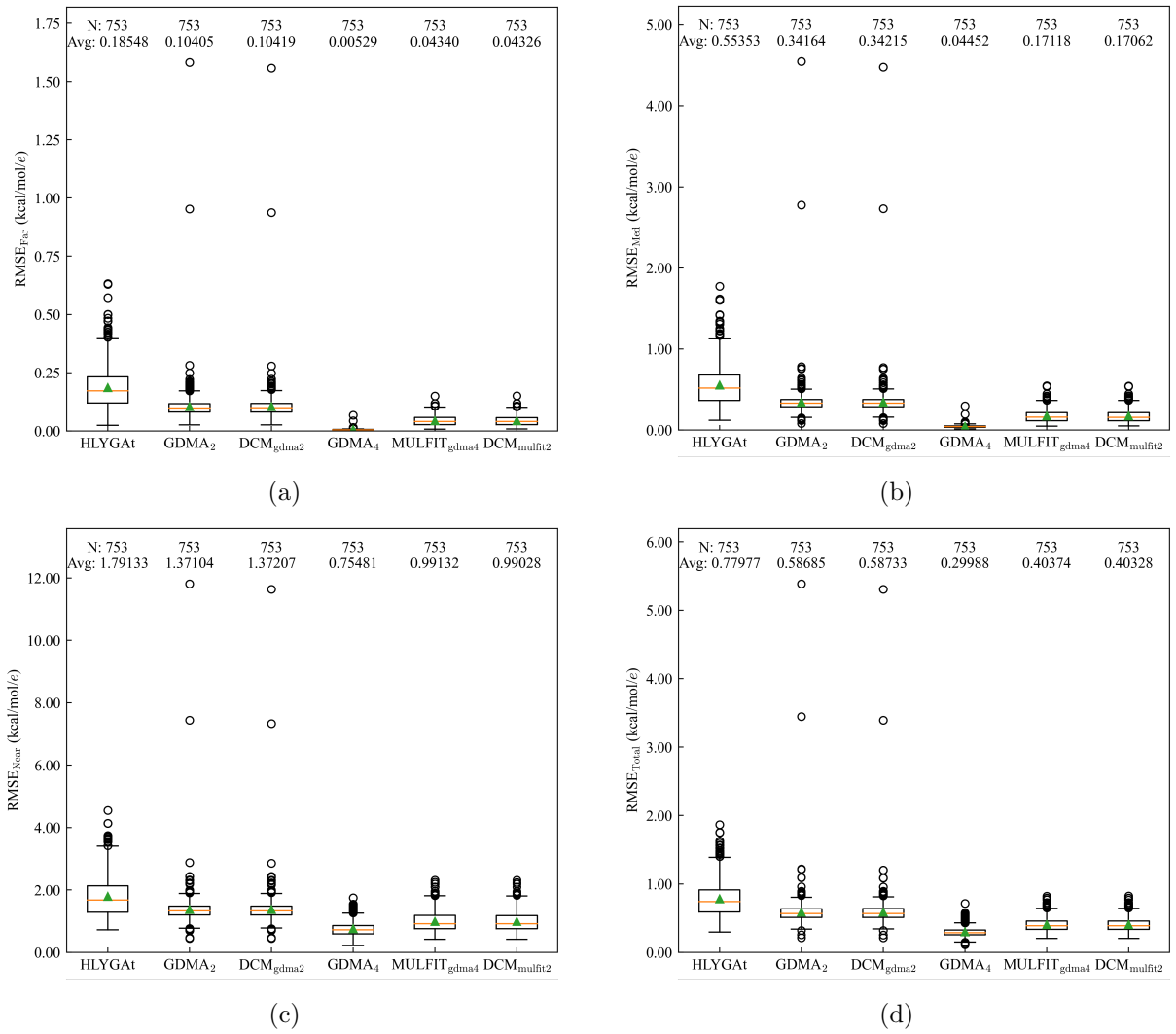


Figure K.2: Distribution of RMSEs in the model-predicted ESPs for structures in the CE–755 dataset: (a) the long-range RMSE, (b) the mid-range RMSE, (c) the short-range RMSE, and (d) the total RMSE. ADAMAN08 and FURANE10 have been omitted due to highly erroneous results with DCM. Over the remaining 753 structures, the bounds of each box give the 1st and 3rd quartile, while whiskers represent the ±1.5x interquartile range. The median is given by the orange ━ in each box, while means are given by the green ▲ (and at the top of the plot). Outliers are represented with ○.

with DCM, I do not believe this reflects a problem with the DCM model. These two outliers have been omitted from the benchmarking results.

As one might expect, the ESPs predicted using the atomic-charge model (HLYGAt) consistently have the worst agreement (highest average RMSE) with the *ab initio* reference ESP. This is observed for ESP points at all ranges. Progressing to the rank 2 multipole model (GDMA$_2$), there is a marked decrease in average RMSE scores. However, from Figure K.3, it can be seen that the improvement in RMSE scores achieved by the GDMA$_2$ model is not ubiquitous to all structures. In fact, there are still a fair number of structures that have $\Delta\text{RMSE} > 0$, as indicated by the upper whisker exceeding $y = 0$ for the GDMA$_2$ model in Figure K.3. The accuracy of

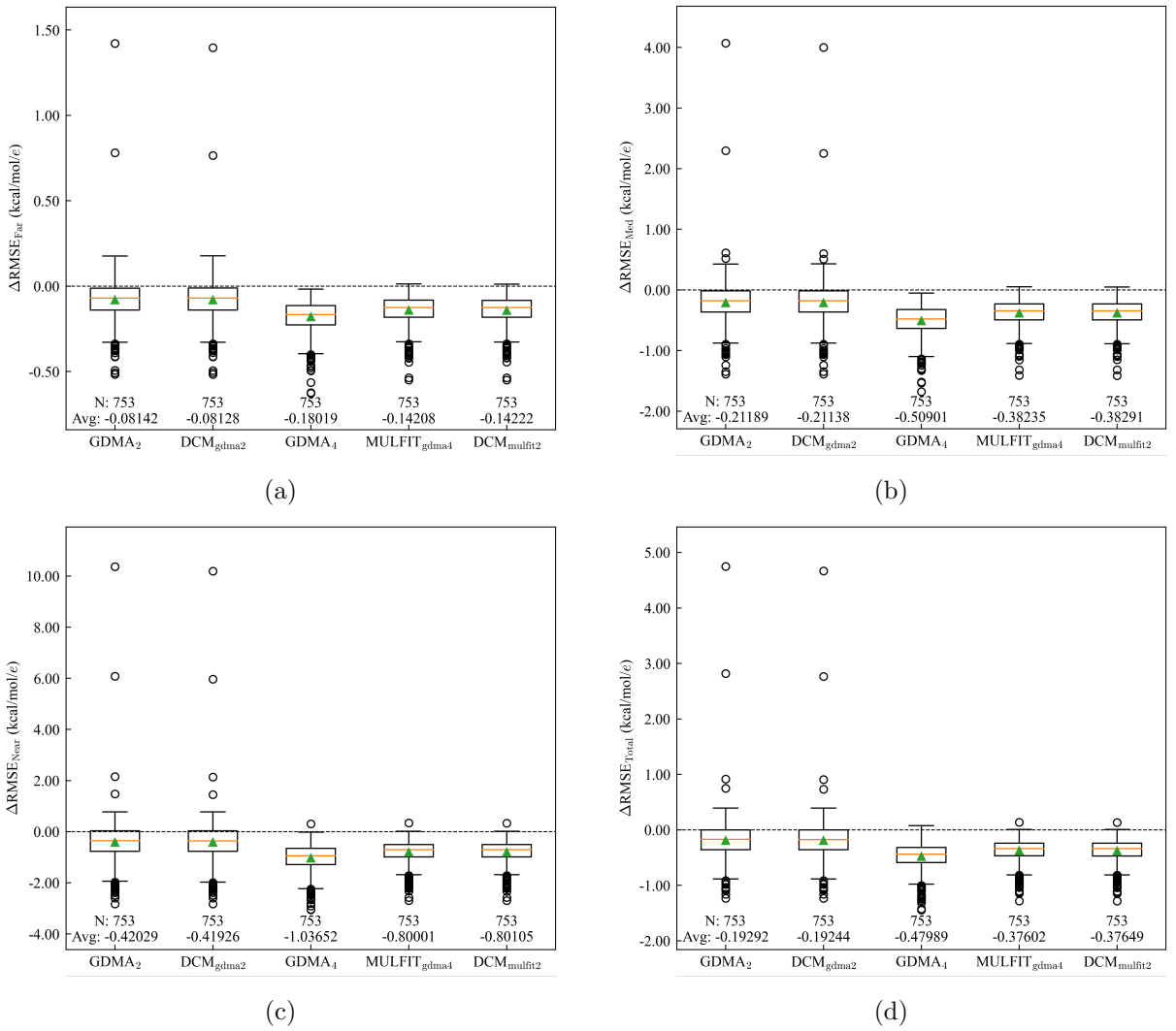


Figure K.3: Distribution of ΔRMSEs in the model-predicted ESPs for structures in the CE–755 dataset: (a) the long-range RMSE, (b) the mid-range RMSE, (c) the short-range RMSE, and (d) the total RMSE. ADAMAN08 and FURANE10 have been omitted due to highly erroneous results with DCM. Over the remaining 753 structures, the bounds of each box give the 1st and 3rd quartile, while whiskers represent the ±1.5x interquartile range. The median is given by the orange ━ in each box, while means are given by the green ▲ (and at the top of the plot). Outliers are represented with ○.

the electrostatic model continue to improve when the multipole model is truncated at $l = 4$ ($GDMA_4$), instead of $l = 2$. For the $GDMA_4$ model, average RMSE scores are lower than the preceding two models, and the distribution of ΔRMSE lie almost exclusively below 0. The latter suggests that for almost all structures evaluated, using the $GDMA_4$ model leads to an improved description of the *ab initio* ESP over the HLYGAt model.

By the DCM methodology, the behavior of any multipole expansion truncated at $l = 2$ should be exactly reproduced by the DCM-derived satellite-charges. This is demonstrated by the DCM-charge model derived from the $GDMA_2$ multipole expansion (DCM_{gdma2} in Figures K.2 and K.3). Comparing the distribution of RMSE/ΔRMSE between $GDMA_2$ and DCM_{gdma2},

the latter is almost an exact duplicate of the former, verifying that the DCM approach [398] correctly transmutes the rank 2 multipole model to an equivalent rank 0 satellite-charge model. Given that the $GDMA_2$ model already offers some improvement over the atomic-charge model, it follows that the DCM_{gdma2} model would already be a step forward for SCERP. However, as previously noted, a non-negligible proportion of structures do not actually exhibit improved ESP-predictions using $GDMA_2$ versus HLYGAt. The ability of HLYGAt-charges to outperform $GDMA_2$-multipoles reflects that the former are fit directly to the *ab initio* ESP. In contrast, GDMA-multipoles are determined based on AIM-theory, and thus their accuracy is strongly tied to the truncation rank of the expansion. To summarize:

1. I have validated that the DCM-charge model reproduces the input $l = 2$ multipole model.
2. SCERP could use charge-embedding based on the DCM_{gdma2} charge model, but due to the underlying $GDMA_2$ model having only modest accuracy, DCM_{gdma2} may not consistently lead to improved accuracy over HLYGAt, for a diverse range of structures.

For greater certainty of improved performance with a DCM-charge model, the corollary is that a more accurate $l = 2$ multipole model must be provided to the DCM algorithm. Driven by this goal, I subject the $GDMA_4$ model to MULFIT [177], so as to obtain an empirically fitted $l = 2$ multipole model that (hopefully) has closer fidelity to the $GDMA_4$ model. From the figures, it can be seen that the resulting $MULFIT_{gdma4}$ model does successfully achieve this: the distribution of (and average) RMSE scores for the $MULFIT_{gdma4}$ model more closely resembles the $GDMA_4$ model, despite being truncated at the same rank as $GDMA_2$ (Figure K.2). Also like the $GDMA_4$ model, when using the $MULFIT_{gdma4}$ model, almost all structures exhibit more-accurately predicted-ESPs than the HLYGAt model (Figure K.3). Leveraging the exact nature of the DCM transformation, the $DCM_{mulfit2}$ model retains the superior performance of $MULFIT_{gdma4}$, outperforming both $GDMA_2$ and DCM_{gdma2}.

To conclude, the charge-embedding constraint of Gaussian has motivated exploration of the DCM method to obtain an accurate satellite-charge model that could be used with SCERP. Although this can be straightforwardly done by $GDMA_2 \rightarrow DCM_{gdma2}$, to extract greater model improvement, I instead propose using $GDMA_4 \rightarrow MULFIT_{gdma4} \rightarrow DCM_{mulfit2}$. This latter workflow provides a satellite-charge model that is closer in quality to the superior $GDMA_4$ model and outperforms the earlier two models. It should also be noted that the MULFIT algorithm is extremely fast, hence there is no noticeable cost difference between the two workflows.

K.3 DCM in SCERP

K.3.1 Algorithmic Implementation

In light of the benchmarking results, there is justification to develop a satellite-charge SCERP implementation using DCM. I have added this as an option for SCERP within CSO–RMv1.5.2. Briefly, with reference to Figure 7.1, wherever $\boldsymbol{Q}_{I,\text{field}}^{\text{total,pc}}$ needs to be updated, a set of $DCM_{mulfit2}$

charges are derived through a series of calls to the relevant external programs. First, GDMAv2.2 [181, 182] is called to generate a set of distributed rank 4 multipoles from an SCF charge density (e.g., from the SCERP QM calculation). Using this $l = 4$ multipole expansion as an input, MULFIT [177] is called to fit a corresponding $l = 2$ multipole expansion. Finally, the $\text{MULFIT}_{\text{gdma4}}$ multipoles are passed to the DCM program [398], which generates the required set of $\text{DCM}_{\text{mulfit2}}$ satellite-charges. The coordinates of these satellite-charges are stored in the same body-fixed coordinate system as the central molecule. In this way, when the crystallographic positions of periodic images in the SCERP-field are determined through any relevant symmetry/translation operations, the corresponding positions of the satellite-charges are defined through identical operations. An example of this is shown in Figure K.4 for CYCYPR.

In principle, the rest of the SCERP process shown in Figure 7.1 should remain unchanged.

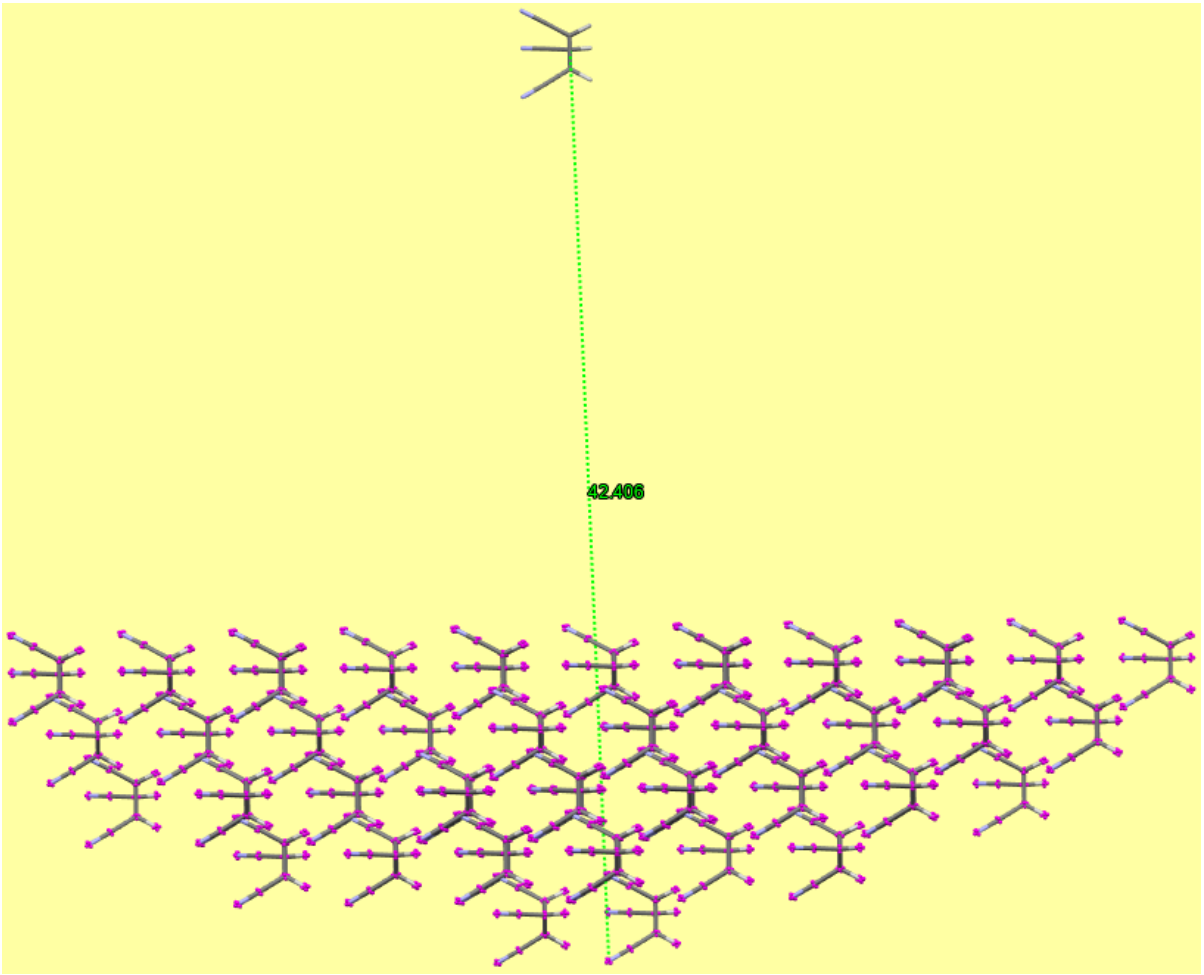


Figure K.4: The 40 Å SCERP embedding-sphere around the central CYCYPR molecule (top) using: atomic-charges (represented with molecules) and DCM-charges (represented with magenta spheres). The charges associated with the first 100 molecules of the embedding-sphere are shown. The DCM-charges overlay exactly with the atomic-positions of their corresponding molecular periodic images, indicating that the transformation of DCM-charge coordinates to their SCERP-field positions has been correctly implemented.

Unfortunately, the DCM-SCERP implementation is afflicted by a few challenges, stemming from an observation that the DCM-charges are not always determined in the same order (for a given site). This in turn stymies our ability to compare DCM-charge values, in a consistent manner, between SCERP iterations, leading to a few critical consequences. Firstly, without any guarantee of the charge-ordering, the convergence of the SCERP process cannot be reliably inferred from the DCM-charges. As such, when DCM-SCERP is used, HLYGAt atomic-charges continue to be derived and used as the arbiter of the process's convergence. Secondly, the DCM-charges cannot be subject to the deceleration protocol, since this risks 'blending' charge values between unrelated satellite-charges.

K.3.2 DCM-SCERP Performance

Despite the listed challenges, I continue to assess the DCM-SCERP model performance. The single-point energy evaluations done in Section 7.3.2 are repeated here, but using DCM-SCERP in place of atomic-charge SCERP. All other SCERP parameters are taken from Table 7.4.

To begin, I consider the convergence and speed of the SCERP process. In Figure K.5, I present the value of $\Delta Q_{Ii,t}^{\text{total,pc}}$ over the SCERP process. For a fair comparison with the atomic-charge SCERP results (Figure 7.6), I have kept $t_{\text{slow}} = 6$ during these computations. Readers are reminded that this affects the value of only the atomic-charges used in the assessment of SCERP process convergence, but the actual DCM-charges used in the SCERP-field are *not decelerated* in any way. From the 790 molecules present across the CE–755 dataset, only 752 successfully converged the SCERP process. The remaining molecules either failed to converge within the allocated number of iterations ($t_{\text{max}} = 15$), or were disqualified due to divergent behavior. The latter is almost certainly the result of the DCM-charges not being subject to the deceleration protocol. Of the structures that do converge, (on average) ~1 fewer iteration is expended compared to atomic-charge SCERP, although the DCM-SCERP process has more

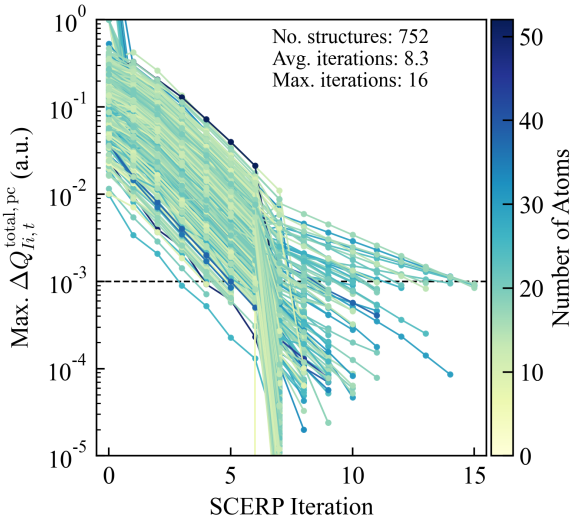


Figure K.5: The maximum values of $\Delta Q_{Ii,t}^{\text{total,pc}}$ during the single-point DCM-SCERP evaluation of structures from the CE–755 dataset. The dashed line indicates $\varepsilon_{\text{scrp}} = 1 \times 10^{-3}$ a.u.

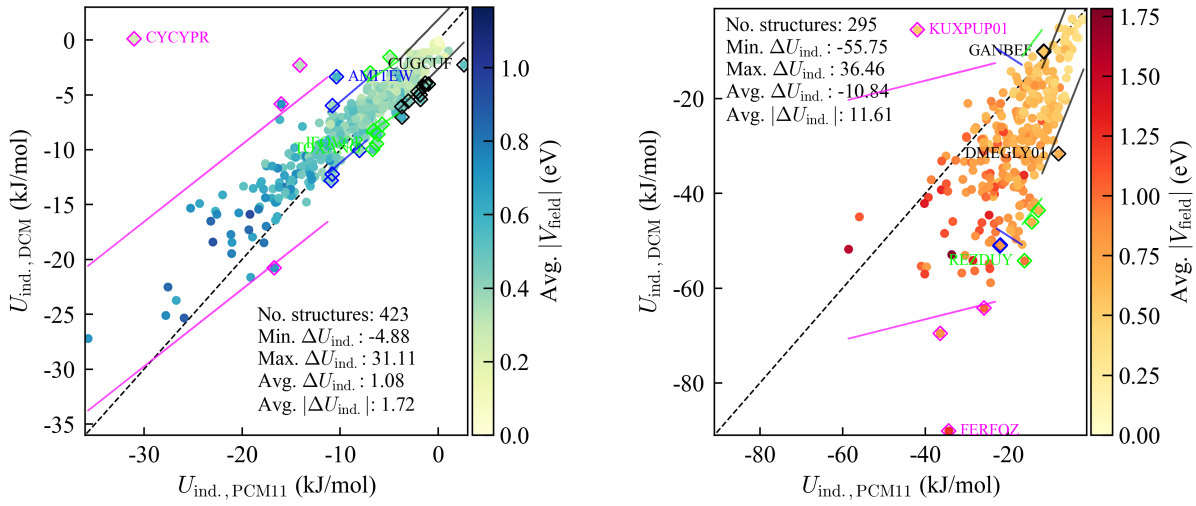
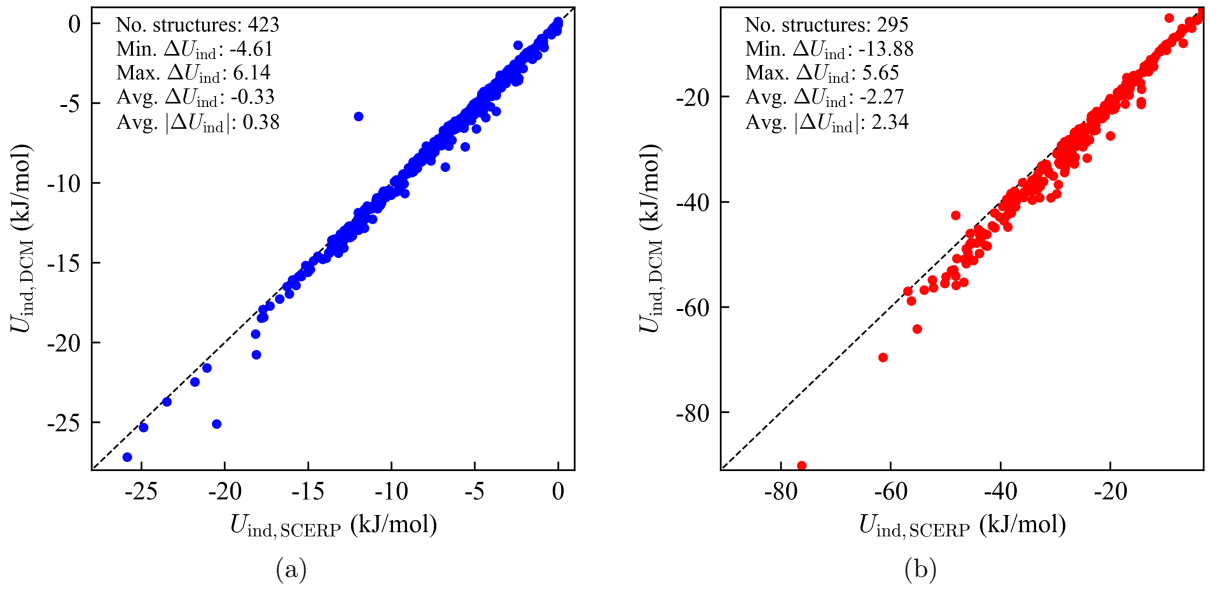


Figure K.6: Predicted U_{ind} for the PCM- and SCERP-embedded (with DCM-charges) models in (a) NHB, and (b) HB structures. The colors of each point reflect the Avg. $|V_{\text{field}}|$ values, defined by Equations 7.3.7 and 7.3.8. ΔU_{ind} is given by Equation 7.3.6. The solid lines represent a $\pm 2\sigma_{\text{std.dev}}$ bound from the best fit line in each quantile. Points outside these bounds are flagged as outliers ($\diamond$), with the top-5 outliers labeled by their CSD Refcodes.

structures that approach the iteration limit. In practice, $t_{\text{slow}} = 0$ would probably be a more sensible choice to use with DCM-SCERP (since the deceleration is not actually doing anything). In which case, I would expect a lower average number of SCERP iterations to reach convergence, and fewer structures would fail due to reaching $t_{\text{max}} = 15$. Unfortunately, any cost savings that might be gained from this are offset by the increased CPU costs for each QM calculation. This is likely due to the greater number of embedded-charges. Using DCM, the number of SCERP-field charges have increased by 6x, while each QM calculation is ~2–3x more expensive than those involving only atomic-charges, a considerable increase in cost.

The deterioration in SCERP process reliability and cost could be justifiable if the DCM-SCERP model contributed to a significant change in predicted U_{ind}. From Figures K.6 and K.7, the application of DCM-SCERP generally leads to more stabilization from U_{ind}. The reason for this can be intuited from the FF training results. From Section 6.2.2, it is known that using a multipole electrostatic model tends to promote greater stabilization from U_{elec}. It follows that the electric field generated by a multipole model should be *more polarizing* than one based solely on atomic-charges. Here, the satellite-charge model is designed to more closely resemble the multipole model. Therefore, it is reasonable that the DCM-SCERP model should induce greater polarization than the atomic-charge SCERP model. However, the extent to which U_{ind} is enhanced by DCM-SCERP is quite marginal, usually not exceeding a few kJ/mol (Figure K.7). This is true even for the HB structures, which tend to be more sensitive towards changes in the electrostatic/induction model. Additionally, from Figure K.6, the use of DCM-SCERP has not led to any meaningful improvement in the outliers, CYCYPR and KUXPUP01.



(a) (b)

Figure K.7: Parity in U_{ind} for the SCERP-embedded models using atomic-charges (SCERP) and DCM-charges (DCM), in (a) NHB, and (b) HB structures. ΔU_{ind} is given by $\Delta U_{ind} = U_{ind}^{DCM} - U_{ind}^{SCERP}$.

K.4 Conclusions

Zhang [71] and Welch [183, 189] have raised valid concerns about the physical fidelity of a SCERP model based solely on atomic-charges. However, due to the inherent limitations of the Gaussian package, an improved SCERP-field description cannot be straightforwardly achieved by multipole-embedding. Forced to stay within the paradigm of charge-embedding, I have considered the possibility of using a satellite-charge model generated via the DCM method [398]. With some simple 'workarounds', it is possible to derive a DCM satellite-charge model that predicts molecular ESPs to an accuracy between a rank 2 and a rank 4 GDMA-multipole model. Unfortunately, integrating this satellite-charge model into the SCERP process is shown to cause difficulties in converging the field, whilst also inflating computation costs. In return for this deterioration in performance, it does not seem that much is gained in the accuracy of the model predictions. Ultimately, in its current state, it is difficult to justify the routine use of DCM-SCERP in the SCERP process, thus I will continue to use atomic-charge SCERP in the rest of this work.

Appendix L

Atom-pair Specific Tang–Toennies Damping Parameters

Atom-pair specific Tang–Toennies damping parameters are computed by solving Equation 8.2.4 for $\beta_{\text{atm}}^{tt'}$, using the Bondi radii [3] tabulated in Table 8.3. The damping parameters for $\%f_6$ = 70–90% are tabulated below.

Table L.1: Atom-pair specific Tang–Toennies damping parameters. All damping parameters are reported in units of bohr^{-1}

t	t'	$r_{\text{Bondi}}^{tt'}$ (Å)	$\%f_6$ = 70%	$\%f_6$ = 80%	$\%f_6$ = 90%
C	C	3.400	1.26240669	1.41249630	1.63921557
C	H	2.900	1.48006302	1.65603014	1.92183895
C	O	3.220	1.33297601	1.49145572	1.73084874
C	N	3.250	1.32067162	1.47768844	1.71487167
C	F	3.170	1.35400087	1.51498026	1.75814919
C	Cl	3.450	1.24411094	1.39202534	1.61545882
C	S	3.500	1.22633793	1.37213926	1.59238084
C	H_N	2.700	1.58969732	1.77869904	2.06419739
C	H_o	2.700	1.58969732	1.77869904	2.06419739
H	H	2.400	1.78840948	2.00103642	2.32222206
H	O	2.720	1.57800837	1.76562037	2.04901946
H	N	2.750	1.56079373	1.74635906	2.02666652
H	F	2.670	1.60755909	1.79868443	2.08739062
H	Cl	2.950	1.45497721	1.62796184	1.88926540
H	S	3.000	1.43072759	1.60082914	1.85777765
H	H_N	2.200	1.95099216	2.18294883	2.53333316
H	H_o	2.200	1.95099216	2.18294883	2.53333316
O	O	3.040	1.41190222	1.57976560	1.83333321
O	N	3.070	1.39810513	1.56432815	1.81541790
O	F	2.990	1.43551263	1.60618308	1.86399095
O	Cl	3.270	1.31259412	1.46865059	1.70438316
O	S	3.320	1.29282613	1.44653235	1.67871474
O	H_N	2.520	1.70324713	1.90574897	2.21164006

Continued on next page

Table L.1 – continued from previous page

t	t'	$r^{tt'}_{\text{Bondi}}$ (Å)	$\%f_6$ = 70%	$\%f_6$ = 80%	$\%f_6$ = 90%
O	H_o	2.520	1.70324713	1.90574897	2.21164006
N	N	3.100	1.38457508	1.54918949	1.79784934
N	F	3.020	1.42125257	1.59022762	1.84547448
N	Cl	3.300	1.30066144	1.45529922	1.68888877
N	S	3.350	1.28124858	1.43357833	1.66368148
N	H_N	2.550	1.68320892	1.88332840	2.18562076
N	H_o	2.550	1.68320892	1.88332840	2.18562076
F	F	2.940	1.45992611	1.63349912	1.89569148
F	Cl	3.220	1.33297601	1.49145572	1.73084874
F	S	3.270	1.31259412	1.46865059	1.70438316
F	H_N	2.470	1.73772581	1.94432689	2.25641010
F	H_o	2.470	1.73772581	1.94432689	2.25641010
Cl	Cl	3.500	1.22633793	1.37213926	1.59238084
Cl	S	3.550	1.20906557	1.35281336	1.56995294
Cl	H_N	2.750	1.56079373	1.74635906	2.02666652
Cl	H_o	2.750	1.56079373	1.74635906	2.02666652
S	S	3.600	1.19227299	1.33402428	1.54814804
S	H_N	2.800	1.53292241	1.71517408	1.99047605
S	H_o	2.800	1.53292241	1.71517408	1.99047605
H_N	H_N	2.000	2.14609138	2.40124371	2.78666647
H_N	H_o	2.000	2.14609138	2.40124371	2.78666647
H_o	H_o	2.000	2.14609138	2.40124371	2.78666647
Average			1.52374094	1.70490100	1.97855405

Appendix M

SCERP-Embedded Parameter Estimation

M.1 Training Sets for Parameter Estimation

Table M.1: Training set structures partitioned from the CE–755 to form TP442. This comprises 442 crystal structures: 317 non-hydrogen bonding (NHB) and 125 hydrogen bonding (HB) structures. No multicomponent crystals are included in the dataset, hence any structure with $Z' > 1$ has repeated units of the same molecule. The NHB47, HB53, and TP100 datasets are subsets of this larger set.

Refcode	HB?	Z'	nAtms	Chemical Formula	NHB47	HB53	TP100
ADAMAN08	NHB	1	26	$C_{10}H_{16}$	✓		✓
ANTCEN16	NHB	1	24	$C_{14}H_{10}$	✓		✓
BADAMN10	NHB	1	50	$C_{20}H_{30}$	✓		✓
BAPOCM10	NHB	1	48	$C_{18}H_{30}$	✓		✓
BCPROP02	NHB	1	16	C_6H_{10}	✓		✓
BENZEN06	NHB	1	12	C_6H_6	✓		✓
CRYSEN	NHB	1	30	$C_{18}H_{12}$	✓		✓
CUBANE	NHB	1	16	C_8H_8	✓		✓
HEXANE01	NHB	1	20	C_6H_{14}	✓		✓
MEYCEY	NHB	1	21	$C_{11}H_{10}$	✓		✓
MEYCIC	NHB	1	11	C_5H_6	✓		✓
NAPHTA31	NHB	1	18	$C_{10}H_8$	✓		✓
OCTANE10	NHB	1	26	C_8H_{18}	✓		✓
PENTAN01	NHB	1	17	C_5H_{12}	✓		✓
PERLEN06	NHB	1	32	$C_{20}H_{12}$	✓		✓
PHENAN08	NHB	1	24	$C_{14}H_{10}$	✓		✓
PYRENE03	NHB	1	26	$C_{16}H_{10}$	✓		✓
TRIPHE13	NHB	1	30	$C_{18}H_{12}$	✓		✓
BACJUN	NHB	2	10/10	$C_4H_4N_2$	✓		✓
BAYJUG	NHB	1	30	$C_{10}H_{18}N_2$	✓		✓

Continued on next page

Table M.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula	NHB47	HB53	TP100
BEBXOV	NHB	1	19	$C_9H_6N_4$	✓		✓
BIPYRL04	NHB	1	20	$C_{10}H_8N_2$	✓		✓
BITVUV	NHB	1	12	C_4H_7N	✓		✓
BOXRUE	NHB	1	25	$C_7H_{11}N_7$	✓		✓
COSNAZ	NHB	1	22	$C_6H_{12}N_4$	✓		✓
CTCDEC	NHB	1	25	$C_{11}H_{13}N$	✓		✓
CUVWEV	NHB	1	19	C_9H_9N	✓		✓
CUXXAW	NHB	1	26	$C_{14}H_9N_3$	✓		✓
CYTCOD10	NHB	1	17	C_9H_7N	✓		✓
DAZNAP	NHB	1	16	$C_8H_6N_2$	✓		✓
DAZNAQ	NHB	1	16	$C_8H_6N_2$	✓		✓
DAZNAR	NHB	1	16	$C_8H_6N_2$	✓		✓
DMDZCD	NHB	1	32	$C_{10}H_{20}N_2$	✓		✓
HXMTAM10	NHB	1	22	$C_6H_{12}N_4$	✓		✓
NAPTYR11	NHB	1	16	$C_8H_6N_2$	✓		✓
PHENAZ04	NHB	1	22	$C_{12}H_8N_2$	✓		✓
PRMDIN01	NHB	1	10	$C_4H_4N_2$	✓		✓
PYRAZI01	NHB	1	10	$C_4H_4N_2$	✓		✓
PYRDNA07	NHB	1	11	C_5H_5N	✓		✓
QORPAR	NHB	1	18	$C_8H_5N_5$	✓		✓
TCYCPR01	NHB	1	13	$C_7H_2N_4$	✓		✓
TCYETY	NHB	1	10	C_6N_4	✓		✓
TCYETY01	NHB	1	10	C_6N_4	✓		✓
TELPIJ	NHB	1	11	$C_3H_5N_3$	✓		✓
TETDAM03	NHB	1	20	$C_6H_{12}N_2$	✓		✓
TRIZIN01	NHB	1	9	$C_3H_3N_3$	✓		✓
XAKPAC01	NHB	1	20	$C_{10}H_5N_5$	✓		✓
ABEWAG	NHB	1	9	C_2H_6O			
AMITEW	NHB	1	19	$C_{10}H_6O_3$			
ASUKAB	NHB	1	23	$C_9H_{12}O_2$			
AVIJEW	NHB	1	23	$C_8H_{12}O_3$			
AXOREM	NHB	1	18	$C_9H_6O_3$			
AXOSOW	NHB	1	8	C_3H_4O			
AXUDED	NHB	1	18	$C_8H_8O_2$			
AYOHEB	NHB	1	18	C_9H_8O			
BCYOAN10	NHB	1	23	$C_{10}H_{10}O_3$			
BEYDEO01	NHB	1	18	$C_6H_8O_4$			
BEZWIM	NHB	1	21	$C_8H_{10}O_3$			

Continued on next page

Table M.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula	NHB47	HB53	TP100
BICVIS	NHB	1	18	$C_6H_8O_4$			
BNZQUI03	NHB	1	12	$C_6H_4O_2$			
BONNEA	NHB	1	17	$C_7H_8O_2$			
BURXAN	NHB	1	24	$C_{10}H_{10}O_4$			
BUTFUT	NHB	1	25	$C_9H_{12}O_4$			
CAMPQU03	NHB	1	26	$C_{10}H_{14}O_2$			
CEGPOV	NHB	1	20	$C_9H_8O_3$			
COUMAR01	NHB	1	17	$C_9H_6O_2$			
COUMAR13	NHB	1	17	$C_9H_6O_2$			
COUMAR14	NHB	2	17/17	$C_9H_6O_2$			
CUKCIU03	NHB	1	14	$C_4H_8O_2$			
CYHEXO	NHB	1	16	$C_6H_8O_2$			
DEXBEN01	NHB	1	26	$C_{10}H_{14}O_2$			
FURANE10	NHB	1	9	C_4H_4O			
PTOXEC	NHB	1	20	$C_5H_{10}O_5$			
QOTDEL	NHB	1	23	$C_8H_{10}O_5$			
SUCANH12	NHB	1	11	$C_4H_4O_3$			
TOXOCN	NHB	1	16	$C_4H_8O_4$			
TROXAN	NHB	1	12	$C_3H_6O_3$			
ABORER01	NHB	1	27	$C_{14}H_{10}N_2O$			
ACSKAT01	NHB	1	24	$C_{11}H_{11}NO$			
ADIVUF	NHB	1	30	$C_{14}H_8N_6O_2$			
AKEFIH	NHB	1	23	$C_9H_9N_3O_2$			
AMIXPO	NHB	1	25	$C_{10}H_{10}N_2O_3$			
AMUVEL	NHB	1	22	$C_{10}H_9NO_2$			
APINIY	NHB	1	29	$C_{14}H_{11}NO_3$			
APUHOJ	NHB	1	28	$C_{10}H_{14}N_2O_2$			
ARIDUB	NHB	1	20	$C_8H_8N_2O_2$			
AVABII	NHB	1	25	$C_{10}H_{13}NO$			
BAGRAC	NHB	1	26	$C_{11}H_{10}N_2O_3$			
BAKQUA	NHB	1	26	$C_{10}H_{10}N_2O_4$			
BAPBUQ	NHB	1	27	$C_{14}H_{10}N_2O$			
BAPBUQ03	NHB	1	27	$C_{14}H_{10}N_2O$			
BARKAH	NHB	1	21	$C_9H_9NO_2$			
BAZPUO	NHB	1	26	$C_9H_{12}N_2O_3$			
BEKLIP	NHB	1	25	$C_{12}H_8N_2O_3$			
BERPIX	NHB	1	31	$C_{13}H_{13}N_3O_2$			
ADAREE	NHB	1	18	$C_6H_6F_6$			

Continued on next page

Table M.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula	NHB47	HB53	TP100
AXUBUR01	NHB	1	14	C_8H_5F			
BAZSEA	NHB	1	16	C_8F_8			
CAXNUL01	NHB	1	18	$C_{10}H_4F_4$			
FACFOE	NHB	1	12	$C_6H_4F_2$			
VEHCET	NHB	1	18	$C_6H_8F_4$			
XAPCAS	NHB	1	9	C_3F_6			
XOGJEK	NHB	1	15	$C_7H_6F_2$			
YICFEX	NHB	1	15	C_7H_7F			
YICFEX01	NHB	1	15	C_7H_7F			
AQOZUB01	NHB	1	18	$C_{10}H_5FN_2$			
BOQSIJ	NHB	1	34	$C_{18}H_{14}FN$			
BOXQEK	NHB	1	31	$C_{17}H_{12}FN$			
BUCNEU	NHB	1	25	$C_{13}H_7F_4N$			
CATXEC	NHB	1	23	$C_{13}H_5F_4N$			
CAXPAT	NHB	1	17	$C_9H_4F_3N$			
DAHXAO	NHB	1	28	$C_{12}H_{12}F_3N$			
DATLAN	NHB	1	11	$C_5H_3F_2N$			
DATLER	NHB	1	11	$C_5H_2F_3N$			
GOZXIF	NHB	1	18	$C_8H_6FN_3$			
LOTFAE	NHB	1	25	$C_{13}H_9FN_2$			
XOYDIB	NHB	1	22	$C_{10}H_7F_3N_2$			
AKOYAA	NHB	1	25	$C_{12}H_6F_6O$			
AYOHUR	NHB	1	18	$C_9H_4F_4O$			
BIKCII	NHB	1	24	$C_{13}H_4F_6O$			
DFNAPQ	NHB	1	18	$C_{10}H_4F_2O_2$			
DUCTAZ	NHB	1	23	$C_{11}H_8F_2O_2$			
FOQJIH	NHB	1	27	$C_{12}H_9F_3O_3$			
MOWDIO	NHB	1	22	$C_{12}H_4F_4O_2$			
VUCJAH	NHB	1	23	$C_{11}H_6F_4O_2$			
ADIGEA	NHB	1	25	$C_{10}H_9F_2N_3O$			
AKUNUQ	NHB	1	22	$C_8H_7F_3N_2O_2$			
AYUZAW	NHB	1	32	$C_{13}H_{15}F_2NO$			
BETMET	NHB	1	15	$C_8H_3F_2NO$			
BONWAE	NHB	1	21	$C_7H_7F_3N_2O_2$			
DUBPUO	NHB	1	32	$C_{13}H_{13}F_3N_2O$			
ELULIH	NHB	1	22	$C_{10}H_5F_3N_2O_2$			
EYAVEF	NHB	1	26	$C_8H_7F_6N_3O_2$			
FADBEU	NHB	1	25	$C_{10}H_8F_4N_2O$			

Continued on next page

Table M.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula	NHB47	HB53	TP100
GOSJEG	NHB	1	22	$C_{10}H_7F_2NO_2$			
MOWDAG	NHB	1	24	$C_{14}H_4F_2N_2O_2$			
SUDYOI	NHB	1	22	$C_{10}H_5F_3N_2O_2$			
SUGRUK	NHB	1	27	$C_{12}H_{10}F_3NO$			
ABUMIT	NHB	1	12	$C_6H_4Cl_2$			
AHCHEX01	NHB	1	18	$C_6H_6Cl_6$			
BAHCUI	NHB	1	23	$C_9H_6Cl_8$			
CLFORM03	NHB	1	5	$CHCl_3$			
CLFULV01	NHB	2	18/18	$C_{10}Cl_8$			
CLMETH03	NHB	2	5/5	CH_3Cl			
DCLANT10	NHB	1	24	$C_{14}H_8Cl_2$			
DCLETH02	NHB	1	8	$C_2H_4Cl_2$			
DECPYR10	NHB	1	26	$C_{16}Cl_{10}$			
FEMKUD	NHB	1	19	$C_7H_8Cl_4$			
HCCYHD	NHB	1	18	$C_6H_6Cl_6$			
HCCYHG02	NHB	1	18	$C_6H_6Cl_6$			
HCLBNZ13	NHB	1	12	C_6Cl_6			
HCLTRL	NHB	1	22	$C_{10}H_6Cl_6$			
MCBENZ	NHB	2	12/12	C_6H_5Cl			
MECHLF12	NHB	1	8	$C_2H_3Cl_3$			
PIRPEL	NHB	1	18	$C_6H_6Cl_6$			
PNCLBZ	NHB	1	12	C_6HCl_5			
RIVDOQ	NHB	1	8	$C_2H_4Cl_2$			
TCCYHX10	NHB	2	18/18	$C_6H_9Cl_3$			
TCHLBZ	NHB	1	12	$C_6H_3Cl_3$			
XAXCOQ03	NHB	1	8	C_2H_5Cl			
YAFHUI	NHB	1	18	$C_{10}H_2Cl_6$			
ZZZBLP02	NHB	1	24	$C_{14}H_8Cl_2$			
ZZZBLP03	NHB	1	24	$C_{14}H_8Cl_2$			
BAWNIW	NHB	1	14	$C_8Cl_4N_2$			
BAWNIW02	NHB	3	14/14/14	$C_8Cl_4N_2$			
BOPHUJ	NHB	2	18/18	$C_8H_7ClN_2$			
BOTYIT	NHB	1	10	$C_4H_2Cl_2N_2$			
CACTAD	NHB	1	29	$C_{11}H_{12}Cl_3N_3$			
CIPWOO	NHB	1	19	$C_9H_7ClN_2$			
CLAZBZ02	NHB	1	24	$C_{12}H_8Cl_2N_2$			
CLBECN03	NHB	1	13	$C_7H_2Cl_3N$			
DAYFIV	NHB	1	13	$C_7H_3Cl_2N$			

Continued on next page

Table M.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula	NHB47	HB53	TP100
DCPYAZ	NHB	1	10	$C_4H_2Cl_2N_2$			
DOJFAL	NHB	1	20	$C_8H_6Cl_3N_3$			
EWUZAZ	NHB	1	13	$C_7H_3Cl_2N$			
GETVEH	NHB	1	11	$C_5H_2Cl_3N$			
GOLYAI	NHB	1	16	$C_8H_5ClN_2$			
GUJLAZ	NHB	1	15	$C_7H_4ClN_3$			
HCZCHP	NHB	1	16	$C_6H_3Cl_6N$			
IBUZIP	NHB	1	22	$C_{12}H_6Cl_2N_2$			
AHUGUH	NHB	1	19	$C_7H_7Cl_3O_2$			
AHUHAO	NHB	1	29	$C_{11}H_{13}Cl_3O_2$			
AZIBER	NHB	1	14	$C_7H_4Cl_2O$			
BAGYIR	NHB	1	14	$C_4H_5Cl_3O_2$			
CDCDOX	NHB	1	14	$C_4H_6Cl_2O_2$			
CLBENQ	NHB	1	12	$C_6H_3ClO_2$			
DCLBQN	NHB	1	12	$C_6H_2Cl_2O_2$			
GISXEM	NHB	1	16	$C_7H_5Cl_3O$			
HUGQOQ	NHB	1	14	$C_7H_4Cl_2O$			
LEJDAH	NHB	1	22	$C_{12}H_7ClO_2$			
LEJDEL	NHB	1	22	$C_{12}H_7ClO_2$			
WERNUF01	NHB	1	17	$C_8H_6Cl_2O$			
XOZFOK	NHB	1	26	$C_{12}H_{11}ClO_2$			
AKETAN	NHB	1	29	$C_{13}H_{10}ClN_3O_2$			
BAFLEZ	NHB	1	19	$C_7H_5ClN_2O_4$			
BEGCOF	NHB	1	30	$C_{12}H_{13}ClN_2O_2$			
BEGPEJ	NHB	1	24	$C_7H_9ClN_4O_3$			
BEGPIN	NHB	2	24/24	$C_7H_9ClN_4O_3$			
CETLAQ	NHB	1	16	$C_8H_4ClNO_2$			
COHSAV	NHB	1	24	$C_{10}H_9Cl_2NO_2$			
FICFON	NHB	1	24	$C_{10}H_9Cl_2NO_2$			
FOXWEW	NHB	1	23	$C_{11}H_8ClNO_2$			
GOQJUU	NHB	1	20	$C_9H_6Cl_2N_2O$			
FAVTAX	NHB	1	22	$C_{12}H_4F_4Cl_2$			
IHERES	NHB	1	5	CH_2FCl			
NAFCUV	NHB	1	12	C_6H_4FCl			
NAFDAC	NHB	1	12	C_6H_4FCl			
XEZHAM	NHB	1	22	$C_{12}H_8FCl$			
XEZHEQ	NHB	2	22/22	$C_{12}H_8FCl$			
AQUSEM	NHB	2	21/21	$C_9H_6F_3ClN_2$			

Continued on next page

Table M.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula	NHB47	HB53	TP100
AWACAE	NHB	2	16/16	$C_8H_4FClN_2$			
AWOBIZ	NHB	1	13	C_7H_3FClN			
JEHYIF	NHB	1	12	$C_6HFCl_2N_2$			
LAPLAR01	NHB	1	16	$C_8H_3FCl_2N_2$			
RAGPEX	NHB	2	17/17	$C_9H_4F_2ClN$			
RAGPEX01	NHB	1	17	$C_9H_4F_2ClN$			
AJOYEF	NHB	1	23	$C_{11}H_7F_2ClO_2$			
BAYXAD	NHB	2	26/26	$C_{13}H_6F_5ClO$			
ILAMOZ	NHB	1	14	C_7H_4FClO			
ILAMUF	NHB	1	14	$C_7H_3F_2ClO$			
ILANAM	NHB	1	14	$C_7H_3F_2ClO$			
KUSQOG	NHB	1	20	$C_9H_6F_3ClO$			
BIRQOL	NHB	1	34	$C_{18}H_{10}FCl_2NO_2$			
IYIXUJ	NHB	1	23	$C_{11}H_7FClNO_2$			
IYIYAQ	NHB	1	20	$C_{10}H_4F_2ClNO_2$			
JIHNOG	NHB	1	17	$C_7H_2F_3Cl_2NO_2$			
QUZTAH	NHB	1	22	$C_{10}H_2F_3Cl_4NO_2$			
ACUWAZ	NHB	1	30	$C_{18}H_{10}S_2$			
AKOZOR	NHB	1	30	$C_{18}H_8S_4$			
ASOPAB	NHB	1	29	$C_{14}H_{10}S_5$			
AYOJAZ	NHB	1	18	C_9H_8S			
CUGCUF	NHB	1	30	$C_{16}H_{12}S_2$			
DITHAN02	NHB	1	14	$C_4H_8S_2$			
DTENYL02	NHB	1	16	$C_8H_6S_2$			
FABPON11	NHB	1	18	$C_9H_6S_3$			
KOZXIJ	NHB	1	34	$C_{18}H_{14}S_2$			
LEVMUU	NHB	1	15	C_8H_6S			
LIQWEN	NHB	1	12	C_3H_8S			
NPDTOL02	NHB	2	30/30	$C_{18}H_8S_4$			
PADQAN	NHB	1	9	C_2H_6S			
THIANT05	NHB	1	22	$C_{12}H_8S_2$			
TRITAN03	NHB	1	12	$C_3H_6S_3$			
ZAQFAB	NHB	1	18	$C_8H_4S_6$			
AFEWAK	NHB	1	18	$C_9H_5N_3S$			
AXORIQ	NHB	1	15	$C_4H_7N_3S$			
AYAVOM	NHB	1	21	$C_{10}H_8N_2S$			
DEWYOV	NHB	1	28	$C_{11}H_{14}N_2S$			
DIZVOZ	NHB	1	18	$C_7H_6N_4S$			

Continued on next page

Table M.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula	NHB47	HB53	TP100
DMITDZ	NHB	1	17	$C_6H_7N_3S$			
DOSZUG	NHB	2	15/15	$C_8H_4N_2S$			
FANJOT	NHB	1	16	$C_6H_6N_2S_2$			
FANJOT01	NHB	1	16	$C_6H_6N_2S_2$			
POZYUB02	NHB	1	19	$C_{10}H_6N_2S$			
POZYUB04	NHB	1	19	$C_{10}H_6N_2S$			
QUCGAZ	NHB	1	24	$C_{13}H_8N_2S$			
YOTTOT	NHB	1	26	$C_{15}H_9NS$			
ZUCMUI	NHB	1	27	$C_{12}H_{12}N_2S$			
ZUCNAP	NHB	1	30	$C_{13}H_{14}N_2S$			
ZUJSIJ	NHB	1	19	$C_{10}H_7NS$			
AJEVOB	NHB	1	20	$C_8H_8O_3S$			
AKOXUV	NHB	1	19	$C_8H_8OS_2$			
ASIBOT	NHB	1	27	$C_{10}H_{14}OS_2$			
ATOMAZ	NHB	2	30/30	$C_{14}H_8O_6S_2$			
BOZFEC	NHB	1	26	$C_{12}H_{10}O_2S_2$			
CABYAG	NHB	1	17	C_9H_6OS			
CANRAM	NHB	1	30	$C_{15}H_{12}O_2S$			
CHPDTP	NHB	1	20	$C_{11}H_6OS_2$			
DMETSO	NHB	1	10	C_2H_6OS			
DMETSO05	NHB	1	10	C_2H_6OS			
FORROU	NHB	1	14	$C_5H_6OS_2$			
GOWPUG	NHB	1	26	$C_{11}H_{10}O_3S_2$			
GUFJAS	NHB	1	13	C_5H_6OS			
GUFJIA	NHB	1	18	$C_7H_8O_2S$			
RIQKAE	NHB	1	22	$C_8H_{10}O_2S_2$			
TOVDIU	NHB	1	29	$C_{14}H_{12}O_2S$			
COXKAE	NHB	1	34	$C_{12}H_{16}N_2O_3S$			
CUBTEB	NHB	1	29	$C_{16}H_9NOS_2$			
KUGKIJ	NHB	1	36	$C_{16}H_{15}NO_3S$			
NOTYAZ	NHB	1	23	$C_9H_9NO_3S$			
NOVSUP	NHB	1	33	$C_{17}H_{13}NOS$			
PUHBIG	NHB	1	40	$C_{19}H_{17}NO_2S$			
SUFKUC	NHB	1	24	$C_{12}H_9NOS$			
XORSIJ	NHB	1	33	$C_{13}H_{15}NO_2S_2$			
XOYJED	NHB	1	35	$C_{17}H_{15}NOS$			
XOZNIM	NHB	1	43	$C_{20}H_{17}N_3O_2S$			
XUMBIT	NHB	1	33	$C_{14}H_{15}NO_2S$			

Continued on next page

Table M.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula	NHB47	HB53	TP100
XUMCIU	NHB	1	30	$C_{13}H_{13}NO_2S$			
ZUJRII	NHB	1	23	$C_{11}H_9NOS$			
ZUJWEJ	NHB	2	33/33	$C_{17}H_{13}NOS$			
FEBROU	NHB	1	27	$C_{13}H_4F_8S_2$			
IJIYAD	NHB	1	27	$C_{13}H_4F_8S_2$			
LINRUV	NHB	1	24	$C_{10}H_{10}F_2S_2$			
MAKQIZ	NHB	1	24	$C_{14}H_4F_4S_2$			
VAFFUF	NHB	1	26	$C_{10}H_8F_6S_2$			
APATER	NHB	1	20	$C_{11}H_4F_3NS$			
COZXAT	NHB	1	28	$C_{13}H_7F_6NS$			
GAXKAU	NHB	1	26	$C_{11}H_{11}F_2NS$			
GOCPEW	NHB	1	24	$C_{12}H_6F_2N_2S_2$			
IJEPEU	NHB	1	29	$C_{14}H_5F_8NS$			
LIQJUS	NHB	1	15	$C_6H_5F_2NS$			
NOSNOY	NHB	1	25	$C_{10}H_9F_3N_2S$			
OPAJIA	NHB	1	24	$C_{13}H_4F_5NS$			
CEXVAF	NHB	1	22	$C_{11}H_5F_3O_2S$			
COWTIT	NHB	1	21	$C_{10}H_5F_3O_2S$			
DFPSLO	NHB	1	25	$C_{12}H_8F_2O_2S$			
FOCJOZ	NHB	1	25	$C_{12}H_9FOS_2$			
LUZXUA	NHB	1	28	$C_{14}H_9F_3OS$			
NUQGUD	NHB	1	19	$C_8H_7FO_2S$			
UKUNAR	NHB	1	28	$C_{14}H_9F_3OS$			
XUPPEE	NHB	1	19	$C_7H_5FO_4S_2$			
YELLIN	NHB	1	19	$C_{10}H_2F_4O_2S$			
HABVIQ	NHB	1	27	$C_{13}H_{12}N_2$		✓	✓
JOSLOU	NHB	1	25	$C_{13}H_{11}N$		✓	✓
FIRLAU	NHB	1	10	$C_4H_3F_2N$			
IDAHUR	NHB	1	14	C_6H_6FN			
CEBZIU	NHB	1	30	$C_{10}H_{11}F_3N_2O_4$			
QOFGAV	NHB	1	27	$C_{11}H_{12}O_2S_2$			
AMMEPY01	HB	1	16	$C_6H_8N_2$		✓	✓
AMPYRE	HB	1	13	$C_5H_6N_2$		✓	✓
AMPYRM11	HB	1	12	$C_4H_5N_3$		✓	✓
BIGUAN01	HB	1	14	$C_2H_7N_5$		✓	✓
BZDMAZ02	HB	1	15	$C_7H_6N_2$		✓	✓
DAMTRZ11	HB	1	12	$C_2H_5N_5$		✓	✓
FITZEO01	HB	1	26	$C_{14}H_9N_3$		✓	✓

Continued on next page

Table M.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula	NHB47	HB53	TP100
FULPIM	HB	1	12	$C_4H_6N_2$		✓	✓
HIFCIJ	HB	1	20	$C_8H_{10}N_2$		✓	✓
IMAZOL06	HB	1	9	$C_3H_4N_2$		✓	✓
ITOBAU	HB	1	17	$C_5H_{11}N$		✓	✓
MELAMI06	HB	1	15	$C_3H_6N_6$		✓	✓
OCAPUE	HB	1	11	$C_5H_4N_2$		✓	✓
PIWXEY	HB	1	22	$C_6H_{14}N_2$		✓	✓
PURINE	HB	1	13	$C_5H_4N_4$		✓	✓
TRAZOL03	HB	1	8	$C_2H_3N_3$		✓	✓
ABINOS02	HB	1	20	$C_5H_{10}O_5$		✓	✓
ACETAC07	HB	1	8	$C_2H_4O_2$		✓	✓
BENZAC12	HB	1	15	$C_7H_6O_2$		✓	✓
BESGEL	HB	1	23	$C_{13}H_8O_2$		✓	✓
BESKAL04	HB	1	17	$C_7H_6O_4$		✓	✓
FORMAC02	HB	1	5	CH_2O_2		✓	✓
JUBBIR	HB	1	24	$C_{12}H_8O_4$		✓	✓
MALIAC12	HB	1	12	$C_4H_4O_4$		✓	✓
MALIAC13	HB	1	12	$C_4H_4O_4$		✓	✓
METHOL04	HB	1	6	CH_4O		✓	✓
NAPOAC01	HB	1	21	$C_{11}H_8O_2$		✓	✓
SUCACB03	HB	1	14	$C_4H_6O_4$		✓	✓
TELXAJ	HB	1	12	$C_4H_6O_2$		✓	✓
ZEFPUX	HB	1	15	$C_4H_{10}O$		✓	✓
ZZZKPE03	HB	1	20	$C_6H_{12}O_2$		✓	✓
AZUDIJ	HB	1	24	$C_8H_{11}N_3O_2$		✓	✓
EDOMUI	HB	1	21	$C_8H_9NO_3$		✓	✓
EGIRUJ	HB	1	22	$C_{10}H_8N_2O_2$		✓	✓
FUMRAM01	HB	1	13	$C_4H_5NO_3$		✓	✓
HIHNIV	HB	1	15	$C_5H_5N_3O_2$		✓	✓
JEFTEU	HB	1	18	$C_6H_7N_3O_2$		✓	✓
KUFVAI03	HB	1	20	$C_9H_8N_2O$		✓	✓
KUXPUP	HB	1	15	$C_6H_5NO_3$		✓	✓
MALEHY01	HB	1	12	$C_4H_4N_2O_2$		✓	✓
MALEHY11	HB	1	12	$C_4H_4N_2O_2$		✓	✓
MALEHY14	HB	1	12	$C_4H_4N_2O_2$		✓	✓
NMALAM01	HB	1	15	$C_3H_5N_3O_4$		✓	✓
OVUKAS	HB	1	27	$C_{10}H_{10}N_4O_3$		✓	✓
QICVAZ01	HB	2	11/11	$C_3H_3N_3O_2$		✓	✓

Continued on next page

Table M.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula	NHB47	HB53	TP100
REZDUY	HB	1	15	$C_5H_5N_3O_2$		✓	✓
SALMID04	HB	1	17	$C_7H_7NO_2$		✓	✓
SALMID07	HB	1	17	$C_7H_7NO_2$		✓	✓
SEPNUX	HB	1	15	$C_5H_4N_2O_4$		✓	✓
XAGWIM	HB	1	13	$C_5H_5NO_2$		✓	✓
ZAPGII	HB	1	15	$C_3H_5N_5O_2$		✓	✓
APUREK	HB	1	16	$C_6H_3F_4N_3$			
BENAFP	HB	1	16	C_7H_6FNO			
CAFVAG	HB	1	16	$C_6H_4F_4N_2$			
GANBEF	HB	1	23	$C_9H_7F_4NO_2$			
GEHSIW	HB	1	19	$C_9H_4F_3NO_2$			
GEMRIB	HB	1	22	$C_9H_9F_2NO$			
HOJZUB	HB	1	14	$C_4H_4F_3NO_2$			
JEGTUN	HB	1	18	$C_7H_5F_3N_2O$			
KUSJAL	HB	1	14	$C_6H_5F_2N$			
MUMSET	HB	1	11	$C_4HF_3N_2O$			
RAGQIC	HB	1	19	$C_9H_5F_3N_2$			
UTODOX	HB	1	20	$C_8H_5F_4NO_2$			
VEPXOG	HB	1	13	$C_5H_2F_4N_2$			
ABENEB	HB	1	16	$C_7H_5FO_3$			
COVJIG	HB	1	15	$C_7H_5FO_2$			
DUTREQ	HB	1	15	$C_7H_4F_2O_2$			
FACETC01	HB	1	8	$C_2H_3FO_2$			
GISZIS	HB	1	15	$C_7H_4F_2O_2$			
IKAYEZ	HB	1	15	$C_7H_3F_3O_2$			
IKESIC	HB	1	16	$C_7H_2F_4O_3$			
ODUSEM	HB	1	18	$C_8H_2F_4O_4$			
PFBZAC	HB	1	15	$C_7HF_5O_2$			
UDUXEY	HB	1	12	C_5HF_4NO			
WESBEB	HB	1	19	$C_8H_7FO_3$			
WESBEB01	HB	1	19	$C_8H_7FO_3$			
ABEROP	HB	1	18	$C_8H_6ClN_3$			
AFIGUS	HB	1	19	$C_8H_7Cl_2NO$			
AGIDOL	HB	1	23	$C_{11}H_8ClNO_2$			
AMCLPY	HB	1	13	$C_5H_5ClN_2$			
ASATET	HB	1	16	C_7H_6ClNO			
ASUWAN	HB	1	16	$C_6H_6ClNO_2$			
AZCFOM01	HB	1	14	$C_2H_4Cl_2N_6$			

Continued on next page

Table M.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula	NHB47	HB53	TP100
BAWVAW	HB	1	21	$C_{10}H_8ClNO$			
BEXHOE	HB	1	17	$C_8H_4ClNO_3$			
BOMSED	HB	1	21	$C_7H_9ClN_2O_2$			
CANAPQ01	HB	2	20/20	$C_{10}H_6ClNO_2$			
CLTRZL	HB	1	8	$C_2H_2ClN_3$			
FARMIW	HB	1	15	$C_7H_5ClN_2$			
GAPPIY	HB	1	12	$C_4H_4ClN_3$			
GOTJUW	HB	1	15	$C_7H_5ClN_2$			
AZERUU	HB	2	17/17	$C_8H_5ClO_3$			
BETXOQ	HB	1	18	$C_8H_4Cl_2O_4$			
CHPYRD	HB	1	12	C_5H_4ClNO			
CHXIQL	HB	1	18	C_9H_6ClNO			
CLACET02	HB	1	8	$C_2H_3ClO_2$			
CLBZAC01	HB	2	15/15	$C_7H_5ClO_2$			
COTFEX	HB	1	26	$C_{12}H_8Cl_2O_4$			
DCDHNQ	HB	1	20	$C_{10}H_4Cl_2O_4$			
DOQQEH	HB	2	14/14	$C_6H_3Cl_2NO_2$			
GADHAU	HB	1	9	$C_3HCl_3O_2$			
HOFHEQ	HB	1	14	$C_6H_4Cl_2O_2$			
AJIVIZ	HB	1	13	$C_5H_4ClNO_2$			
DECZEQ	HB	1	21	$C_9H_8ClNO_2$			
ABALAS	HB	1	21	$C_{10}H_8N_2S$			
ADAHET	HB	1	20	$C_8H_7N_3S_2$			
AGAKUP	HB	2	27/27	$C_{14}H_9N_3S$			
AGAKUP01	HB	1	27	$C_{14}H_9N_3S$			
AKOVOL	HB	1	12	C_5H_5NS			
AKOVOL02	HB	2	12/12	C_5H_5NS			
BONVOS	HB	1	20	$C_9H_7N_3S$			
MOVJAL	HB	2	10/10	$C_3H_4N_2S$			
PHESAZ08	HB	1	23	$C_{12}H_9NS$			
AQIGUC	HB	1	30	$C_{14}H_{12}O_2S_2$			
CAFKEZ	HB	1	26	$C_{10}H_{12}O_3S$			
CAPMSU	HB	1	20	$C_8H_8O_3S$			
DERBEH	HB	2	16/16	$C_4H_6O_4S_2$			
DITFIX	HB	1	38	$C_{14}H_{18}O_2S_4$			
GUMFUS	HB	2	18/18	$C_5H_{10}O_2S$			
XOPSED	HB	1	18	$C_7H_6O_4S$			
APOXOS01	HB	1	26	$C_{10}H_{10}N_4OS$			

Continued on next page

Table M.1 – continued from previous page

Refcode	HB?	Z'	nAtms	Chemical Formula	NHB47	HB53	TP100
AYUBUS	HB	1	23	$C_6H_{12}N_2O_2S$			
HEVZAL	HB	1	25	$C_{10}H_7N_3O_4S$			
HTDZDX10	HB	1	15	$C_3H_3N_3O_5S$			
LELVIK	HB	2	18/18	$C_4H_9NO_3S$			

M.2 TP442 Parameter Estimation Performance Metrics

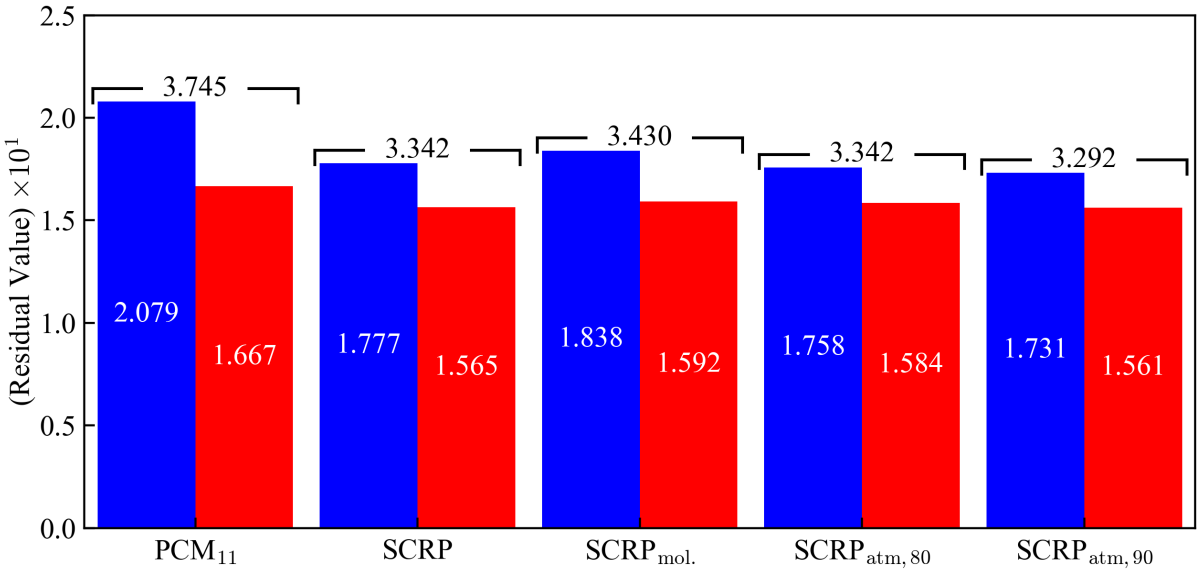


Figure M.1: Sum of geometry (blue) and energy (red) residual scores for the TP442 training set. The total objective function value is summed above each pair.

Table M.2: Error metrics for the models trained on TP442. The metrics are reported for the minimized structures post-CE, and (for SCERP-embedded models) post-CSORM. The latter values are given after the '/'. Root-mean-square-deviations (between CE and CSORM) in $RMSD_{15}$ and dU_{inter} scores are also reported for the SCERP-embedded models. The numbers of structures with $|\Delta(RMSD_{15})_{CE\rightarrow CSORM}| \geq 0.2$ Å and $|\Delta(dU_{inter})_{CE\rightarrow CSORM}| \geq 1$ kJ/mol are given in parenthesis. $dU_{inter} = U_{inter}^{calc} - U_{inter}^{ref}$, such that $dU_{inter} > 0$ indicates underbinding.

Model	Avg. $RMSD_{15}$ (Å)	AD (kJ/mol)	AAD (kJ/mol)	SD (kJ/mol)	$RMSD_{RMSD_{15}}$ (Å)	$RMSD_{dU_{inter}}$ (kJ/mol)
NHB Structures (317 crystals)						
PCM_{11}	0.1897	−0.1163	2.8502	3.9647	-	-
SCRP	0.1770/0.1939	−0.0419/0.0160	2.8048/2.8697	3.8518/3.9365	0.0805 (3)	0.3220 (7)
$SCRP_{mol.}$	0.1848/0.1977	−0.1869/−0.1682	2.8272/2.8808	3.8872/3.9699	0.0438 (3)	0.2911 (5)
$SCRP_{atm,80}$	0.1853/0.1965	−0.0034/0.0080	2.8468/2.9111	3.8975/3.9862	0.0349 (1)	0.2819 (4)
$SCRP_{atm,90}$	0.1822/0.1925	−0.0880/−0.0293	2.8148/2.8788	3.8552/3.9369	0.0281 (1)	0.2888 (5)
HB Structures (125 crystals)						
PCM_{11}	0.1916	−0.0257	3.4837	4.8703	-	-
SCRP	0.1768/0.1969	0.0013/0.8904	3.0786/3.4576	4.3262/4.8304	0.0984 (1)	1.5767 (57)
$SCRP_{mol.}$	0.1799/0.1967	−0.1501/−0.7836	3.1959/3.4506	4.2718/4.6324	0.0565 (1)	1.3112 (55)
$SCRP_{atm,80}$	0.1746/0.2018	−0.1220/−0.3672	3.0810/3.2692	4.2549/4.6124	0.1427 (2)	1.1518 (50)
$SCRP_{atm,90}$	0.1728/0.1980	−0.0480/0.2358	3.0238/3.2311	4.2143/4.5832	0.1383 (2)	1.1394 (41)

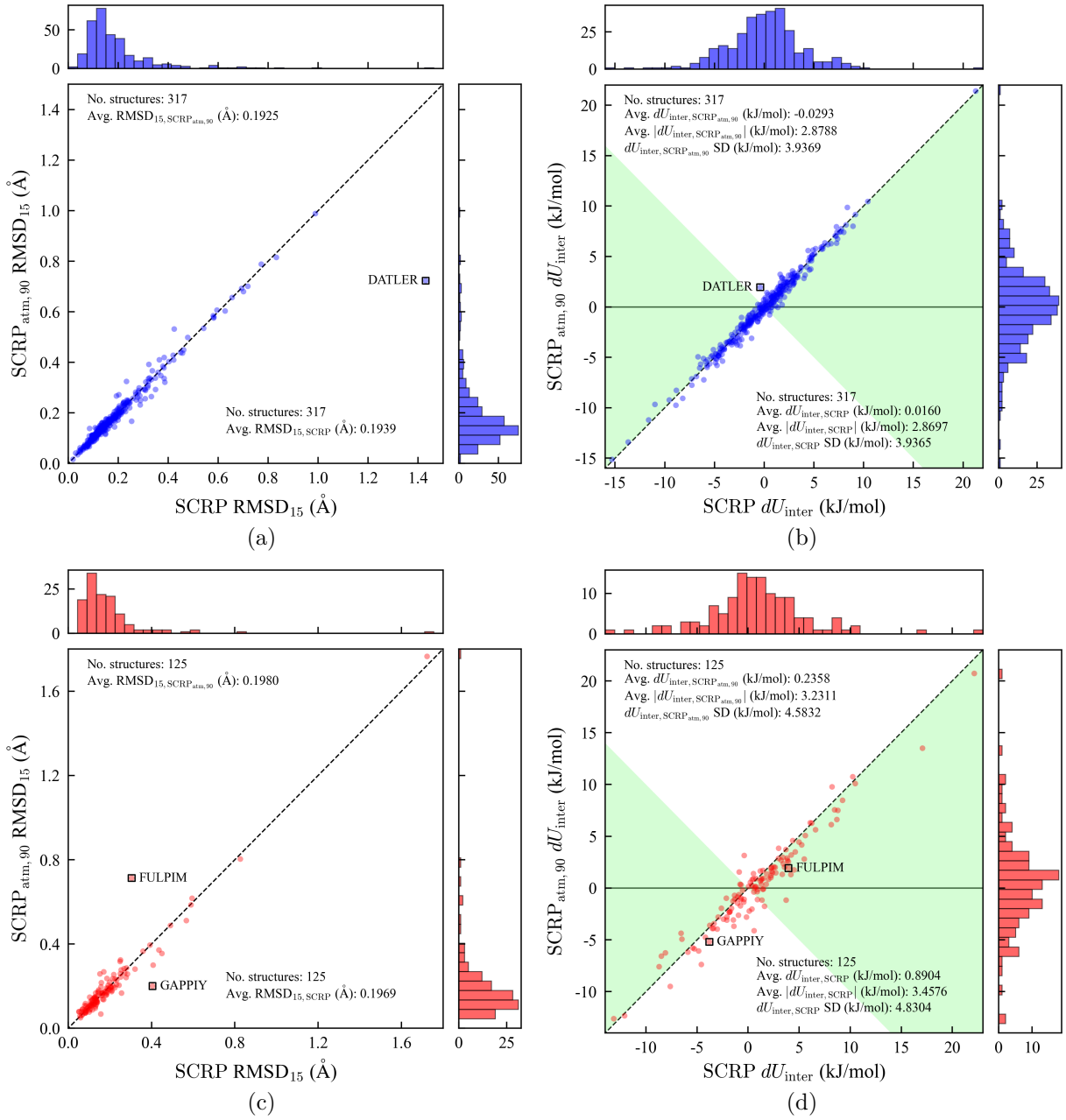


Figure M.2: Post-CSORM TP442 errors for SCRP and $\text{SCRP}_{\text{atm,90}}$: (a) NHB crystals RMSD_{15}, (b) NHB crystals dU_{inter}, (c) HB crystals RMSD_{15}, and (d) HB crystals dU_{inter}. Structures with $|\Delta\text{RMSD}_{15}| \geq 0.20$ Å are marked with □. Green areas in (b) and (d) denote where the y-axis model has a lower $|dU_{\text{inter}}|$. Histograms of each model's errors are shown on the corresponding axes.

M.3 Optimized Force-field Parameters for SCERP-embedded Models

Table M.3: Optimized FF for multipole electrostatics with $\mathrm{SCRP}_{\mathrm{atm},90}$, by Scheme 5.

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
C	C	3037.7688993985	0.2777777778	23.1179318902
C	H	399.6385477589	0.2724795640	4.8893329028
C	O	2852.4053422542	0.2645502646	17.3127808588
C	N	2474.8662233812	0.2710027100	18.4521350132
C	F	2688.0307044656	0.2577319588	9.4101066929
C	Cl	6042.9911775207	0.2812939522	45.7796387519
C	S	4255.2873622197	0.2898550725	50.3398393068
C	$\mathrm{H_N}$	674.2651430571	0.2421307506	3.4398150923
C	$\mathrm{H_O}$	0.0000000000	0.2421307506	0.7927721003
H	H	106.6203654214	0.2673796791	1.7311263912
H	O	291.5648900150	0.2597402597	0.0000000000
H	N	417.5840293000	0.2659574468	3.1610616928
H	F	265.8446129875	0.2531645570	0.7176465536
H	Cl	605.2471476570	0.2758620690	7.0195406128
H	S	512.4580493561	0.2840909091	7.8250428367
H	$\mathrm{H_N}$	5.9097251015	0.2380952381	0.2228508096
H	$\mathrm{H_O}$	439.7887697614	0.2380952381	3.0617754504
O	O	6320.0058976888	0.2525252525	17.4019373215
O	N	6335.2947127260	0.2583979328	19.8130509063
O	F	4005.3696424178	0.2463054187	8.1645968550
O	Cl	3918.7035963797	0.2677376171	17.9527607839
O	S	2192.6274093769	0.2754820937	3.1868814121
O	$\mathrm{H_N}$	151.9752213280	0.2320185615	0.0000000000
O	$\mathrm{H_O}$	62.1134690919	0.2320185615	0.0000000000
N	N	3113.7279528446	0.2645502646	12.7545518283
N	F	4962.3088435520	0.2518891688	13.7458369825
N	Cl	4651.3741155798	0.2743484225	29.1583468695
N	S	2680.0477444931	0.2824858757	24.6293885889
N	$\mathrm{H_N}$	347.2541275568	0.2369668246	0.6535656605
N	$\mathrm{H_O}$	257.2932500021	0.2369668246	3.6689943636
F	F	4177.4569450888	0.2403846154	7.7420995960
F	Cl	5674.6158059019	0.2607561930	16.3835246198
F	S	2650.7596612799	0.2680965147	8.9256811128
F	$\mathrm{H_N}$	527.8927500098	0.2267573696	2.0314699475
F	$\mathrm{H_O}$	0.0000000000	0.2267573696	0.0244830957
Cl	Cl	10947.8559017447	0.2849002849	84.0858477346

Continued on next page

Table M.3 – continued from previous page

t	t'	$A^{t,t'}$ (eV)	$B^{t,t'}$ (Å)	$C^{t,t'}$ (eV Å^{-6})
Cl	S	4312.2835756029	0.2936857562	65.0366904278
Cl	H_N	9640.2463772602	0.2447980416	28.5980345985
Cl	H_O	12313.8027958264	0.2447980416	20.0609464358
S	S	6078.3301301092	0.3030303030	144.4798033120
S	H_N	1163.3720399020	0.2512562814	10.7439158566
S	H_O	1103.2316637525	0.2512562814	14.8096181223
H_N	H_N	0.0000000000	0.2145922747	0.2083557948
H_N	H_O	114.7803877684	0.2145922747	0.0000000000
H_O	H_O	0.0842175988	0.2145922747	0.0000000000

Table M.4: Optimized FF for multipole electrostatics with $\mathrm{SCRP}_{\mathrm{atm},90}$, by Scheme 7.

t	t'	$\varepsilon^{t,t'}$ (eV)	$\sigma^{t,t'}$ (Å)	$\lambda_{\mathrm{r}}^{t,t'}$ (-)	$\lambda_{\mathrm{a}}^{t,t'}$ (-)
C	C	0.0046434220	3.4034953513	17.1665275238	6.0000000000
C	H	0.0018150206	2.9576762385	11.1989052478	6.0000000000
C	N	0.0050320799	3.2733138805	16.1557997755	6.0000000000
C	O	0.0049367098	3.1638456104	12.6142711681	6.0000000000
C	F	0.0030623724	3.2005009657	21.0899283359	6.0000000000
C	S	0.0094671553	3.4454778968	17.0972826340	6.0000000000
C	Cl	0.0079304388	3.4708065173	14.2114454946	6.0000000000
C	$\mathrm{H_N}$	0.0013219779	2.8623928432	7.5310991975	6.0000000000
C	$\mathrm{H_O}$	0.0024345784	2.2168691547	56.5352614396	6.0000000000
H	H	0.0004615709	3.0631239477	6.2845886208	6.0000000000
H	N	0.0003968797	3.4608905040	7.9315828719	6.0000000000
H	O	0.0001455825	4.0482194884	6.2479768428	6.0000000000
H	F	0.0003091847	3.5322282332	6.3951287859	6.0000000000
H	S	0.0018927900	3.1315888788	11.1223071148	6.0000000000
H	Cl	0.0012118064	3.5011527187	6.1658857476	6.0000000000
H	$\mathrm{H_N}$	0.0003517874	2.1646367484	28.5480104335	6.0000000000
H	$\mathrm{H_O}$	0.0045154003	2.3235708153	22.3440593313	6.0000000000
N	N	0.0028674925	3.3797072013	14.9968568463	6.0000000000
N	O	0.0037848422	3.3560615045	11.1415991216	6.0000000000
N	F	0.0023215069	3.3863397101	11.8380561277	6.0000000000
N	S	0.0049612324	3.3664697323	12.7431017263	6.0000000000
N	Cl	0.0047535446	3.3553404206	13.8278065555	6.0000000000
N	$\mathrm{H_N}$	0.0000125458	4.5014901927	7.4151720210	6.0000000000
N	$\mathrm{H_O}$	0.0014753817	2.4484153734	7.4341711120	6.0000000000
O	O	0.0030624783	3.3392267256	13.0528243032	6.0000000000
O	F	0.0015684121	3.2059881967	21.7449724257	6.0000000000
O	S	0.0002118330	3.9219037975	15.5054903084	6.0000000000
O	Cl	0.0021551228	3.6231496173	8.4228726776	6.0000000000
O	$\mathrm{H_N}$	0.0000019831	5.6685276876	7.0056285748	6.0000000000
O	$\mathrm{H_O}$	0.0000983727	1.5860214915	12.0574759253	6.0000000000
F	F	0.0021508331	3.2062949540	11.4168453764	6.0000000000
F	S	0.0027653598	3.3285809439	23.0120735495	6.0000000000
F	Cl	0.0032504993	3.4315753409	19.7294053514	6.0000000000
F	$\mathrm{H_N}$	0.0040095564	2.3094928198	19.2413773728	6.0000000000
F	$\mathrm{H_O}$	0.0010736749	2.8614290612	22.5448696179	6.0000000000
S	S	0.0272202832	3.3215272294	13.9454953402	6.0000000000
S	Cl	0.0132850727	3.4157965711	13.9681487874	6.0000000000
S	$\mathrm{H_N}$	0.0055037009	2.4917664510	32.2043219879	6.0000000000
S	$\mathrm{H_O}$	0.0069729289	2.6859416803	6.1969297200	6.0000000000
Cl	Cl	0.0099117077	3.6182322572	10.2986972395	6.0000000000

Continued on next page

Table M.4 – continued from previous page

t	t'	$\varepsilon^{t,t'}$ (eV)	$\sigma^{t,t'}$ (Å)	$\lambda_{\mathrm{r}}^{t,t'}$ (-)	$\lambda_{\mathrm{a}}^{t,t'}$ (-)
Cl	H_N	0.0049653434	3.2929869855	6.1998351951	6.0000000000
Cl	H_O	0.0047855751	3.3222693598	12.1703122539	6.0000000000
H_N	H_N	0.0001150831	2.0453016379	27.5970172796	6.0000000000
H_N	H_O	0.0001134616	1.8385526392	40.6234542037	6.0000000000
H_O	H_O	0.0026019139	1.5223061237	42.8505947200	6.0000000000

M.4 CE–755 Energy Accuracy without Outliers

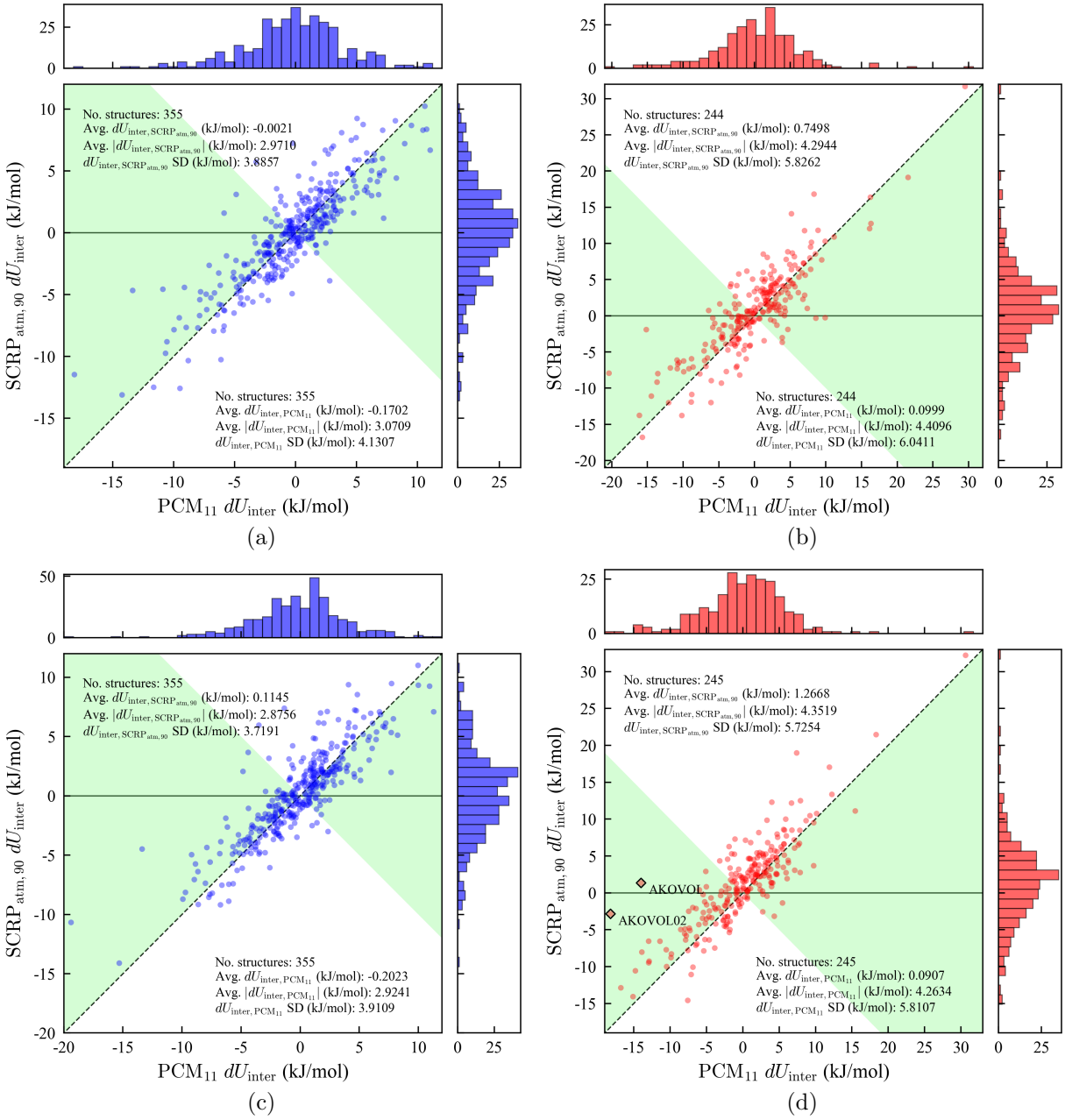


Figure M.3: CE–755 TRN set dU_{inter} for PCM_{11} and post-CSORM $SCRP_{atm,90}$: (a) NHB crystals with Scheme 5 FF, (b) HB crystals with Scheme 5 FF, (c) NHB crystals with Scheme 7 FF, and (d) HB crystals with Scheme 7 FF. In (b), OCAPUE is omitted from the HB statistics, due to its fusion with $SCRP_{atm,90}$. The outlier structures from Figures 8.7 and 8.8 (CYCYPR, KUXPUP01, CLTRZL, and GEYSEI) have been omitted from these plots. Green areas denote where the y-axis model has a lower $|dU_{inter}|$. Histograms of each model's errors are shown on the corresponding axes.

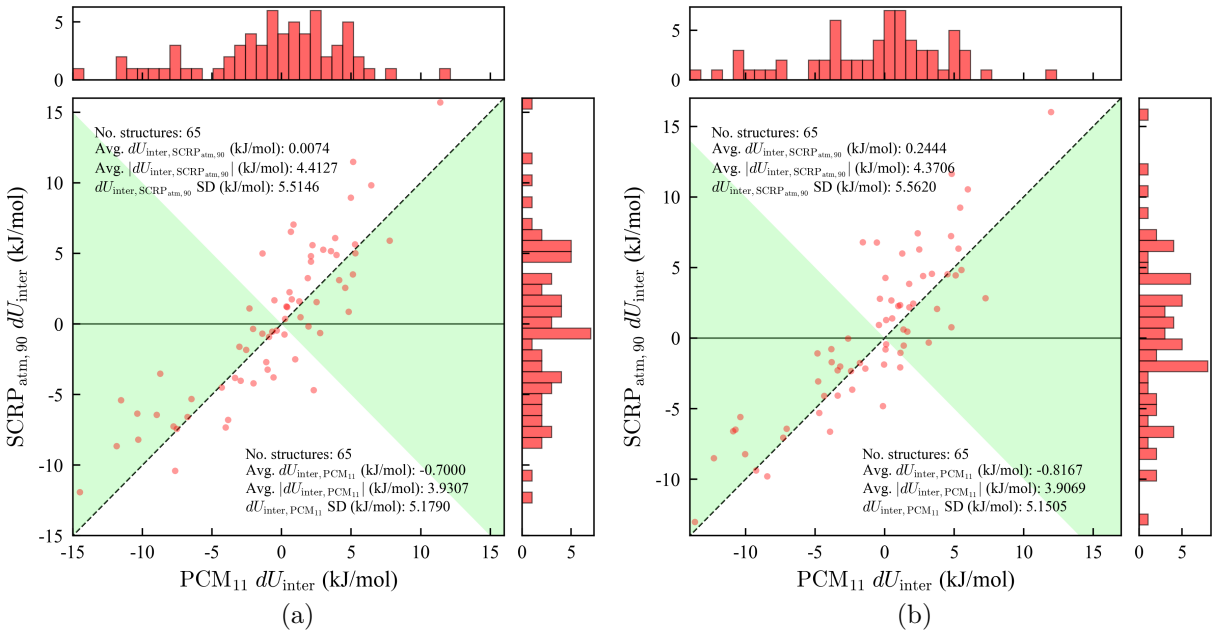


Figure M.4: CE–755 TST set dU_{inter} for PCM_{11} and post-CSORM $\text{SCRP}_{\text{atm,90}}$: (a) HB crystals with Scheme 5 FF, and (b) HB crystals with Scheme 7 FF. The outlier structures from Figures 8.9d and 8.10d (EREFEP) have been omitted from these plots. Green areas denote where the y-axis model has a lower $|dU_{\text{inter}}|$. Histograms of each model's errors are shown on the corresponding axes.

Appendix N

CSP Case Studies: Supplementary Information

N.1 2,6-Diamino-4(3H)-Pyrimidinone Monohydrate CSP

N.1.1 Molecular Definition & Global Search LAM Construction

In the CSP study of 2,6-Diamino-4(3H)-Pyrimidinone monohydrate (MH-DAP), independent gas-phase minimizations of the water and DAP molecules are done at the PBE0/6-311G(d,p) level of theory. To enhance the description of intermolecular hydrogen bonding in the crystal, all QM calculations are conducted in the presence of PCM, with $\varepsilon_{\mathrm{PCM}}$ = 11. Using Gaussian (G09-d01), the *in vacuo* conformations and associated conformational energies are found from the geometry optimization (Table N.1), and atomic-charges are determined at the minimized conformation by Gaussian's HLYGAt method.

Table N.1: Gas-phase conformational energies for the MH-DAP system. All computations are done at the PBE0/6-311G(d,p) level of theory, with $\varepsilon_{\mathrm{PCM}}$ = 11, in G09-d01.

Molecule	$\boldsymbol{U}_{\mathbf{intra}}$ **(Hartrees)**
Water	−76.3688650310
DAP	−449.942747051

Since MH-DAP is approximated as rigid, no additional QM computations are necessary to construct LAM models for the Global Search.

N.1.2 Local Minimization of Experimental Forms: Global Search Stage

To select an appropriate HAIEFF FF to employ during the Global Search, local minimizations of the MH-DAP experimental structures are conducted using CrystalPredictorv2.5.2.1. The *in vacuo* conformations are used to represent the molecules in the crystal, and their associated HLYGAt-charges are employed in the electrostatic description. This is paired with three different FFs to conduct local minimizations, initialized from the experimental reference structures. The results of this are summarized in Table N.2.

For all three FFs, reasonable matches to MH-I and MH-III are achieved. However, neither the FIT FF, nor the $\mathrm{DB14}_{\mathrm{pc}}$ FF yield good geometry predictions for MH-II; they are tenuous matches at best. The ($\mathrm{H_p}$-modified) Scheme 5 dBHP FF parametrized in this work (denoted

Table N.2: Local minimization of the MH-DAP experimental structures with the Global Search energy model, using CrystalPredictorv2.5.2.1.

FF	**Metric**	**MH-I (SEYDIJ)**	**MH-II (SEYDIJ01)**	**MH-III (SEYDIJ03)**
FIT[a]	$RMSD_{20}$ (Å)	0.3198	1.2154	0.4027
	U_{latt} (kJ/mol)	−100.5186	−101.5433	−102.3649
$DB14_{pc}$[b]	$RMSD_{20}$ (Å)	0.2692	0.9758	0.5337
	U_{latt} (kJ/mol)	−96.5703	−99.8220	−96.8947
$BT^{PCM}_{dBhp,pc}$[c]	$RMSD_{20}$ (Å)	0.2814	0.3036	0.5302
	U_{latt} (kJ/mol)	−109.5971	−107.2974	−105.8818

[a] This corresponds to the H_p variant of FIT [166], wherein *both* H_O and H_N are described with the H_N FF parameters.

[b] This FF is tabulated in Table G.14 of Ref [2], and uses H_p to describe *both* H_O and H_N.

[c] This corresponds to the FF tabulated in Table J.11 of this work. For this system, I have used the H_N FF parameters to describe *both* H_O and H_N, similar to FIT.

'$BT^{PCM}_{dBhp,pc}$') is the only FF that gives satisfactory geometry predictions of MH-II. In terms of relative rankings, only the $DB14_{pc}$ FF correctly predicts MH-II as the most stable form, but this is a slightly contrived interpretation of the results, since the geometry of the optimized structure barely corresponds to MH-II.

At the Global Search stage, I am less concerned about relative rankings, since these can still be corrected in the Refinement stages. However, if the predicted geometries are too imprecise at this juncture, there is a risk that the refinement model cannot recover a good match to the reference structure. To avoid this, it is most prudent to prioritize geometry accuracy, hence I select the (H_p-modified) $BT^{PCM}_{dBhp,pc}$ FF. With this FF, the geometric matches to the three forms are all within a tolerable energy window (~4 kJ/mol between the forms), such that there is a low chance that one form will be significantly higher than the rest on the landscape.

N.1.3 Global Search

With the fully specified Global Search energy model, I progress to the main CSP workflow. The details of the Global Search are summarized in Table N.3, and the resulting Global Search landscape is presented in Figure N.1.

Within a 10 kJ/mol energy window from the GM structure, 890 unique clusters are identified from the search. This is a relatively small number of clusters compared to what is typically expected from a Global Search (several thousand unique minima), but this is likely an artefact of the rigid-molecule approximation limiting the search space. Critically, amongst these clusters, multiple matches to each experimental form are found, with $RMSD_{20} \leq 0.60$ Å. On top of the minima identified from the preceding local minimizations, additional matches are inadvertently found as a result of the multistart search protocol. For the purpose of providing starting points to the CSP refinement, the acquired Global Search landscape is adequate.

Table N.3: MH-DAP Global Search, conducted with CrystalPredictorv2.5.2.1 and G09-d01.

Energy Model	
QM Model[a]	PBE0/6-311G(d,p), PCM with $\varepsilon_{\mathrm{PCM}} = 11$
Electrostatic Model	HLYGAt Atomic-charges
HAIEFF FF	$\mathrm{BT}^{\mathrm{PCM}}_{\mathrm{dBhp,pc}}$ FF (Table J.11), modified with $\mathrm{H_O} = \mathrm{H_N}$
Candidate Generation Validity Range	
Max. $\Delta U_{\mathrm{intra},0}$ (kJ/mol)[b]	N/A for rigid system
Max. $U_{\mathrm{inter},0}$ (kJ/mol)[b]	+50.0
Max. $\Delta U^{\mathrm{GM}}_{\mathrm{latt}}$ (kJ/mol)[c]	+10.0
Search Statistics	
Minimizations Attempted	1,000,000
Stable Crystals Accepted	492,150
Landscape Analysis Statistics	
Stable Crystals Analyzed	136,599 (only +10.00 kJ/mol window from GM considered)
Unique Minima Identified[d]	52,008
Minima Written to .res	52,008
Clustering Statistics	
Res Available	52,008
Res Clustered	52,008
Unique Clusters Identified	890
Energy Window (kJ/mol)	$[-110.0161, -100.0062]$

[a] The same QM settings are used for the intramolecular and electrostatic model.

[b] Initial crystal structures must satisfy these criteria to qualify for U_{latt} minimization.

[c] Stable minima are only accepted if they are within this energy window from the incumbent GM.

[d] Minima with *identical* geometry and energy are clustered. This is a stringent form of clustering.

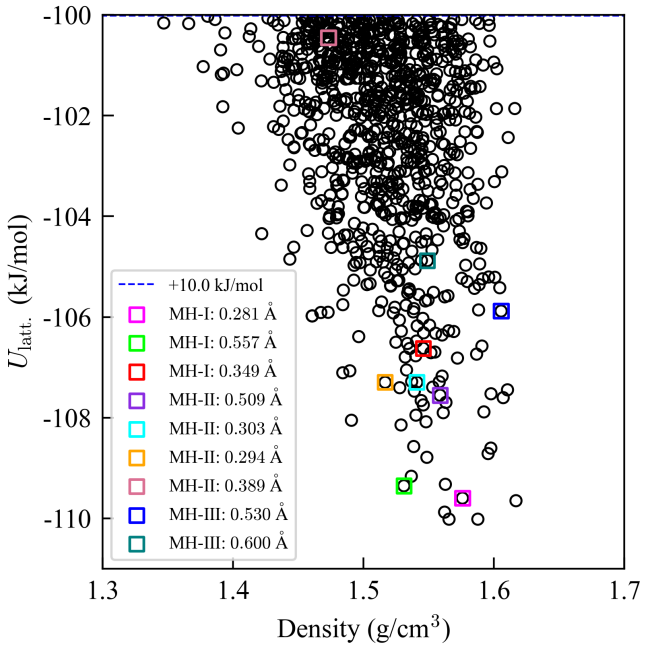
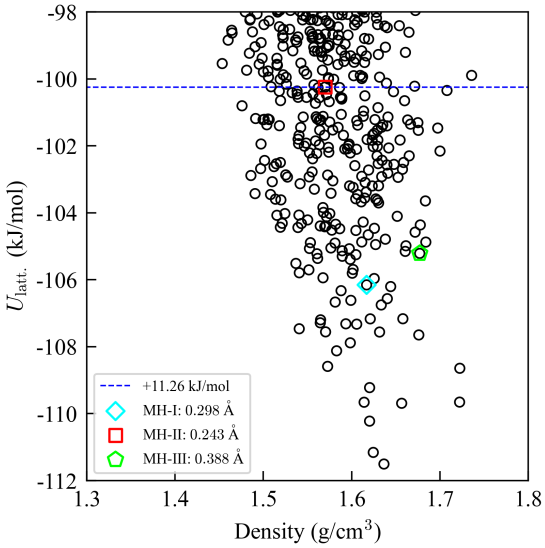


Figure N.1: Global Search landscape for MH-DAP. Matches to MH-I, MH-II, and MH-III are indicated with □, and labeled with their RMSD_{20} scores.

N.1.4 Refinement Landscapes

The 890 structures from the Global Search stage are refined using the models in Table 9.2, and their refined landscapes (post-clustering) are presented below. The lowest U_{latt} match found for each form is shown, with RMSD_{20} scores in Å and $\Delta U_{\text{latt}}^{\text{GM}}$ in kJ/mol.

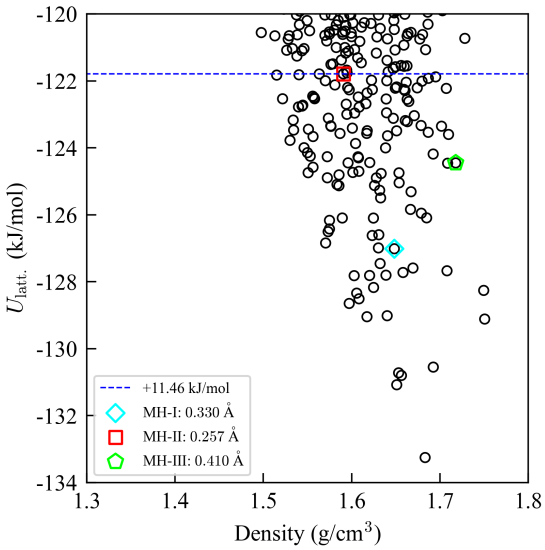
N.1.4.1 MH-DAP Refinement: Static + FIT



Form	RMSD_{20} (Å)	$\Delta U_{\text{latt}}^{\text{GM}}$ (kJ/mol)	Rank
MH-I	0.2983	+5.3510	28
MH-II	0.2433	+11.2608	215
MH-III	0.3883	+6.2919	41

Figure N.2: Refined landscape for Static + FIT. Crystals post-clustering: 773.

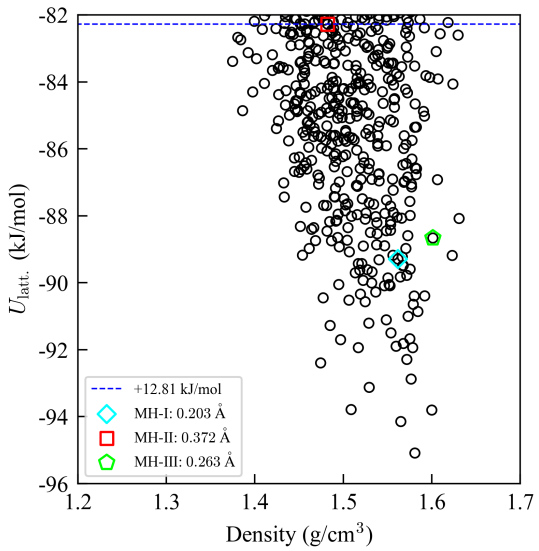
N.1.4.2 MH-DAP Refinement: PCM_{11} + FIT



Form	RMSD_{20} (Å)	$\Delta U_{\text{latt}}^{\text{GM}}$ (kJ/mol)	Rank
MH-I	0.3303	+6.2360	21
MH-II	0.2565	+11.4598	123
MH-III	0.4104	+8.8004	51

Figure N.3: Refined landscape for PCM_{11} + FIT. Crystals post-clustering: 772.

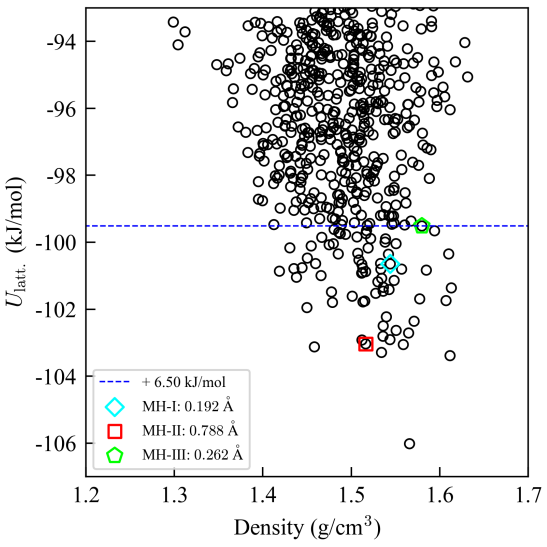
N.1.4.3 MH-DAP Refinement: Static + DB1$_{mp}$



Form	RMSD$_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
MH-I	0.2031	+5.7944	48
MH-II	0.3717	+12.8128	391
MH-III	0.2631	+6.4232	63

Figure N.4: Refined landscape for Static + DB1$_{mp}$. Crystals post-clustering: 736.

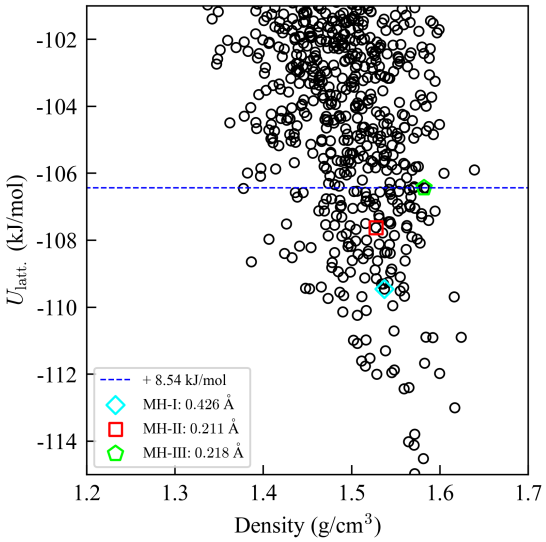
N.1.4.4 MH-DAP Refinement: PCM$_{11}$ + DB14$_{mp}$



Form	RMSD$_{20}$ (Å)	ΔU_{latt}^{GM} (kJ/mol)	Rank
MH-I	0.1918	+5.3714	44
MH-II	0.7881	+2.9833	6
MH-III	0.2615	+6.5040	71

Figure N.5: Refined landscape for PCM$_{11}$ + DB14$_{mp}$. Crystals post-clustering: 762.

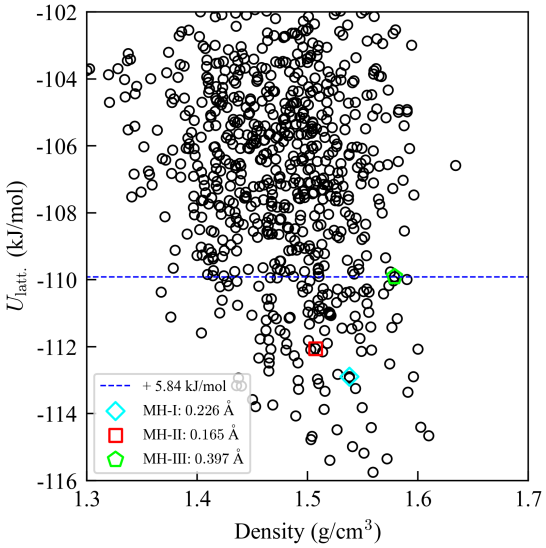
N.1.4.5 MH-DAP Refinement: Static + $BT^{Static}_{dBhp,mp}$



Form	$RMSD_{20}$ (Å)	ΔU^{GM}_{latt} (kJ/mol)	Rank
MH-I	0.4259	+5.5156	36
MH-II	0.2109	+7.3501	90
MH-III	0.2176	+8.5381	148

Figure N.6: Refined landscape for Static + $BT^{Static}_{dBhp,mp}$. Crystals post-clustering: 812.

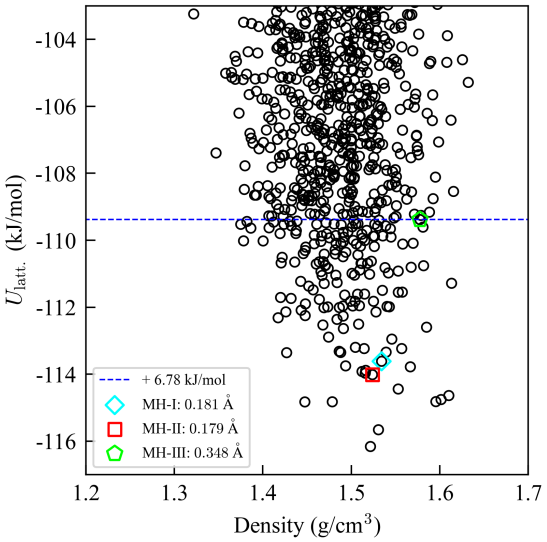
N.1.4.6 MH-DAP Refinement: PCM_{11} + $BT^{PCM}_{dBhp,mp}$



Form	$RMSD_{20}$ (Å)	ΔU^{GM}_{latt} (kJ/mol)	Rank
MH-I	0.2260	+2.8547	30
MH-II	0.1652	+3.6962	46
MH-III	0.3968	+5.8401	122

Figure N.7: Refined landscape for PCM_{11} + $BT^{PCM}_{dBhp,mp}$. Crystals post-clustering: 770.

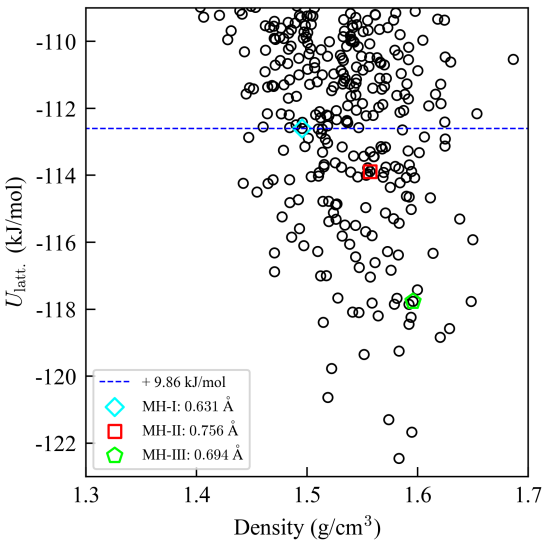
N.1.4.7 MH-DAP Refinement: SCRP + $BT^{SCRP}_{dBhp,mp}$



Form	$RMSD_{20}$ (Å)	ΔU^{GM}_{latt} (kJ/mol)	Rank
MH-I	0.1805	+2.5458	15
MH-II	0.1792	+2.1470	9
MH-III	0.3485	+6.7779	152

Figure N.8: Refined landscape for SCRP + $BT^{SCRP}_{dBhp,mp}$. Crystals post-clustering: 828.

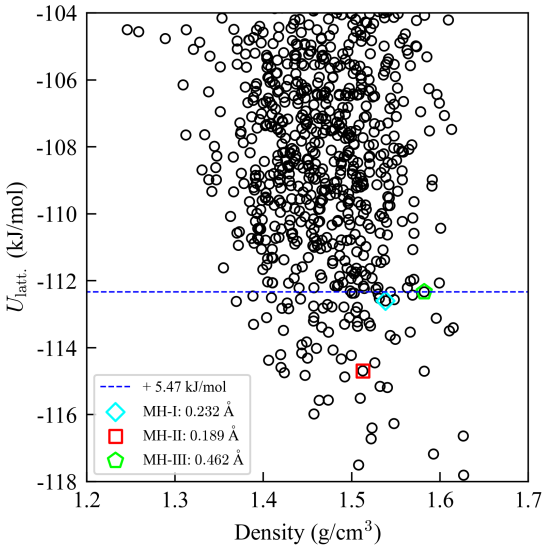
N.1.4.8 MH-DAP Refinement: Static + $BT^{Static}_{Mie,mp}$



Form	$RMSD_{20}$ (Å)	ΔU^{GM}_{latt} (kJ/mol)	Rank
MH-I	0.6309	+9.8553	120
MH-II	0.7564	+8.5728	85
MH-III	0.6942	+4.6908	20

Figure N.9: Refined landscape for Static + $BT^{Static}_{Mie,mp}$. Crystals post-clustering: 793.

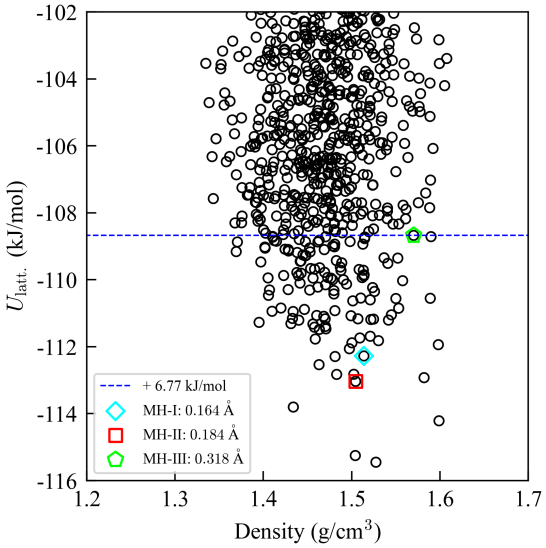
N.1.4.9 MH-DAP Refinement: PCM_{11} + $BT^{PCM}_{Mie,mp}$



Form	$RMSD_{20}$ (Å)	ΔU^{GM}_{latt} (kJ/mol)	Rank
MH-I	0.2323	+5.2061	62
MH-II	0.1886	+3.1166	18
MH-III	0.4620	+5.4747	73

Figure N.10: Refined landscape for PCM_{11} + $BT^{PCM}_{Mie,mp}$. Crystals post-clustering: 801.

N.1.4.10 MH-DAP Refinement: SCRP + $BT^{SCRP}_{Mie,mp}$



Form	$RMSD_{20}$ (Å)	ΔU^{GM}_{latt} (kJ/mol)	Rank
MH-I	0.1638	+3.1741	11
MH-II	0.1836	+2.4149	5
MH-III	0.3183	+6.7747	121

Figure N.11: Refined landscape for SCRP + $BT^{SCRP}_{Mie,mp}$. Crystals post-clustering: 836.

N.1.5 MH-DAP Refinement: CSO–RM Convergence Behavior

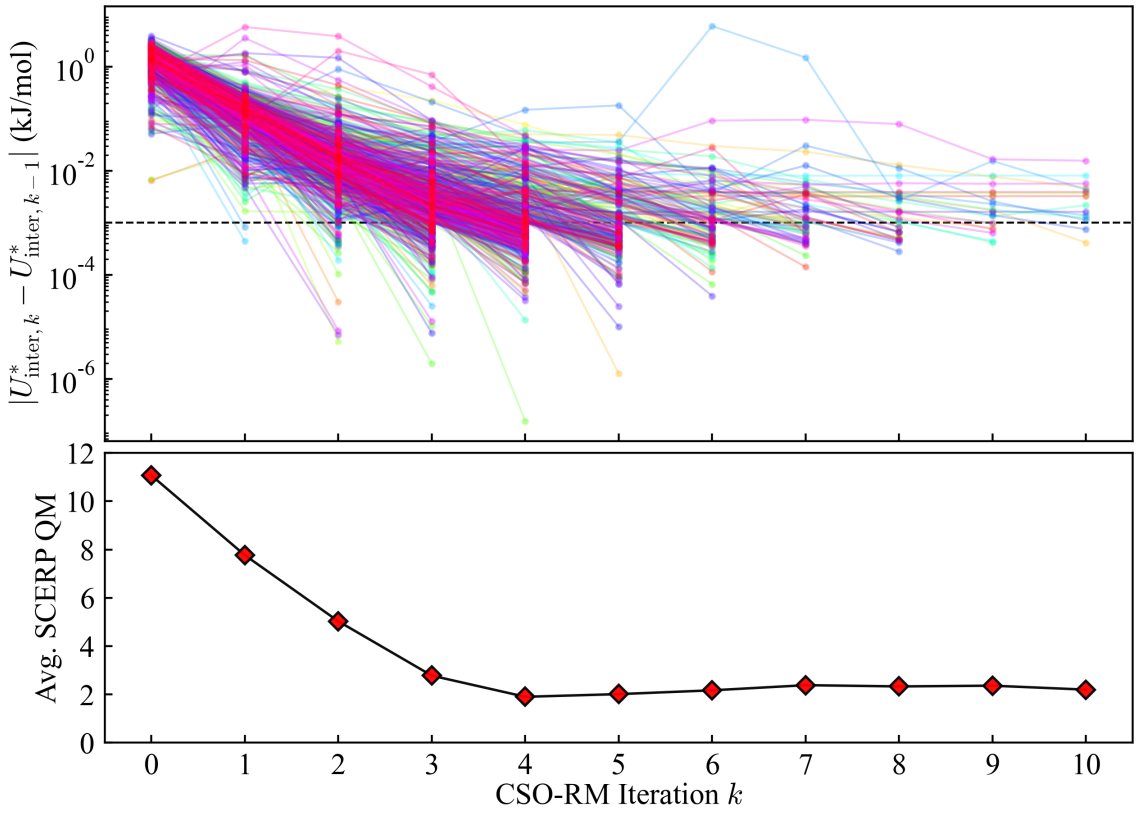


Figure N.12: Top: convergence of the SCERP-embedded CSO–RM minimization ($\varepsilon_{\text{scrpRM}} = 1 \times 10^{-3}$ kJ/mol (dashed line) and $k_{\max} = 10$) for the SCRP + $\text{BT}^{\text{SCRP}}_{\text{Mie,mp}}$ model. Bottom: the average number of QM computations (normalized by $Z' = 2$) to converge the SCERP process at iteration k.

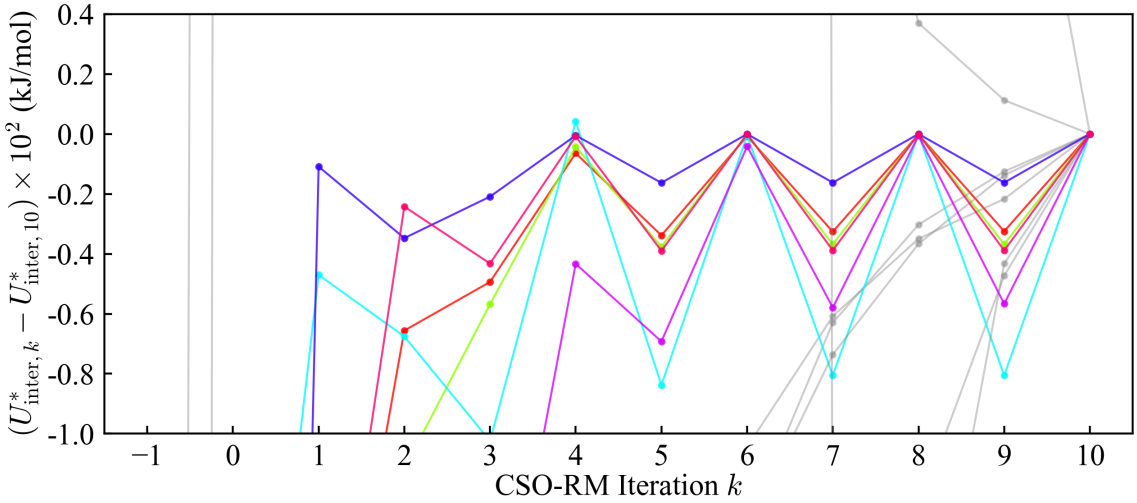
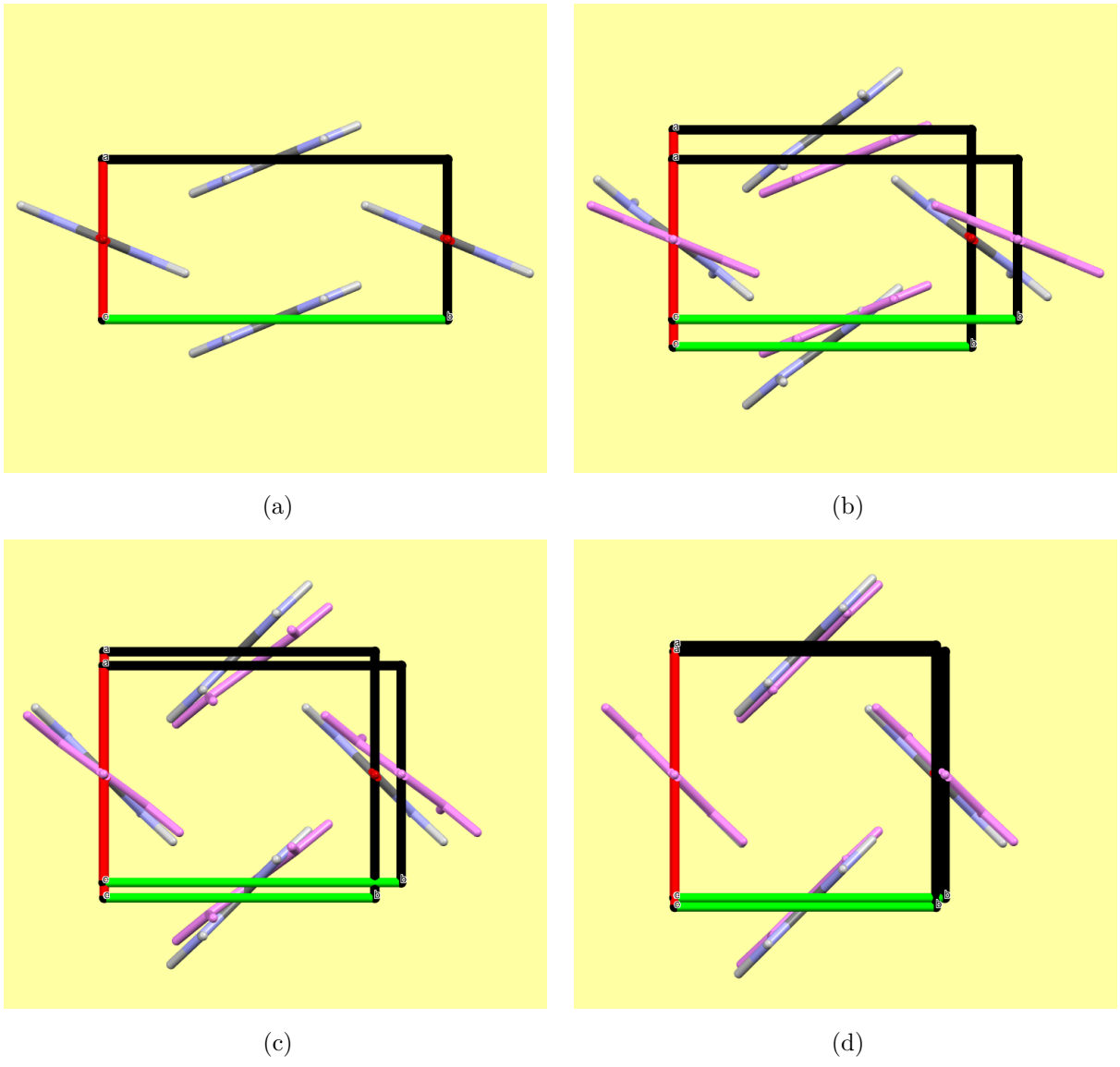


Figure N.13: The difference in $U^*_{\text{inter},k}$ relative to $U^*_{\text{inter},10}$ for structures that failed to converge in Figure N.12, using the SCRP + $\text{BT}^{\text{SCRP}}_{\text{Mie,mp}}$ model. Structures with oscillatory non-convergence are shown in color, while other failures are shown in gray.

N.2 Urea CSP

N.2.1 Local Minimization of Experimental Forms: Periodic DFT-D

Prior to the CSP of urea, the experimental forms (F-I, III, and IV) are locally minimized by periodic DFT-D, to assess their stabilities at static conditions. F-IV is a high pressure (> 7 GPa [28]) form of urea, which is closely related in geometry to F-I [381]. Without the application of pressure during the periodic DFT-D computation, F-IV's distorted crystal structure is found to be unstable, and it instead interconverts to F-I (Figure N.14).



(a) (b)

(c) (d)

Figure N.14: F-IV's crystal structure over the periodic DFT-D minimization. (a) Initial structure, UREAXX26. (b) UREAXX26 (pink) overlaid with its counterpart after one round of periodic DFT-D optimization. (c) The first periodic DFT-D optimization structure (pink) overlaid with the final converged DFT-D structure. (d) The converged DFT-D structure overlaid with F-I, UREAXX25.

N.2.2 Molecular Definition & Global Search LAM Construction

Gas-phase minimization of the urea molecule is done at the PBE0/6-311G(d,p) level of theory, with and without PCM, using Gaussian (G09-d01). *In vacuo* conformations and conformational energies are found from the geometry optimization (Table N.4); atomic-charges are also fit at the minimized geometry by Gaussian's HLYGAt method. To enhance the description of intermolecular hydrogen bonding in the crystal, the PCM-embedded model is used during the Global Search stage. However, the ground-state conformational energy is defined with respect to the 'Static' environment, i.e., ΔU_{intra} are computed as $U_{\text{intra}} - U_{\text{intra}}^{\text{Static}}$

Table N.4: Gas-phase conformational energies for urea. All computations are done at the PBE0/6-311G(d,p) level of theory in G09-d01.

Environment	$\boldsymbol{U_{\text{intra}}}$ **(Hartrees)**
Static	−225.085673522
PCM (ε_{PCM} = 11)	−225.097652059

Urea is treated flexibly throughout the CSP study, to facilitate a better description of the crystalline hydrogen bonding network. For the Global Search, a LAM database is constructed by the adaptive LAM algorithm [98], with the flexible molecule definition summarized in Table N.5. Briefly, Dih4 and 7 describe the rotation of each $-NH_2$ with respect to the C=O axis; this involves the rotation of one hydrogen per $-NH_2$. The position of the second $-NH_2$ hydrogen is dependent on the 'leading' hydrogen, as reflected by the improper torsions, Dih5 and 8. However, if Dih5 and 8 are treated as dependent dofs, there exist multiple local minimum positions for the secondary hydrogen (relative to the leading hydrogen). For instance, depending on where H_2 is initialized from, the gas-phase optimization may find a locally minimum value of Dih5 at either 109° or −109° (assuming a tetragonal hybridization of the nitrogen). In short, the values of Dih5 and 8 are not uniquely defined by the values of Dih4 and 7 (respectively). This 'flip-flopping' behavior can lead to spurious errors in the surrogate model when the LAMs are smoothed, and as such it is critical to treat the improper torsions appropriately as flexible dofs. The chosen flexible ranges can be rationalized by chemical intuition.

The adaptive LAM algorithm is ran with the settings shown in Table N.6, and converged in three rounds of LAM generation: one round of uniform LAM placement, and two rounds of

Table N.5: Flexible definition of the urea molecule used in Global Search LAM construction. Site labels correspond to the labeling shown in Figure 9.3.

Dof	**Participating Sites**	**In Vacuo (°)**	**Lower Bound (°)**	**Upper Bound (°)**
Dih4	H_1−N_1−C_1−O_1	−12.9	−180.0	180.0
Dih5	H_2−N_1−C_1−H_1	223.1	90.0	270.0
Dih7	H_3−N_2−C_1−O_1	−12.9	−180.0	180.0
Dih8	H_4−N_2−C_1−H_3	−126.9	90.0	270.0

Table N.6: Adaptive LAM generation settings applied in the Global Search LAMs for urea.

Setting	Value
Gaussian Version	G16-c01-avx2
Reference U_{intra} [a]	−225.085673522 Hartree.
LAM QM Level of Theory	PBE0/6-311G(d,p), ε_{PCM} = 11
U_{acc}	5.0 kJ/mol
U_{cut} [b]	+20.0 kJ/mol

[a] This corresponds to the gas-phase conformation in the Static environment.

[b] By defining the reference U_{intra} at the Static environment, the chosen U_{cut} is approximately equivalent to a +40.0 kJ/mol ΔU_{intra} window in the PCM model.

adaptive LAM placement. A total of 2,588 LAMs are generated, and 2D projections of the final landscape are shown in Figure N.15.

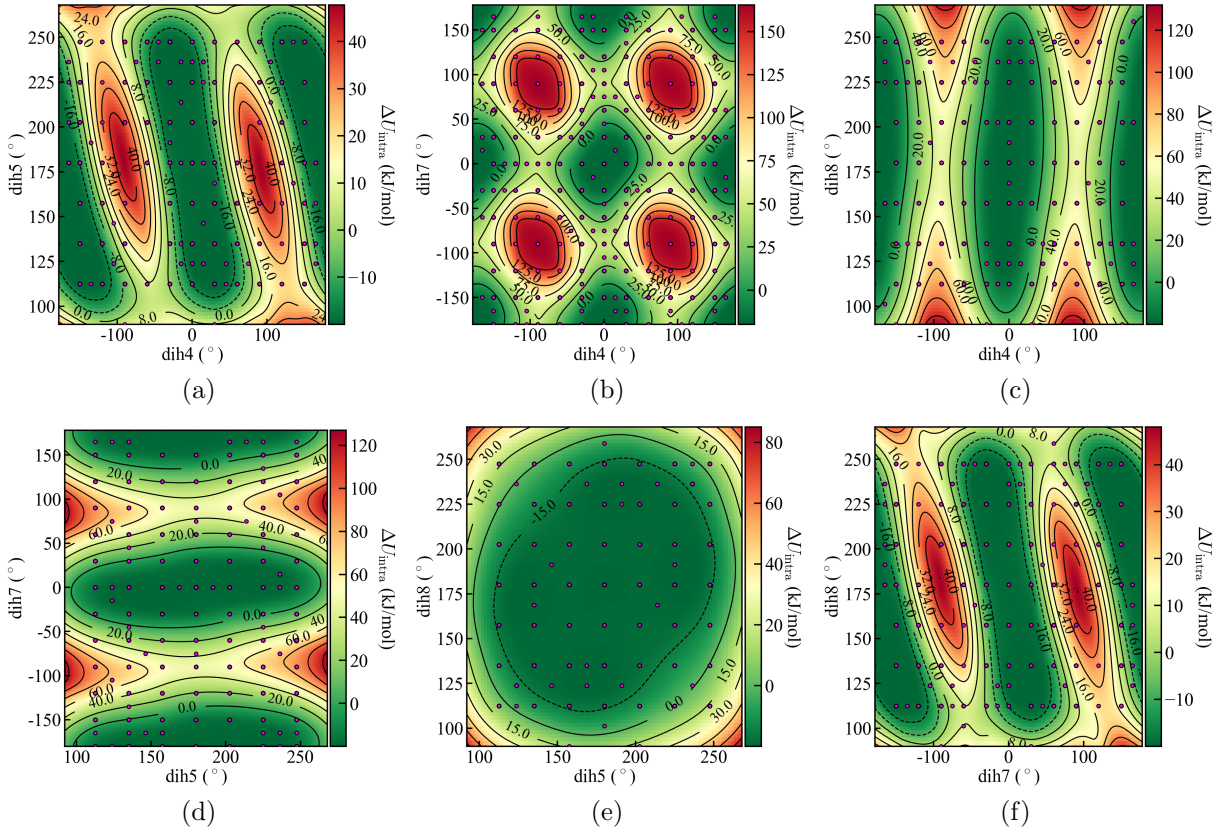


Figure N.15: Converged LAM landscape for the Global Search of urea (2,588 LAMs). When evaluating DihX/DihY, all other dihedrals are kept at the midpoint of their respective search ranges shown in Table N.5. Magenta points show the LAM point placements.

N.2.3 Local Minimization of Experimental Forms: Global Search Stage

Given that urea is a relatively simple molecule, I do not anticipate any major difficulties in obtaining good geometry matches to its experimental forms, as long as a sensible FF is paired

with the rest of the HAIEFF model. For this reason, I choose to use the $DB14_{pc}$ FF (Table G.14 of Ref [2]) during the urea Global Search. From the MH-DAP CSP study, it is known that the DB FFs work reasonably well when paired with PCM. Moreover, since the DB FFs are parametrized according to Scheme 1 of Table 6.2, they inherently favor geometry accuracy in HB structures, like urea. No further local minimization of F-I, III, and/or IV was done using the Global Search energy model to validate the suitability of the $DB14_{pc}$ FF.

N.2.4 Global Search

With the fully specified Global Search energy model, I progress to the main CSP workflow. The details of the Global Search are summarized in Table N.7, and the resulting Global Search landscape is presented in Figure N.16.

From the Global Search, 8,969 unique minima are identified, of which 5,110 were written out to .res files. This subset of 5,110 comprises an energy window of +16.67 kJ/mol from the global minimum structure ($U_{\text{latt}}^{\text{GM}}$ = −131.0509 kJ/mol). Following clustering, only 1,909 unique clusters remain in the +16.67 kJ/mol energy window. Relatively good geometry matches to all three experimental forms are found within +15 kJ/mol of the GM ($RMSD_{20} \leq 0.55$ Å), and the highest ranked experimental match (F-IV: ⬠) is placed 1,079th with $\Delta U_{\text{latt}}^{\text{GM}}$ = +14.11 kJ/mol.

Table N.7: Urea Global Search, conducted with CrystalPredictorv2.5.2.0 and G16-c01-avx2.

Energy Model	
QM Model[a]	PBE0/6-311G(d,p), PCM with ε_{PCM} = 11
Electrostatic Model	HLYGAt Atomic-charges
HAIEFF FF	$DB14_{pc}$ FF (Table G.14 of Ref [2])
Candidate Generation Validity Range	
Max. $\Delta U_{\text{intra,0}}$ (kJ/mol)[b]	+25.0
Max. $U_{\text{inter,0}}$ (kJ/mol)[b]	+50.0
Max. $\Delta U_{\text{latt}}^{\text{GM}}$ (kJ/mol)[c]	+20.0
Search Statistics	
Minimizations Attempted	500,000
Stable Crystals Accepted	323,128
Landscape Analysis Statistics	
Stable Crystals Analyzed	323,128
Unique Minima Identified[d]	8,969
Minima Written to .res	5,110 (only +16.67 kJ/mol window from GM considered)
Clustering Statistics	
Res Available	5,110
Res Clustered	5,110
Unique Clusters Identified	1,909
Energy Window (kJ/mol)	$[-131.0509, -114.3801]$

[a] The same QM settings are used for the intramolecular and electrostatic model.

[b] Initial crystal structures must satisfy these criteria to qualify for U_{latt} minimization.

[c] Stable minima are only accepted if they are within this energy window from the incumbent GM.

[d] Minima with *identical* geometry and energy are clustered. This is a stringent form of clustering.

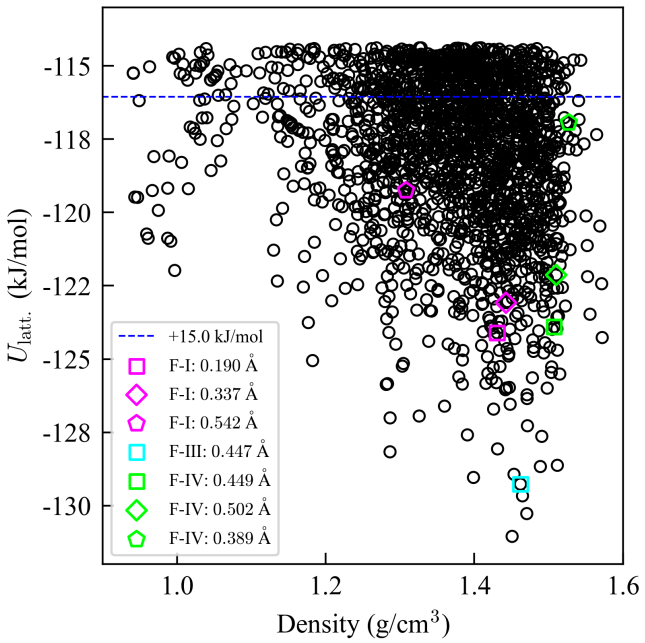


Figure N.16: Global Search landscape for urea. Matches to F-I, F-III, and F-IV are indicated in magenta, cyan, and green respectively, and labeled with their $RMSD_{20}$ scores.

Although not all unique minima were written to .res files following the search, good geometry matches to the experimental forms are already present amongst the available structures. From a practical standpoint, there is also no interest to refine more structures than are already available from clustering. As such, for the purpose of providing starting points to the CSP refinement, the acquired Global Search landscape is adequate.

N.2.5 Refinement Overview

Table N.8: Refinement I of urea, conducted with CSO–FMv1.3.1, G09-d01, and GDMAv2.2. Model numbering corresponds to those shown in Table 9.5.

Setting	Model 1	Model 2	Model 3
Global Search Cutoffs			
Structures Refined	1,342		
$\Delta U_{\text{latt}}^{\text{GM}}$ (kJ/mol)	+15.00		
Intramolecular Model			
Flexible Dofs	Dihedrals: 4, 5, 7, and 8		
Intra. LAM Δ_{tol} Bond Length (Å)	0.05		
Intra. LAM Δ_{tol} Bond Angle (°)	3.50		
Intra. LAM Δ_{tol} Torsion Angle (°)	5.00		
Level of Theory	PBE0/6-311G(d,p)		
PCM?	$\varepsilon_{\text{PCM}} = 11$	Static	$\varepsilon_{\text{PCM}} = 11$
Intermolecular Electrostatic Model			
Elec. LAM Δ_{tol} Bond Length (Å)	0.05		
Elec. LAM Δ_{tol} Bond Angle (°)	3.50		
Elec. LAM Δ_{tol} Torsion Angle (°)	5.00		
Level of Theory	PBE0/6-311G(d,p)		
PCM?	$\varepsilon_{\text{PCM}} = 11$	Static	$\varepsilon_{\text{PCM}} = 11$
Electrostatic Model	Rank 4, GDMA Multipoles ($r_{\text{H}}^{\text{GDMA}} = 0.650$ Å)		
Electrostatic Damping	None		
Intermolecular Force-field Model			
Force-field Label	DB14_{mp}	$\text{BT}_{\text{Mie,mp}}^{\text{Static}}$	$\text{BT}_{\text{Mie,mp}}^{\text{PCM}}$
Force-field Parameters	Ref [2], Table I.14	Table J.20	Table J.27
Refined Landscape Statistics			
No. of Crystals Post-Clustering	659	720	615
$U_{\text{latt}}^{\text{GM}}$ (kJ/mol)[a]	−133.87	−94.92	−139.03
F-I Found?	✓	✓	✓
F-III Found?	✓	✓	✓
F-IV Found?	✓	×	✓

[a] Note, all ΔU_{intra} are defined with respect to the Static gas-phase energy in Table N.4. This explains the large gap in U_{latt} between calculations done with/without PCM in the intramolecular model.

Table N.9: Refinement II of urea, conducted with CSO–FMv1.4.0.2, G09-d01, and GDMAv2.2. These settings correspond to Model 4 in Table 9.5.

Setting	**Model 4**		
Refinement I Cutoffs			
Refinement I Landscape	Model 3 from Table N.8		
Structures Refined	200		
$\Delta U_{\text{latt}}^{\text{GM}}$ (kJ/mol)	+13.41		
Intramolecular Model			
Flexible Dofs	Dihedrals: 4, 5, 7, and 8		
Intra. LAM $\boldsymbol{\Delta}_{\text{tol}}$ Bond Length (Å)	0.05		
Intra. LAM $\boldsymbol{\Delta}_{\text{tol}}$ Bond Angle (°)	3.50		
Intra. LAM $\boldsymbol{\Delta}_{\text{tol}}$ Torsion Angle (°)	5.00		
Level of Theory	PBE0/6-311G(d,p)		
PCM?	Static		
Intermolecular Electrostatic Model			
Elec. LAM $\boldsymbol{\Delta}_{\text{tol}}$ Bond Length (Å)	0.05		
Elec. LAM $\boldsymbol{\Delta}_{\text{tol}}$ Bond Angle (°)	3.50		
Elec. LAM $\boldsymbol{\Delta}_{\text{tol}}$ Torsion Angle (°)	5.00		
Level of Theory	PBE0/6-311G(d,p)		
PCM?	Static		
Electrostatic Model	Rank 4, GDMA Multipoles ($r_{\text{H}}^{\text{GDMA}} = 0.650$ Å)		
Electrostatic Damping	Using $\beta_{\text{atm}}^{tt'}$ based on $\%f_6 = 90\%$ in Table L.1		
Intermolecular SCERP Model			
SCERP-Field Representation	HLYGAt Atomic-Charges		
r_{scrp} (Å)	40.0		
t_{slow}	6		
$\varepsilon_{\text{scrp}}$ (a.u.)	1×10^{-5}		
t_{max}	30		
$\varepsilon_{\text{scrpRM}}$ (kJ/mol)	1×10^{-5}		
k_{max}	18		
Intermolecular Force-field Model			
Force-field Label	$\text{BT}_{\text{Mie,mp}}^{\text{SCRP}}$		
Force-field Parameters	Table M.4		
Refined Landscape Statistics			
CSO–FM Optimality Tolerance[a]	3.0×10^{-3}		
FD Mode (per Table 7.6)	2	3	4
No. of Crystals Post-Clustering	182	186	179
$U_{\text{latt}}^{\text{GM}}$ (kJ/mol)[b]	−99.45	−99.39	−99.61
F-I Found?	✓	✓	✓
F-III Found?	✓	✓	✓
F-IV Found?	×	×	×

[a] This is decided by Equation 7.4.14. Without SCERP, the default τ_{nag} is 1×10^{-8}.

[b] Note, all ΔU_{intra} are defined with respect to the Static gas-phase energy in Table N.4.

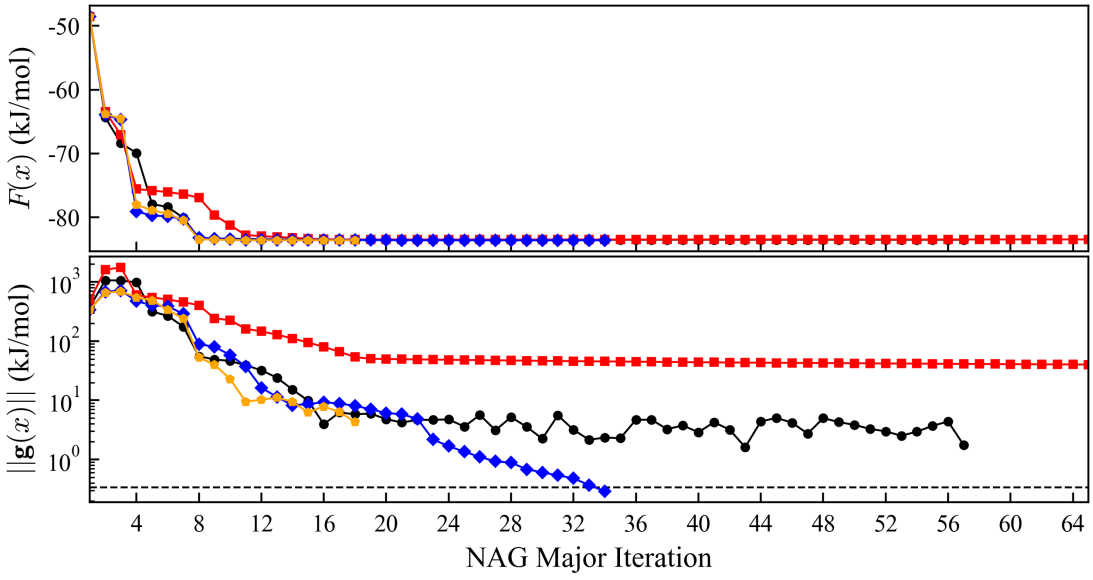
N.2.6 Final Periodic DFT-D Re-ranking

Table N.10: COMPACK comparison between the Model 4 structures before and after local optimization by periodic DFT-D. Comparisons of the optimized structures to the experimental forms of urea are also presented. Blank entries indicate no valid $RMSD_{20}$ score could be generated.

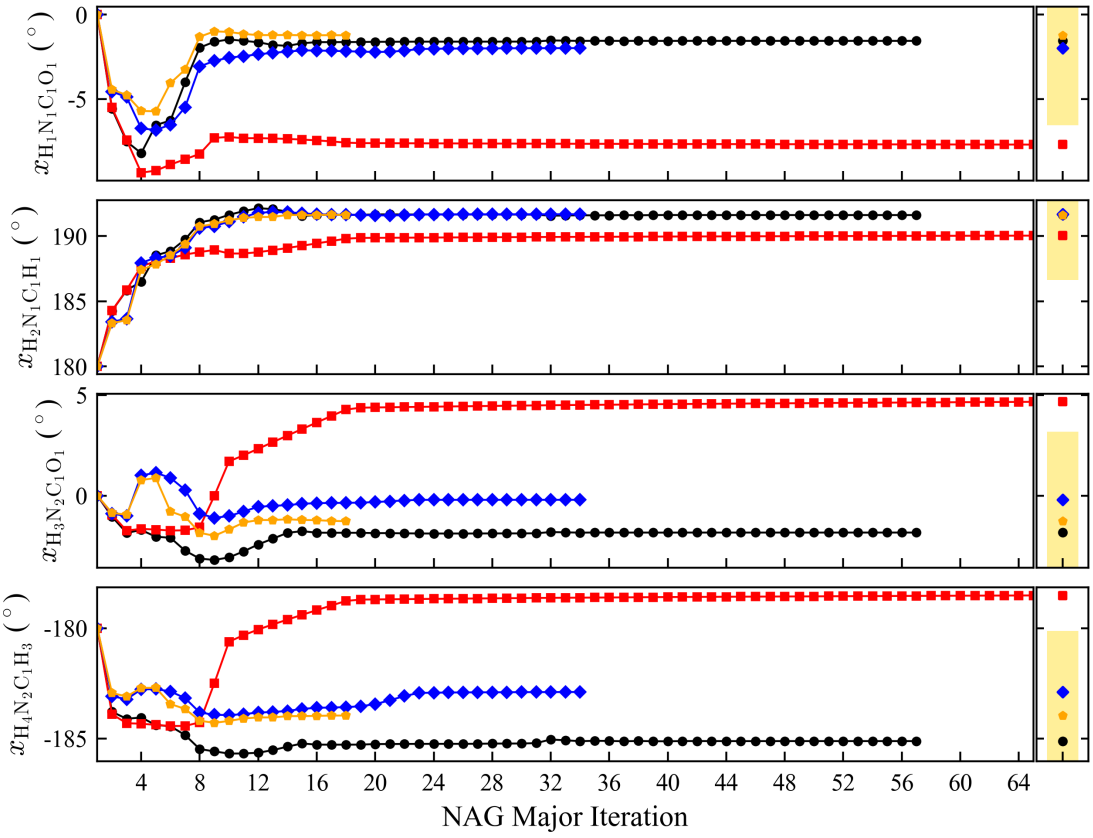
$RMSD_{20}$ (Å)		**Reference Structure**													
		F-I	F-III	F-IV	1	2	3	4	5	6	7	8	9	10	124
Optimised Structure	1				0.257		0.218								
	2					0.176									
	3				0.240		0.199								
	4							0.421	0.330						
	5							0.421	0.331						
	6									0.202					
	7										0.158				
	8		0.219									0.246			
	9												0.166		
	10													0.168	
	124	0.056													0.150

N.2.7 Assessment of SCERP Gradient Approximations

N.2.7.1 Local Minimization of Urea Polymorphs

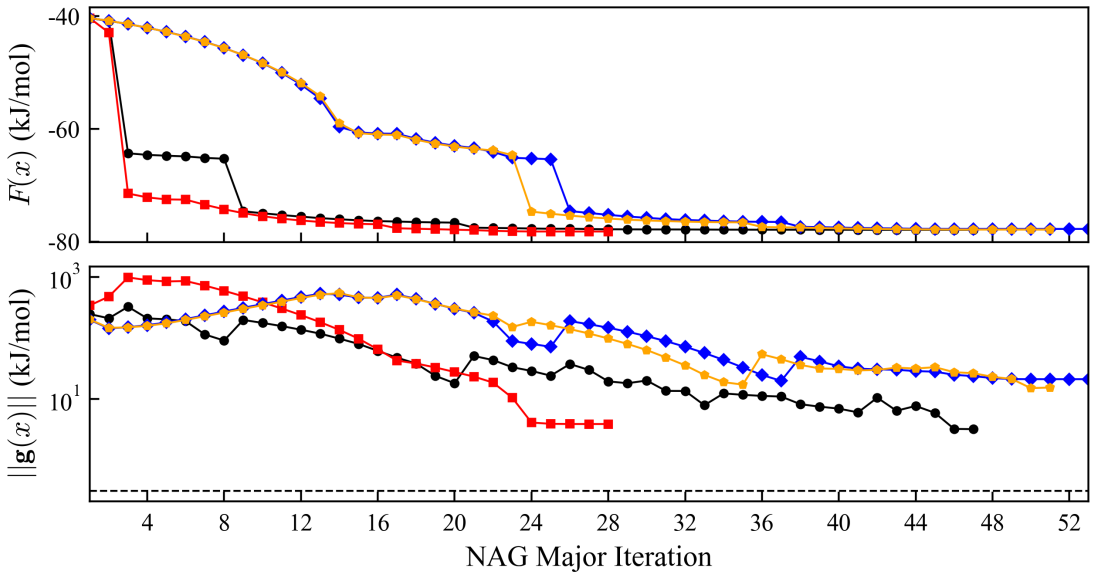


(a)

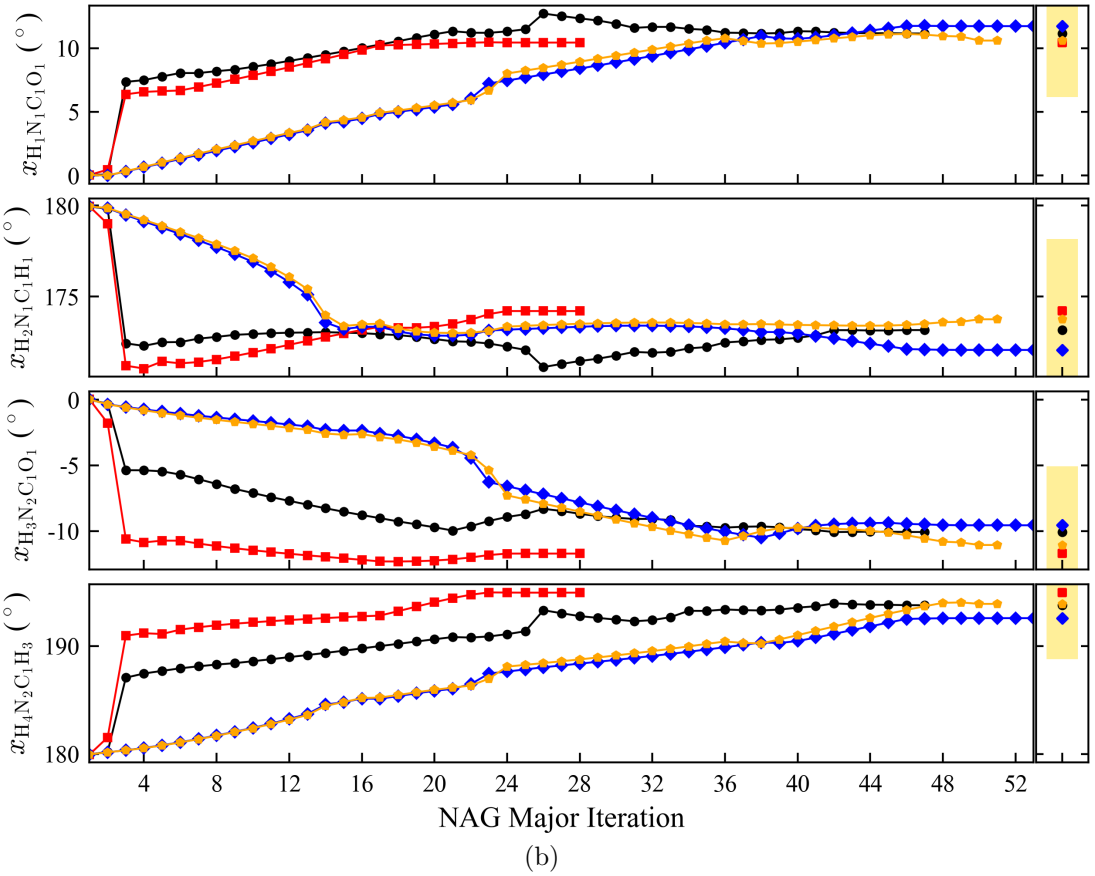


(b)

Figure N.17: Local minimization trajectories of urea F-I using Mode 1 (red), Mode 2 (blue), Mode 3 (orange), and Mode 4 (black) FD gradients: (a) the function value and gradient norm, and (b) the dof values. The yellow region in (b) indicate ±5° from the final Mode 4 dof value.

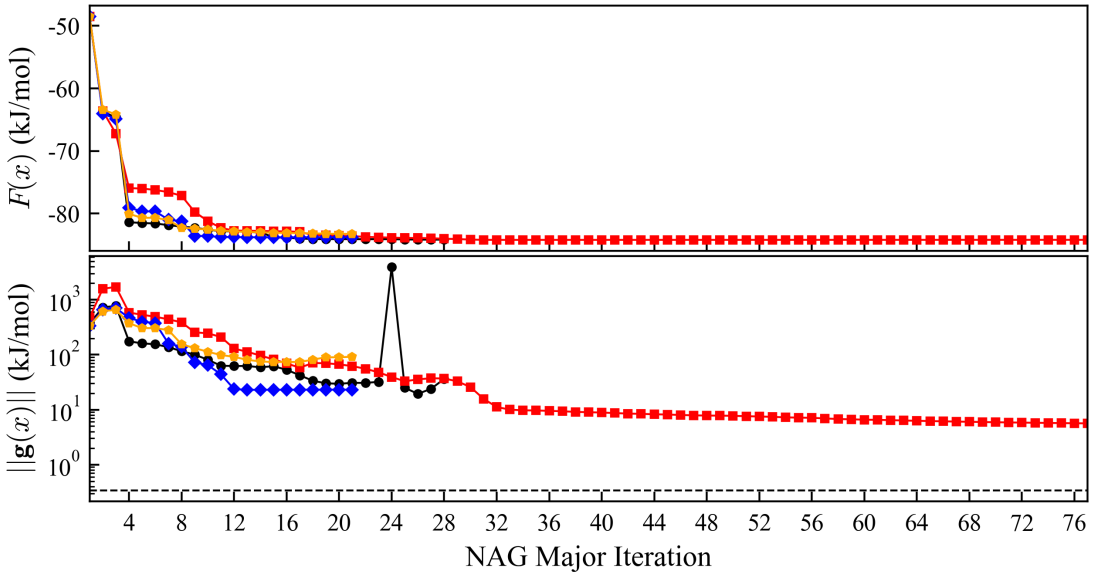


(a)

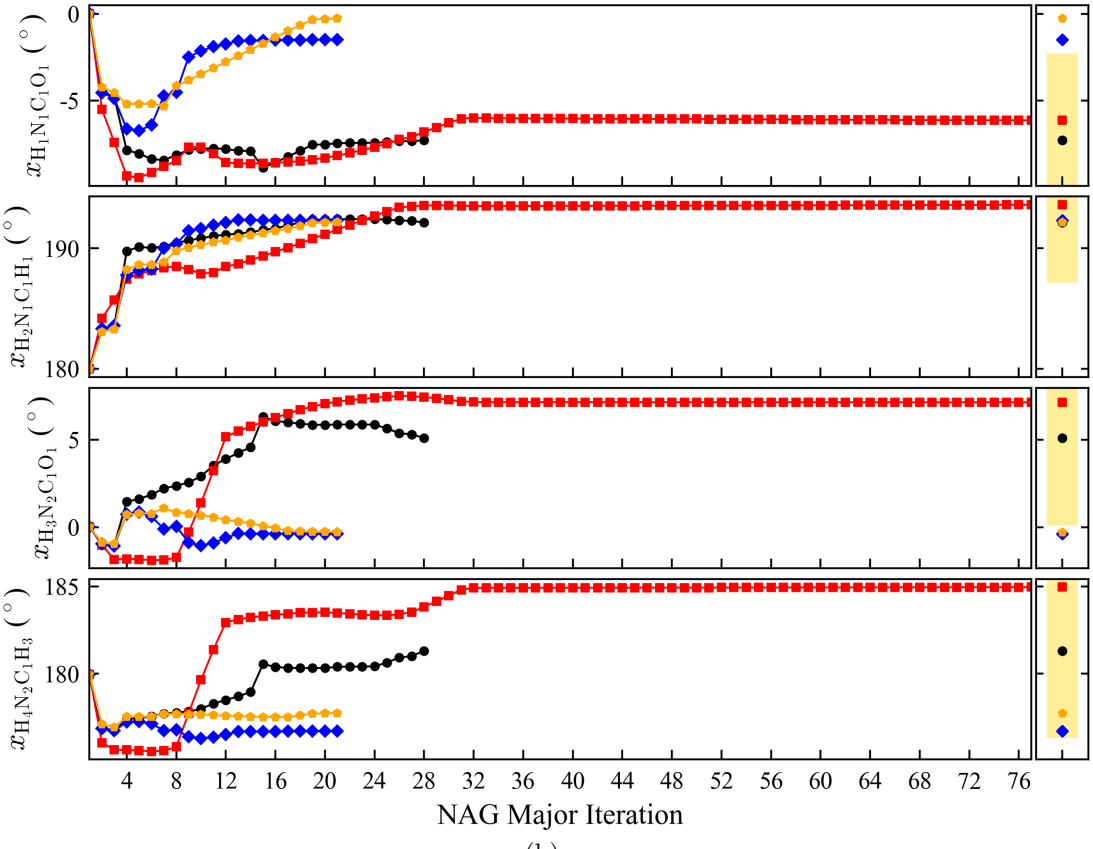


(b)

Figure N.18: Local minimization trajectories of urea F-III using Mode 1 (red), Mode 2 (blue), Mode 3 (orange), and Mode 4 (black) FD gradients: (a) the function value and gradient norm, and (b) the dof values. The yellow region in (b) indicate ±5° from the final Mode 4 dof value.



(a)



(b)

Figure N.19: Local minimization trajectories of urea F-IV using Mode 1 (red), Mode 2 (blue), Mode 3 (orange), and Mode 4 (black) FD gradients: (a) the function value and gradient norm, and (b) the dof values. The yellow region in (b) indicate ±5° from the final Mode 4 dof value.

N.2.7.2 Landscape Minimization by FD Modes 2 to 4

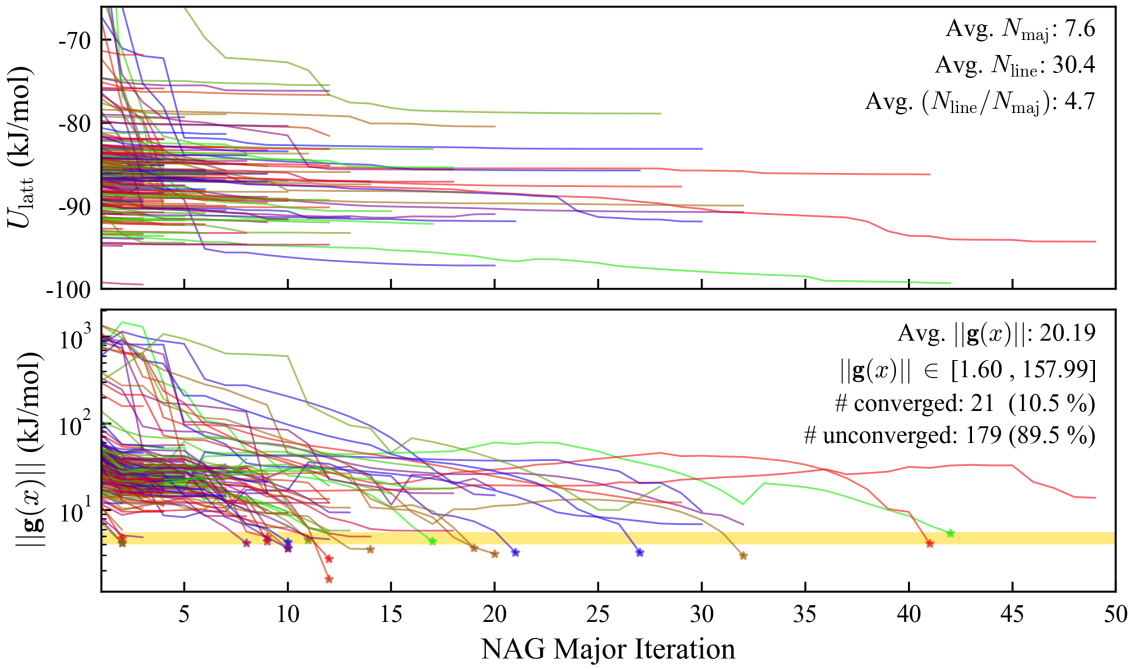


Figure N.20: Minimization trajectories of urea Refinement II by FD Mode 2: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $||\boldsymbol{g}(x)||$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★.

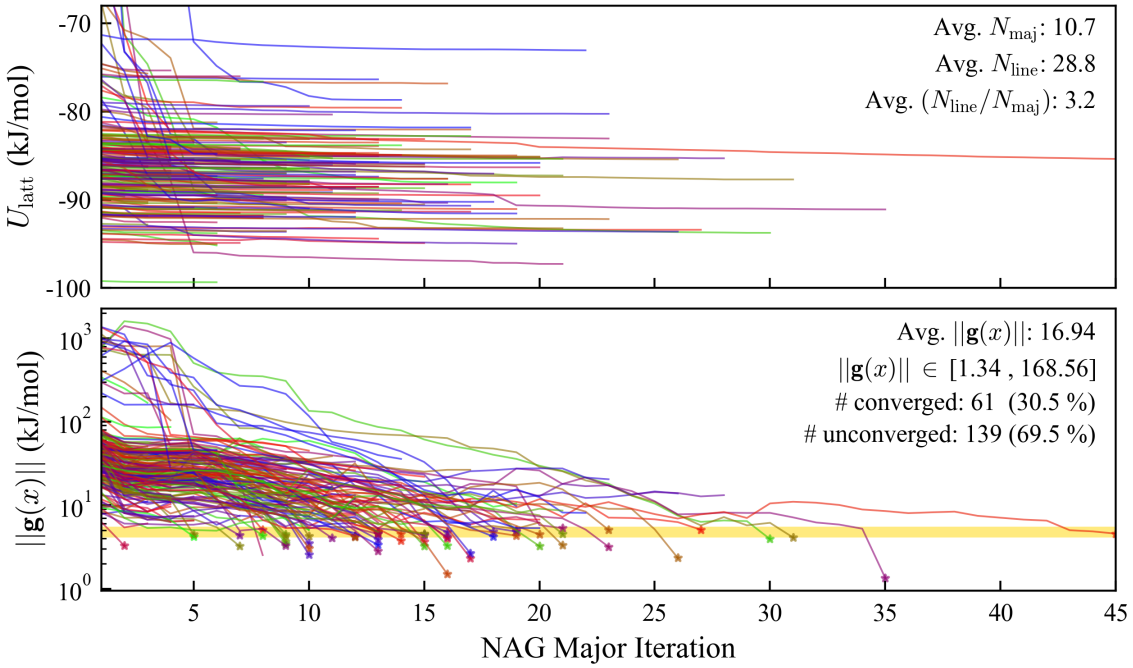


Figure N.21: Minimization trajectories of urea Refinement II by FD Mode 3: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $||\boldsymbol{g}(x)||$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★.

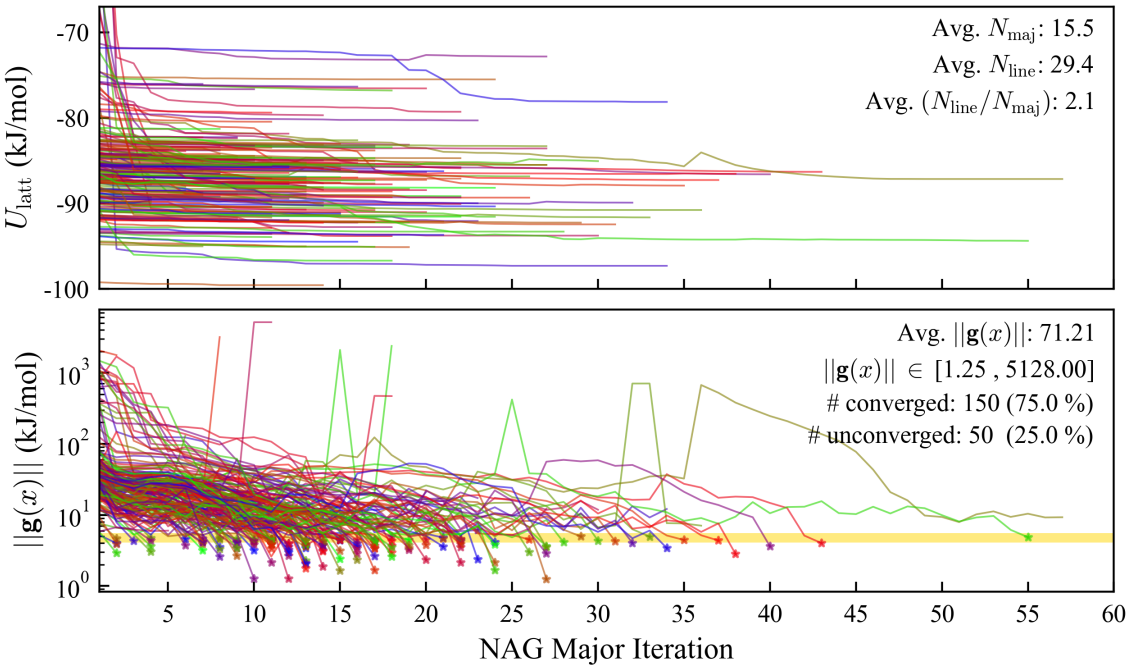


Figure N.22: Minimization trajectories of urea Refinement II by FD Mode 4: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $||\boldsymbol{g}(x)||$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★.

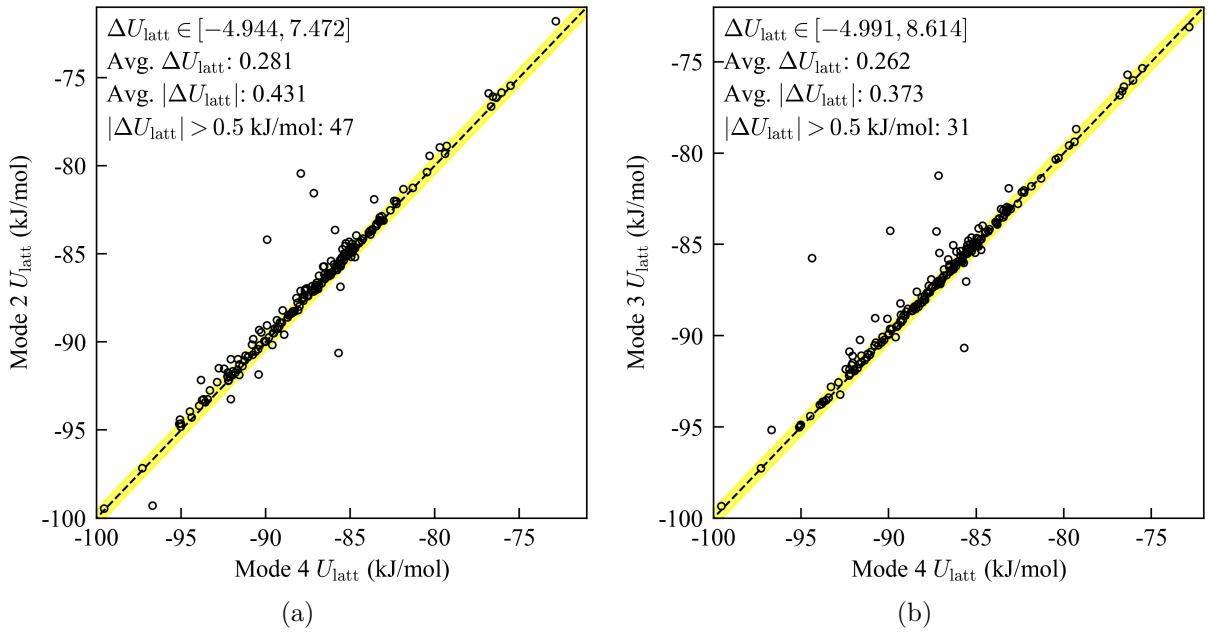


Figure N.23: Comparison of optimized urea U_{latt} obtained using (a) Mode 2 and (b) Mode 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$. All comparisons are on a structure-by-structure basis. The yellow region indicates $|\Delta U_{\text{latt}}| \leq 0.5$ kJ/mol.

N.3 Glycine CSP

N.3.1 Molecular Definition & Global Search LAM Construction

Gas-phase minimization of the glycine molecule is done at the PBE0/6-311G(d,p) level of theory, with different PCM settings, using Gaussian (G16-c01-avx2). *In vacuo* conformations and conformational energies are found from the geometry optimization (Table N.11); atomic-charges are also fit at the minimized geometry by Gaussian's HLYGAt method. Critically, the glycine molecule is able to persist in its zwitterionic state only when $\varepsilon_{\text{PCM}} = 11$ is used, while reverting back to a neutral state otherwise. As such, $\varepsilon_{\text{PCM}} = 11$ is used during the Global Search stage, and the ground-state conformational energy is defined with respect to the same environment, i.e., ΔU_{intra} are computed as $U_{\text{intra}} - U_{\text{intra}}^{\text{PCM}_{11}}$

Table N.11: Gas-phase conformational energies for glycine. All computations are done at the PBE0/6-311G(d,p) level of theory in G16-c01-avx2.

Environment	$\boldsymbol{U_{\text{intra}}}$ **(Hartrees)**	**Final State**
Static	−284.200301413	Neutral
PCM ($\varepsilon_{\text{PCM}} = 3$)	−284.208495691	Neutral
PCM ($\varepsilon_{\text{PCM}} = 7$)	−284.211958245	Neutral
PCM ($\varepsilon_{\text{PCM}} = 11$)	−284.202843345	Zwitterionic

Glycine is treated flexibly throughout the CSP study, particularly to facilitate a better description of the crystalline hydrogen bonding. For the Global Search, a LAM database is constructed by the adaptive LAM algorithm [98], with the flexible molecule definition summarized in Table N.12. Dih5 and 8 correspond to the rotation of the carboxylate and amino group respectively, relative to the amino acid backbone. By allowing these groups to rotate, greater variation in hydrogen bonding can be explored during the Global Search. The flexible ranges have been chosen to simply allow full rotation of both groups. In principle, symmetry about the amino and carboxylate group could be used to limit these search ranges, but given the small molecular size and low dimensionality of $\boldsymbol{\theta}^{\text{ind}}$, I have not done this here.

The adaptive LAM algorithm is ran with the settings shown in Table N.13, and converged in three rounds of LAM generation: one round of uniform LAM placement, and two rounds of adaptive LAM placement. A total of 548 LAMs are generated and a 2D projection of the converged landscape is shown in Figure N.24. The periodicity associated with a 180° rotation

Table N.12: Flexible definition of the glycine molecule used in Global Search LAM construction. Site labels correspond to the labeling shown in Figure 9.7.

Dof	**Participating Sites**	**In Vacuo (°)**	**Lower Bound (°)**	**Upper Bound (°)**
Dih5	O_1−C_1−C_2−H_1	−119.37	−300.0	60.0
Dih8	H_3−N_1−C_2−C_1	−242.85	−300.0	60.0

Table N.13: Adaptive LAM generation settings applied in the Global Search LAMs for glycine.

Setting	Value
Gaussian Version	G16-c01-avx2
Reference U_{intra} [a]	−284.202843345 Hartree.
LAM QM Level of Theory	PBE0/6-311G(d,p), ε_{PCM} = 11
U_{acc} [b]	1.0 kJ/mol
U_{cut} [b]	+60.0 kJ/mol

[a] This corresponds to the gas-phase conformation in the PCM_{11} environment.

[b] Given the small molecular size and low dimensionality of $\boldsymbol{\theta}^{\text{ind}}$, I have opted to use extensive LAM generation cutoffs.

of the carboxylate group (Dih5), and a 120° rotation of the amino group (Dih8) is properly reflected in this landscape.

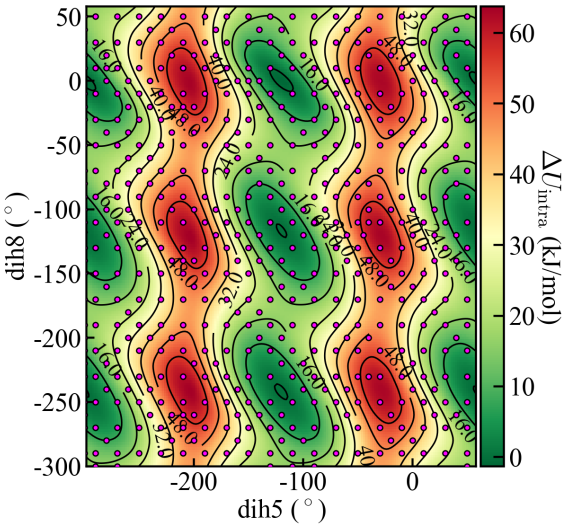


Figure N.24: Converged LAM landscape for the Global Search of glycine (548 LAMs). Magenta points show the LAM point placements.

N.3.2 Local Minimization of Experimental Forms: Global Search Stage

To select an appropriate HAIEFF FF to employ during the Global Search, local minimizations of the known glycine structures are conducted using CrystalPredictorv2.5.2.1. The LAM model depicted in Figure N.24 is paired with four different FFs to conduct local minimizations, initialized from the experimental reference structures (Table N.14).

For all four models, a valid geometry match to β could not be identified following local minimization. Upon closer investigation, it was found that this is due to CrystalPredictor 'pasting' the LAM-predicted conformation incorrectly into the asymmetric unit cell. Consequently, the starting point for local minimization fails to correspond properly to the β form, leading to a different minimized structure. The results for this form can be neglected. For the remaining four forms (α, γ, δ, and ε), geometry predictions are equally good using all four models. The

Table N.14: Local minimization of the glycine experimental structures with the Global Search energy model, using CrystalPredictorv2.5.2.1.

FF	**Metric**	***α*** **(GLYCIN86)**	***β*** **(GLYCIN32)**	***γ*** **(GLYCIN15)**	***δ*** **(GLYCIN67)**	***ε*** **(GLYCIN68)**
FIT[a]	$RMSD_{20}$ (Å)	0.4619	5/20 match	0.1098	0.3506	0.4035
	U_{latt} (kJ/mol)	−215.1859	−205.4209	−212.2848	−213.2258	−209.4505
$DB14_{pc}$[b]	$RMSD_{20}$ (Å)	0.8093	5/20 match	0.1079	0.5784	0.3889
	U_{latt} (kJ/mol)	−212.4074	−213.3673	−204.4161	−209.6048	−206.9818
$BT^{PCM}_{dBhp,pc}$[c]	$RMSD_{20}$ (Å)	0.2185	5/20 match	0.1341	0.5564	0.3709
	U_{latt} (kJ/mol)	−207.7952	−211.2774	−202.5575	−207.2520	−206.2057
$BT^{PCM}_{Mie,pc}$[d]	$RMSD_{20}$ (Å)	0.1494	8/20 match	0.1070	0.4487	0.3532
	U_{latt} (kJ/mol)	−206.7101	−209.1303	−205.7330	−206.8989	−207.3490

[a] This corresponds to the H_p variant of FIT [166], wherein *both* H_O and H_N are described with the H_N FF parameters.

[b] This FF is tabulated in Table G.14 of Ref [2], and uses H_p to describe *both* H_O and H_N.

[c] This corresponds to the FF tabulated in Table J.11 of this work.

[d] This corresponds to the FF tabulated in Table J.13 of this work.

exception to this, is the α form predicted using the $DB14_{pc}$ FF, which returns a poor match to the experimental structure ($RMSD_{20}$ ~0.80 Å). For this reason, I exclude the $DB14_{pc}$ FF for use in the Global Search. The remaining three models are all viable choices, but ultimately, I select the FIT model since it correctly predicts the relative stabilities of the ambient forms. The other two models overstabilize the β form, over α and γ.

N.3.3 Global Search

With the fully specified Global Search energy model, I progress to the main CSP workflow. The details of the Global Search are summarized in Table N.15, and the resulting Global Search landscape is presented in Figure N.25.

From the Global Search, 25,862 unique minima are identified, of which 10,173 were written out to .res files. This subset of 10,173 comprises an energy window of +19.85 kJ/mol from the global minimum structure (U^{GM}_{latt} = −220.9732 kJ/mol). Following clustering, only 4,702 unique clusters remain in the +19.85 kJ/mol energy window. Relatively good geometry matches to all five experimental forms are found within +10 kJ/mol of the GM ($RMSD_{20} \leq 0.64$ Å), and the highest ranked experimental match (ε: $\diamond$) is placed 701st with ΔU^{GM}_{latt} = +9.86 kJ/mol. Although not all unique minima were written to .res files following the search, good geometry matches to the experimental forms are already present amongst the available structures. From a practical standpoint, there is also no interest to refine more structures than are already available from clustering. As such, for the purpose of providing starting points to the CSP refinement, the acquired Global Search landscape is adequate.

Table N.15: Glycine Global Search, conducted with CrystalPredictorv2.5.2.1 and G16-c01-avx2.

Energy Model	
QM Model[a]	PBE0/6-311G(d,p), PCM with $\varepsilon_{\mathrm{PCM}} = 11$
Electrostatic Model	HLYGAt Atomic-charges
HAIEFF FF	FIT (Table I.1, using H_p)
Candidate Generation Validity Range	
Max. $\Delta U_{\mathrm{intra},0}$ (kJ/mol)[b]	+60.0
Max. $U_{\mathrm{inter},0}$ (kJ/mol)[b]	+50.0
Max. $\Delta U_{\mathrm{latt}}^{\mathrm{GM}}$ (kJ/mol)[c]	+30.0
Search Statistics	
Minimizations Attempted	1,000,000
Stable Crystals Accepted	661,264
Landscape Analysis Statistics	
Stable Crystals Analyzed	661,264
Unique Minima Identified[d]	25,862
Minima Written to .res	10,173 (only +19.85 kJ/mol window from GM considered)
Clustering Statistics	
Res Available	10,173
Res Clustered	10,173
Unique Clusters Identified	4,702
Energy Window (kJ/mol)	$[-220.9732, -201.1256]$

[a] The same QM settings are used for the intramolecular and electrostatic model.

[b] Initial crystal structures must satisfy these criteria to qualify for U_{latt} minimization.

[c] Stable minima are only accepted if they are within this energy window from the incumbent GM.

[d] Minima with *identical* geometry and energy are clustered. This is a stringent form of clustering.

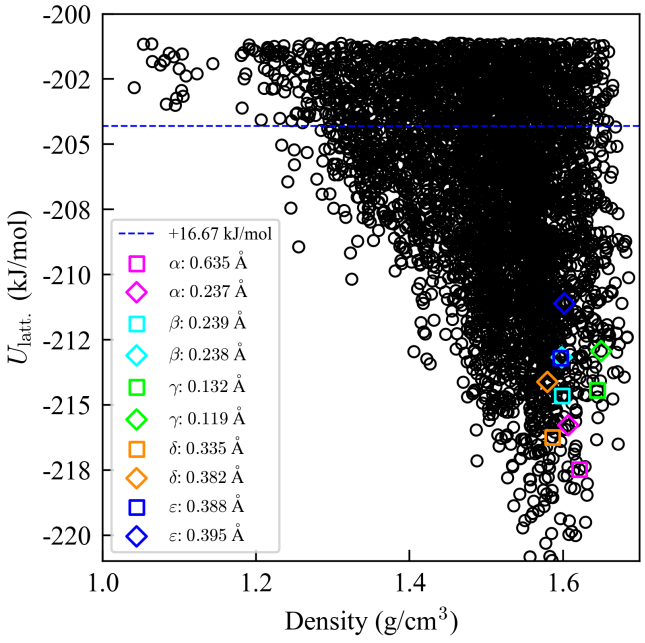


Figure N.25: Global Search landscape for glycine. Matches to α, β, γ, δ and ε are indicated in magenta, cyan, green, orange, and blue respectively, and labeled with their RMSD_{20} scores.

N.3.4 Refinement Overview

Table N.16: Refinement I of glycine, conducted with CSO–FMv1.3.1, G09-d01, and GDMAv2.2. Model numbering corresponds to those shown in Table 9.12.

Setting	Model 1	Model 2	Model 3
Global Search Cutoffs			
Structures Refined	3,000		
$\Delta U_{\text{latt}}^{\text{GM}}$ (kJ/mol)	+16.67		
Intramolecular Model			
Flexible Dofs	Dihedrals: 5 and 8 + Bond Angle: 7		
Intra. LAM Δ_{tol} Bond Length (Å)	0.05		
Intra. LAM Δ_{tol} Bond Angle (°)	1.50		
Intra. LAM Δ_{tol} Torsion Angle (°)	2.50		
Level of Theory	PBE0/6-311G(d,p)		
PCM?	$\varepsilon_{\text{PCM}} = 11$		
Intermolecular Electrostatic Model			
Elec. LAM Δ_{tol} Bond Length (Å)	0.05		
Elec. LAM Δ_{tol} Bond Angle (°)	1.50		
Elec. LAM Δ_{tol} Torsion Angle (°)	2.50		
Level of Theory	PBE0/6-311G(d,p)		
PCM?	$\varepsilon_{\text{PCM}} = 11$		
Electrostatic Model	Rank 4, GDMA Multipoles ($r_{\text{H}}^{\text{GDMA}} = 0.650$ Å)		
Electrostatic Damping	None		
Intermolecular Force-field Model			
Force-field Label	DB14_{mp}	$\text{BT}_{\text{dBhp,mp}}^{\text{PCM}}$	$\text{BT}_{\text{Mie,mp}}^{\text{PCM}}$
Force-field Parameters	Ref [2], Table I.14	Table J.25	Table J.27
Refined Landscape Statistics			
No. of Crystals Post-Clustering	1,212	1,158	1,173
$U_{\text{latt}}^{\text{GM}}$ (kJ/mol)[a]	−232.81	−230.67	−229.40
α Found?	✓	✓	✓
β Found?	✓	✓	✓
γ Found?	✓	✓	✓
δ Found?	✓	✓	✓
ε Found?	✓	✓	✓

[a] Note, all ΔU_{intra} are defined with respect to the PCM_{11} gas-phase energy in Table N.11.

Table N.17: Refinement II of glycine, conducted with CSO–FMv1.4.0.2, G09-d01, and GDMAv2.2. These settings correspond to Model 4 in Table 9.12.

Setting	Model 4		
Refinement I Cutoffs			
Refinement I Landscape	Model 3 from Table N.16		
Structures Refined	200		
$\Delta U_{\text{latt}}^{\text{GM}}$ (kJ/mol)	+8.29		
Intramolecular Model			
Flexible Dofs	Dihedrals: 5 and 8 + Bond Angle: 7		
Intra. LAM Δ_{tol} Bond Length (Å)	0.05		
Intra. LAM Δ_{tol} Bond Angle (°)	1.50		
Intra. LAM Δ_{tol} Torsion Angle (°)	2.50		
Level of Theory	PBE0/6-311G(d,p)		
PCM?	$\varepsilon_{\text{PCM}} = 11$		
Intermolecular Electrostatic Model			
Elec. LAM Δ_{tol} Bond Length (Å)	0.05		
Elec. LAM Δ_{tol} Bond Angle (°)	1.50		
Elec. LAM Δ_{tol} Torsion Angle (°)	2.50		
Level of Theory	PBE0/6-311G(d,p)		
PCM?	Static		
Electrostatic Model	Rank 4, GDMA Multipoles ($r_{\text{H}}^{\text{GDMA}} = 0.650$ Å)		
Electrostatic Damping	Using $\beta_{\text{atm}}^{tt'}$ based on $\%f_6 = 90\%$ in Table L.1		
Intermolecular SCERP Model			
SCERP-Field Representation	HLYGAt Atomic-Charges		
r_{scrp} (Å)	40.0		
t_{slow}	6		
$\varepsilon_{\text{scrp}}$ (a.u.)	1×10^{-5}		
t_{max}	30		
$\varepsilon_{\text{scrpRM}}$ (kJ/mol)	1×10^{-5}		
k_{max}	16		
Intermolecular Force-field Model			
Force-field Label	$\text{BT}_{\text{Mie,mp}}^{\text{SCRP}}$		
Force-field Parameters	Table M.4		
Refined Landscape Statistics			
CSO–FM Optimality Tolerance[a]	1.0×10^{-3}		
FD Mode (per Table 7.6)	2	3	4
No. of Crystals Post-Clustering	186	186	186
$U_{\text{latt}}^{\text{GM}}$ (kJ/mol)[b]	−219.57	−219.53	−219.49
α Found?	✓	✓	✓
β Found?	✓	✓	✓
γ Found?	✓	✓	✓
δ Found?	✓	✓	✓
ε Found?	✓	✓	✓

[a] This is decided by Equation 7.4.14. Without SCERP, the default τ_{nag} is 1×10^{-8}.

[b] Note, all ΔU_{intra} are defined with respect to the PCM_{11} gas-phase energy in Table N.11.

N.3.5 Polar Cells and SCERP

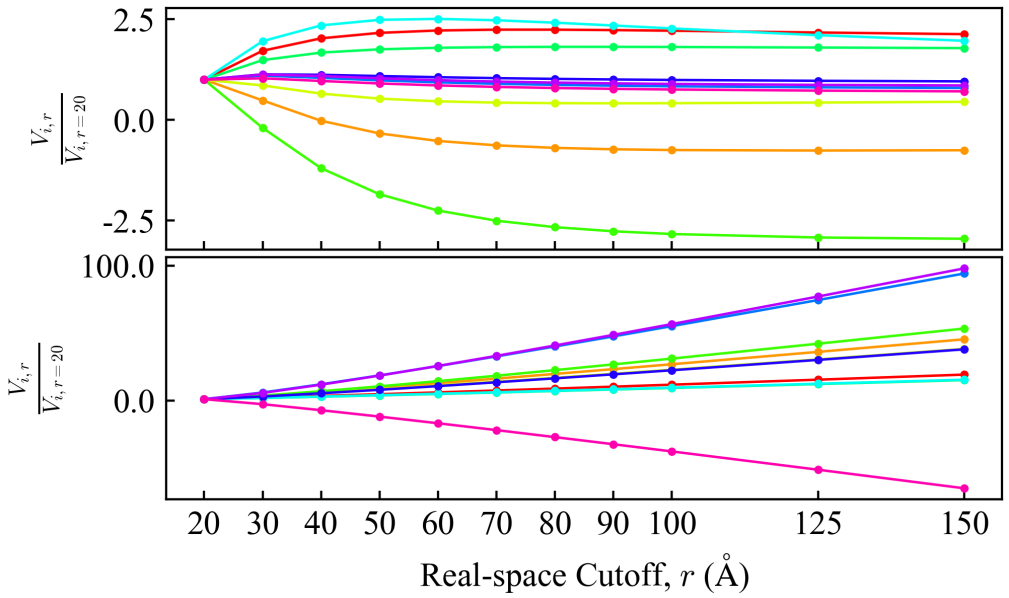


Figure N.26: ESP exerted on atom i of δ- (top) and ε-glycine (bottom) at different real-space cutoffs. All ESPs are reported as a ratio with respect to the ESP at r = 20 Å.

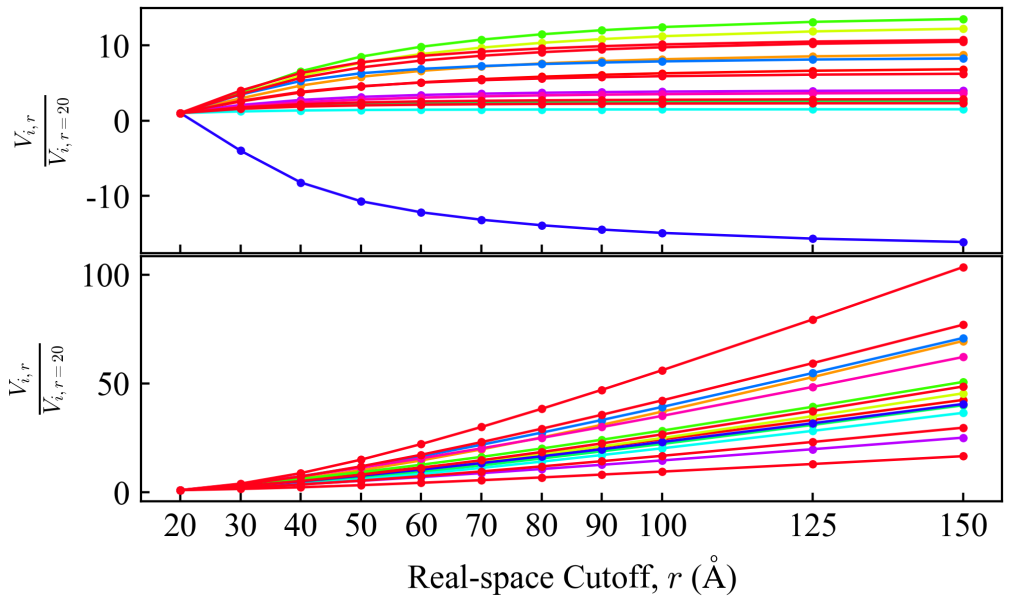


Figure N.27: ESP exerted on atom i of KUXPUP (top), and KUXPUP01 (bottom) at different real-space cutoffs. All ESPs are reported as a ratio with respect to the ESP at r = 20 Å.

N.3.6 Final Periodic DFT-D Re-ranking

Table N.18: COMPACK comparison between the Model 4 structures before and after local optimization by periodic DFT-D. Comparisons of the optimized structures to the experimental forms of glycine are also presented. Blank entries indicate no valid $RMSD_{20}$ score could be generated.

$RMSD_{20}$ (Å)		**Reference Structure**														
		α	β	γ	δ	ε	1	2	3	4	5	6	21	29	174	185
Optimised Structure	1						0.207									
	2							0.246								
	3								0.179							
	4									0.192						
	5										0.166					
	6	0.065										0.197				
	21				0.188								0.145			
	29		0.042											0.158		
	174			0.060											0.171	
	185					0.256										0.165

N.3.7 Assessment of SCERP Gradient Approximations

N.3.7.1 Landscape Minimization by FD Modes 2 to 4

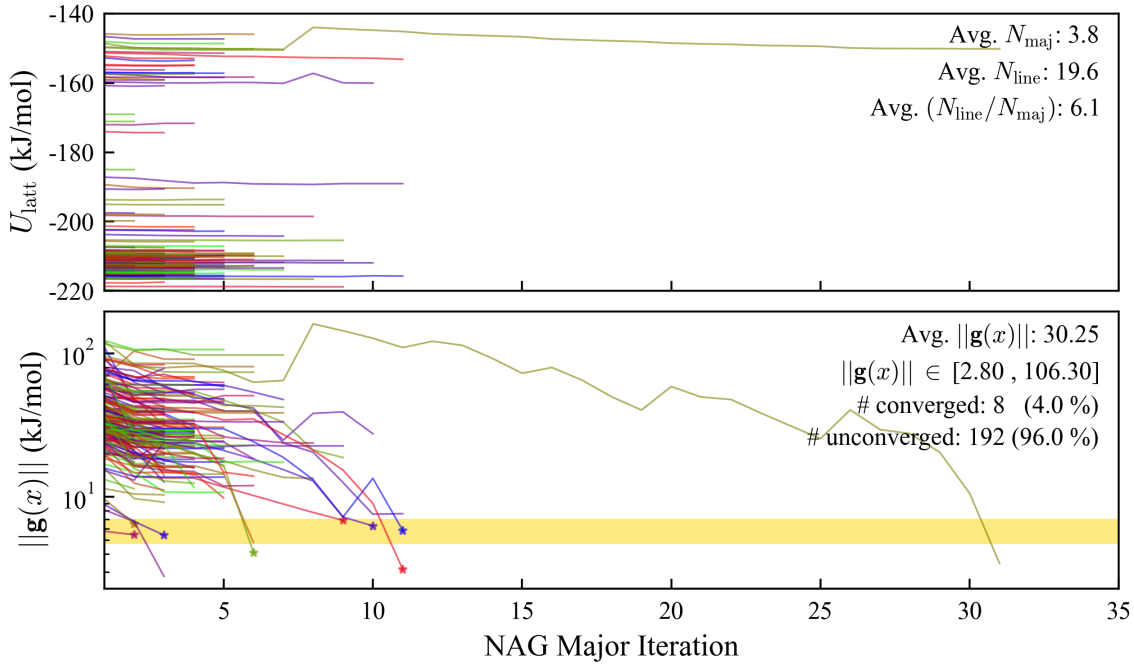


Figure N.28: Minimization trajectories of glycine Refinement II by FD Mode 2: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $||\boldsymbol{g}(x)||$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★.

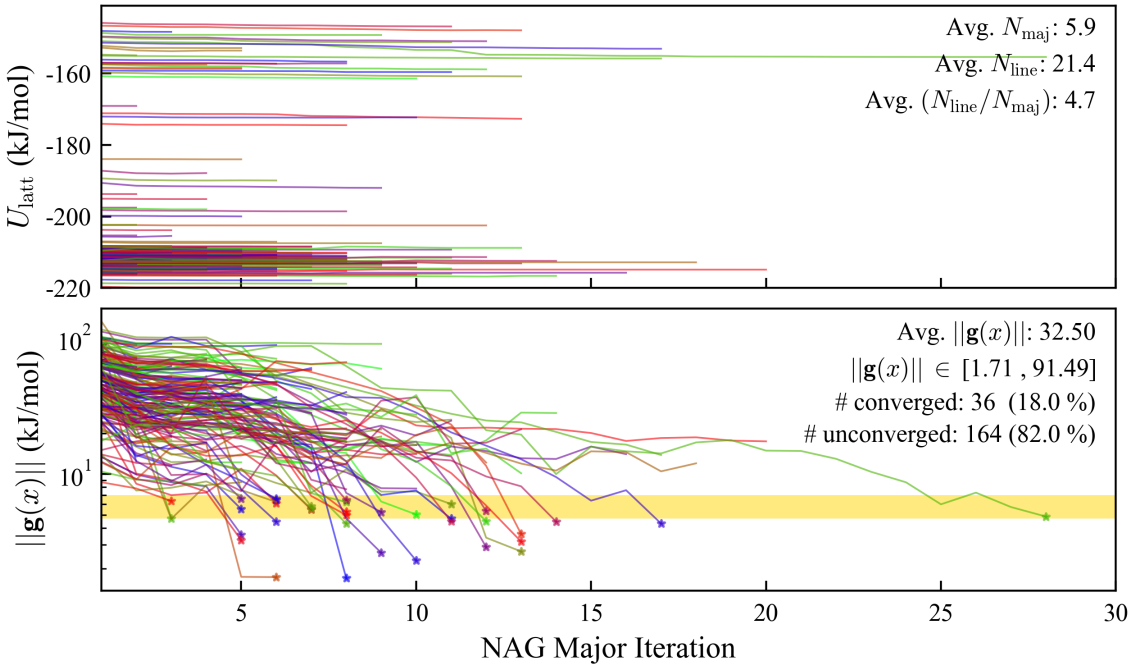


Figure N.29: Minimization trajectories of glycine Refinement II by FD Mode 3: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $||\boldsymbol{g}(x)||$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★.

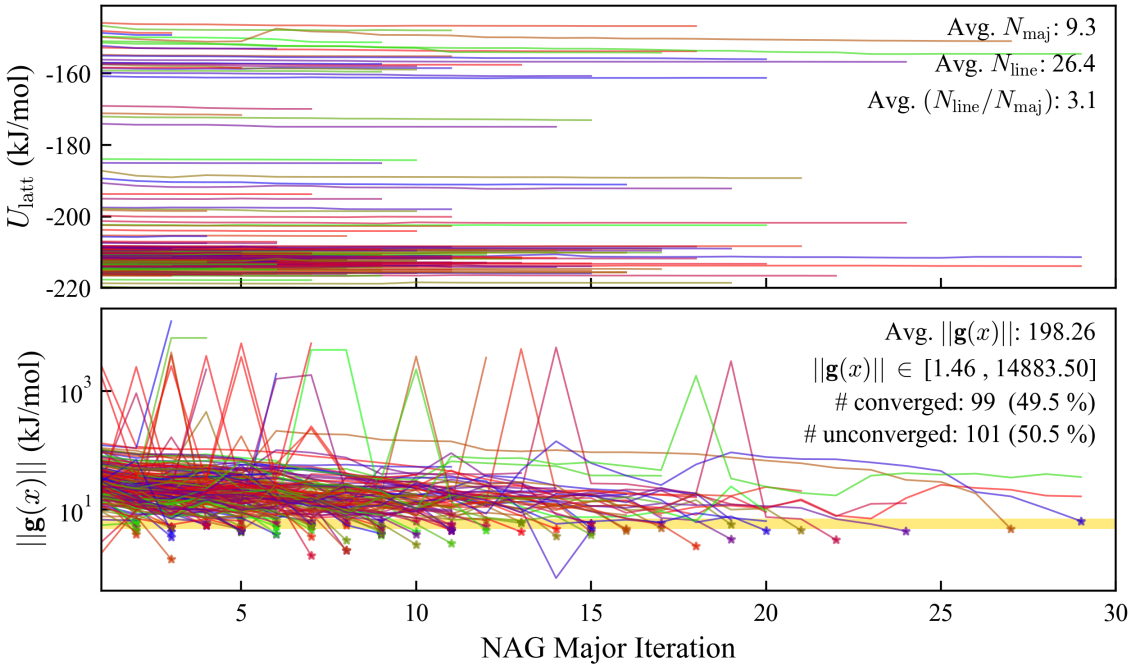


Figure N.30: Minimization trajectories of glycine Refinement II by FD Mode 4: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $||\boldsymbol{g}(x)||$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★.

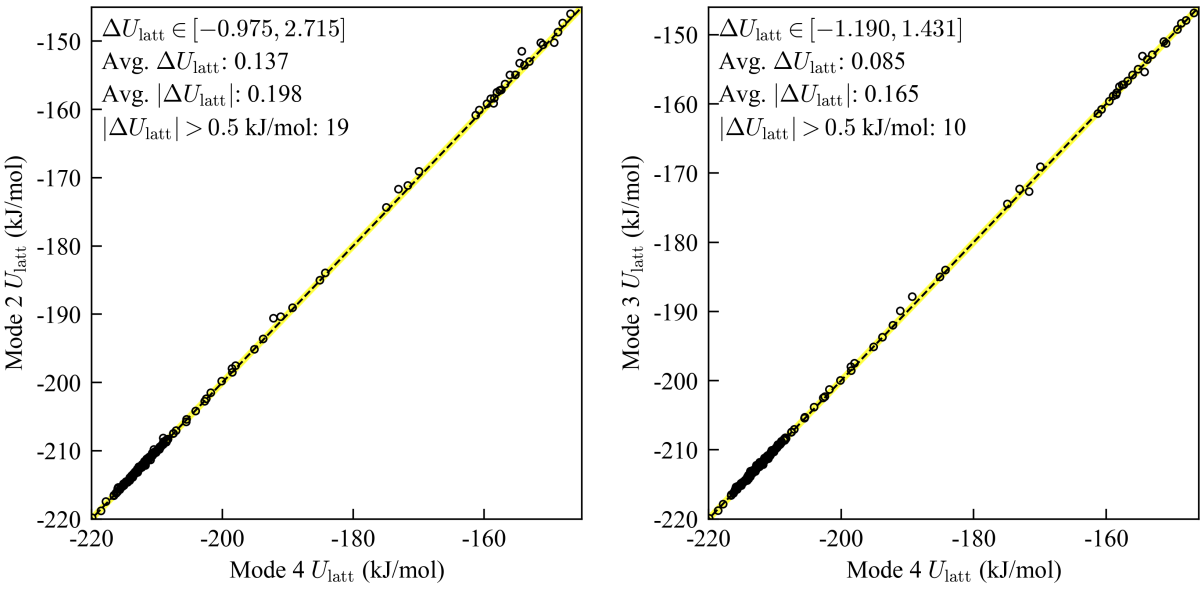


Figure N.31: Comparison of optimized glycine U_{latt} obtained using (a) Mode 2 and (b) Mode 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$. All comparisons are on a structure-by-structure basis. The yellow region indicates $|\Delta U_{\text{latt}}| \leq 0.5$ kJ/mol.

N.4 L-/DL-Proline CSP

N.4.1 Molecular Definition & Global Search LAM Construction

Gas-phase minimization of the L-prolin molecule is done at the PBE0/6-311G(d,p) level of theory, with different PCM settings, using Gaussian (G16-c01-avx2). *In vacuo* conformations and conformational energies are found from the geometry optimization (Table N.19); atomic-charges are also fit at the minimized geometry by Gaussian's HLYGAt method. Critically, the glycine molecule persists in its zwitterionic state only when $\varepsilon_{\text{PCM}} \geq 7$ is used, while reverting back to a neutral state otherwise. For consistency with the glycine CSP, $\varepsilon_{\text{PCM}} = 11$ is used during the proline Global Search stage, and the ground-state conformational energy is defined with respect to the same environment, i.e., ΔU_{intra} are computed as $U_{\text{intra}} - U_{\text{intra}}^{\text{PCM}_{11}}$

Table N.19: Gas-phase conformational energies for L-proline. All computations are done at the PBE0/6-311G(d,p) level of theory in G16-c01-avx2.

Environment	$\boldsymbol{U_{\text{intra}}}$ **(Hartrees)**	**Final State**
Static	−400.820118743	Neutral
PCM ($\varepsilon_{\text{PCM}} = 3$)	−400.827253892	Neutral
PCM ($\varepsilon_{\text{PCM}} = 7$)	−400.823665602	Zwitterionic
PCM ($\varepsilon_{\text{PCM}} = 11$)	−400.826025343	Zwitterionic

Proline is treated flexibly throughout the CSP study, particularly to facilitate a better description of the crystalline hydrogen bonding. For the Global Search, a LAM database is constructed by the adaptive LAM algorithm [98], with the flexible molecule definition summarized in Table N.20. Dih4 corresponds to a rotation of the carboxylate group relative to the five-membered ring, while Dih14 corresponds to a rotation of the amino group hydrogen within the ring's plane. By permitting these groups to rotate, greater variation in hydrogen bonding can be explored during the Global Search. Dih4 has been allowed to rotate 360°, while Dih14 has a limited range of rotation due to its participation in the highly-strained ring. In principle, symmetry about the carboxylate group can be used to limit its flexible search range, but given the moderate molecular size and low dimensionality of $\boldsymbol{\theta}^{\text{ind}}$, I have not done this here.

The adaptive LAM algorithm is ran with the settings shown in Table N.21, and converged in two rounds of LAM generation: one round of uniform LAM placement, and one round of adaptive LAM placement. A total of 226 LAMs are generated and a 2D projection of the

Table N.20: Flexible definition of the L-proline molecule used in Global Search LAM construction. Site labels correspond to the labeling shown in Figure 9.12.

Dof	**Participating Sites**	**In Vacuo (°)**	**Lower Bound (°)**	**Upper Bound (°)**
Dih4	$O_1{-}C_1{-}C_2{-}C_3$	114.70	−30.0	330.0
Dih14	$H_6{-}N_1{-}C_2{-}C_1$	−1.530	−90.0	90.0

Table N.21: Adaptive LAM generation settings applied in the Global Search LAMs for L-proline.

Setting	Value
Gaussian Version	G16-c01-avx2
Reference U_{intra} [a]	−400.826025343 Hartree.
LAM QM Level of Theory	PBE0/6-311G(d,p), ε_{PCM} = 11
U_{acc}	5.0 kJ/mol
U_{cut} [b]	+60.0 kJ/mol

[a] This corresponds to the gas-phase conformation in the PCM_{11} environment.

[b] Given the moderate molecular size and low dimensionality of $\boldsymbol{\theta}^{\text{ind}}$, I have opted to use extensive LAM generation cutoffs.

converged landscape is shown in Figure N.32. The periodicity associated with a 180° rotation of the carboxylate group (Dih4) is correctly reflected in this landscape. Moreover, the chosen flexibility range of the amino-hydrogen is justified since near the bounds (−90°/90°), ΔU_{intra} exceed 60.0 kJ/mol, which should be energetically infeasible in observable crystals [21].

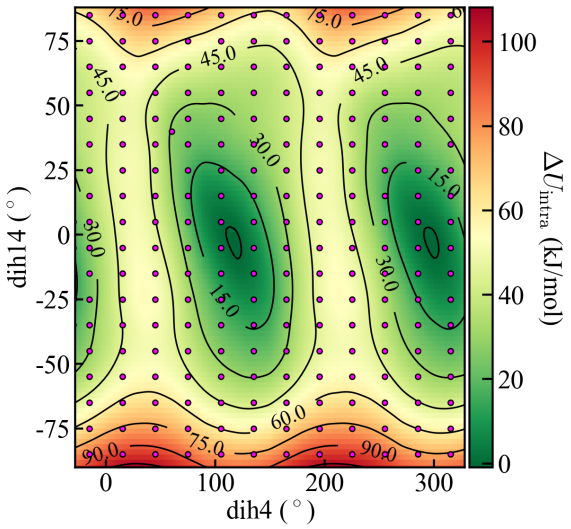


Figure N.32: Converged LAM landscape for the Global Search of L-proline (226 LAMs). Magenta points show the LAM point placements.

N.4.2 Local Minimization of Experimental Forms: Global Search Stage

To select an appropriate HAIEFF FF for the Global Search, local minimizations of the known (Z' = 1) proline structures are conducted using CrystalPredictorv2.5.2.1. The LAM model depicted in Figure N.32 is paired with the DB14_{pc} FF (Ref [2], Table G.14) to conduct local minimizations initialized from the experimental reference structures (Table N.22). Following local minimization with the Global Search HAIEFF model, valid geometry matches to both known forms (L-I and DL-I) are found, with RMSD_{20} < 0.47 Å. With this FF, the relative stability of L-I > DL-I is also correctly captured, and the two forms are in a reasonably tight

energy window (relative to each other) of about 4 kJ/mol. On this basis, this FF should be adequate for the Global Search stage.

Table N.22: Local minimization of the L-/DL-proline experimental structures with the Global Search energy model, using CrystalPredictorv2.5.2.1.

FF	**Metric**	**L-I (PROLIN03)**	**DL-I (QANRUT)**
$DB14_{pc}$ [a]	$RMSD_{20}$ (Å)	0.4651	0.1498
	U_{latt} (kJ/mol)	−173.5455	−169.5067

[a] This FF is tabulated in Table G.14 of Ref [2], and uses H_p to describe *both* H_O and H_N.

N.4.3 Global Search

With the fully specified Global Search energy model, I progress to the main CSP workflow. The details of the Global Search are summarized in Table N.23, and the resulting Global Search landscape is presented in Figure N.33.

From the Global Search, 11,366 unique minima are identified, of which 3,730 were written out to .res files. This subset of 3,730 comprises an energy window of +17.56 kJ/mol from the global minimum structure (U_{latt}^{GM} = −177.3280 kJ/mol). Following clustering, only 2,739 unique clusters remain in the +17.56 kJ/mol energy window. Relatively good geometry matches to both experimental forms are found within +10 kJ/mol of the GM ($RMSD_{20} \leq 0.78$ Å), and the highest ranked experimental match (DL-I: ⬠) is placed 315th with ΔU_{latt}^{GM} = +8.80 kJ/mol. Matches to L-II were not searched for here, since it is a Z' = 2 structure, and is not expected to exist on the Z' = 1 CSP landscape. Although not all unique minima were written to .res files following the search, good geometry matches to the experimental forms are already present amongst the available structures. From a practical standpoint, there is also no interest to refine more structures than are already available from clustering. As such, for the purpose of providing starting points to the CSP refinement, the acquired Global Search landscape is adequate.

Table N.23: Proline Global Search, conducted with CrystalPredictorv2.5.2.1 and G16-c01-avx2.

Energy Model	
QM Model[a]	PBE0/6-311G(d,p), PCM with ε_{PCM} = 11
Electrostatic Model	HLYGAt Atomic-charges
HAIEFF FF	$DB14_{pc}$ (Ref [2], Table G.14)
Candidate Generation Validity Range	
Max. $\Delta U_{intra,0}$ (kJ/mol)[b]	+60.0
Max. $U_{inter,0}$ (kJ/mol)[b]	+50.0
Max. ΔU_{latt}^{GM} (kJ/mol)[c]	+25.0
Search Statistics	
Minimizations Attempted	1,000,000
Stable Crystals Accepted	577,886
Landscape Analysis Statistics	
Stable Crystals Analyzed	577,886
Unique Minima Identified[d]	11,366
Minima Written to .res	3,730 (only +17.56 kJ/mol window from GM considered)
Clustering Statistics	
Res Available	3,730
Res Clustered	3,730
Unique Clusters Identified	2,739
Energy Window (kJ/mol)	$[-177.3280, -159.7703]$

[a] The same QM settings are used for the intramolecular and electrostatic model.

[b] Initial crystal structures must satisfy these criteria to qualify for U_{latt} minimization.

[c] Stable minima are only accepted if they are within this energy window from the incumbent GM.

[d] Minima with *identical* geometry and energy are clustered. This is a stringent form of clustering.

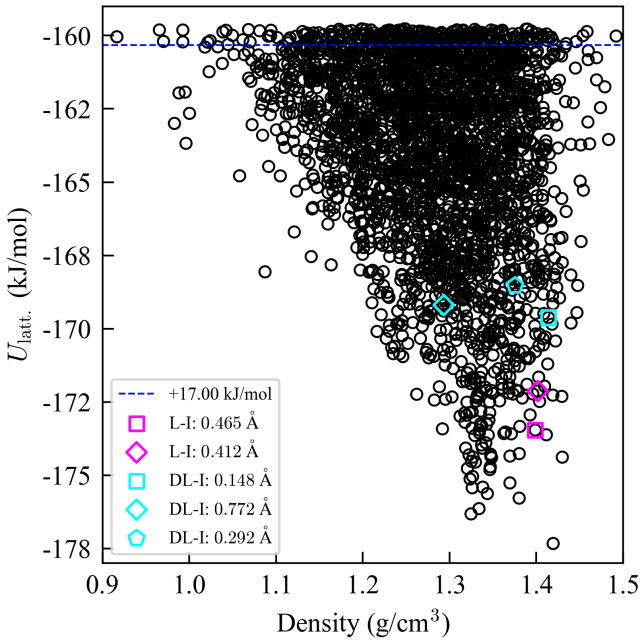


Figure N.33: Global Search landscape for proline. Matches to L-I and DL-I are indicated in magenta and cyan respectively, and labeled with their $RMSD_{20}$ scores.

N.4.4 Refinement Overview

Table N.24: Refinement I of proline, conducted with CSO–FMv1.3.1, G09-d01, and GDMAv2.2. Model numbering corresponds to those shown in Table 9.17.

Setting	Model 1	Model 2	Model 3
Global Search Cutoffs			
Structures Refined	2,471		
$\Delta U_{\text{latt}}^{\text{GM}}$ (kJ/mol)	+17.00		
Intramolecular Model			
Flexible Dofs	Dihedrals: 4, 12, and 14		
Intra. LAM Δ_{tol} Bond Length (Å)	0.05		
Intra. LAM Δ_{tol} Bond Angle (°)	1.50		
Intra. LAM Δ_{tol} Torsion Angle (°)	2.50		
Level of Theory	PBE0/6-311G(d,p)		
PCM?	$\varepsilon_{\text{PCM}} = 11$		
Intermolecular Electrostatic Model			
Elec. LAM Δ_{tol} Bond Length (Å)	0.05		
Elec. LAM Δ_{tol} Bond Angle (°)	1.50		
Elec. LAM Δ_{tol} Torsion Angle (°)	2.50		
Level of Theory	PBE0/6-311G(d,p)		
PCM?	$\varepsilon_{\text{PCM}} = 11$		
Electrostatic Model	Rank 4, GDMA Multipoles ($r_{\text{H}}^{\text{GDMA}} = 0.650$ Å)		
Electrostatic Damping	None		
Intermolecular Force-field Model			
Force-field Label	DB14_{mp}	$\text{BT}_{\text{dBhp,mp}}^{\text{PCM}}$	$\text{BT}_{\text{Mie,mp}}^{\text{PCM}}$
Force-field Parameters	Ref [2], Table I.14	Table J.25	Table J.27
Refined Landscape Statistics			
No. of Crystals Post-Clustering	1,797	1,760	1,740
$U_{\text{latt}}^{\text{GM}}$ (kJ/mol)[a]	−182.05	−179.84	−180.30
L-I Found?	✓	✓	✓
L-II Found?	✓	✓	✓
DL-I Found?	✓	✓	✓

[a] Note, all ΔU_{intra} are defined with respect to the PCM_{11} gas-phase energy in Table N.19.

Table N.25: Refinement II of proline, conducted with CSO–FMv1.4.0.2, G09-d01, and GDMAv2.2. These settings correspond to Model 4 in Table 9.17.

Setting	**Model 4**		
Refinement I Cutoffs			
Refinement I Landscape	Model 3 from Table N.24		
Structures Refined	200		
$\Delta U_{\text{latt}}^{\text{GM}}$ (kJ/mol)	+8.87		
Intramolecular Model			
Flexible Dofs	Dihedrals: 4, 12, and 14		
Intra. LAM $\boldsymbol{\Delta}_{\text{tol}}$ Bond Length (Å)	0.05		
Intra. LAM $\boldsymbol{\Delta}_{\text{tol}}$ Bond Angle (°)	1.50		
Intra. LAM $\boldsymbol{\Delta}_{\text{tol}}$ Torsion Angle (°)	2.50		
Level of Theory	PBE0/6-311G(d,p)		
PCM?	$\varepsilon_{\text{PCM}} = 11$		
Intermolecular Electrostatic Model			
Elec. LAM $\boldsymbol{\Delta}_{\text{tol}}$ Bond Length (Å)	0.05		
Elec. LAM $\boldsymbol{\Delta}_{\text{tol}}$ Bond Angle (°)	1.50		
Elec. LAM $\boldsymbol{\Delta}_{\text{tol}}$ Torsion Angle (°)	2.50		
Level of Theory	PBE0/6-311G(d,p)		
PCM?	Static		
Electrostatic Model	Rank 4, GDMA Multipoles ($r_{\text{H}}^{\text{GDMA}} = 0.650$ Å)		
Electrostatic Damping	Using $\beta_{\text{atm}}^{tt'}$ based on $\%f_6 = 90\%$ in Table L.1		
Intermolecular SCERP Model			
SCERP-Field Representation	HLYGAt Atomic-Charges		
r_{scrp} (Å)	40.0		
t_{slow}	6		
$\varepsilon_{\text{scrp}}$ (a.u.)	1×10^{-5}		
t_{max}	30		
$\varepsilon_{\text{scrpRM}}$ (kJ/mol)	1×10^{-5}		
k_{max}	16		
Intermolecular Force-field Model			
Force-field Label	$\text{BT}_{\text{Mie,mp}}^{\text{SCRP}}$		
Force-field Parameters	Table M.4		
Refined Landscape Statistics			
CSO–FM Optimality Tolerance[a]	1.5×10^{-3}		
FD Mode (per Table 7.6)	2	3	4
No. of Crystals Post-Clustering	190	191	191
$U_{\text{latt}}^{\text{GM}}$ (kJ/mol)[b]	−173.36	−173.32	−173.37
L-I Found?	✓	✓	✓
L-II Found?	✓	✓	✓
DL-I Found?	✓	✓	✓

[a] This is decided by Equation 7.4.14. Without SCERP, the default τ_{nag} is 1×10^{-8}.

[b] Note, all ΔU_{intra} are defined with respect to the PCM_{11} gas-phase energy in Table N.19.

N.4.5 Final Periodic DFT-D Re-ranking

Table N.26: COMPACK comparison between the Model 4 structures before and after local optimization by periodic DFT-D. Comparisons of the optimized structures to the experimental forms of proline are also presented. Blank entries indicate no valid $RMSD_{20}$ score could be generated.

$RMSD_{20}$ (Å)		**Reference Structure**																				
		L-I	DL-I	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	86
Optimised Structure	1			0.226																		
	2				0.119																	
	3					0.112																
	4						0.159															
	5							0.233				0.221										
	6								0.117													
	7									0.191												
	8										0.202											
	9							0.201				0.190										
	10												0.133									
	11		0.073											0.193								
	12														0.074							
	13															0.111						
	14																0.273					
	15																	0.147				
	16																		0.167			
	17																			0.342		
	18																				0.218	
	86	0.065																				0.348

N.4.6 Assessment of SCERP Gradient Approximations

N.4.6.1 Landscape Minimization by FD Modes 2 to 4

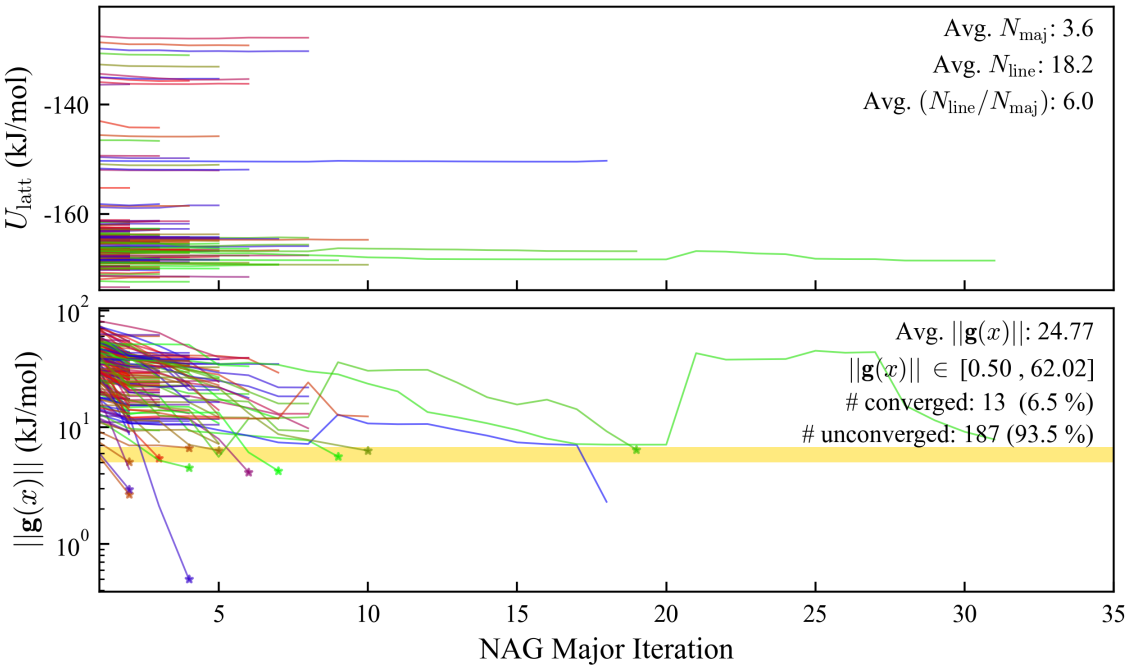


Figure N.34: Minimization trajectories of proline Refinement II by FD Mode 2: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $||\boldsymbol{g}(x)||$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★.

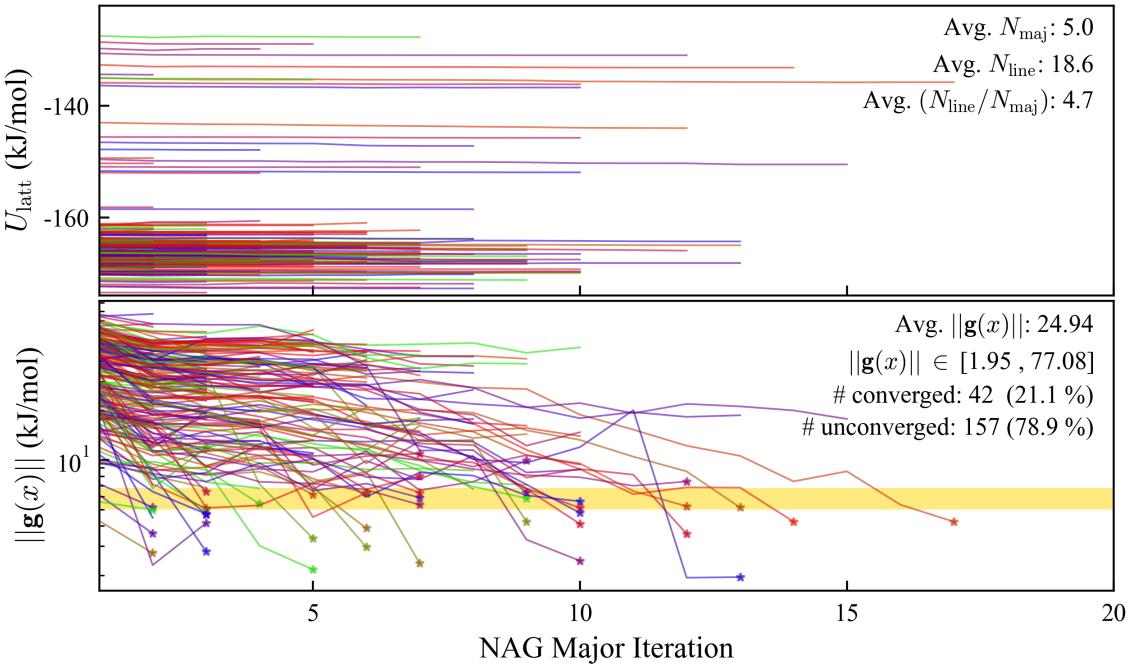


Figure N.35: Minimization trajectories of proline Refinement II by FD Mode 3: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $||\boldsymbol{g}(x)||$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★.

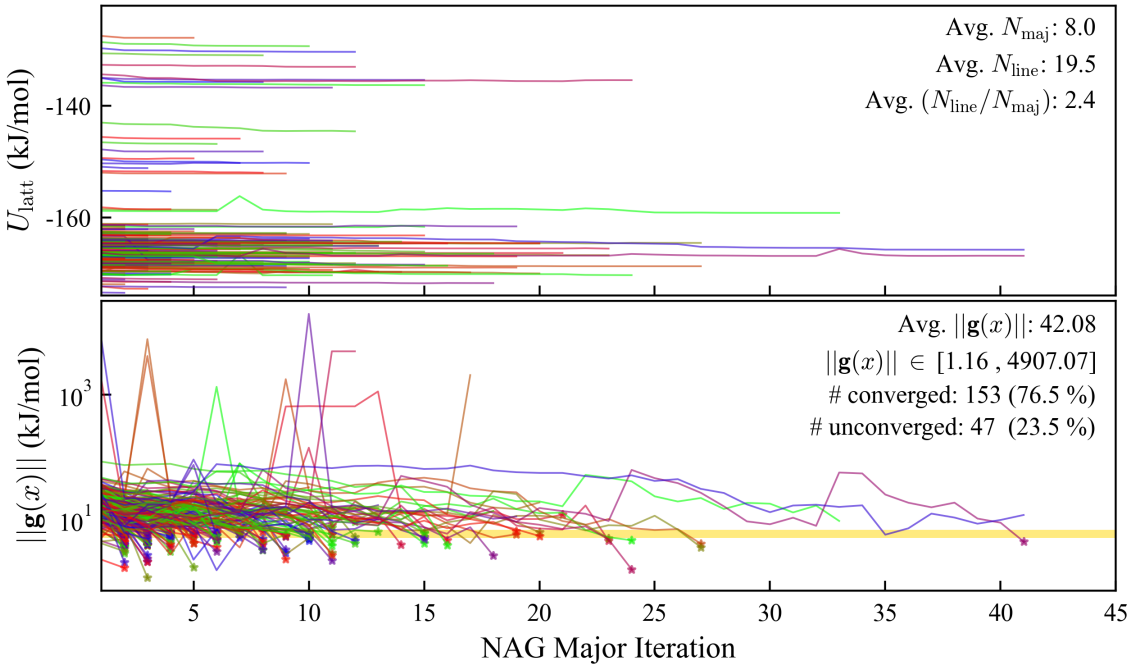


Figure N.36: Minimization trajectories of proline Refinement II by FD Mode 4: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $||\boldsymbol{g}(x)||$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★.

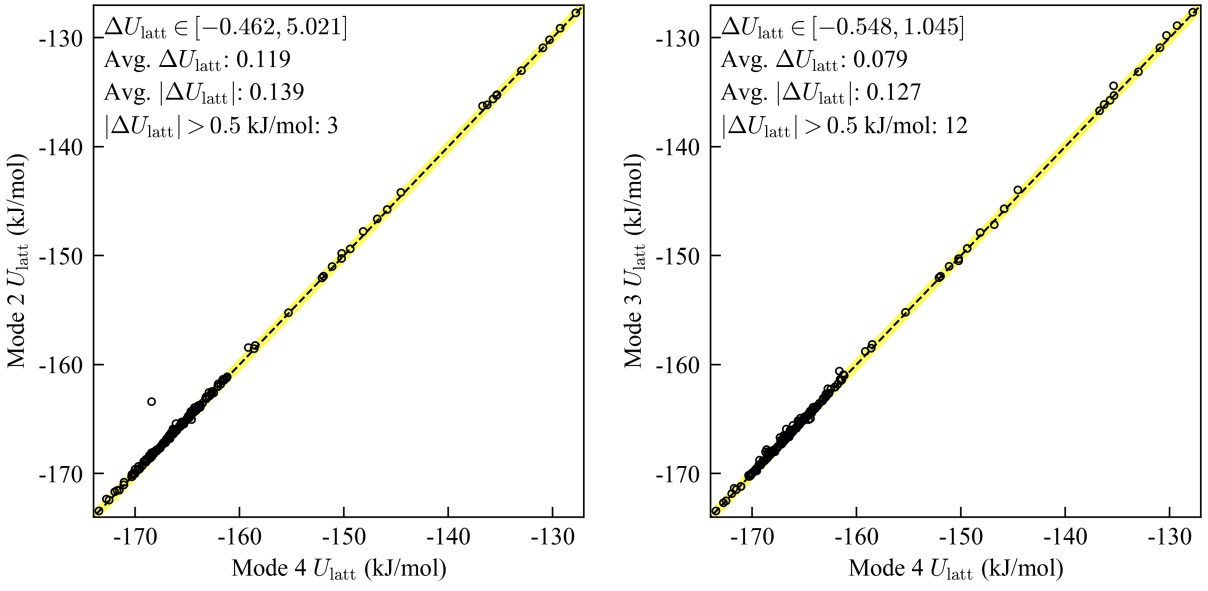


Figure N.37: Comparison of optimized proline U_{latt} obtained using (a) Mode 2 and (b) Mode 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$. All comparisons are on a structure-by-structure basis. The yellow region indicates $|\Delta U_{\text{latt}}| \leq 0.5$ kJ/mol.

N.5 1:1 Caffeine + *m*-Nitrobenzoic Acid CSP

N.5.1 Molecular Definition & Global Search LAM Construction

Independent gas-phase minimization of the CAFF and mNBA molecule are done at the PBE0/6-311G(d,p) level of theory, with and without PCM, using Gaussian (G16-c01-avx2). *In vacuo* conformations and conformational energies are found from the geometry optimization (Table N.27); atomic-charges are also fit at the minimized geometry by Gaussian's HLYGAt method. To enhance the description of intermolecular hydrogen bonding in the crystal, the PCM-embedded model is used during the Global Search stage, and the ground-state conformational energy is defined with respect to the same environment, i.e., ΔU_{intra} are computed as $U_{\text{intra}} - U_{\text{intra}}^{\text{PCM}_{11}}$

Table N.27: Gas-phase conformational energies for CAFF and mNBA. All computations are done at the PBE0/6-311G(d,p) level of theory in G16-c01-avx2.

Caffeine	
Environment	U_{intra} **(Hartrees)**
Static	−679.787684921
PCM ($\varepsilon_{\text{PCM}} = 11$)	−679.799029211

***m*-Nitrobenzoic Acid**	
Environment	U_{intra} **(Hartrees)**
Static	−624.802954112
PCM ($\varepsilon_{\text{PCM}} = 11$)	−624.812844591

Throughout the CSP, CAFF is approximated as rigid, whilst mNBA is treated flexibly. For CAFF, no additional QM computations are necessary to construct LAMs in the Global Search. Conversely, for mNBA, a Global Search LAM database is constructed by the adaptive LAM algorithm [98], with the flexible molecule definition summarized in Table N.28. Dih9 and Dih11 describe the rotation of −COOH and $-NO_2$ relative to the aromatic ring, while Dih13 permits the −COOH hydrogen to rotate internally within the group. For simplicity, I have allowed all three torsions to rotate 360°, although Dih11 is expected to exhibit periodicity.

The adaptive LAM algorithm is ran with the settings shown in Table N.29, and converged in four rounds of LAM generation: one round of uniform LAM placement, and three rounds of adaptive LAM placement. A total of 1,578 LAMs are placed, and 2D projections of the final landscapes are shown in Figure N.38. Periodicity associated with a 180° rotation of the $-NO_2$ group is correctly reflected in the landscapes involving Dih11. Both Dih9 and Dih11 favor values around 0° and 180°, reflecting the preference of −COOH and $-NO_2$ to preserve a planar configuration with respect to the aromatic ring. Finally, Dih13 favors a value of 180°, which

Table N.28: Flexible definition of the mNBA molecule used in Global Search LAM construction. Site labels correspond to the labeling shown in Figure 9.16.

Dof	**Participating Sites**	**In Vacuo (°)**	**Lower Bound (°)**	**Upper Bound (°)**
Dih9	$O_3-C_{15}-C_{11}-C_{10}$	−180.0	−22.50	337.5
Dih11	$O_5-N_5-C_9-C_{10}$	0.0	−22.50	337.5
Dih13	$H_{11}-O_3-C_{15}-C_{11}$	−180.0	−22.50	337.5

Table N.29: Adaptive LAM generation settings applied in the Global Search LAMs for mNBA.

Setting	**Value**
Gaussian Version	G16-c01-avx2
Reference U_{intra} [a]	−624.812844591 Hartree.
LAM QM Level of Theory	PBE0/6-311G(d,p), $\varepsilon_{\text{PCM}} = 11$
U_{acc} [b]	3.0 kJ/mol
U_{cut} [b]	+60.0 kJ/mol

[a] This corresponds to the gas-phase conformation in the PCM_{11} environment.

[b] Given the moderate molecular size and low dimensionality of $\boldsymbol{\theta}^{\text{ind}}$, I have opted to use extensive LAM generation cutoffs.

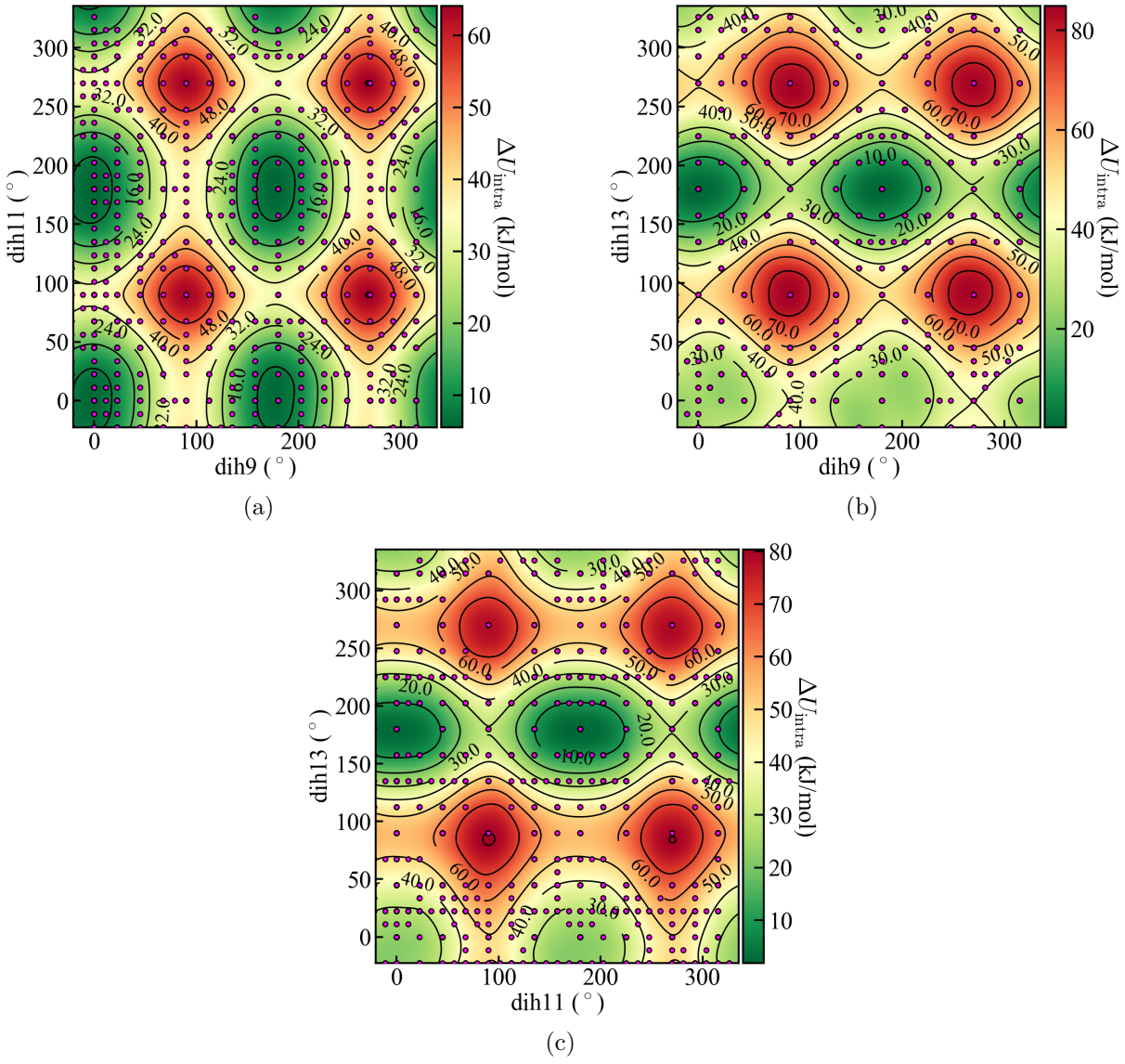


Figure N.38: Converged mNBA LAM landscape (1,578 LAMs) for the Global Search of CAFF-mNBA. When evaluating DihX/DihY, the third dihedral is kept at its search range's midpoint, as given in Table N.28. Magenta points show the LAM point placements.

corresponds to −COOH hydrogen pointing away from the mNBA aromatic ring (as illustrated in Figure 9.16). Overall, these trends align well with chemical intuition. Importantly, around the expected values of each dihedral, there is a non-negligible range of torsion values wherein ΔU_{intra} is reasonably low (≤ 30 kJ/mol), justifying their flexible treatment.

N.5.2 Local Minimization of Experimental Forms: Global Search Stage

To select an appropriate HAIEFF FF for the Global Search, local minimizations of the known CAFF-mNBA structures are conducted using CrystalPredictorv2.5.2.1. The LAM model depicted in Figure N.38 is paired with the $DB14_{pc}$ FF (Ref [2], Table G.14) to conduct local minimizations initialized from the experimental reference structures (Table N.30). Following local minimization with the Global Search HAIEFF model, valid geometry matches to both known forms (F-I and F-II) are found, with $RMSD_{20} < 0.30$ Å. With this FF, the relative stability of F-I > F-II is correctly predicted, and the two forms are in a reasonably tight energy window (relative to each other) of about 1 kJ/mol. These results are congruent with the findings from local minimization by periodic DFT-D (Table 9.20). On this basis, this FF should be adequate for the Global Search stage.

Table N.30: Local minimization of the CAFF-mNBA experimental structures with the Global Search energy model, using CrystalPredictorv2.5.2.1.

FF	**Metric**	**F-I (DIJVUN)**	**F-II (DIJVUN01)**
$DB14_{pc}$[a]	$RMSD_{20}$ (Å)	0.1998	0.2981
	U_{latt} (kJ/mol)	−106.1650	−105.1784

[a] This FF is tabulated in Table G.14 of Ref [2], and uses H_p to describe *both* H_O and H_N.

N.5.3 Global Search

With the fully specified Global Search energy model, I progress to the main CSP workflow. The details of the Global Search are summarized in Table N.31, and the resulting Global Search landscape is presented in Figure N.39.

From the Global Search, 146,353 unique minima are identified, all of which were written out to .res files. This subset of 146,353 comprises an energy window of +15.00 kJ/mol from the global minimum structure ($U_{\text{latt}}^{\text{GM}}$ = −109.3443 kJ/mol). Following clustering, only 52,666 unique clusters remain in the +15.00 kJ/mol energy window. Relatively good geometry matches to both experimental forms are found well within +5 kJ/mol of the GM ($RMSD_{20} \leq 0.37$ Å), and the highest ranked experimental match (F-II: □) is placed 203rd with $\Delta U_{\text{latt}}^{\text{GM}}$ = +3.94 kJ/mol. For the purpose of providing starting points to the CSP refinement, the acquired Global Search landscape is more than adequate.

Table N.31: CAFF-mNBA Global Search with CrystalPredictorv2.5.2.1 and G16-c01-avx2.

Energy Model	
QM Model[a]	PBE0/6-311G(d,p), PCM with $\varepsilon_{\mathrm{PCM}} = 11$
Electrostatic Model	HLYGAt Atomic-charges
HAIEFF FF	$\mathrm{DB14_{pc}}$ (Ref [2], Table G.14)
Candidate Generation Validity Range	
Max. $\Delta U_{\mathrm{intra,0}}$ (kJ/mol)[b]	+50.0
Max. $U_{\mathrm{inter,0}}$ (kJ/mol)[b]	+50.0
Max. $\Delta U_{\mathrm{latt}}^{\mathrm{GM}}$ (kJ/mol)[c]	+30.0
Search Statistics	
Minimizations Attempted	1,200,000
Stable Crystals Accepted	577,185
Landscape Analysis Statistics	
Stable Crystals Analyzed	152,888 (only +15.00 kJ/mol window from GM considered)
Unique Minima Identified[d]	146,353
Minima Written to .res	146,353
Clustering Statistics	
Res Available	146,353
Res Clustered	146,353
Unique Clusters Identified	52,666
Energy Window (kJ/mol)	$[-109.3443, -94.3443]$

[a] The same QM settings are used for the intramolecular and electrostatic model.

[b] Initial crystal structures must satisfy these criteria to qualify for U_{latt} minimization.

[c] Stable minima are only accepted if they are within this energy window from the incumbent GM.

[d] Minima with *identical* geometry and energy are clustered. This is a stringent form of clustering.

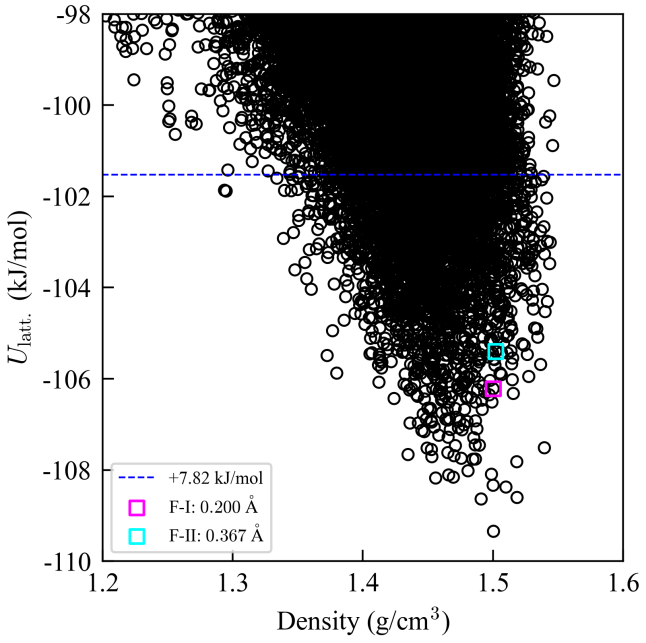


Figure N.39: Global Search landscape for CAFF-mNBA. Matches to F-I and F-II are indicated in magenta and cyan respectively, and labeled with their $\mathrm{RMSD_{20}}$ scores.

N.5.4 Refinement Overview

Table N.32: Refinement I of CAFF-mNBA, conducted with CSO–FMv1.3.2, G09-d01, and GDMAv2.2. Model numbering corresponds to those shown in Table 9.21.

Setting	Model 1	Model 2	Model 3
Global Search Cutoffs			
Structures Refined		3,000	
$\boldsymbol{\Delta U_{\text{latt}}^{\text{GM}}}$ **(kJ/mol)**		+7.82	
Intramolecular Model			
CAFF Flexible Dofs		Rigid	
mNBA Flexible Dofs		Dihedrals: 9, 11, and 13	
Intra. LAM $\boldsymbol{\Delta_{\text{tol}}}$ **Bond Length (Å)**		0.05	
Intra. LAM $\boldsymbol{\Delta_{\text{tol}}}$ **Bond Angle (°)**		1.50	
Intra. LAM $\boldsymbol{\Delta_{\text{tol}}}$ **Torsion Angle (°)**		2.50	
Level of Theory		PBE0/6-311G(d,p)	
PCM?	$\varepsilon_{\text{PCM}} = 11$	Static	$\varepsilon_{\text{PCM}} = 11$
Intermolecular Electrostatic Model			
Elec. LAM $\boldsymbol{\Delta_{\text{tol}}}$ **Bond Length (Å)**		0.05	
Elec. LAM $\boldsymbol{\Delta_{\text{tol}}}$ **Bond Angle (°)**		1.50	
Elec. LAM $\boldsymbol{\Delta_{\text{tol}}}$ **Torsion Angle (°)**		2.50	
Level of Theory		PBE0/6-311G(d,p)	
PCM?	$\varepsilon_{\text{PCM}} = 11$	Static	$\varepsilon_{\text{PCM}} = 11$
Electrostatic Model		Rank 4, GDMA Multipoles ($r_{\text{H}}^{\text{GDMA}} = 0.650$ Å)	
Electrostatic Damping		None	
Intermolecular Force-field Model			
Force-field Label	DB14_{mp}	$\text{BT}_{\text{Mie,mp}}^{\text{Static}}$	$\text{BT}_{\text{Mie,mp}}^{\text{PCM}}$
Force-field Parameters	Ref [2], Table I.14	Table J.20	Table J.27
Refined Landscape Statistics			
No. of Crystals Post-Clustering	2,449	2,479	2,397
$\boldsymbol{U_{\text{latt}}^{\text{GM}}}$ **(kJ/mol)**[a]	−107.92	−120.15	−120.26
F-I Found?	✓	✓	✓
F-II Found?	✓	✓	✓

[a] This is decided by Equation 7.4.14. Without SCERP, the default τ_{nag} is 1×10^{-8}.

[b] Note, ΔU_{intra} for Models 1 and 3 are defined with respect to the PCM_{11} gas-phase energy in Table N.27, while for Model 2 the Static gas-phase energy in Table N.27 is used.

Table N.33: Refinement II of CAFF-mNBA, conducted with CSO–FMv1.4.0.2, G09-d01, and GDMAv2.2. These settings correspond to Model 4 in Table 9.21.

Setting	**Model 4**		
Refinement I Cutoffs			
Refinement I Landscape	Model 3 from Table N.32		
Structures Refined	120		
$\boldsymbol{\Delta U_{\text{latt}}^{\text{GM}}}$ **(kJ/mol)**	+8.87		
Intramolecular Model			
CAFF Flexible Dofs	Rigid		
mNBA Flexible Dofs	Dihedrals: 9, 11, and 13		
Intra. LAM $\boldsymbol{\Delta_{\text{tol}}}$ **Bond Length (Å)**	0.05		
Intra. LAM $\boldsymbol{\Delta_{\text{tol}}}$ **Bond Angle (°)**	1.50		
Intra. LAM $\boldsymbol{\Delta_{\text{tol}}}$ **Torsion Angle (°)**	2.50		
Level of Theory	PBE0/6-311G(d,p)		
PCM?	Static		
Intermolecular Electrostatic Model			
Elec. LAM $\boldsymbol{\Delta_{\text{tol}}}$ **Bond Length (Å)**	0.05		
Elec. LAM $\boldsymbol{\Delta_{\text{tol}}}$ **Bond Angle (°)**	1.50		
Elec. LAM $\boldsymbol{\Delta_{\text{tol}}}$ **Torsion Angle (°)**	2.50		
Level of Theory	PBE0/6-311G(d,p)		
PCM?	Static		
Electrostatic Model	Rank 4, GDMA Multipoles ($r_{\text{H}}^{\text{GDMA}} = 0.650$ Å)		
Electrostatic Damping	Using $\beta_{\text{atm}}^{tt'}$ based on $\%f_6 = 90\%$ in Table L.1		
Intermolecular SCERP Model			
SCERP-Field Representation	HLYGAt Atomic-Charges		
$\boldsymbol{r_{\text{scrp}}}$ **(Å)**	40.0		
$\boldsymbol{t_{\text{slow}}}$	6		
$\boldsymbol{\varepsilon_{\text{scrp}}}$ **(a.u.)**	1×10^{-5}		
$\boldsymbol{t_{\text{max}}}$	30		
$\boldsymbol{\varepsilon_{\text{scrpRM}}}$ **(kJ/mol)**	1×10^{-5}		
$\boldsymbol{k_{\text{max}}}$	16		
Intermolecular Force-field Model			
Force-field Label	$\text{BT}_{\text{Mie,mp}}^{\text{SCRP}}$		
Force-field Parameters	Table M.4		
Refined Landscape Statistics			
CSO–FM Optimality Tolerance[a]	2.5×10^{-3}		
FD Mode (per Table 7.6)	2	3	4
No. of Crystals Post-Clustering	119	119	119
$\boldsymbol{U_{\text{latt}}^{\text{GM}}}$ **(kJ/mol)**[b]	−119.90	−119.92	−119.90
F-I Found?	✓	✓	✓
F-II Found?	✓	✓	✓

[a] This is decided by Equation 7.4.14. Without SCERP, the default τ_{nag} is 1×10^{-8}.

[b] Note, all ΔU_{intra} are defined with respect to the Static gas-phase energy in Table N.27.

N.5.5 Assessment of SCERP Gradient Approximations

N.5.5.1 Landscape Minimization by FD Modes 2 to 4

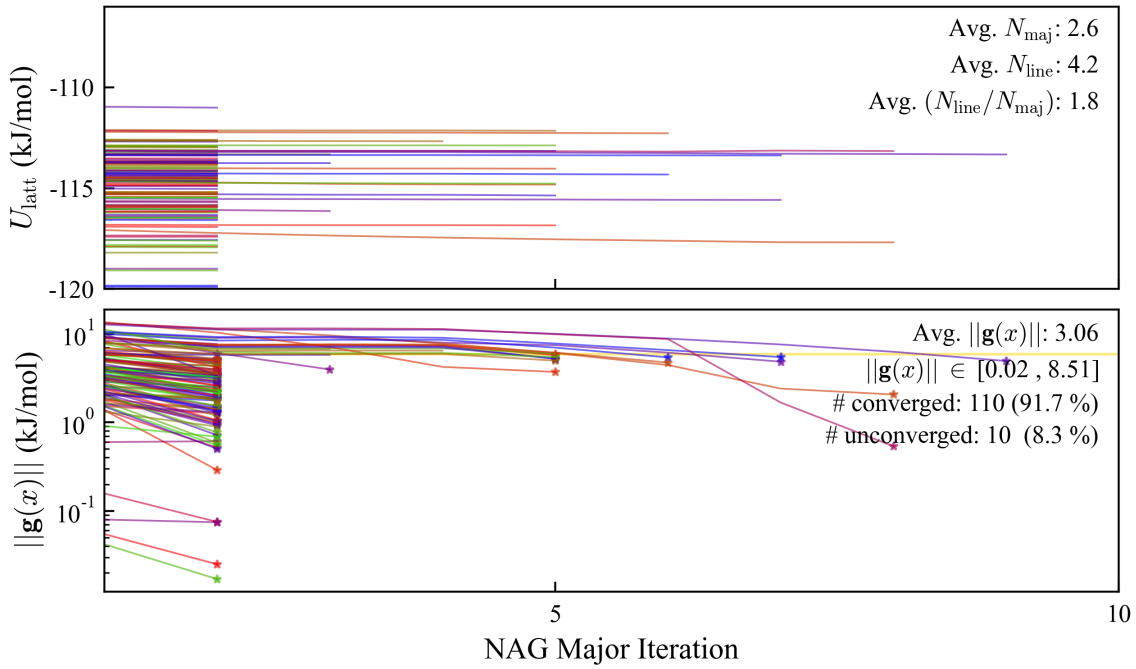


Figure N.40: Minimization trajectories of CAFF-mNBA Refinement II by FD Mode 2: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $||\boldsymbol{g}(x)||$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★.

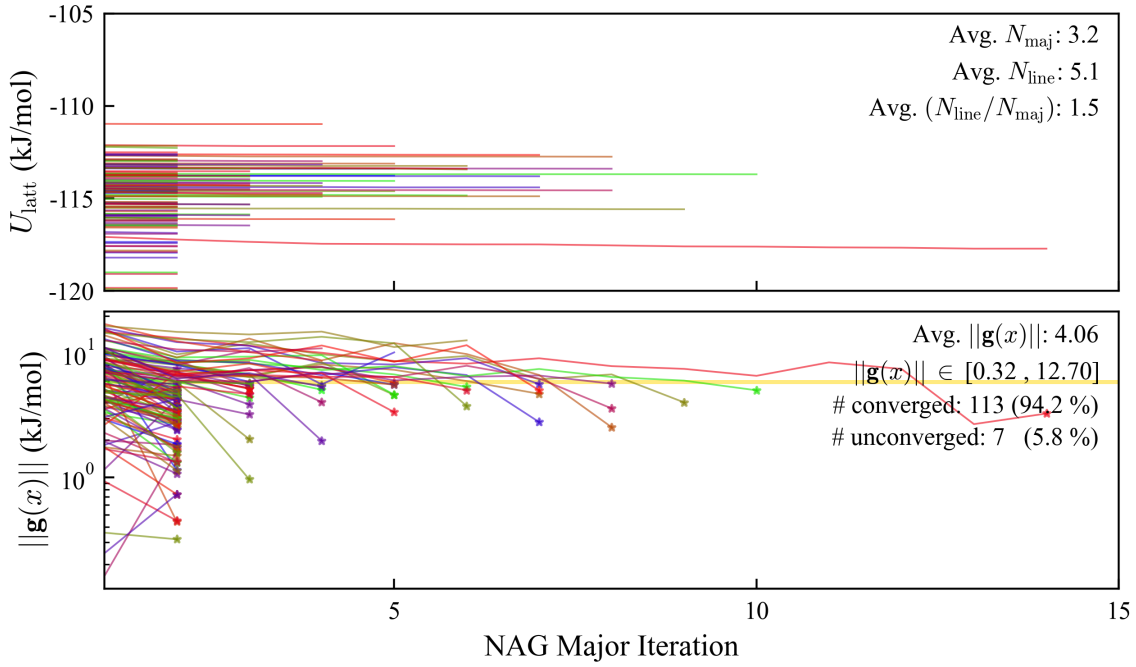


Figure N.41: Minimization trajectories of CAFF-mNBA Refinement II by FD Mode 3: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $||\boldsymbol{g}(x)||$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★.

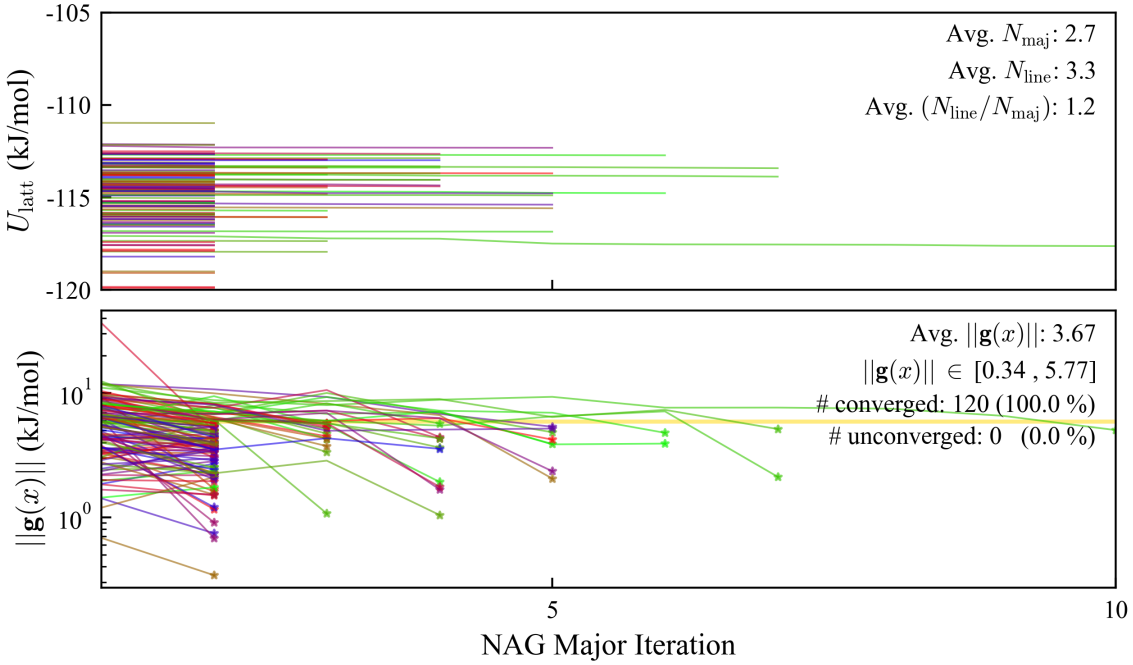


Figure N.42: Minimization trajectories of CAFF-mNBA Refinement II by FD Mode 4: U_{latt} values (top), and gradient vector norms (bottom). The yellow region in the $||\boldsymbol{g}(x)||$ plot approximates convergence, according to Equation 7.4.11. Converged structures are denoted by ★.

Table N.34: Optimised U_{latt} and geometries for CAFF-mNBA predicted by Modes 2 and 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$, while RMSD_{20} scores are generated by the COMPACK algorithm. All comparisons are on a structure-by-structure basis.

Mode	AD U_{latt} (kJ/mol)	AAD U_{latt} (kJ/mol)	No. $\lvert\Delta U_{\text{latt}}\rvert$ > 0.5 kJ/mol	Avg. RMSD_{20} (Å)	Max. RMSD_{20} (Å)	No. RMSD_{20} > 0.5 Å
2	0.005	0.011	0	0.008	0.068	0
3	0.000	0.011	0	0.011	0.085	0

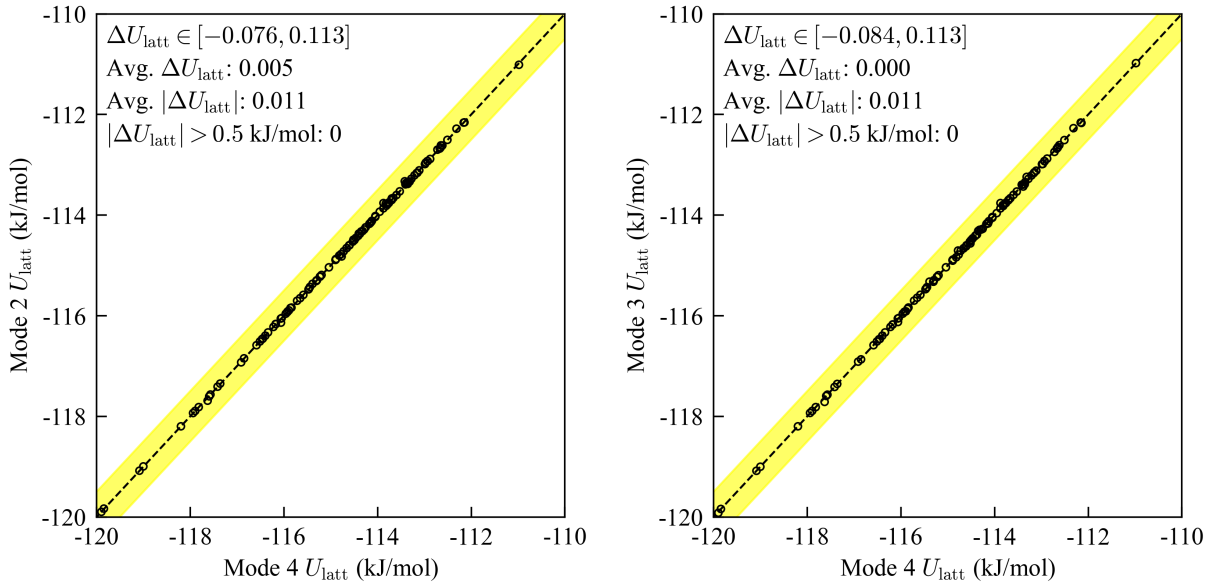
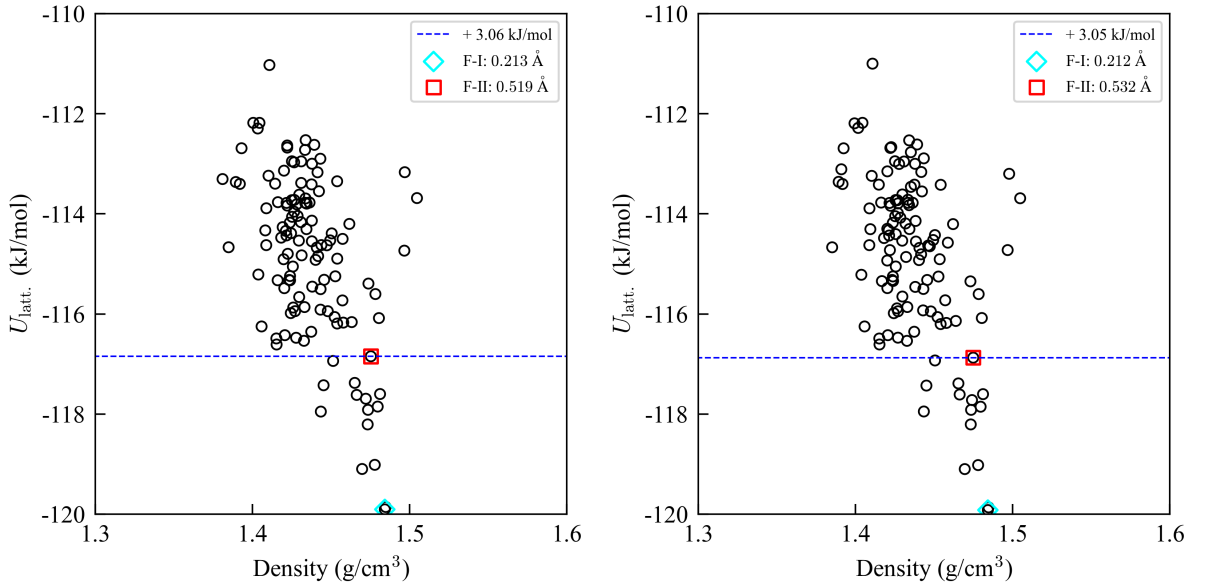


Figure N.43: Comparison of optimized CAFF-mNBA U_{latt} obtained using (a) Mode 2 and (b) Mode 3, relative to Mode 4. ΔU_{latt} are determined as $U_{\text{latt}}^{\text{Mode 2/3}} - U_{\text{latt}}^{\text{Mode 4}}$. All comparisons are on a structure-by-structure basis. The yellow region indicates $|\Delta U_{\text{latt}}| \leq 0.5$ kJ/mol.



Form	RMSD_{20} (Å)	$\Delta U_{\text{latt}}^{\text{GM}}$ (kJ/mol)	Rank
F-I	0.2132	+0.0000	1
F-II	0.5189	+3.0586	15

Form	RMSD_{20} (Å)	$\Delta U_{\text{latt}}^{\text{GM}}$ (kJ/mol)	Rank
F-I	0.2122	+0.0000	1
F-II	0.5316	+3.0455	15

Figure N.44: Refined landscape of CAFF-mNBA using Model 4, with Mode 2 (left) and Mode 3 (right) FD gradient approximations. The lowest-energy match to F-I and F-II are labeled as ◇ and □ respectively, with their corresponding RMSD_{20} scores.