

Molecular Flexibility in Crystal Structure Prediction

Andrey Kazantsev

2011

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

Centre for Process Systems Engineering
Department of Chemical Engineering
Imperial College London
London SW7 2AZ
United Kingdom

Abstract

The packing of molecules in solids greatly affects the properties of the bulk materials. This is particularly important for the pharmaceutical industry, where the discovery of crystal forms at a late stage of process development can have disastrous consequences. As a result, the importance of polymorphism in crystal structures of organic molecules has been recognised for many years. This thesis presents computational developments that can complement experimental form screening of molecules for which conformational flexibility is significant.

Current methods for crystal structure prediction are limited by the extent of molecular flexibility that can be practically handled due to the prohibitive computational cost associated with quantum mechanical calculations integrated in most of the successful approaches. In order to reduce the number of quantum mechanical evaluations, local approximate models can be defined for the estimation of the intramolecular energy, molecular geometry and the conformationally dependent intermolecular electrostatic model.

A novel algorithm, CrystalOptimizer, for the accurate local minimisation of the lattice energy of crystals involving flexible organic molecules is presented. The main novelty of the algorithm is the use of dynamically constructed and updated local approximate models which essentially make available the full accuracy of quantum mechanical models at each and every iteration of the minimisation algorithm, requiring only a small number of explicit quantum mechanical calculations. This has made possible the accurate treatment of molecules involving a relatively large numbers of atoms with significant flexibility in torsional and bond angles and even bond lengths. The performance of the algorithm is critically assessed and demonstrated on a set of single and multi-component crystals.

An extension of an existing algorithm for the identification of low energy crystal structures of flexible molecules, CrystalPredictor, is also described. In the proposed modification, the intramolecular energy and the molecular conformation are modelled using local approximate models. This provides a more realistic model for the effects of the flexible degrees of freedom on the molecular geometry and lattice energy. The use of deterministic low-discrepancy sequences ensures an extensive and uniform coverage

of the multivariable search space. A parallelised implementation of the algorithm allows minimisations from several hundreds of thousands of initial guesses to be carried out in reasonable time. A further computational benefit is derived by the storage of the information used to construct the local approximate models in databases, which can be re-used in subsequent re-minimisation of structures with more accurate models for the lattice energy. The usefulness of these modifications is demonstrated on the ROY molecule, for which the structures of all experimentally known polymorphs are identified by the algorithm.

By combining the above algorithms, a comprehensive multi-stage methodology for *ab initio* determination of the crystal structure of a given molecule based solely on its atomic connectivity is presented. The application of the methodology to two large and flexible molecules of pharmaceutical interest is also demonstrated.

Declaration

I confirm the work presented in this thesis is my own. All the information derived from other sources is appropriately referenced.

Andrey Kazantsev

Acknowledgements

Foremost, I would like to thank my supervisors Professor Costas Pantelides and (now) Professor Claire Adjiman for their guidance and support throughout my studies. They introduced to me and allowed me to explore the academic world (literally), and have always been perceptive and accommodating to my needs and requirements making my PhD experience seamless. By sharing their vision and expertise during many stimulating discussions, they taught me a new way to view research and deal with complex problems. These lessons will be invaluable in my future life.

A key person in my PhD has been Panos Karamertzanis. Not only did he prove to be the first point of reference for any query (a walking encyclopaedia indeed!), but his drive and dedication were the motivating factors behind my own work. There are no challenges that he is not willing (and able) to tackle and I am grateful to him for this stimulating attitude.

The close contacts with Sally Price's group at UCL demonstrated the value of academic collaboration. I have relied on their expertise to gain insight into many fundamental aspects of crystal structure prediction and crystallography.

The development and validation of the work presented in this thesis required considerable computational resources. Having access to the Imperial College High Performance Computing Cluster, which I actively used (and sometimes abused) proved instrumental in completing this work.

The completion of the PhD requires a creative environment where ideas can flourish. As a result, I am most grateful to the whole of the Molecular Systems Engineering group for the warm and stimulating atmosphere that they allowed me to be part of.

Finally, this work would not be possible without the financial support from the Engineering and Physical Sciences Research Council (EPSRC) under the Molecular Systems Engineering grant (EP/E0163340) and the Overseas Research Students Awards Scheme (ORSAS).

Конечно же, не имею права не поблагодарить свою семью, и особенно родителей, за безвозмездную поддержку, оказанную в течение этих лет, проведённых в научной среде. Я обязан вам всем, что у меня есть и чего сейчас добился.

Ещё, я очень благодарен своему другу, Женьке, с которым мы, почти каждый день, встречались для того, чтобы позабыв о рабочей суете, обсудить наболевшее или поскрипеть мозгами над очередной головоломкой. Наши перерывы на обед (и кофе) надолго останутся в моей памяти.

Напоследок, мне остаётся поблагодарить самого мне дорогого человека - мою жену Алёночку. Она никогда не давала мне падать духом и преобразила мою жизнь, огородив меня заботой и создав уют, необходимый для творческой (и технической) деятельности.

Все плоды моей работы, в частности, ваши тоже!

Contents

Abstract.....	2
Declaration.....	3
Acknowledgements.....	4
Contents....	6
List of Publications	10
Conference Communications.....	11
List of Tables	14
List of Figures	17
Chapter 1. Introduction.....	22
1.1. Background.....	22
1.2. Scope and outline of the work	24
Chapter 2. Literature Review.....	27
2.1. Crystal lattices.....	27
2.1.1. Crystal structure representation	27
2.1.2. Thermodynamic versus kinetic considerations	29
2.2. Lattice energy calculation	31
2.2.1. Intramolecular energy calculation – ΔE^{intra}	32
2.2.2. Intermolecular energy calculation – U^{inter}	33
<i>Electrostatic interactions</i> – U^{elec}	34
<i>Repulsion-dispersion interactions</i> – U^{vdw}	35
<i>Other contributions to the intermolecular energy</i> – U^{other}	37
2.2.3. Infinite summation of intermolecular energy	38
2.2.4. Alternative methods for lattice energy calculation	39
2.3. Molecular flexibility	41
2.4. Crystal structure prediction methods for flexible molecules	43
2.4.1. Crystal structure generation	45
<i>Construction of molecular clusters</i>	46

<i>Unit cell generation</i>	48
2.4.2. Local minimisation of lattice energy	53
2.4.3. <i>A posteriori</i> analysis of low-energy crystal structures.....	56
2.4.4. CCDC Blind Tests of Crystal Structure Prediction	58
2.5. Conclusions.....	63
Chapter 3. Local Approximate Models (LAMs)	65
3.1. Molecular Model.....	65
3.2. Treatment of Molecular Flexibility.....	67
3.3. Intramolecular Energy LAM.....	68
3.4. Electrostatic Model LAM	71
3.5. Range of LAM validity	73
3.6. Re-usability of QM calculations (LAM/QM databases).....	77
Chapter 4. Local minimisation of lattice energy – the CrystalOptimizer algorithm	80
4.1. Formulation of the lattice energy minimisation problem	80
4.2. DMACRYS – inner minimisation algorithm.....	82
4.3. CrystalOptimizer – outer minimisation algorithm.....	82
4.4. Algorithm validation, Results and Discussion.....	89
4.4.1. Systems studied.....	89
4.4.2. Molecular modelling and computational considerations	91
4.4.3. Use of linear updates for multipole LAMs	92
4.4.4. Application to single component crystals	94
4.4.5. Application to multicomponent crystals	102
4.4.6. Application to ab initio Crystal Structure Prediction.....	104
4.5. Conclusions.....	107
Chapter 5. Global minimisation of lattice energy – extension to CrystalPredictor	109
5.1. Description and limitations of CrystalPredictor	109
5.1.1. Algorithm overview	110
5.1.2. Representation of molecular flexibility	112
5.2. Geometric modelling of flexible molecules.....	115
5.3. Lattice energy minimisation of flexible molecules.....	119
5.3.1. Minimisation variables.....	119
5.3.2. Intramolecular energy and molecular conformation.....	121

5.3.3. Generation of LAMs for CrystalPredictor	124
5.3.4. Conformationally-dependent electrostatic models	127
5.4. Illustrative example.....	128
5.4.1. Model system ROY.....	128
5.4.2. Computational model and search strategies.....	129
5.4.3. Results of crystal structure search.....	132
5.5. Conclusions.....	136
Chapter 6. Crystal structure prediction of large flexible molecules	139
6.1. Crystal Structure Prediction methodology.....	139
6.1.1. Step 1: Computational model selection and conformational analysis .	140
6.1.2. Step 2: Crystal structure generation	143
6.1.3. Step 3: Crystal structure refinement	145
6.1.4. Step 4: Identification of relevant minima on the lattice energy surface	146
6.2. Crystal structure prediction of a candidate active pharmaceutical ingredient	147
6.2.1. Introduction.....	147
6.2.2. Step 1: Computational model selection and conformational analysis .	148
6.2.3. Step 2: Crystal structure generation	153
Case 1 – multiple rigid-body searches.....	153
Case 2 – explicit treatment of flexible torsions as search variables.....	155
6.2.4. Steps 3 & 4: Crystal structure refinement and relative lattice energies	158
6.2.5. Approximate computational cost	161
6.2.6. Conclusions.....	162
6.3. Target molecule XX in the Fifth Blind Test of Crystal Structure Prediction	164
6.3.1. Introduction.....	164
6.3.2. Step 1: Computational model selection and conformational analysis .	165
6.3.3. Step 2: Crystal structure generation	168
6.3.4. Steps 3 & 4: Crystal structure refinement and relative lattice energies	172
6.3.5. Approximate computational cost	177
6.3.6. Conclusions.....	178
6.4. Conclusions.....	179

Chapter 7. Conclusions and directions for future work.....	181
Bibliography	184
Appendix A Proposed Z-matrix generation procedure	204
Appendix B Derivation of the approximation of the “rigid” degrees of freedom as an explicit linear function of the flexible degrees of freedom.....	209
Appendix C An alternative derivation of the approximation of the “rigid” degrees of freedom as an explicit linear function of the flexible degrees of freedom	212
Appendix D Derivation of the intramolecular energy LAM as an explicit function of the flexible degrees of freedom	214
Appendix E Description of molecular rotation through quaternions	219
Appendix F Several low-energy hypothetical structures for the BMS molecule	221

List of Publications

1. Vasileiadis, M., Kazantsev, A. V., Adjiman, C. S., and Pantelides, C. C.; "Ab Initio Prediction of ROY polymorphs", **2011**, in preparation.
2. Bardwell, D. A, Adjiman, C. S., Ammon, H. L., Arnautova, E. A., Bartashevich, E., Boerrigter, S. X. M., Braun, D. E., Cruz-Cabeza, A. J., Day, G. M., Della Valle, R. G., Desiraju, G. R., van Eijck, B. P., Facelli, J. C., Ferraro, M. B., Grillo, D., Habgood, M, Hofmann, D. W. M., Hofmann, F., Jose, J., Karamertzanis, P. G., Kazantsev, A. V., Kendrick, J., Kuleshova, L. N., Leusen, F. J. J., Maleev, A., Misquitta, A. J., Mohamed, S., Needs, R. J., Neumann, M. A., Nikylov, D., Orendt, A. M., Pal, R., Pantelides, C. C., Pickard, C. J., Price, L. S., Price, S. L., Scheraga, H. A., van de Streek, J., Thakur, T. S., Tiwari, S., Venuti, E., and Zhitkov, I.; "Towards crystal structure prediction of pharmaceutically relevant molecules - a report on the fifth blind test.", *Acta Crystallographica Section B - Structural Science*, **2011**, submitted.
3. Kazantsev, A. V., Karamertzanis, P. G., Adjiman, C. S., Pantelides, C. C., Price, S. L., Galek, P. T. A, Day, G. M., and Cruz-Cabeza, A. J.; "Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction", *International Journal of Pharmaceutics*, **2011**, in press.
4. Kazantsev A. V., Karamertzanis, P. G., Adjiman, C. S., and 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.
5. Kazantsev, A. V., Karamertzanis, P. G., Pantelides, C. C., and Adjiman, C. S.; "An Efficient Algorithm for Lattice Energy Minimization of Organic Crystals Using Isolated-Molecule Quantum Mechanical Calculations", in *Process Systems Engineering: Volume 6: Molecular Systems Engineering*, Adjiman, C. S. and Galindo, A., Wiley-VCH, Hamburg, **2010**, 1-42.

6. Kazantsev A. V., Karamertzanis, P. G., Adjiman, C. S., and Pantelides, C. C.; "Ab Initio Crystal Structure Prediction for Flexible Molecules", *20th European Symposium on Computer Aided Process Engineering – ESCAPE20*, **2010**, 28, 817-822.
7. Karamertzanis, P. G., Kazantsev, A. V., Issa, N., Welch, G. W. A., Adjiman, C. S., Pantelides, C. C., and Price, S. L.; "Can the Formation of Pharmaceutical Cocrystals Be Computationally Predicted? II. Crystal Structure Prediction.", *Journal of Chemical Theory and Computation*, **2009**, 5, 5, 1432-1448.

Conference Communications

1. Kazantsev A. V., Vasileiadis M., Karamertzanis P.G., Adjiman, C. S., and Pantelides, C. C.; "Predicting Crystal Structures for Flexible Organic Molecules", *CPSE Consortium Meeting*, London, UK, May **2011**.
2. Vasileiadis M., Kazantsev A. V., Karamertzanis P.G., Adjiman, C. S., and Pantelides, C. C.; "Polymorph Prediction for a Challenging Molecule", *CAPE-FORUM*, Bradford, UK, March **2011**.
3. Kazantsev A. V., Karamertzanis P.G., Adjiman, C. S., and Pantelides, C. C.; "Recent Advances in Computational Techniques for Crystal Structure Prediction", *CAPE-FORUM*, Bradford, UK, March **2011**.
4. Kazantsev A. V., Karamertzanis P.G., Adjiman, C. S., and Pantelides, C. C.; "Achieving QM Accuracy at a Fraction of Computational Cost: Lattice Energy Minimisation of Organic Crystals Containing highly Flexible Molecule", *AIChE Annual Meeting*, Salt Lake City, UT, USA, November **2010**.

5. Kazantsev A. V., Karamertzanis P.G., Adjiman, C. S., and Pantelides, C. C.; "Ab Initio Crystal Structure Prediction of Flexible Molecules", *PPEPPD Conference*, Suzhou, China, May **2010**.
6. Kazantsev A. V., Karamertzanis P.G., Adjiman, C. S., and Pantelides, C. C.; "Accurate Lattice Energy Minimisation of Organic Crystals Containing Flexible Molecules", *CPOSS Open Day*, London, UK, March **2010**.
7. Kazantsev A. V., Karamertzanis P.G., Adjiman, C. S., and Pantelides, C. C.; "Accurate Lattice Energy Minimisation of Organic Crystals Containing Flexible Molecules", *Chemical Engineering PhD Research Symposium*, Imperial College London, London, UK, March **2010**.
8. Kazantsev A. V., Karamertzanis P.G., Adjiman, C. S., and Pantelides, C. C.; "Ab Initio Crystal Structure Prediction of Flexible Molecules", *AIChE Annual Meeting*, Nashville, TN, USA, November **2009**.
9. Kazantsev A. V., Karamertzanis P.G., Adjiman, C. S., and Pantelides, C. C.; "Accurate Lattice Energy Minimisation of Cocrystals, Salts and Solvates Containing Flexible Molecules", *AIChE Annual Meeting*, Nashville, TN, USA, November **2009**.
10. Kazantsev A. V., Karamertzanis P. G., Adjiman, C. S., and Pantelides, C. C.; "Crystal Structure Prediction of Pharmaceutical Cocrystals", *Thermodynamics Conference*, London, UK, September **2009**.
11. Kazantsev A. V., Karamertzanis P. G., Adjiman, C. S., and Pantelides, C. C.; "A Computationally Efficient Algorithm for Accurate Local Energy Minimisation of Crystal Structures Containing Flexible Molecules", *FOMMS Conference*, Blaine, WA, USA, July **2009**.
12. Kazantsev A. V., Karamertzanis P. G., Adjiman, C. S., and Pantelides, C. C.; "Computational Effects of Pressure on Organic Crystal Structure", *FOMMS Conference*, Blaine, WA, USA, July **2009**.

13. Kazantsev A. V., Karamertzanis P. G., Adjiman, C. S., and Pantelides, C. C.;
"Computational Prediction of Effects of Pressure on Organic Crystal Structure",
CPOSS Annual Meeting, London, UK, March **2009**.
14. Kazantsev A. V., Karamertzanis, P. G., Adjiman, C. S., and Pantelides, C. C.;
"A Computationally Efficient Algorithm for Accurate Local Energy
Minimisation of Crystal Structures Containing Flexible Molecules", *AIChE
Annual Meeting*, Philadelphia, PA, USA, November **2008**.
15. Kazantsev A. V., Karamertzanis, P. G., Adjiman, C. S., and Pantelides, C. C.;
"Computational Prediction of Effects of Pressure on Organic Crystal Structure",
AIChE Annual Meeting, Philadelphia, PA, USA, November **2008**.
16. Karamertzanis P.G., Kazantsev A. V., Adjiman, C. S., and Pantelides, C. C.;
"Modelling and Prediction of the Crystal Structure of Pharmaceutical
Cocrystals", *AIChE Annual Meeting*, Philadelphia, PA, USA, November **2008**.

List of Tables

Table 1.1 Typical structure complexity that a) has been considered in crystal structure prediction studies as shown by the four targets in the fourth blind test of crystal structure prediction, CSP2007 (Day et al. 2009); b) needs to be considered for pharmaceutical targets.	25
Table 2.1 Reviewed methods for crystal structure generation.	44
Table 2.2 Reviewed methods for the accurate evaluation of the lattice energy (refinement) of crystal structures of flexible molecules.	45
Table 2.3 Diagrams and reference numbers (in Roman numerals) of the target molecules in the blind tests of crystal structure prediction. The three values below the target reference numbers indicate, from left to right, the total number of participants, the number of participants who obtained the correct experimental structure in their extended list and among their three best predictions for each target respectively.....	60
Table 3.1 Sample Z-matrix for the gas-phase geometry of formamide computed at MP2/6-31G(d,p) level of theory. Refer to Figure 3.1 for molecular diagram and atom labels.	66
Table 4.1 List of optimisation parameters in CrystalOptimizer.....	85
Table 4.2 Systems considered for CrystalOptimizer refinement.	90
Table 4.3 Detailed lattice energy minimisation results for the experimental structure of Xylitol.	97
Table 4.4 Detailed lattice energy minimisation results for the experimental structure of α -D-glucose.....	98
Table 4.5 Detailed lattice energy minimisation results for the experimental structure of piracetam Form II.	99
Table 4.6 Lattice energy minimisation of additional crystal structures containing α -D-glucose or piracetam in the asymmetric unit. Refer to Table 4.2 for molecular diagrams. The number of flexible degrees of freedom considered for each molecule during the minimisation is given in square brackets.	103
Table 4.7 Lattice energy minimisation of selected crystal structures of varying size and complexity. Refer to Table 4.2 for molecular diagrams. The number of flexible degrees of freedom considered for each molecule during the minimisation is given in square brackets.	104

Table 5.1 Structure reproduction and rank of ROY polymorphs obtained after search Case 2 and Case 3 respectively. Structures are shown in order of decreasing stability at a temperature of 40° C (Yu et al. 2000, Chen et al. 2005).....	137
Table 6.1 Crystal structure prediction methodology adopted in this work.	141
Table 6.2 Algorithm characteristics for crystal structure generation and crystal structure refinement steps of the proposed multistage methodology (Table 6.1).	144
Table 6.3 Feasible ranges of torsion values for the BMS molecule. (Refer to Figure 6.1 for torsion definitions). The conformations of the experimental forms are also shown.	153
Table 6.4 Molecular conformations considered for the preliminary rigid-body search. The molecular conformations of the two minimised experimental forms are also provided (Refer to Figure 6.1 for torsion definitions).	155
Table 6.5 Search definition and ranges of torsional angles considered explicitly during the (Refer to Figure 6.1 for torsion definitions).	156
Table 6.6 Unit cell parameters and structure similarity between experimental and the predicted structures. For the space groups shown $\alpha = \gamma = 90^\circ$	161
Table 6.7 Outline of the computational cost associated with structure generation and refinement (steps 2 and 3 in the crystal structure prediction methodology described in section 6.1) for the specific case of the BMS molecule.	162
Table 6.8 The unit cell parameters, rank and structure similarity between the experimental and predicted structures of target XX by three groups participating in the fifth blind test of crystal structure prediction (Bardwell et al. 2011). For the space group shown $\alpha = \gamma = 90^\circ$	165
Table 6.9 Energetically feasible torsion ranges for molecule XX. The torsion values in the experimental conformation are also given. Refer to Figure 6.9 for torsion definitions.	168
Table 6.10 Search enumeration and ranges of torsional angles considered explicitly during the (Refer to Figure 6.9 for torsion definitions).	170
Table 6.11 The unit cell parameters for the experimental and predicted structures. For the space group shown $\alpha = \gamma = 90^\circ$	175
Table 6.12 Outline of the computational cost associated with structure generation and refinement (steps 2 and 3 in the crystal structure prediction methodology described in section 6.1) for the specific case of molecule XX in the fifth blind test of crystal structure prediction.	178
Table A.1 Gas-phase minimum conformation, computed at the MP2/6-31G(d,p) level of theory, used to generate the Z-matrix for ethyl ethanoate (refer to Figure A.1 for atom labels).	205

Table A.2 Generated Z-matrix for the gas-phase geometry of ethyl ethanoate (Table A.1). Refer to Figure A.1 for molecular diagram and atom labels. 205

Table F.1 Crystal structures and molecular conformations of several plausible single-component crystal forms of the BMS molecule. Rank, lattice energy, density, space group and unit cell parameters are also provided for every structure. 221

List of Figures

- Figure 1.1** Spatial arrangements of different polymorphs of piracetam (2-oxo-pyrrolidine-acetamide) in the crystal: a) Form III (Galdecki and Glowka 1983); b) Form IV (Fabbiani et al. 2005). 22
- Figure 1.2** Seven ROY (5-methy-2-[(2-nitrophenyl) amino]-3-thiophenecarbonitrile) polymorphs for which the crystalline structure has been solved experimentally. Adapted from (Yu et al. 2000, Chen et al. 2005). 23
- Figure 2.1** Lattice vectors (a , b , c) and angles (α , β , γ) defining the unit cell in piracetam Form II crystal (Admiraal et al. 1982). 28
- Figure 2.2** Lattice energy (top right), intermolecular energy (bottom left) and intramolecular energy (bottom right) as a function of two torsional angles for the experimental crystal structure of Xylitol (1,2,3,4,5-pentapentanol, top left) (Kim and Jeffrey 1969). Calculations performed using the methodology and algorithm described in Chapter 4. 42
- Figure 3.1** Molecular diagram and atom labels of formamide. 66
- Figure 3.2** One flexible torsion, θ^f , shown as a black arrow captures most of the flexibility in 4,4'-bipyridine. The geometry of the aromatic rings can be assumed to be “rigid” for the purpose of lattice energy minimisation. 67
- Figure 3.3** Molecular diagram and atom labelling for piracetam (2-oxo-pyrrolidine-acetamide). 73
- Figure 3.4** (a) QM intramolecular energy, $\Delta E^{\text{intra, QM}}$ at the HF/6-31G(d,p) level of theory and (b) LAM error for the intramolecular energy defined as $\Delta E^{\text{intra, QM}} - \Delta E^{\text{intra, LAM}}$ as a function of two “flexible” dihedral angles (blue arrows in Figure 3.3) for piracetam. Open circles correspond to the reference point for the LAM. 74
- Figure 3.5** Difference between the quantum mechanical and estimated LAM values for selected “rigid” degrees of freedom (orange arrows top left) as a function of two “flexible” dihedral angles (blue arrows in Figure 3.3) for piracetam (2-oxo-pyrrolidine-acetamide). QM calculations performed at the HF/6-31G(d,p) level of theory. Open circles correspond to the reference point for the LAM. 75
- Figure 3.6** Error in the electrostatic potential energy (in eV; $1\text{eV} = 96.5\text{ kJ mol}^{-1}$) on the surface of xylitol (1,2,3,4,5-pentapentanol) at twice the van der Waals radii as a function of the H1-O1-C1-C2 torsional angle. LAM used to estimate the rigid degrees of freedom and rotate the atomic multipole moments with their local environment. The quantum mechanical electrostatic potential energy in eV for the reference molecular conformation is also shown for comparison (centre). All electrostatic potential energy surfaces were computed with atomic multipoles up to the hexadecapole level, at the

PBE/6-31G(d,p) level of theory, using ORIENT (Stone et al. 2006). The van der Waals radius for hydroxyl hydrogen atoms was set to 1 au, the radii for other atoms were taken from (Bondi 1964)..... 76

Figure 4.1 CrystalOptimizer local lattice energy minimisation algorithm flowchart. Refer to Figure 4.2 for the procedures for LAM construction (LAM_Construct), lattice energy (Lattice_Energy) and lattice energy derivative (Lattice_Gradients) calculations. 83

Figure 4.2 Procedures for lattice energy derivative (top left) and lattice energy (right) calculations and LAM construction (bottom left) in CrystalOptimizer. $\dim[\theta^f]$ denotes the dimensionality of vector θ^f 84

Figure 4.3 Pseudo code for the CrystalOptimizer local lattice energy minimisation algorithm. Refer to Table 4.1 for the list of optimisation parameters and Figure 4.2 for the procedures for LAM construction (LAM_Construct), lattice energy (Lattice_Energy) and lattice energy derivatives (Lattice_Gradients) calculation. $\dim[\theta^f]$ denotes the dimensionality of vector θ^f 86

Figure 4.4 Intermolecular energy as a function of the H1-O1-C1-C2 torsional angle (blue arrow in Figure 3.6) for three different computational strategies: 1) molecular geometry and the electrostatic potential (ESP) obtained by full quantum mechanical calculations (black diamonds); 2) LAM used to update rigid degrees of freedom (dof) and multipole rotation (blue dashed line); 3) LAMs used to update the rigid degrees of freedom, multipole rotation and a linear update of the multipole moments (continuous green line). 93

Figure 4.5 Top row: Overlay of the experimental (coloured by element) and minimised (green) structures of xylitol, α -D-glucose and piracetam when full molecular flexibility is allowed (atomistic representation). The root mean square deviation for one molecule (rms_1) and 15-molecule coordination sphere (rms_{15}) (Chisholm and Motherwell 2005) compared to the experimental structure are shown below. Bottom row: The gas-phase molecular conformations for the three molecules..... 100

Figure 4.6 Total number of CrystalOptimizer minimisation iterations (solid), number of LAM updates (dashed) and the actual number of QM evaluation sets necessary to construct LAM/QM database (dotted) as a function of the number of crystal structures minimised..... 106

Figure 5.1 Flowchart of the CrystalPredictor search and minimisation algorithm.... 111

Figure 5.2 Description of flexible molecules used in the original CrystalPredictor algorithm. a) Fragment definition for a molecule with just one flexible torsion (grey arrow). Note that the sulphur atom (yellow) and the nitrogen atom (blue) are part of both fragments (a fact not indicated on the diagram in the interests of clarity). b) Overlay of the molecule geometry before (coloured by element) and after (green) a 60° change in the flexible torsion angle (grey arrow). The geometry of the individual fragments is kept fixed throughout. 113

Figure 5.3 (a) Whole molecule and (b) the terminal conjugated ring (base fragment) overlays (excluding hydrogen atoms) of the molecular geometries obtained using constrained QM geometry optimisation at the PBE0/6-31G(d,p) level of theory (coloured by element) and that obtained by rotating the fragments of the molecule from its gas-phase geometry (also obtained at the PBE0/6-31G(d,p) level of theory). The values of the five flexible torsions (grey arrows) are the same for both molecular models in both cases. Detailed crystal structure prediction results for this molecule are provided in section 6.2..... 114

Figure 5.4 Location of Hermite interpolant grid points (black squares) and LAM points (open circles) necessary to capture the conformational space of two hypothetical torsions (defined by the boundaries of the dashed square). The conformational hypercube (defined by the boundaries of the red square) captured by the evaluation of one LAM at the reference conformation indicated by the open circle at the centre of the diagram is also shown. For this illustrative example, LAM validity (defined as the distance from the reference conformation) is taken to be equal to the grid spacing. .. 126

Figure 5.5 Molecular diagram, atom labels and the definition of the main two torsional angles in the molecule of ROY (5-methy-2-[(2-nitrophenyl) amino]-3-thiophenecarbonitrile)..... 129

Figure 5.6 Intramolecular energy (kJ mol^{-1}) surfaces as a function of two main torsional angles of ROY (refer to Figure 5.5 for torsion definition) computed from a grid of QM calculations (top, Cases 1 and 2) and estimated using local approximate models (bottom, Case 3). Open circles indicate reference grid (top) or LAM (bottom) points at which explicit QM calculations (at the PBE0/6-31G(d,p) level of theory) were performed. The location of the known polymorphs on the intramolecular energy surfaces is also shown as red dots..... 130

Figure 5.7 Two alternative definitions of rigid fragments in the molecule of ROY. a) The middle (blue) fragment defined as containing atoms C6, N1, C7 and H5; b) The middle (blue) fragment is defined without the H5 hydrogen (i.e. C6, N1, C7), whilst H5 is included in the definition of the red fragment..... 131

Figure 5.8 Geometries of low-energy structures obtained by search Case 1 a) expected intermolecular hydrogen bond in the structure of ROY and b) spurious intramolecular hydrogen bond (note that only one oxygen atom of the neighbouring ROY molecule is shown here)..... 133

Figure 5.9 Lattice energy landscapes produced for search a) Case 2 and b) Case 3. The rank of the known experimental forms when minimised with the same computational model is shown in brackets. 134

Figure 5.10 Overlay of the experimental R polymorph of ROY with the conformations (coloured by element) with the conformations obtained (green) after a) original CrystalPredictor (Case 2) and b) CrystalPredictor with LAMs. Similarly, overlay of the experimental structure of the R polymorph of ROY (coloured by element) and the predicted crystal structures (green) using c) Original Crystal Predictor and (d) CrystalPredictor with LAMs. Root mean square deviations in 1- and 15-molecule coordination spheres are also shown..... 135

- Figure 6.1** Molecular connectivity diagram of the BMS target molecule with the definitions of key flexible degrees of freedom (nine torsions). 148
- Figure 6.2** Overlay of experimental structure (solid colours) and the local minimum structure (green) obtained by lattice energy minimisation starting from the experimental form and allowing 9 flexible torsions to vary (Figure 6.1). Structure similarity, given by rmsd_{15} , is 0.256 Å. 149
- Figure 6.3** Three fragments of the BMS used for a) quantum mechanical scans and b) statistical search of the CSD during the conformational analysis. 151
- Figure 6.4** Comparison of the CSD MOGUL geometry analysis (bars, left axis) with the gas-phase intramolecular energy scan at the PBE0/6-31G(d,p) level of theory (lines, right axis) for the dih20 torsion in the BMS molecule (Figure 6.1). Results of the gas-phase calculations shown for a fragment of the BMS molecule defined at the top of Figure 6.3a (continuous black line) and the entire molecule (dashed line). 152
- Figure 6.5** Comparison of the CSD MOGUL geometry analysis (bars, left axis) with the gas-phase intramolecular energy scan at the PBE0/6-31G(d,p) level of theory (black line, right axis) performed on a fragment of the BMS molecule defined at the top of Figure 6.3a for the dih21 torsion (Figure 6.1). 152
- Figure 6.6** Lattice energy versus density plot for the generated crystal structures of the BMS molecule. Results shown for the four independent flexible searches outlined in Table 6.5. The most stable structure is that with the lowest lattice energy. The minima obtained when the experimental polymorphs are minimised with the same computational model are also shown. 157
- Figure 6.7** Final lattice energy landscape produced after refinement of 196 structures from the search using CrystalOptimizer algorithm and clustering. 126 structures remain as independent minima. The minima obtained when the experimental polymorphs are minimised with the same computational model are also shown. 159
- Figure 6.8** Overlay of the experimental conformations (coloured by element) with the final conformations obtained after CrystalOptimizer for experimental form 1 (a) and form 2 (b). Overlay of the experimental structures (coloured by element) and the predicted crystal structures (green) for experimental form 1 (c) and form 2 (d). Root mean square deviations in 1- and 15-molecule coordination spheres are also shown. 160
- Figure 6.9** Molecular diagram of target molecule XX with the definitions of key torsion angles and atom labels. 165
- Figure 6.10** Comparison of the CSD MOGUL geometry analysis (bars, right axis) with the gas-phase intramolecular energy scan at the B3LYP/6-31G(d,p) level of theory (black line, left axis) for the torsion PhCH₂-OCO in molecule XX. N is the total number of similar fragments obtained from the CSD. The dashed line indicates the value as observed in the crystal structure of molecule XX. Note that the diagram is symmetric about 180°. (Assistance of A. J. Cruz-Cabeza in the preparation of this figure is gratefully acknowledged.) 166

Figure 6.11 Overlay of the experimental conformation (coloured by element) with the closest local gas-phase minimum conformation obtained with the B3LYP/6-31G(d,p) level of theory (green).....	167
Figure 6.12 Space group frequency for $Z' = 1$ and $Z' = 2$ structures similar to molecule XX reported in the CSD.....	169
Figure 6.13 Lattice energy landscape generated by eight independent searches for molecule XX. Refer to Table 6.10 for search enumeration. The most stable structure is that with the lowest lattice energy. The minimum which most closely corresponds to the experimental structure is also shown.	171
Figure 6.14 Overlay of the experimental conformation (coloured by element) with the closest conformation obtained from the flexible crystal structure generation step (green). Root mean square deviation in 1-molecule coordination sphere (rmsd_1) is 0.167 Å. .	172
Figure 6.15 Molecular flexibility (14 non-aromatic torsions and 5 chain bond angles) considered explicitly during the CrystalOptimizer refinement stage for molecule XX.	173
Figure 6.16 Final lattice energy landscape produced after the refinement of 1500 lowest energy structures from the search stage using the CrystalOptimizer algorithm and clustering. The minimum corresponding to the experimental structure is also shown. All 553 independent minima are shown.	174
Figure 6.17 Crystal structures of the 3 candidates selected for the blind test submission. Structure identification number, space group, lattice energy, density and hydrogen bonding type are also shown.....	175
Figure 6.18 Overlay of the experimental conformation (a) and structure (b) (coloured by element) and the predicted lowest energy structure (green). The root mean square deviation in the 1- and 15-molecule coordination sphere is also shown.	176
Figure A.1 Molecular diagram and atom labelling of ethyl ethanoate.	204
Figure A.2 Bond angle defined with respect to three atoms A-B-C.	206
Figure A.3 Torsion angle defined with respect to four atoms A-B-C-D.	207

Chapter 1. Introduction

1.1. Background

A big proportion of commercial organic molecular solids are obtained in crystalline form - either as single crystals or crystalline powders. A **crystal** is a solid in which molecules are packed in a long-range, regularly ordered, repeating pattern extending in all three spatial dimensions (Figure 1.1). Crystallisation of a given compound does not always lead to the same crystal structure. Depending on the crystallisation conditions, different packings of the same compound can often be observed. This phenomenon is called **polymorphism** - the ability of a molecule to crystallise in more than one structural form. The diversity in packing motifs arises from differences in spatial arrangements of molecules and/or conformational changes in molecular geometry within the crystal (Figure 1.1). Polymorphs are found in many classes of molecular materials including and not limited to saturated and unsaturated fatty acids, alkanes, aromatic π -bonded systems, amino acids, amides and carboxylic acids. Currently, the molecule that has the most known polymorphs is 5-methyl-2-[(2-nitrophenyl) amino]-3-thiophenecarbonitrile, which has been crystallised in ten polymorphs (Yu et al. 2000, Chen et al. 2005). This compound has been named ROY for its red, orange and yellow crystal colours (Figure 1.2).

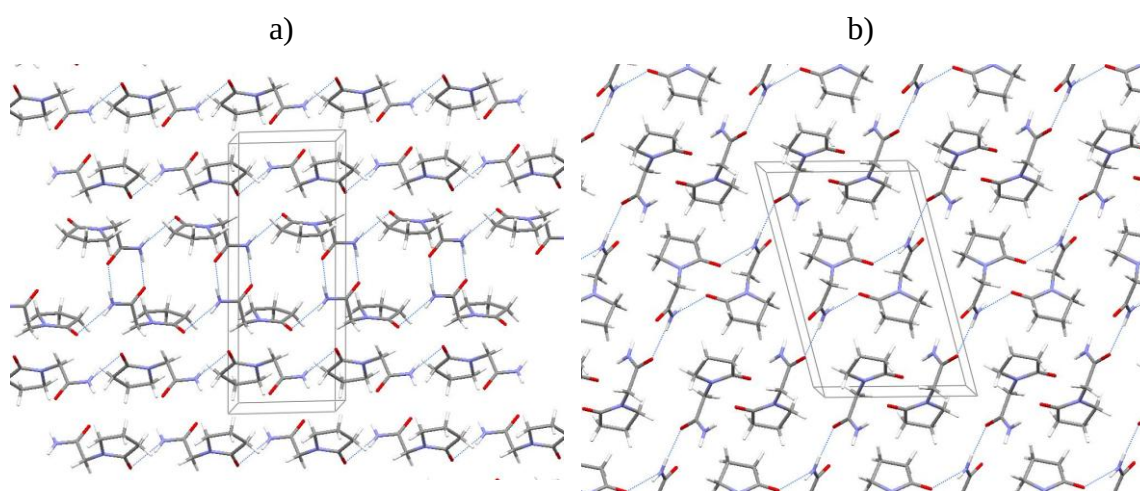


Figure 1.1 Spatial arrangements of different polymorphs of piracetam (2-oxo-pyrrolidine-acetamide) in the crystal: a) Form III (Galdecki and Glowka 1983); b) Form IV (Fabbiani et al. 2005).

Because the underlying chemistry of the molecules in the crystals remains unchanged, the chemical properties are identical for all polymorphs. However, differences in molecular packing between polymorphs lead to differences in physical behaviour. The specific stacking of molecules in the crystal affects important properties of the material including colour, density, morphology, stability, solubility and dissolution rate. As a result, polymorphism is of great importance across a wide range of industries including, pharmaceuticals, health-care, agrochemicals, pigments, dyestuffs and foods.

Therefore, knowledge of the three-dimensional atomic structure of a crystal is the basis towards understanding the properties of the material. For crystals of sufficient quality and size, the structure can be determined reliably using single crystal X-ray diffraction. Nevertheless, suitable crystals of the compound can not always be grown. In this case, structure determination for powders, needles and plates is much more problematic.

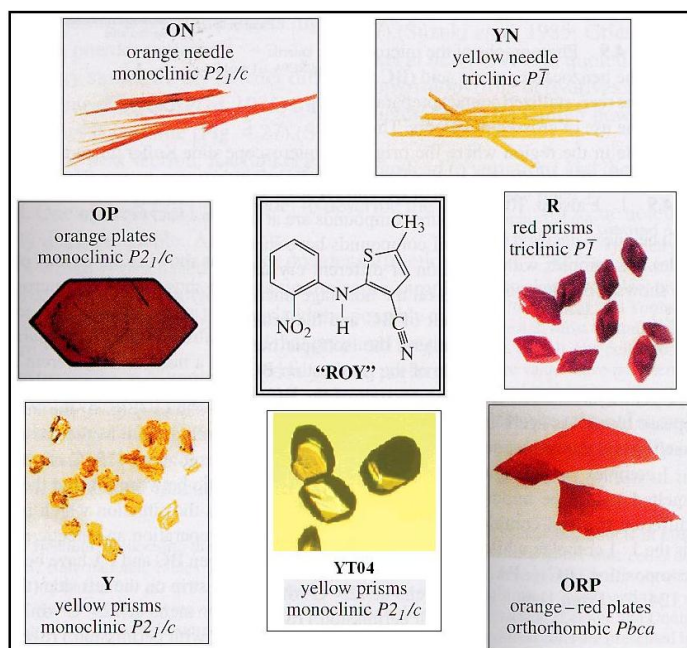


Figure 1.2 Seven ROY (5-methy-2-[(2-nitrophenyl) amino]-3-thiophenecarbonitrile) polymorphs for which the crystalline structure has been solved experimentally. Adapted from (Yu et al. 2000, Chen et al. 2005).

Even when a crystal structure is known, a method of predicting possible polymorphism would have great practical value to avoid problems in manufacture and patenting of the crystalline product (Bernstein 2002). For instance, ritanovir is a protease inhibitor for the HIV virus, marketed as Norvir (Chemburkar et al. 2000,

Bauer et al. 2001). From the discovery of ritanovir until the new drug application (NDA) filing, only one crystalline form was known to exist. Attempts to identify additional polymorphs were unsuccessful. Nevertheless, two years after the launch of the drug on the market, a new polymorph appeared in the manufacturing line. The new polymorph was more stable (i.e. less soluble) than the original form leading to a shift in bioavailability of the drug, creating a crisis in the market and leading to large economic losses for the manufacturer. If the polymorphic landscape of ritanovir could have been modelled, it would have been possible to avoid the problem.

Crystal structure prediction by computer simulation can be used to propose possible structures of organic molecules and their properties prior to their synthesis. Once reliable tools exist for this purpose, they will allow the tailored design of high performance materials with the desired properties. They will also provide grounds for crystallisation and manufacturing process development, since solid-liquid separation, comminution, particle flow and formulation characteristics are polymorph dependent (Davey 1994).

1.2. Scope and outline of the work

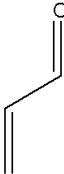
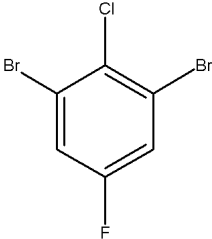
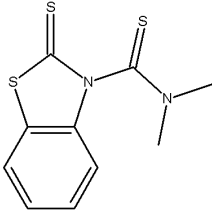
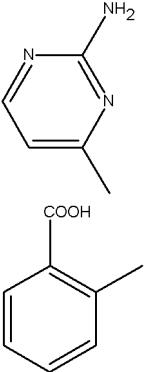
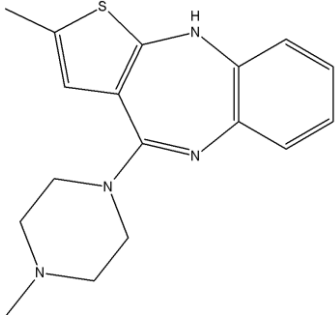
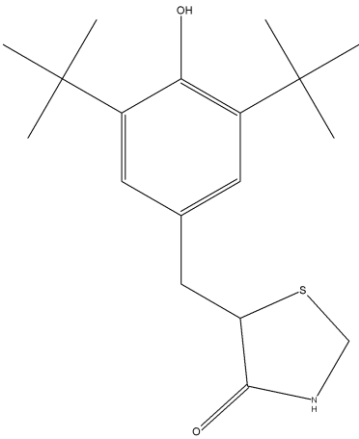
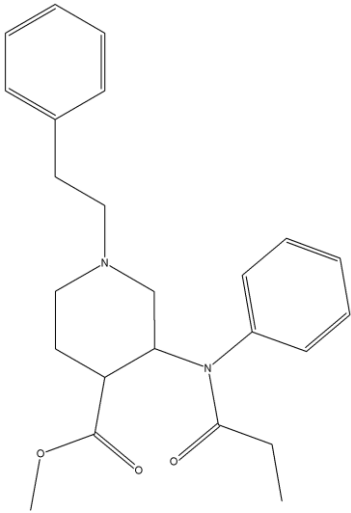
It has been observed that the number of experimentally-determined crystalline forms is proportional to the time and the resources spent searching for them (McCrone 1965). Nevertheless, as was highlighted by the ritanovir incident (Chemburkar et al. 2000), no experimental protocol can guarantee that *all* forms are found. The calculation of the crystal energy landscape can complement solid form screening beyond confirming that the most thermodynamically stable form has been found (Lancaster et al. 2006, Price 2008b, Cruz-Cabeza et al. 2009, Braun et al. 2011).

Existing crystal structure prediction methods, reviewed in Chapter 2, have been generally applied to small and relatively rigid molecules (Table 1.1a). However, in order to be of practical use to the pharmaceutical industry, the algorithms need to be routinely able to deal with much larger and significantly more flexible molecules (Table 1.1b).

Chapter 3 presents the mathematical formulation of local approximate models (LAMs), that can be used to provide an estimate of the intramolecular energy and the

conformationally-dependent charge density as a function of molecular geometry. These models are practically as accurate as explicit quantum mechanical calculations, but carry a much smaller computational burden.

Table 1.1 Typical structure complexity that a) has been considered in crystal structure prediction studies as shown by the four targets in the fourth blind test of crystal structure prediction, CSP2007 (Day et al. 2009); b) needs to be considered for pharmaceutical targets.

a) CSP2007			
			
C_3H_4O	$C_6H_2BrCl_2F$	$C_{10}H_{10}N_2S_3$	1:1 cocrystal $C_5H_7N_3 : C_8H_8O_2$
b) Typical molecules of pharmaceutical interest			
			
Olanzapine $C_{17}H_{20}N_4S$	Tazofelone $C_{18}H_{27}NO_2S$	Carfentanil $C_{24}H_{30}N_2O_3$	

These dynamically constructed and updated local approximate models are implemented in the framework of a novel lattice energy minimisation algorithm described in Chapter 4. This has made possible the accurate treatment of molecules involving a relatively large number of atoms with arbitrary flexibility in torsional and

bond angles and even bond lengths. The application of the algorithm is demonstrated on a large set of single and multi-component crystals.

Chapter 5 extends the use of local approximate models to an existing search algorithm used to efficiently sample the multidimensional lattice energy surface to identify favourable packing arrangements. The improvements are demonstrated through the generation of crystal structures for the ROY molecule.

Chapter 6 presents a crystal structure prediction methodology for flexible molecules. This is followed by the detailed description of two examples – a drug candidate and the largest and most flexible target in the latest blind test of crystal structure prediction.

General conclusions of this work and areas which may benefit from further research and development are, finally, identified in Chapter 7.

Chapter 2. Literature Review

The literature is complete with reviews on the subject of crystal structure prediction. There are some reviews that aim to present the overall challenges (Gavezzotti 2002b, Dunitz 2003, Price 2008a, Desiraju 2010, Aakeröy 2010), others are more specific in their description of the underlying methodologies and algorithms (Verwer and Leusen 1998, Karamertzanis 2004, Day 2011). Reviews that focus on the interpretation of the results produced by the computational methods (Price 2009) and surveys of the systems studied with such approaches (Beyer et al. 2001) can also be found in the literature. Furthermore, the debate (Maddox 1988, Gavezzotti 1994, Dunitz 2003, Lehmann 2011) about the progress and successes of the existing methodologies has been objectively evaluated via a series of five blind tests (Lommerse et al. 2000, Motherwell et al. 2002, Day et al. 2005a, Day et al. 2009, Bardwell et al. 2011).

In this chapter, an attempt is made to overview only those subjects that are relevant to the work completed in this thesis. The parameters defining the crystal lattices are introduced first in section 2.1. This is followed by a description of methods used for the calculation of the relative stability of crystal structures in section 2.2. The importance of molecular flexibility is highlighted in section 2.3. The available methods for crystal structure prediction of flexible molecules are reviewed and their performance in the series of blind test of crystal structure prediction is discussed in section 2.4. The reader is referred to the original topical reviews and references therein for more detailed descriptions.

2.1. *Crystal lattices*

2.1.1. Crystal structure representation

A perfect molecular crystal can be described as a three-dimensional compilation of identical building blocks referred to as ***unit cells***. Predicting the crystal structure, therefore, involves the determination of the size and shape of the unit cell as well as the

identification of all atomic positions in it. The unit cell is defined by the magnitudes of the lattice vectors (a , b , c) and the angles between them (α , β , γ) (Figure 2.1).

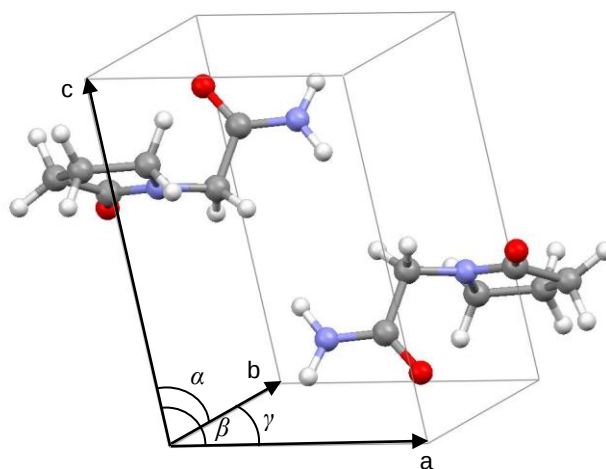


Figure 2.1 Lattice vectors (a , b , c) and angles (α , β , γ) defining the unit cell in piracetam Form II crystal (Admiraal et al. 1982).

The atomic arrangement within the unit cell may also display additional internal symmetry (such as mirror, rotational, inversion and translational). The cell is called '**primitive**' when no translational symmetry is present. However, in certain cases it is preferable to define a larger '**centred**' cell. The specific combination of symmetry elements present in a unit cell defines the **space group** of the crystal. In three dimensions, 230 different space groups can be constructed (Kitaigorodski 1961, Kitaigorodski 1973, Stout and Jensen 1989). Nevertheless, it has been experimentally observed that more than 90% of all crystalline solids in the Cambridge Crystallographic Database (CSD) occur in just a few triclinic, monoclinic and orthorhombic space groups¹ (Allen 2002).

The coordinates of a unique part of the structure – the **asymmetric unit**, along with the space group are sufficient to define the positions of all remaining atoms in the unit cell when symmetry is present. By convention, the number of molecules per unit cell is denoted by Z , while the number of formula units in the asymmetric unit is usually labelled Z' (van Eijck and Kroon 2000).² When a molecule is itself symmetric, the asymmetric unit may contain just a fraction of that molecule in the definition of the

¹ Molecular crystals are known in all 230 space groups, although for many only a few of the structures have been reported (Wilson 1993). The tendency of molecules to crystallise in just a small set of space groups is often used in crystal structure prediction algorithms to reduce computational cost.

² Note that this is not necessarily the same as the number of independent molecules in the asymmetric unit. A cocrystal with 1:1 stoichiometry and $Z' = 1$ will have 2 molecules in the asymmetric unit.

formula unit. However, it is also possible to have $Z' > 1$, as in the case of two or more independent molecules in single component crystals or cocrystals, salts and hydrates with stoichiometries greater than 1:1.

The same lattice can be described by different sets of parameters; therefore, the choice of cell parameters is not unique. Several methods and algorithms (Andrews and Bernstein 1988) have been developed to obtain the unique standard cell parameters of the '*reduced cell*', which is the standard primitive cell used to describe a given lattice.

2.1.2. Thermodynamic versus kinetic considerations

Polymorphism is the ability of a solid material to exist in more than one crystalline form. For one component, the Gibbs phase rule dictates that only one phase is observed when both temperature and pressure remain as independent variables (for triple point and the coexistence curves, these variables are fully constrained or related respectively). In contrast, it has been shown experimentally that various solid forms can coexist at extended temperature and pressure domains (e.g. ROY molecule (Chen et al. 2005)), which, at first sight, would contradict the Gibbs phase rule. At actual thermodynamic equilibrium, one phase must have a lower Gibbs free energy than the other. However, the rate of transformation of the metastable phase(s) is often so slow, that for all intents and purposes, they are experimentally relevant forms, and their existence needs to be modelled and predicted.

Prediction of the relative thermodynamic stability of polymorphs at a given temperature T and pressure P requires the minimisation of the Gibbs free energy (G) with respect to the unit cell dimensions and the positions of all atoms in the unit cell:

$$\min G = \min(U + PV - TS) \quad \text{Eq. 2.1}$$

where U is the internal energy (which includes the zero-point energy contributions (van Eijck 2001)), V the volume and S the entropy. Thus, the free energy variation arises from the difference in packing energy, crystal density and entropy. Since the compressibility of organic solids is low, indicating that the density differences between different polymorphs for organic molecules are small (Gavezzotti and Filippini 1995), the contribution of the PV term can be assumed negligible. Nevertheless, the term may

become significant during simulation of crystal phases at very high pressures (typically above 1 GPa)¹ (Fabbiani and Pulham 2006). Furthermore, the thermal, entropic and zero-point contributions to the Gibbs free energy of flexible molecules cannot, at present, be readily and accurately computed². There is also emerging evidence that most minima on the free energy surface are also minima on the lattice energy surface (Karamertzanis et al. 2008b). Consequently, most of the existing computational approaches for crystal structure prediction focus on the minimisation of lattice energy (Day et al. 2009, Bardwell et al. 2011). The lattice energy, E^{latt} , refers to the internal energy, U , at 0 K and 0 Pa (ignoring the zero-point energy contribution). The assumption underpinning crystal structure prediction is that crystal structures that occur in nature correspond to low-lying minima in the crystal energy landscape, which is usually approximated as the lattice energy.

However, the thermodynamic stability is not necessarily the decisive factor in the packing patterns of molecules. Kinetics plays an important role, which is strongly influenced by crystallisation conditions and solvent environment. In fact, in many cases the solid first formed from solution crystallisation is amongst the least stable polymorphs with more stable forms appearing at later stages - a phenomenon known as the Ostwald's rule of stages (Nývlt 1995).

The nucleation kinetics and macroscopic shape of a crystal can be significantly affected by supersaturation conditions and solvent environment, which may modify the growth rates of different crystal surfaces and, thus, favour growth of a particular polymorph without affecting relative thermodynamic stability of different forms (Weissbuch et al. 1995, Yu 2007). Therefore, the calculation of thermodynamic quantities such as the lattice energy, or even the more exact Gibbs free energy, for a set of predicted polymorphs cannot guarantee that the structure with the lowest energy will be observed experimentally because its nucleation or growth may be inhibited kinetically. It is, however, still possible that an alternative crystallisation method under different experimental conditions (cooling rate, seeding, solvent, pressure etc.) may be able to reduce the nucleation barrier towards the production of the thermodynamically expected structures.

¹ We have studied (computationally) the reversible single-crystal to single-crystal transformation of piracetam Form II $\leftrightarrow$ Form V (Fabbiani et al. 2007) under the application of an isotropic pressure tensor. The results have not been published yet.

² There is ongoing work in our research group to develop free energy minimisation algorithms for flexible molecules.

Crystallisation is a process that is usually complicated, poorly understood and often difficult to reproduce, where numerous factors can play a decisive role. Therefore, the approach involving the identification of the known crystal structure as the global energy minimum is limited because only the thermodynamic properties are taken into account. Nevertheless, the experimental structure is still expected to appear within a narrow energy range (usually taken to be 10 kJ mol^{-1} (Bernstein 2002)) of the global lattice energy minimum. The knowledge of the polymorphic landscape can help to avert problems in manufacturing and guide the experimental search for as yet undiscovered structures.

2.2. Lattice energy calculation

There are a number of different ways to calculate the lattice energy depending on the required balance between the accuracy, ease of implementation and computational cost. Ideally, the lattice energy of a crystal structure should be evaluated by modelling the entire crystal from first principles, i.e. quantum mechanically. Significant progress towards this goal has been made through the development of periodic solid-state density functional theory (DFT) algorithms (e.g. the GRACE method (Neumann and Perrin 2005) described in more detail in section 2.4). In such calculations, all electron density and nuclear positions are optimised simultaneously allowing the molecules to adjust naturally to the crystalline environment (Day 2011). This, however, remains computationally expensive for many molecules of interest. Instead, the lattice energy of a crystal can be partitioned into independent energy contributions. By adding more terms to this energy expansion, the accuracy of the predicted energy can be improved. For instance, the lattice energy is usually split into the intramolecular and intermolecular energy contributions:

$$E^{\text{latt}} = \Delta E^{\text{intra}} + U^{\text{inter}} \quad \text{Eq.2.2}$$

The intramolecular energy, ΔE^{intra} , is the energy required to deform a molecular conformation from its most stable gas-phase geometry. The intermolecular contributions, U^{inter} , describe the interactions between molecules in the crystal. It is

very important that the different contributions to the lattice energy are well balanced and do not artificially favour some particular interactions over others.

The following sections, describe how the different contributions to the lattice energy can be evaluated.

2.2.1. Intramolecular energy calculation – ΔE^{intra}

Neglecting molecular flexibility and treating the molecules as rigid can be a valid approach in some circumstances (e.g. when considering fused aromatic ring systems) (Lommerse et al. 2000, Motherwell et al. 2002, Day et al. 2005a). For these cases the intramolecular energy term simply vanishes. However, for more realistic systems the molecular flexibility needs to be taken into account explicitly.

One of the simplest approaches to incorporate the intramolecular terms is to use molecular mechanics force-fields of the form:

$$\Delta E^{\text{intra}} = E^{\text{s}} + E^{\text{b}} + E^{\text{tor}} \quad \text{Eq. 2.3}$$

where the intramolecular energy of a target flexible molecule is calculated as the sum of several contributions arising from bond stretching (E^{s}), angle bending (E^{b}) and dihedral angle torsions (E^{tor}). Usually these terms are of quadratic form and, for small molecules, are parameterised to quantum mechanical calculations on isolated molecules (Day 2011). A large number of molecular mechanics force-fields have been successfully used in studies including molecular dynamics, reaction mechanisms, proteins, transition metal inorganic and organometallic compounds (e.g. Dewar et al. 1985, Mayo et al. 1990, Cornell et al. 1996).

Despite being computationally efficient, the molecular mechanics description of the intramolecular energies and the molecular geometry is not sufficiently accurate for crystal structure prediction (Day et al. 2007, Day 2011). Isolated-molecule electronic structure calculations can generally provide the accuracy required for modelling the deformation of the molecular structure and energy within the crystal (Gavezzotti 1997). However, QM calculations may fail to provide accurate results for molecules which can adopt folded conformations due to the errors in intramolecular dispersion (van Mourik et al. 2006). Furthermore, the computational cost associated with such calculations is

much greater than when using molecular-mechanics force-fields and the scaling of the computational cost with molecular size is an issue for extending this approach to larger and more flexible molecules. Section 3.3 addresses the latter limitation and extends the use of isolated molecule quantum mechanical intramolecular energy in lattice energy minimisation.

2.2.2. Intermolecular energy calculation – U^{inter}

As mentioned earlier, the intermolecular contributions, U^{inter} , describe the interactions between neighbouring molecules in the crystals and consist primarily of the electrostatic (U^{elec}) and repulsion-dispersion (U^{vdw}) interactions, although other terms (U^{other}) such as intermolecular induction, can also be included:

$$U^{\text{inter}} = U^{\text{elec}} + U^{\text{vdw}} + U^{\text{other}} \quad \text{Eq. 2.4}$$

The most commonly used approach to calculate the different intermolecular energy terms is the atom-atom potential method for pair-wise potentials (Pertsin and Kitaigorodski 1987, Karamertzanis 2004, Day et al. 2005a, Day et al. 2009, Day 2011). The energy of a pair of two sites i and i' at a distance r apart is denoted by $U_{ii'}(r)$ and depends on the nature of the two sites. The energy between two molecules j and j' is then the summation over all sites:

$$U_{jj'} = \sum_{i=1}^{N_j^S} \sum_{i'=1}^{N_{j'}^S} U_{ii'}(\|\mathbf{r}_{ji} - \mathbf{r}_{j'i'}\|) \quad \text{Eq. 2.5}$$

where N_j^S is the number of sites in molecule j and $\mathbf{r}_{ji}$ is the spatial position of site i of molecule j . The intermolecular contribution to the lattice energy can then be calculated by extending the summation over all molecules (N_{mol}) in the crystal:

$$U_{jj'} = \frac{1}{2} \sum_{j=1}^{N_{\text{mol}}} \sum_{j'=1}^{N_{\text{mol}}} \sum_{i=1}^{N_j^S} \sum_{i'=1}^{N_{j'}^S} U_{ii'}(\|\mathbf{r}_{ji} - \mathbf{r}_{j'i'}\|) \quad \text{Eq. 2.6}$$

Electrostatic interactions – U^{elec}

In its simplest form, the electrostatic potential energy can be calculated using Coulomb's law through the pair-wise interaction of isotropic (spherical) point charges placed on each atom:

$$U_{ii'}^{\text{elec}}(r_{ii'}) = \frac{1}{4\pi\epsilon_0} \frac{q_i q_{i'}}{r_{ii'}} \quad \text{Eq. 2.7}$$

where ϵ_0 is the vacuum permittivity constant (also known as the dielectric constant). As a result, the inter-atomic interaction energy $U_{ii'}^{\text{elec}}(r)$ depends only on the separation distance between two atoms, $r_{ii'}$, and their respective charges (q_i). Unfortunately, site charges cannot be measured experimentally or computed directly. A common approach is to assign or fit them to reproduce the molecular electrostatic potential or the electrostatic field (Mulliken 1955, Cox and Williams 1981, Sigfridsson and Ryde 1998). The electrostatic potential can be computed through quantum mechanical calculations, or from the experimental charge density analysis of X-ray diffraction data (Klooster et al. 1992, Ghermani et al. 1993, Espinosa et al. 1996).

Because of the anisotropic nature of the charge distribution, models based exclusively on atomic charges may be unable to describe short-distance directional intermolecular interactions (such as hydrogen bond geometries). As a result, extensions to the isotropic atom model have been proposed as to define off-nuclear charge sites at bond centres and lone pair positions (Williams and Weller 1983, Williams 1984) or at satellite positions that are optimised to provide the best fit to the electrostatic field (Karamertzanis and Pantelides 2004, Karamertzanis 2004).

A more elaborate representation of the molecular charge distribution can be derived through the distributed multipole analysis (Stone and Alderton 1985, Stone 1996, Stone 2005). In this framework, a multipole series expansion (up to the hexadecapole level¹) is placed at each atom centre. The multipoles can be either fitted to the electrostatic potential (Mooij et al. 1999a, Mooij and Leusen 2001) or derived directly from the isolated molecule wavefunction (Stone and Alderton 1985, Stone 2005). Essentially, the atomic multipole moments describe the electrostatic potential with the accuracy of

¹ i.e. multipole series of the form: charge, dipole, quadrupole, octapole, hexadecapole.

the wavefunction and thus offer a significant improvement in the reproduction of the ESP potential.

Unlike the atomic charge models, the distributed multipole moments have been shown to be successful in predicting the highly directional (anisotropic) lone-pair interactions, π - π stacking in aromatic rings and hydrogen bond geometries in molecular organic crystals (Mooij and Leusen 2001, Brodersen et al. 2003, Day et al. 2005b).

Repulsion-dispersion interactions – U^{vdw}

The van der Waals interaction energy typically consists of a repulsive and an attractive contribution. The repulsive component results from the Pauli Exclusion Principle that states that particles (and consequently atoms) exhibit space-occupying behaviour (i.e. excluded volume of atoms). As a result, the component is strongly repulsive at short distances and decays quickly as the inter-atomic separation increases. The weak attractive London dispersion forces arise from the instantaneously induced dipoles in molecules. These forces dominate the interaction of non-polar molecules and increase with the molar mass due to the presence of larger, more dispersed electron clouds. Since the electrostatic term is computed independently in the atom-atom approach to lattice energy calculations (see above), the van der Waals interactions are usually referred to as repulsion-dispersion interactions.

There are many different mathematical forms of the repulsion-dispersion potentials used in molecular simulation (Reed and Gubbins 1973, Gil-Villegas et al. 1997, Gavezzotti 2002c, Potoff and Bernard-Brunel 2009). The two isotropic forms of the van der Waals energy between atoms i and i' commonly used in crystal structure prediction are the 12-6 Lennard-Jones model (Lennard-Jones 1924):

$$U_{ii'}^{LJ}(r_{ii'}) = 4\epsilon_{ii'}^{LJ} \left[\left(\frac{\sigma_{ii'}^{LJ}}{r_{ii'}} \right)^{12} - \left(\frac{\sigma_{ii'}^{LJ}}{r_{ii'}} \right)^6 \right] \quad \text{Eq. 2.8}$$

with two parameters $\epsilon_{ii'}^{LJ}$, $\sigma_{ii'}^{LJ}$ and the exp-6 Buckingham potential (Buckingham 1938):

$$U_{ii'}^{exp-6}(r_{ii'}) = A_{ii'} \exp(-B_{ii'} r_{ii'}) - \frac{C_{ii'}}{r_{ii'}^6} \quad \text{Eq. 2.9}$$

with three parameters, $A_{ii'}$, $B_{ii'}$ and $C_{ii'}$.

In the potentials found in the literature, the parameters for the interactions between similar atom types, i.e. $i' = i$, are usually defined. The interactions between dissimilar sites are found by applying pre-defined mixing rules (for example (Coombes et al. 1996, Williams 2001)). In most cases, the parameters for the interaction between two dissimilar sites i , i' are the arithmetic or geometric mean of the parameters corresponding to the interactions between the similar sites i , i and i' , i' . The atom types are normally chosen according to atomic number, but further classifications are often made to distinguish between different hybridisation states or bonding environment (e.g. aliphatic vs. aromatic carbon atoms; hydrogen bonded to carbon vs. hydrogen bonded to polar atoms such as oxygen or nitrogen) (Coombes et al. 1996, Williams 2001).

The potential model parameters are usually determined by fitting to the experimental structures and/or properties (e.g. sublimation energies) (Cox et al. 1981, Williams and Cox 1984, Coombes et al. 1996, Williams 2001). Despite being successful in many predictions¹, such empirically parameterised models contain several limitations. For parameters to be truly transferable, a large set of known crystal structures is required to efficiently sample a range of different types of interactions at various inter-atomic separations and orientations. In reality, there is often insufficient experimental data to fit reliable transferable parameters for less common atom types (particularly sulphur) and to include anisotropy in the potentials known to be significant for certain atoms (such as polar flattening of halogen atoms (Nyburg and Faerman 1985)). Additionally, the structural data used for the parameterisation is often determined at a range of physical conditions (temperature and pressure), which, strictly speaking, is not suited for the calculation of the lattice energy at 0 K and 0 Pa.

Recently, there has been significant progress in the development of non-empirical atom-atom potentials for specific molecules (Nobeli and Price 1999, Tsui and Price 1999, Mooij et al. 1999a, Misquitta et al. 2008) or small families of molecules (Mooij et al. 1999b, Mitchell and Price 2000, Stone and Misquitta 2007). In these

¹ In many cases the lattice energy is dominated by the strong electrostatic interactions and thus is not sensitive to the errors inherent in the empirical repulsion-dispersion potentials. Furthermore, the success of the simple repulsion-dispersion models seems to be more dependent on the parameterisation of the model rather than the exact functional form (van Eijck 2002).

methodologies, quantum mechanical calculations on isolated molecules and small clusters are used to probe many more interaction orientations and separations than are typically available in the experimental data. As a result, the orientational (anisotropic) dependence of the repulsion-dispersion interactions is incorporated in many of these models. The non-empirically parameterised models usually lead to better reproduction of known crystal structures of small molecules (Nobeli and Price 1999, Mooij et al. 1999b, Tremayne et al. 2004), but still remain computationally expensive to be used routinely.

Other contributions to the intermolecular energy – U^{other}

In most atom-atom electrostatic energy models (see above), the polarisation of the charge density by neighbouring molecules in the crystal is not included (Cox et al. 1981, Price 2004a). The need for explicit energy terms for the induction contribution has been the subject of much debate (Fowler and Stone 1987, Gavezzotti 2002a, Welch et al. 2008). This is partly because polarisability effects are to some degree absorbed in the empirically fitted repulsion-dispersion parameters commonly used for crystal structure prediction¹. However, there is mounting evidence that induction effects may be important within crystal structures, even for non-polar molecules (Gavezzotti 2003a, Gavezzotti 2005b). The explicit inclusion of the induction energy contribution, obtained from high quality wavefunctions and large basis sets, in lattice energy calculation has shown to improve the relative ranking of the structures to be more in line with experimental observation (Welch et al. 2008). This approach, however, remains computationally expensive and has not been widely adopted. A computationally cheaper alternative is to perform isolated-molecule charge density calculations within a continuum dielectric to approximate the molecule's environment in the solid state and the resulting polarisation of the charge density. It has been shown that the use of the Polarizable Continuum Model (PCM) (Tomasi et al. 2005) to introduce the continuum dielectric has an important influence on the relative conformational energies and electrostatic interactions in the crystal structure prediction of polar, flexible molecules (Cooper et al. 2008).

¹ Notable exceptions are the non-empirical repulsion-dispersion potentials described earlier.

2.2.3. Infinite summation of intermolecular energy

In theory, an infinite number of van der Waals and electrostatic interactions needs to be considered in a periodic crystal lattice (equation 2.6). In practice, only the interactions between closest neighbours, typically within a predefined cut-off radius, are evaluated.

For the van der Waals interactions, the dispersion energy vanishes more slowly than the repulsive one (e.g. equations 2.8 and 2.9), but still with a fast rate of decay proportional to $1/r^6$. Practically, this means that it is possible to choose an appropriate cut-off distance (usually 15-30 Å) which will introduce insignificant errors (orders of magnitude lower than the depth of the potential well) to the energy calculation. Furthermore, altering the potential function close to the cut-off distance by means of a spline interpolation can ensure continuity in the potential and its derivatives (Kopsias and Theodorou 1998, Karamertzanis 2004).

For polar crystals, the electrostatic potential (equation 2.7), decreases slowly (at a rate of $1/r$) and is, therefore, only conditionally convergent. The Ewald summation technique was developed to enhance the convergence properties of this sum (Ewald 1921) and can be applied to additive interactions with poor convergence, mainly charge-charge, charge-dipole and dipole-dipole. The basic idea behind the Ewald summation is the superposition principle in electrostatics (Allen and Tildesley 1987, Frenkel and Smit 2002). Because the electrostatic energies consist of both short- and long-range interactions, the interaction potential is split into two main parts. The fast converging short-range component is summed in real space and the long-range component in reciprocal (Fourier) space up to a predefined cut-off distance. Several extensions of the Ewald summation exist for large systems (Toukmaji and Board 1996). For instance the particle-mesh (Hockney and Eastwood 1981, Darden et al. 1993) and multipole-based (Greengard and Rokhlin 1987, Shimada et al. 1994) Ewald methods become efficient for system sizes in the order of 10^4 - 10^5 atoms. For crystalline structures containing organic molecules, the number of charges in the unit cell is fairly small (usually less than 1000), hence, the standard Ewald method remains the default approach.

2.2.4. Alternative methods for lattice energy calculation

The main limitation of the atom-atom approach is that the constituent energy contributions are coupled and thus should not be computed separately. Furthermore, all the contributions to the lattice energy should be well-balanced and computed to the same accuracy to avoid non-physical distortions in lattice parameters and molecular conformations (Gavezzotti 2002c, Brodersen et al. 2003). For instance, the intramolecular energy terms evaluated using QM isolated-molecule calculations include intramolecular polarisation, which is not computed for most of the intermolecular energy models. If intermolecular polarisation is included, then its effects should also be extended to other energy terms, such as the van der Waals interactions. Hence, the error in the lattice energy of crystals using the atom-atom approach can be quite important (van Mourik et al. 2006, Karamertzanis et al. 2008a).

Therefore, a potentially more accurate method of evaluating the lattice energy is by modelling the entire crystal from first principles using solid-state periodic density functional theory (DFT). Such calculations solve the problems encountered in atom-atom models associated with the partitioning of the lattice energy function into constituent terms (equations 2.2 and 2.4). Throughout the DFT calculations, all electron density and nuclear positions are optimised simultaneously allowing the molecules to adjust naturally to the crystalline environment (Day 2011). The possibility of charge density reorganisation enables the accurate inclusion of polarisation effects in lattice energy calculations. Early examples of DFT in crystal structure calculations showed that a more favourable energy rank, in line with the experimental observations, could be achieved for several small molecule crystals than with atom-atom potentials and isolated-molecule models (Rovira and Novoa 2001, Chisholm et al. 2005). However, pure DFT usually underestimates the density of the crystals and thus makes it difficult to reproduce the experimental structures (Byrd et al. 2004). This observation can be attributed to the lack of an adequate description of the long-range dispersion attraction between molecules. As a result an empirically-derived dispersion term is usually used to augment the typical functionals in dispersion-corrected DFT+D methods (Grimme 2004, Neumann and Perrin 2005). The empirical coefficients are fitted to reproduce unit cell parameters of a set of experimental structures and the dispersion term is damped at short distances. These dispersion-corrected methods have

been very successful in both structure reproduction and energy ranking of experimental crystals, but are very demanding in terms of computational resources (Day et al. 2009, Chan et al. 2011).

A semi-empirical alternative to the all-electron quantum mechanics calculations is the fast semi-classical density sum (SCDS) scheme (Gavezzotti 2002a, Gavezzotti 2003a). The approach subdivides the quantum mechanically calculated monomer charge density into a large number of interaction sites (pixels), which interact to give the electrostatic, exchange-repulsion, dispersion and polarisation energies. Because each energy contribution is derived separately, the method provides a better partitioning of the lattice energy than the empirical atom-atom potentials (Gavezzotti 2003b, Gavezzotti 2003c). Despite computational improvements over the DFT approaches, still only a small number (10-100) of energy calculations can be done sequentially with the SCDS method in reasonable amount of time.

As an alternative to explicit energy calculation, the stability of the crystal structures can be estimated using other means. For instance, the wealth of the information contained in the Cambridge Crystallographic Database (CSD) (Allen 2002), now containing more than 500 000 crystal structures, has also been used to judge (and rank) crystal structures. Statistical atom-atom potentials derived by analysing the probabilities of group-based interactions have been proposed and used (Hofmann and Lengauer 1999, Motherwell 2001, Hofmann and Apostolakis 2003). Similarly, statistical models for hydrogen bond propensity prediction trained against data from crystal structures of molecules with similar functional groups are also used in rationalising and refining the rank of crystal structures (Galek et al. 2007, Galek et al. 2009). The Cambridge Crystallographic Data Centre (CCDC) offers a set of tools that provide rapid access to the CSD information and contain exhaustive knowledge bases for the intramolecular geometries (Bruno et al. 2004) and intermolecular interactions (Bruno et al. 1997), which are being used in drug discovery and development (Groom and Allen 2010). These approaches do not require significant computational resources (a desktop computer suffices in most cases) and the tools are usually user-friendly.

The assumption behind all these methods is that the most promising crystals should match existing experimental data of similar structures as closely as possible. However, because the statistical fitness functions are derived in terms of the observed frequency distribution of the geometries and intermolecular distances of similar molecules, they

are bound to fail when close contacts are under-represented in the data and are typically not as accurate as explicit lattice energy calculations.

2.3. *Molecular flexibility*

Early methods of crystal structure prediction focused only on fairly small rigid molecules (Lommerse et al. 2000, Motherwell et al. 2002). The equilibrium geometry of the molecule in a crystal is a balance between the forces within the molecule and the forces between the molecule and its neighbours. The interactions between molecules are much weaker than the covalent bonds that hold molecules together. Large changes in bond lengths and bond angles generally require more energy than can be provided by intermolecular forces (Warshel et al. 1974, Harris and Lammertsma 1997, Hargittai and Levy 1999). However, in many cases (if not most) rotations about single bonds are often possible at a small cost in conformational energy.

The modelling of xylitol (Figure 2.2) clearly illustrates that the intra- and intermolecular energy surfaces are qualitatively different, with local minima occurring on opposite bounds of the selected range of the torsional angles. Furthermore, the changes in both energy components are of the same magnitude. From the molecular point of view, this indicates that the molecular geometry is distorted from the gas-phase conformation in order to obtain a more favourable packing arrangement in the crystal. As a result, an accurate treatment of both intra- and intermolecular energies is paramount in obtaining acceptable structure reproduction and reliable lattice energies.

In the atom-atom approach the intra- and intermolecular models are usually derived separately (see section 2.2), thus there is no guarantee that they will be sufficiently well-balanced to model the deformations of the molecular structure caused by the packing forces within the crystal (Gavezzotti 2002c, Karamertzanis et al. 2008a). For instance, the benefits of accurate anisotropic intermolecular models are reduced when they are combined with empirical intramolecular force fields, which lead to non-physical distortions of the molecular geometry (Brodersen et al. 2003) and poor ranking among hypothetical crystal structures (van Eijck et al. 1995, Day et al. 2007).

Electronic structure calculations can generally provide the accuracy required for modelling the deformations of the molecular structure in the crystal (Gavezzotti 1997). However, the use of quantum mechanics in the computation of the lattice energy is

expensive, as the intramolecular energy and the more realistic electrostatic models (e.g. atomic multipole moments) need to be recalculated after every conformational change. This is due to the charge distribution being only approximately transferable between different conformations and very sensitive to the relative positions of polar and polarisable parts within the molecule (Price and Stone 1992, Koch et al. 1995, Price 2000, Price 2004a).

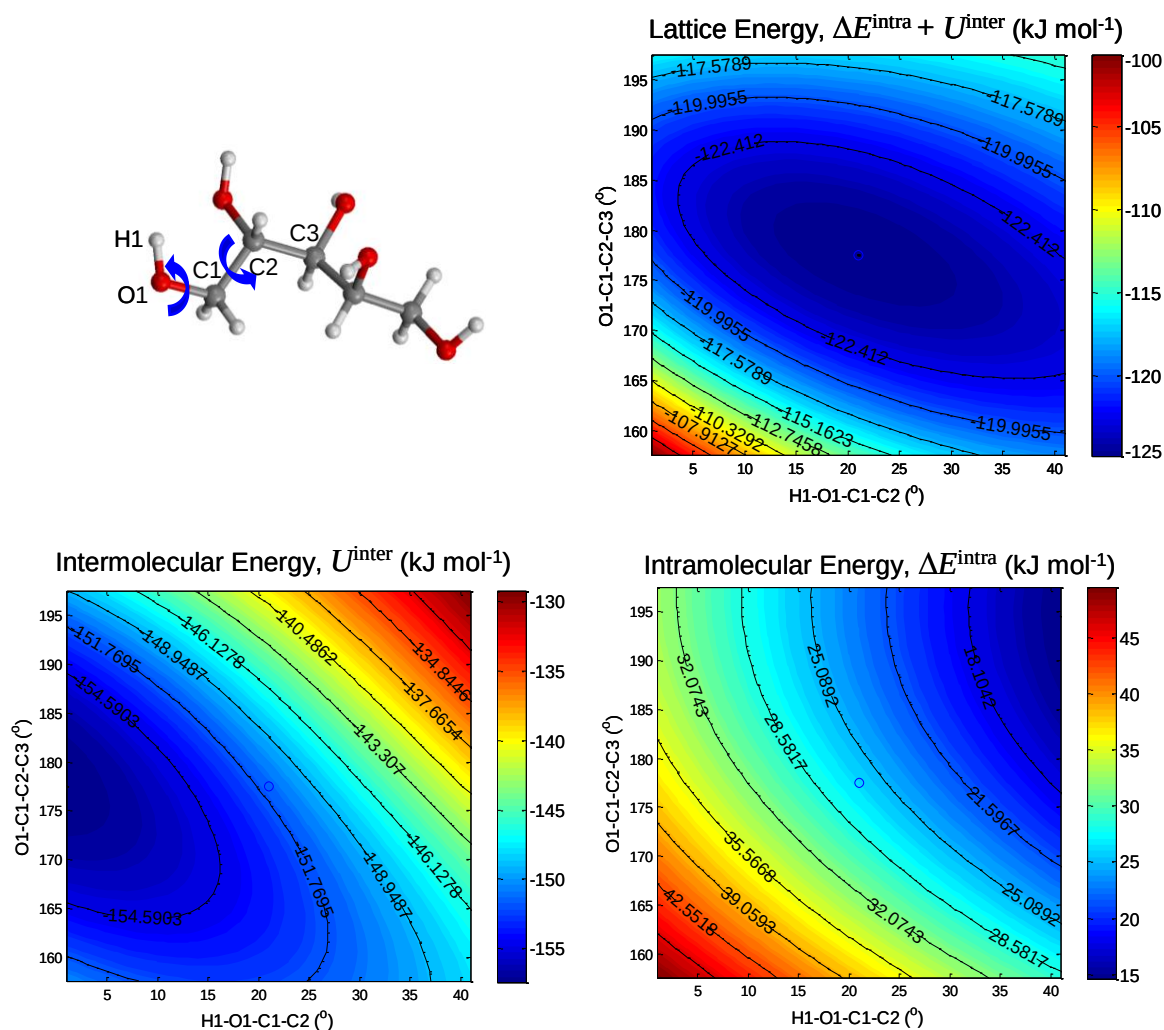


Figure 2.2 Lattice energy (top right), intermolecular energy (bottom left) and intramolecular energy (bottom right) as a function of two torsional angles for the experimental crystal structure of Xylitol (1,2,3,4,5-pentapentanol, top left) (Kim and Jeffrey 1969). Calculations performed using the methodology and algorithm described in Chapter 4.

The additional computational cost associated with evaluating the energy models is not the only obstacle. It is no longer sufficient to define the molecule uniquely by the position of its centre of mass and orientation. All the intramolecular degrees of freedom (bond lengths, angles and torsions) also need to be considered explicitly or implicitly.

This significantly expands the variable space that has to be considered during lattice energy calculation (and minimisation), which has proven to be a major hurdle in the crystal structure prediction methodologies described below.

2.4. Crystal structure prediction methods for flexible molecules

One of the most common approaches to crystal structure prediction (CSP) has been to search for all possible crystal packings and to rank them according to the calculated lattice energy (Lommerse et al. 2000, Motherwell et al. 2002, Day et al. 2004, Day et al. 2005a, Day et al. 2009). The global minimum energy structure is assumed to be the most likely to appear experimentally and other crystals with competitive energies may be potential polymorphs. This approach has been recently successfully applied to systems of increasing complexity including, not only single-component crystals (Day et al. 2007, Bond et al. 2011, Kazantsev et al. 2011b), but also cocrystals (Cruz-Cabeza et al. 2006, Karamertzanis et al. 2009, Görbitz et al. 2010), salts (Gourlay et al. 2007, D'Oria et al. 2010, Antoniadis et al. 2010) and solvates (Hulme and Price 2007, Cruz-Cabeza et al. 2008, Cruz-Cabeza et al. 2010). While there are many variations of the methodology, the main steps in the approach are always similar and consist of some or all of the following steps:

1. selection of computational models (both molecular and interaction potentials described previously in sections 2.2 and 2.3)
2. exhaustive search for all possible packing arrangements in the crystal (section 2.4.1)
3. accurate lattice energy evaluation of the putative crystal structures (section 2.4.2), also termed “refinement”
4. selection of candidate structures based on additional considerations (section 2.4.3)

Each of the steps presents its own challenges and a poor implementation of the methodology at any stage of crystal structure prediction will seriously reduce the chances of locating the likely experimental structures. The precise methodology adopted in this work is described in detail in Chapter 6.

Several thorough reviews (Verwer and Leusen 1998, Karamertzanis 2004, Day 2011) examined and analysed in detail methods and individual algorithms that are currently in use for crystal structure prediction. An attempt here is only made to describe the diversity of methodologies and approaches available for (or that can be extended to) flexible molecules. Brief summaries of the selected methods for crystal structure generation and refinement reviewed in this work are presented in Table 2.1 and Table 2.2 respectively.

Table 2.1 Reviewed methods for crystal structure generation.

Name	Structure search and generation	Molecular Type	Energy function	Selected Reference
PROMET	Construction of aggregates	Rigid	Empirical calculation with pixel method	(Gavezzotti 1991)
FlexCryst	Construction of aggregates	Rigid	Statistical fitness function	(Hofmann and Lengauer 1997)
PMC	Grid-based systematic	Flexible	Empirical repulsion-dispersion with atomic charges and bond dipoles	(Dzyabchenko et al. 1999)
UPACK	Grid-based or random	Flexible	Empirical repulsion-dispersion with atomic charges	(van Eijck and Kroon 1999)
CRYSCA	Random	Flexible	Empirical repulsion-dispersion with atomic charges	(Schmidt and Englert 1996)
TMFF	Random	Flexible	Tailor-made force field derived from DFT+D calculations	(Neumann 2008)
CrystalPredictor	Low-discrepancy sequence	Flexible	Empirical repulsion-dispersion with distributed charges	(Karamertzanis and Pantelides 2007)
Polymorph Predictor	Monte Carlo	Flexible	Empirical	(Gdanitz 1992)
CRYSTALG	Monte Carlo	Flexible	Empirical repulsion-dispersion with atomic charges	(Pillardy et al. 2000)
MDCP	Molecular Dynamics	Rigid	Empirical free energy	(Tajima et al. 1995)
Metadynamics	Molecular Dynamics	Rigid	Empirical free energy potential with distributed atomic multipole model	(Karamertzanis et al. 2008b)
MGAC	Genetic Algorithm	Flexible	Semi-empirical repulsion dispersion potential with atomic charges	(Bazterra et al. 2004a)

Table 2.2 Reviewed methods for the accurate evaluation of the lattice energy (refinement) of crystal structures of flexible molecules.

Name	Lattice energy minimisation procedure	Selected Reference
UPACK	Simultaneous optimisation of molecular geometry and lattice parameters. Transferable QM intramolecular energy and atomic multipole moments computed for reference conformations only.	(van Eijck et al. 2001)
DMAflex	Bi-level optimisation of the molecular geometry (outer problem) and lattice parameters (inner problem). QM intramolecular energy and atomic multipole moments computed for every conformation encountered during the minimisation.	(Karamertzanis and Price 2006)
DMAflex_quick	Bi-level optimisation of the molecular geometry (outer problem) and lattice parameters (inner problem). Intramolecular energy read from a pre-computed grid. Transferable atomic multipole moments computed for one reference conformation only.	(Karamertzanis et al. 2009)
Day et al.	Initial optimisation of molecular geometry using empirical force-fields. Evaluation of the lattice parameters for the optimised geometry using single point QM intramolecular energy and atomic multipole moments.	(Day et al. 2007)
GRACE	Simultaneous optimisation of molecular geometry and lattice parameters using dispersion-corrected periodic solid-state density functional theory	(Neumann and Perrin 2005)

2.4.1. Crystal structure generation

The search for all possible crystal structures of a molecule is a complex multidimensional global optimisation problem. For rigid molecules, there are six degrees of freedom defining the unit cell (unit cell lengths and angles) and a further six variables (position and orientation) for each independent molecule in the unit cell. The conformational flexibility can be included by either considering several stable rigid conformations independently or explicitly varying a few torsional angles during the search. Whenever possible, the latter approach should be preferred as the molecular conformations, positions and unit cell parameters can all be simultaneously sampled for the generation of optimal packing arrangements into the crystal.

To further reduce the dimensionality of the variable space and facilitate the completeness of the search, many approaches rely on space group symmetry. Under such conditions, restrictions are imposed on the relationship between the molecules in

the unit cell and, in most cases, the unit cell is also constrained by the space group. Furthermore, around 90% of the organic and organometallic molecules have been found to crystallise in $Z' \leq 1$ (Steiner 2000). A further 80% of these structures occur in just six space groups ($P\bar{1}$, $P2_1$, $P2_1/c$, $P2_12_12_1$, $C2/c$ and $Pbca$) (Brock and Dunitz 1994). As a result, most of the possible crystal packings for a given molecule can be found by generating structures in a limited set of space groups with just one molecule in the asymmetric unit. However, the number of molecules in the asymmetric unit and a set of likely space groups need to be chosen in advance, and so important structures crystallising with more than one molecule in the asymmetric unit in less common space groups may be missed¹. Due to these risks, there are approaches that do not use symmetry (Tajima et al. 1995, Wawak et al. 1998, Trimarchi and Zunger 2007, Lucia and Gattupalli 2008) and perform searches on unconstrained unit cells, but usually with a fixed number of molecules in the unit cell (Z). Separate searches must then be performed for different Z and the space group of each structure is identified at the end of the calculations.

In the sections that follow, two broad classes of crystal structure generation approaches are presented. The first approach aims to gradually construct feasible unit cells from molecular aggregates. In the second approach, the crystalline periodicity is assumed from the beginning and the methods focus on the formation of many different trial unit cells.

Construction of molecular clusters

It is reasonable to assume that in order to be able to predict the crystal structure, it is necessary to simulate the initial stages of the crystallisation process by forming low-energy molecular clusters, assuming that the final crystal will grow from the constructed nucleus. Strong molecular interactions, such as hydrogen bonds can be identified before the construction of trial cells and are often decisive in the formation of the final crystal structure (Desiraju 1995, Aakeröy 1997, Sarma and Desiraju 2002).

¹ About 10% of homomolecular organic crystals have more than one molecule in the asymmetric unit (Steiner 2000). Molecular crystals are known in all 230 space groups, although for many only a few of the structures have been reported (Wilson 1993).

Therefore, in some cases, the candidate structure can be based on stable configurations of just a few molecules or clusters.

This concept is implemented in the PROMET algorithm (Gavezzotti 1991) and its subsequent extension (Gavezzotti 1999). The approach is based on the construction of stable one-dimensional aggregates that serve as building blocks for two-dimensional clusters, from which the three-dimensional structures can be generated. The one-dimensional dimers are used to construct chains, ribbons and subsequently layers of molecules that are related by the most common symmetry operators – translation, inversion, glide and screw. Many trial crystal structures can be generated from each stable cluster by systematically varying translations vectors. As a result, the six most common space groups with up to eight equivalent positions are allowed ($P\bar{1}$, $P2_1$, $P2_1/c$, $P2_12_12_1$, $C2/c$ and $Pbca$) (Brock and Dunitz 1994). The rank of the most stable candidates is further refined using the accurate semi-classical density sums (SCDS) PIXEL method (Gavezzotti 2005a) described in section 2.2 in more detail.

A similar approach is adopted in the statistical FlexCryst approach (Hofmann and Lengauer 1997, Hofmann and Lengauer 1998, Hofmann and Lengauer 1999). Initially, interaction centres are identified and evaluated by analysing possible intermolecular contacts of a given conformation. Once stable dimers have been generated, the translational vectors spanning the crystal are determined by matching the interaction centres of neighbouring unit cells. Once a large number of trial structures have been generated, the statistical scoring function (refer to section 2.2) is used to rank the candidates. FlexCryst is capable of generating crystal structures in four of the most common space groups ($P1$, $P\bar{1}$, $P2_1$ and $P2_12_12_1$). It was found that the success of the algorithm depends on the particular space group being considered as the statistical scoring function is heavily reliant on the nature of the experimental entries in the CSD.

The crystal structure generation methods described above are based purely on geometrical crystallography. The strength of these procedures is that plausible crystal structures are obtained much more quickly than through random searches of the crystal potential energy. However, since the crystal is gradually built by translation of molecular clusters, the methods are bound to fail when there are no particular stable substructures identified. Furthermore, for flexible molecules a systematic search over all possible conformations is computationally expensive due to the exceptionally large number of possibilities that has to be considered.

Unit cell generation

In this category, there are two main methods of generating crystal structures, either by systematically constructing lattices on a predefined grid of the multidimensional search space or generating crystal structures with random values for the parameters. In both cases the ranges of the variables to be sampled need to be provided at the start of the calculations. Additional methods originating from molecular and biological simulation are also described.

The PMC program (Packing of Molecules in Crystals) (Dzyabchenko et al. 1999) originated from pioneering work to predict crystal structures of benzene based on global energy minimisation (Dzyabchenko 1984). The approach generates candidate structures in a grid-based search, which are then energy-minimised using partial charges to represent the electrostatic interactions and an empirical repulsion-dispersion potential. Several conformational degrees of freedom can also be defined in the search. For many systems of interest, the size of the grid search can become unmanageable with the increasing number of degrees of freedom (e.g. torsional angles, additional molecules in the asymmetric unit etc.). As a result, generating structures randomly is usually a better approach for the higher-dimensionality problems.

Several algorithms, such as UPACK (Utrecht Crystal Packer) (van Eijck and Kroon 1999, van Eijck and Kroon 2000), allow the user to perform searches by either assigning random values or using a systematic grid based on the molecular system under the investigation. UPACK was initially developed for crystal structure prediction of monosaccharides, but later extended to cover flexible molecules in general. During the search, unreasonable structures (based on their density or repulsion thresholds) are discarded from the calculations and computationally efficient force-fields are used to perform preliminary lattice energy minimisation. The most promising crystal structures are kept and re-minimised with more accurate and elaborate computational models (described in greater detail in section 2.4.2).

The Crystal Structure Calculation algorithm (CRYSCA) (Schmidt and Englert 1996) is also based on lattice energy minimisation of random packings of flexible molecules. For *ab initio* predictions, the CSD is usually used to obtain some information on the preferred molecular conformations in the crystal structures of

similar compounds. Different conformations can then be treated separately or a simple intramolecular force-field incorporated into the search procedure. The algorithm randomly assigns lattice parameters, molecular positions, orientations and intramolecular degrees of freedom to generate candidate structures, which are then energy-minimised using a simple atom-atom potential.

The computationally prohibitive dispersion-corrected periodic density functional theory (DFT+D) cannot be readily used for structure generation. Instead, a method has been proposed (Neumann 2008) that uses reference DFT+D calculations to derive a molecule-specific atom-atom force field that includes both non-bonded and bonding terms. This tailor-made force field (TMFF) can then be used to generate candidate structures starting from random unit cells (Neumann and Perrin 2005, Neumann 2007). The purpose of the TMFF is that it is sufficiently cheap to be used during the search for crystal structures, but is still accurate enough so that only a small set of structures have to be refined with more accurate methods at a later stage.

A considerable improvement to the pure random number generators is the implementation of Sobol' deterministic low-discrepancy sequences (Sobol' 1967) during the generation of crystal structures (Della Valle et al. 2003, Karamertzanis and Pantelides 2005). In this approach, decision variables are generated simultaneously in a multi-dimensional variable domain as binary fractions from a set of special direction numbers¹. A characteristic of the approach is that the projection of *each* Sobol' point onto each of the axes corresponds to a *distinct* combination of every decision variable. Such quasi-random sequences avoid the clusters and gaps of points that are often seen in multidimensional random sampling. Consequently, a more uniform coverage of the search domain is guaranteed. Additional benefits are that the final number of trial structures does not have to be defined in advance and that the search could be continued (or restarted) from any point.

This approach has been successfully implemented in CrystalPredictor (Karamertzanis and Pantelides 2005, Karamertzanis and Pantelides 2007). The algorithm has the potential to generate structures with more than one molecule in the asymmetric unit in the most common space groups. The feasible candidates (as determined by their initial density, energy and lattice dimensions) are energy minimised. It is assumed that each molecule consists of a number of sites that can participate in

¹ Refer to the original literature for the details of the methodology and the generation of direction numbers (Sobol' 1967).

repulsion-dispersion and electrostatic interactions. The interaction sites do not have to coincide with atomic positions, allowing the determination of the distributed charges at “satellite” positions for better reproduction of the electrostatic potential (Karamertzanis and Pantelides 2004). A parallel implementation of the algorithm allows a reliable, yet relatively fast, exploration and location of deep energy minima in the lattice energy. Flexible molecules are modelled as a set of fully rigid fragments connected by flexible torsional angles. Quantum mechanical calculations are used, in an initial phase, to compute and tabulate the intramolecular energy and distributed charges as a function of the flexible degrees of freedom. This is computationally demanding and may become infeasible for computing grids with four or more torsions¹. Furthermore the distinction between flexible and fixed intramolecular degrees of freedom in the molecule introduces a degree of approximation which miscalculates the actual positions of the atoms in the molecule within the crystal lattice. As a result, the low-energy structures often have to be further refined with one of the more elaborate local lattice energy minimisation algorithms presented in section 2.4.2.

Random, quasi-random and grid methods aim to sample the entire multidimensional search space. However, it is only the low-energy regions that are of practical interest as they will yield the most likely stable trial structures. The difficulty lies in that these smaller regions cannot be easily identified in advance (hence the requirement of a complete search). Nevertheless, there are methods, borrowed from molecular simulation, that aim to preferentially sample those specific areas of the search space that exhibit a low energy.

One such method is the Monte Carlo approach where random changes (moves) are made to the structural degrees of freedom. The move is then accepted or rejected based on a probability that is related to the change in energy associated with that move according to a standard Metropolis algorithm (Metropolis et al. 1953). For instance, the step is always accepted if its energy has been reduced². The simulated annealing Monte Carlo algorithm has been implemented in the commercially available Polymorph Predictor (Gdanitz 1992, Karfunkel and Gdanitz 1992, Karfunkel et al. 1993). In order

¹ Note that CrystalPredictor is applicable to molecules with more than four torsions. Computational cost during grid construction is reduced by splitting the torsions into separate torsional groups. The grids are then computed for each torsional group assuming that there are no significant interactions between the torsional groups.

² If the relative energy has been increased, the trial structure has a probability of $\exp(-\Delta E/kT)$ being accepted, where ΔE is the associated change in energy with the move, k is the Boltzmann constant and T the temperature of the calculation (Metropolis et al. 1953).

to ensure that the energy surface is explored in great detail and to avoid the system being trapped in local minima, the simulation is performed in the presence of an externally imposed heating and cooling cycle (simulated annealing). Consequently, the energy barriers are easily overcome at the start of the simulation as the system temperature is gradually decreased. This reduces the time spent searching areas of high potential energy surface, while the use of a variable step size improves the search for good candidate structures. Each accepted step through the heating and cooling cycle is further refined by lattice energy minimisation. Due to the stochastic nature of the process, the simulated annealing procedure must be repeated several times to ensure a complete sampling of packing possibilities. Furthermore, the search needs to be repeated for every relevant space group which may become computationally expensive.

A modification to the above is the conformational-family Monte Carlo (CFMC) method, which was originally proposed for the problem of protein folding (Pillardy et al. 2000), but was later adapted to crystal structure prediction within the CRYSTALG package (Pillardy et al. 2001). In contrast, to the classical Monte Carlo procedure, the method uses a whole family of structures (not single structures) and only moves between these families are accepted or rejected. The structures are improved iteratively using a Metropolis-type Monte Carlo algorithm (Metropolis et al. 1953) and energy minimisation, in which the search procedure is biased towards the regions of the lowest energy families. No symmetry (space group) constraints are imposed and the calculations are repeated for a number of molecules in the unit cell. The method has been shown to perform well in identifying experimental structure in the case of rigid and simple flexible molecules (Pillardy et al. 2001). However, the search is unable to efficiently cover the conformational space for relatively flexible molecules containing a large number of torsional degrees of freedom.

Another tool borrowed from molecular simulation is molecular dynamics (MD) where atoms move according to Newton's equations of motion. The search for possible packing patterns is performed by allowing a small number of molecules to reach stable configurations while moving around in a unit cell of variable dimensions under periodic conditions. One such algorithm, Molecular Dynamics for Crystal Packing (MDCP) (Hirano et al. 1995, Arikawa et al. 1995, Tajima et al. 1997), starts with 8 molecules that are initially loosely packed within the flexible simulation box (unit cell). Identification of possible crystal structures is achieved by selecting points with low potential energy on the molecular trajectory. This heuristic molecular-dynamics

approach manages to efficiently sample the conformational space and identifies promising candidate structures that are used as initial points for subsequent energy minimisation. However, standard MD runs are inefficient at sampling all possible minima on the energy surface since high-energy barriers between structures tend to confine the simulation to the neighbourhood of a particular structure for long simulation times. The metadynamics simulation method (Martonak et al. 2005, Martonak et al. 2007) forces quicker transitions by a dynamic addition of a biasing potential which drives the simulation away from all the minima that have already been considered. It should be noted that considerably longer molecular dynamics runs have to be implemented in order to identify successfully all possible crystal structures for large and more complicated (non-symmetric) molecules due to the long equilibration times. This makes the MD approaches ineffective at producing a diverse set of hypothetical crystal structures and computationally infeasible for systems of practical interest.

An advantage of MD, however, is that being dynamic simulations, these methods sample the free energy which tends to be smoother than the temperature-free lattice energy surface, allowing low-energy barriers between related structures to be overcome. This results in a smaller number of candidate structures in a given neighbourhood of the search space (Raiteri et al. 2005, Karamertzanis et al. 2008b). A different approach is to artificially, mathematically, transform the multidimensional lattice energy surface by removing all insignificant energy minima (for instance through the application of the diffusion equation (Wawak et al. 1998)). The few lattice energy wells that remain after the transformation are expected to be traces of the deepest minima of the original surface that can be tracked back by applying a reversing procedure.

Genetic algorithms, based on the ideas of biological evolution, have also been employed for crystal structure prediction. Usually, an initial population of structures, represented by descriptors of the crystal structure (genomes), is randomly generated. A genetic algorithm then assigns random mutations that evolve the structures. A fraction of the most stable individual structures are copied directly to the next generation and the mutation process is recursively repeated until certain convergence criteria are met. MGAC (Modified Genetic Algorithm for Crystal and Cluster structures) (Bazterra et al. 2002a, Bazterra et al. 2002b, Bazterra et al. 2004a, Bazterra et al. 2004b, Kim et al. 2009) was proposed to minimise the lattice energy using a semi-empirical potential function and was later extended to allow the simultaneous relaxation of molecular geometry during the optimisation. Other methods (e.g. (Motherwell 2001)) control the

selection process by the means of penalty functions derived from structural information in the CSD. The genetic algorithm tries to reproduce a fitness function of intermolecular contacts of similar molecules (or fragments) present in the databases. These penalty functions can also be defined to eliminate or avoid generating structures without certain structural features.

A host of diverse search methodologies has been applied to crystal structure prediction. To assess their relative strengths, several direct comparisons of the algorithms have been made via a series of blind tests (Lommerse et al. 2000, Motherwell et al. 2002, Day et al. 2005a, Day et al. 2009, Bardwell et al. 2011) described in more detail in section 2.4.4. A recent general trend in search methods development has been to address multi-component crystals and larger flexible molecules typical of the pharmaceutical industry.

2.4.2. Local minimisation of lattice energy

In order to be able to search what is a very large multidimensional variable space of possible crystal structures, search methods have to rely on computationally efficient (though not exceptionally accurate) estimates for the intra- and intermolecular energy contributions. As a result, the absolute energy and, consequently, the ranking of low-energy crystal structures identified by the search methods are not reliable. They, therefore, have to be corrected by re-minimisation of the crystal lattice energy using more accurate models. An overview of the available methods for flexible molecules is provided below.

One of the first algorithms to take into account molecular flexibility in lattice energy minimisation is the extended version of UPACK (van Eijck et al. 2001) presented in the previous section. In this approach, the intramolecular energy is calculated using a quadratic approximation constructed from the results of an *ab initio* molecular geometry optimisation. The intermolecular potential is fitted to high-level quantum mechanical calculations of alkanes, alcohols and ethers (Mooij et al. 1999a, Mooij et al. 1999b) and involves terms for atomic multipole moments, dipole polarisabilities and repulsion-dispersion contributions. In order to model the conformational dependence of the electrostatic model, the atomic multipole moments are defined in terms of their local-axis system and rotated with the local environment.

Following a significant change in conformation during lattice energy minimisation, the intramolecular potential and the electrostatic model are recalculated to maintain accuracy. A feature of this approach is that the use of Cartesian coordinates for the representation of the molecular structure during lattice energy minimisation forces the user either to neglect flexibility altogether (rigid-body approach) or to account for full flexibility (atomistic representation). In the latter case, the computational cost becomes prohibitive for any molecule of non-trivial size (more than 20 atoms).

An algorithm that allows the optimisation of crystal structures with user-defined flexibility is DMAflex (Karamertzanis and Price 2006). This consists of a bi-level optimisation problem. The inner minimisation determines the minimum crystal structure for rigid molecular entities, whose conformation is determined by the outer minimisation which manipulates several flexible degrees of freedom (usually only torsions). The rigid-body inner minimisation is solved using the DMACRYS algorithm (Willock et al. 1995, Price et al. 2010) which incorporates an accurate distributed multipole model (Stone 2005) for the calculation of the electrostatic contributions to the lattice energy. DMAflex incorporates a full isolated-molecule quantum mechanical molecular geometry optimisation and charge density calculation at every outer iteration, as the intramolecular energy and atomic multipoles need to be recalculated after any conformational change. This results in very high computational cost. The computational burden is further compounded by the use of a gradient-free (simplex) minimisation algorithm (Nelder and Mead 1965) in the outer minimisation that limits the extent of molecular flexibility that can be practically handled to a small number (less than 10) of torsional angles.

A recent modification of the algorithm, DMAflex_quick (Karamertzanis et al. 2009), was proposed that substantially reduces the computational cost associated with quantum mechanical calculations. During the minimisation, the molecule is treated as a set of fully rigid fragments connected by flexible torsions for which the intramolecular energy is interpolated from a pre-computed grid (usually available from the search stage). The atomic multipoles are computed just once for a reference conformation and are expressed with respect to the local axis system of each atom. The multipoles are then used to model the conformationally dependent intermolecular electrostatic interactions by assuming they remain invariant with respect to the local axis system (as in UPACK). Despite its computational efficiency, the approximations in the estimation of the molecular geometry, the intramolecular energy and the charge density lead to

significant errors in terms of structure reproduction and energy rank. As a result, this tool should only be used as an intermediate pre-processor of many search structures prior to the more accurate lattice energy minimisations with DMAflex (or other methods).

Another computationally efficient approach to lattice energy minimisation of flexible molecules is a two-stage procedure proposed by Day and co-workers (Day et al. 2007, Day and Cooper 2010). First, the crystal structure is minimised by allowing the molecular geometry to relax using a molecular mechanics description of energies associated with the changes in the flexible torsion angles. The method trusts the molecular mechanics force field to provide the adequate molecular geometries, but discards the energy, which is not of sufficient accuracy for the final ranking of crystal structures. The resulting crystal structures are then re-optimised without further changes to molecular conformations using DMACRYS (Willock et al. 1995, Price et al. 2010) to model the electrostatic interactions using the atomic multipole model. The intramolecular energy and the atomic multipole moments for each conformation are obtained from a single point isolated-molecule DFT calculation which also usually incorporates the polarisable continuum model (PCM) (Tomasi et al. 2005) with a constant dielectric to reflect the polarisation effects in the crystalline environment.

One of the most accurate lattice energy minimisation methods available for the minimisation of crystal structures is the GRACE package (Generation, Ranking and Characterisation Engine) (Neumann and Perrin 2005, Neumann 2007). The method incorporates a hybrid approach by combining the periodic solid-state density functional theory calculations with an empirical dispersion correction (DFT+D approach). During the minimisation, all nuclear positions and electron density are optimised simultaneously, naturally taking in account molecular flexibility and other charge density reorganisation effects such as polarisation. However, the computationally expensive hybrid DFT+D method can only be used for the final energy ranking of a few selected structures. Despite the computational cost, the approach has shown to reproduce experimental structures very accurately and rank them favourably amongst other hypothetical structures in many, but not all, cases.

2.4.3. *A posteriori* analysis of low-energy crystal structures

Crystal structure prediction techniques usually identify a very large number of distinct competitive low energy minima. This number is definitely higher than the number of polymorphs that can be expected to be found in the real systems. It is therefore necessary to complement the lattice energy minimisation algorithms with post-processing analysis that will sieve and reject those structures that are similar or shown to be unlikely to grow under experimental conditions.

Lattice energy minimisation algorithms typically produce thousands of structures many of which are identical and indistinguishable if thermal effects are taken into account. Therefore clustering algorithms, which classify similar structures into groups, are used at final and/or intermediate stages to remove duplicate structures. Since many thousands of structures have to be compared, the clustering procedure must be both quick and discriminating so that distinct minima are identified rapidly, and similar structures discarded. The most frequently used approaches include the comparison of the inter-atomic distances (Karamertzanis and Pantelides 2005), molecular positions and orientations (Dzyabchenko 1994, van Eijck and Kroon 1997), molecular coordination spheres (Chisholm and Motherwell 2005) or reduced cells (Krivy and Gruber 1976, Spek 2003). Most of the clustering algorithms may be further tuned to achieve the desired extent of structure separation.

The thermodynamically most favourable experimental structures are the structure with the lowest Gibbs free energy. However, the lattice energy calculated in most crystal structure prediction algorithms neglect any vibrational motion taking place in the crystal. Initial attempts to include molecular vibrations have been made through the use of lattice dynamics (Dunitz et al. 2000, van Eijck 2001, Anghel et al. 2002, Day et al. 2003). These studies have shown that the ranking of the crystals is affected when the vibrational contributions to the free energy are taken into account. However, the accuracy to which the frequencies can be calculated is very sensitive to the potentials, which, in most cases, are not parameterised for these purposes. A further limitation of the lattice dynamics approach is that many crystals are quite close to their melting point at ambient conditions. At these temperatures the presence of wide-amplitude anharmonic motions violates the theoretical basis behind lattice dynamics. A more comprehensive way to consider the kinetic effects is, therefore, by molecular dynamics

(MD). Not only do MD simulations allow an evaluation of dynamic contributions to the energies of crystal structure, but the presence of temperature allows the investigation of transitions between structures separated by low-energy barriers.

It appears so far that most of the minima on the lattice energy surface of hydrogen bonded systems remain independent minima on the free energy surface (Karamertzanis et al. 2008b)¹. Additional considerations, such as the calculation of crystal properties and morphology can also be used to assist in the reduction of possible polymorphs predicted using typical lattice energy minimisation algorithms. For instance, the mechanical stability of different crystals can be evaluated through the calculation of the elastic constants, which can be readily done for rigid molecules (Day et al. 2001). Several models for predicting crystal morphology have also been reported (Hartman and Perdok 1955, Hartman and Bennema 1980, Bisker-Leib and Doherty 2001, Bennema et al. 2004, Snyder and Doherty 2009). The comparison of the relative growth rates of different faces in the crystal is usually used to describe the relative growth rates of competing crystals and thus distinguish which crystals are likely to be observed experimentally.

In addition (or as an alternative) to calculating physical properties, the likelihood of particular low-energy crystals may be assessed by analysing the wealth of the information available in crystal structure databases such as the CSD. Common patterns or interactions in the database may reflect thermodynamic and kinetic preferences and highlight the most likely candidates in the list of low-energy crystal structures. Statistical models for hydrogen bond propensity prediction trained against data from crystal structures of molecules with similar functional groups have been used in the past to help rationalise and refine the rank of crystal structures (Galek et al. 2007, Galek et al. 2009). Statistical frequencies of particular intramolecular geometries and intermolecular interactions are also used to rank (and re-rank) hypothetical crystal structures and form a basis in several fitness functions used for crystal structure prediction (Motherwell 2001, Hofmann and Apostolakis 2003).

¹ However, earlier study of benzene (for which the van der Waals dominated energy surface exhibits much lower energy barriers) showed that many lattice energy minima converge to a very small number of free energy minima (Raiteri et. al. 2005).

2.4.4. CCDC Blind Tests of Crystal Structure Prediction

It is difficult to assess the progress in general predictive techniques since the published literature mainly contains success stories for known (selected) polymorphic systems. For example, in a survey of 50 small rigid organic molecules (Day et al. 2005b), in the majority of cases, the experimentally observed structures were found to be either the global minimum in lattice energy (i.e. lowest known energy level) or a competitive low-energy structure. In an independent survey of 253 studies published in the academic literature (Beyer et al. 2001), there were only 26 reported failures, which almost certainly underestimates the difficulties and challenges involved in *ab initio* crystal structure prediction.

The Cambridge Crystallographic Data Centre (CCDC) has arranged for a more objective assessment of different crystal structure prediction methodologies by carrying out a series of blind tests roughly every three years since 1999 (Lommerse et al. 2000, Motherwell et al. 2002, Day et al. 2005a, Day et al. 2009, Bardwell et al. 2011). In these international collaborations, the molecular diagrams of a set of target structures (Table 2.3) are disseminated to the participants, while the experimental information is withheld. Each participant is required to submit three possible crystal structures for each molecule, along with an extended list (usually 100) of low-energy structures within a certain timeframe (usually around 6 months). The target molecules are selected according to different categories of perceived difficulty:

1. Small rigid molecules containing less than 25 atoms with C, H, N and O atoms only.
2. Larger rigid molecules containing less common functional groups or elements (e.g. halogens, S, P, Se, B but not organometallics).
3. Molecules with several (2-4) torsional degrees of flexibility.
4. Multiple independent molecules in the unit cell (e.g. solvates, cocrystals, and molecular salts).
5. Large molecules (50-60 atoms) with significant conformational flexibility.

In the first two blind test (CSP1999 and CSP2001), the participants were informed that the crystals were in the most common space groups with $Z' \leq 1$. The space group

restrictions were lifted in subsequent blind tests, but all monomolecular crystals were guaranteed to adopt structures with $Z' \leq 2$ in CSP 2004 and $Z' \leq 1$ in CSP2007 and CSP2010 (except molecule XX which was specified as $Z' \leq 2$). Two new categories were introduced in CSP2007 and CSP2010 to test predictions specifically for problems of pharmaceutical interest – structures with more than one molecule in the asymmetric unit (category 4) and larger flexible molecules (category 5) respectively.

The promising results for categories 1 and 2 (Table 2.3) have shown that the generation of correct structures (at least somewhere in the extended list of hypothetical crystal structures) for small rigid molecules is no longer a major issue for many of the current prediction methods. The shortcomings mainly arise in the treatment of the anisotropic intermolecular interactions of the less common functional groups, in particular the halogen atoms.

The success for the multi-component crystals and flexible molecules in terms of both generation and ranking has usually been poor. Improving the predictability of these categories of molecules requires substantial advances in the search procedures for the packing of several independent and conformationally flexible molecules, and a step increase in the accuracy of models for the intra- and intermolecular interactions. The development of novel methodologies and algorithms have made it possible for two groups to successfully predict the crystal structure of molecule XX, the largest and most flexible molecule considered in the blind tests to date (Kazantsev et al. 2011b). One of these methods is due to Day and co-workers (Day et al. 2007, Cooper et al. 2008) already described in the sections above, the other approach will be presented in greater detail in this thesis in Chapter 6.

The blind tests provide an objective evaluation of the possibilities and limitations of current methods involved in crystal structure prediction. However, the degree of difficulty of each target can never be known in advance and can have an important influence on the success (or otherwise) of the methods. There are still technical and fundamental obstacles to reliable crystal structure prediction, but significant progress has been made over the last decade. It is encouraging to see that one method, based on dispersion-corrected periodic density functional theory (Neumann and Perrin 2005), managed to identify all four experimental crystal structures in CSP2007 (Neumann et al. 2008) and the small molecules in the first three categories in CSP2010 as the most stable structures. It should also be acknowledged that the comparison and analysis of

the “failed” predictions has served to highlight the immediate areas requiring attention and possible methodological innovation.

Table 2.3 Diagrams and reference numbers (in Roman numerals) of the target molecules in the blind tests of crystal structure prediction. The three values below the target reference numbers indicate, from left to right, the total number of participants, the number of participants who obtained the correct experimental structure in their extended list and among their three best predictions for each target respectively.

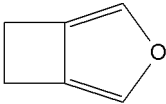
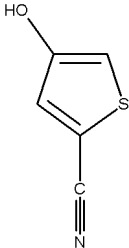
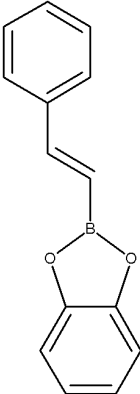
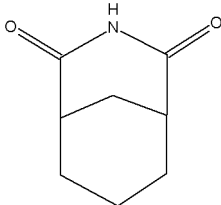
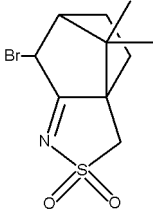
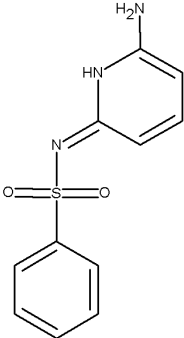
Category 1			Category 2			Category 3			Category 4			Category 5		
<i>CSP1999</i>			(Lommerse et al. 2000)											
									-			-		
I			II			III			-			-		
11	-	4	8	-	1	11	-	1	-	-	-	-	-	-
<i>CSP2001</i>			(Motherwell et al. 2002)											
									-			-		
IV			V			VI			-			-		
16	10	3	15	11	4	11	4	0	-	-	-	-	-	-

Table 2.3 continued.

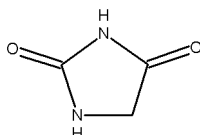
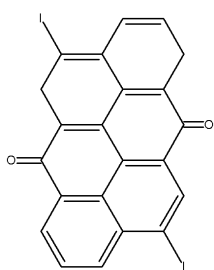
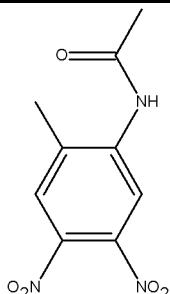
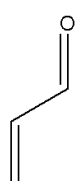
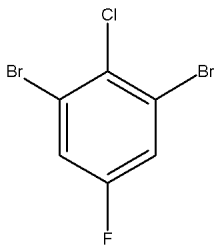
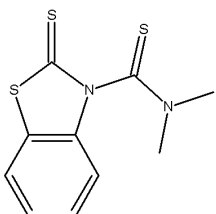
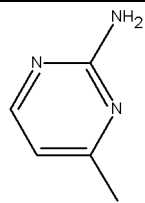
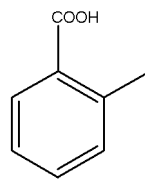
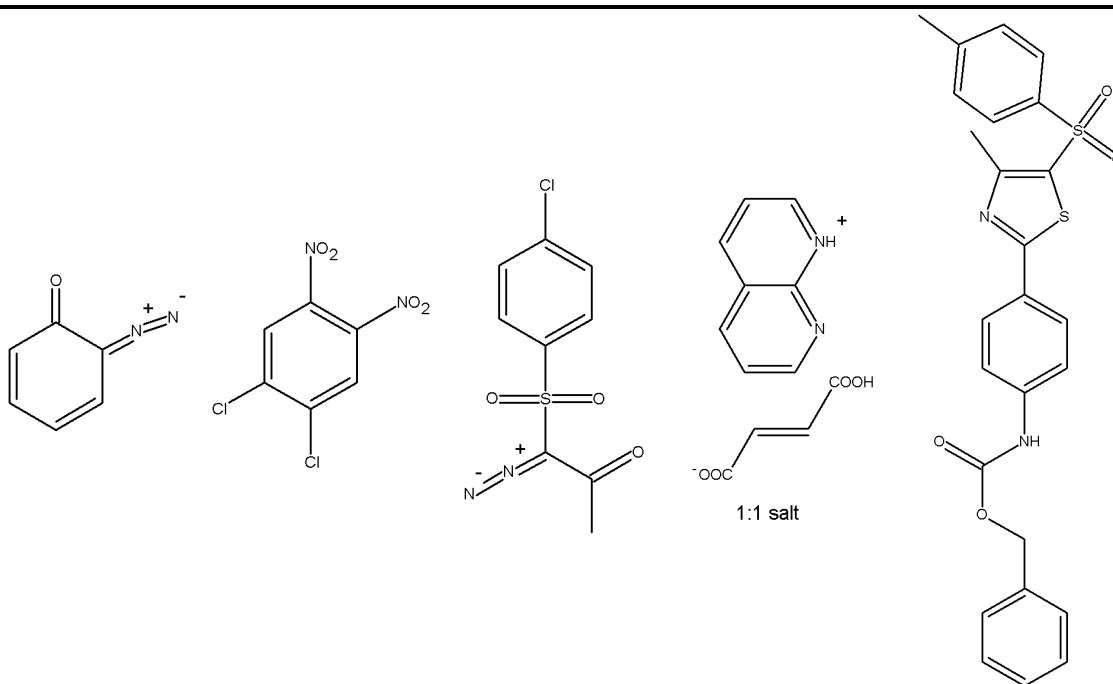
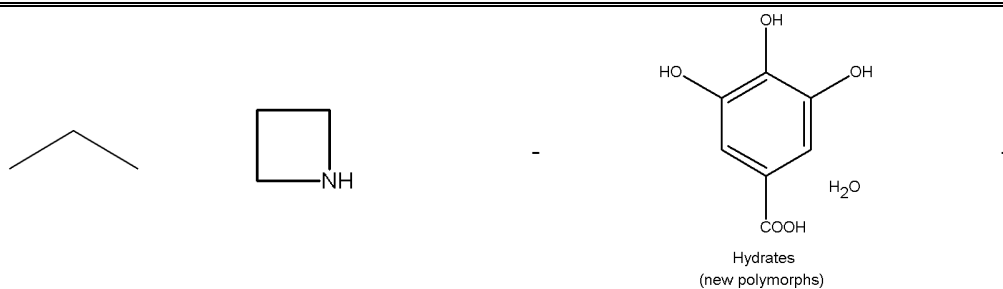
Category 1			Category 2			Category 3			Category 4			Category 5		
CSP2004			(Day et al. 2005a)											
									-			-		
VIII			IX			X			-			-		
14	11	5	16	9	1	15	7	0	-	-	-	-	-	-
CSP2007			(Day et al. 2009)											
												-		
														
1:1 cocrystal														
XII			XIII			XIV			XV			-		
13	10	4	14	9	4	12	9	3	12	5	2	-	-	-

Table 2.3 continued.

Category 1	Category 2	Category 3	Category 4	Category 5
<i>CSP2010</i> (Bardwell et al. 2011)				



XVI			XVII			XVIII			XIX			XX		
15	7	2	13	6	2	13	4	1	11	4	2	10	3	2

Extra molecules

VII (<i>CSP1999</i>)			XI (<i>CSP2001</i>)			-			XXI (<i>CSP2010</i>)			-		
6	-	1	18	4	0	-	-	-	10	5	0	-	-	-

2.5. Conclusions

A large number of crystal structure prediction methods, with varying capabilities and characteristics have been developed in the academic community. However, despite recent progress, there is currently no reliable methodology that can be consistently used for crystal structure prediction of a diverse range of molecules and multi-component systems.

The accuracy of the lattice energy model is of primary importance in crystal structure prediction. It is frequently found that the use of more accurate force fields leads to better ranking of the minimised experimental structure. Several promising anisotropic non-empirical approaches based on quantum mechanical calculations of isolated molecules or periodic systems are being pursued as an alternative to the standard isotropic atom-atom potential.

Conformational flexibility, exhibited by most molecules of interest, presents and remains one of the biggest challenges for crystal structure prediction. The lattice energy landscape becomes particularly complex for such systems. The use of accurate and balanced intra- and intermolecular potentials, usually derived from quantum mechanical calculations, along with an extensive search of the solution space, increases the computational cost substantially and limits the size of molecules that can be handled practically. A critical assessment of the effects of various approximations introduced to reduce the computational cost is required.

The complexity of the crystal structure generation step is further increased when more than one chemically independent molecule is included in the search (as is the case for cocrystals, salts and solvates). However, constraining the search to statistically significant combinations of space groups (to reduce computational cost) must be done with caution in order not to miss any relevant structures.

The issue of the complexity of crystal structure prediction due to molecular flexibility is explored in detail in this thesis. Other aspects, which also play an important role, are not tackled in this work. In particular, number of low-energy minima, predicted by most crystal structure prediction algorithms, is almost certainly larger than the number of attainable experimental crystals. Additional criteria based on statistical comparison to structural databases (such as the CSD), the calculation of the free energies, thermal stability, elastic properties and morphologies can be further used to

distinguish between the predicted structures. However, at present there are no definite conclusions on their merits and drawbacks. It is, therefore, necessary to apply the criteria based on physical properties with caution, until sufficiently accurate methods, validated on a large number of systems of both rigid and flexible molecules, are widely available.

It should not be forgotten that the thermodynamic model for crystal structure prediction is only the starting point. Crystallisation is a kinetic process where nucleation and relative crystal growth play an important part in guiding the formation of a specific crystal form (Yu 2007). Molecular simulations of crystal growth are advancing in accuracy and sophistication. However, the modelling of nucleation of molecular crystals remains a largely unexplored area.

Chapter 3. Local Approximate Models (LAMs)

When considering molecular flexibility during lattice energy minimisation explicitly, it is necessary to have a representation of the intra- and intermolecular energy contributions at every iteration in the minimisation procedure. In order to calculate the lattice energy accurately, quantum mechanical calculations for the intramolecular energy and charge density have to be employed. Consequently, the incorporation of a set of QM calculations within an optimisation loop results in prohibitive computational cost for most systems of interests.

Local approximate models (LAMs), based on Taylor series, can be used as an alternative method to provide an estimate of the intramolecular energy and the conformationally-dependent charge density. These models are practically as accurate as explicit quantum mechanical calculations, but carry a much smaller computational burden. The exact mathematical formulation and the applicability of LAMs to crystal structure prediction have been published previously in Kazantsev et al. (2010) and Kazantsev et al. (2011a), but are presented in greater detail in this chapter.

3.1. Molecular Model

As an alternative representation to the Cartesian coordinates, the position of a molecule can be completely defined by its Z-matrix consisting of $3N-6$ intramolecular degrees of freedom (bond lengths, bond and torsional angles), where N is the number of atoms in the molecule, and further 6 variables defining the Cartesian coordinate of a reference atom and the orientation of three reference atoms (Dunitz 1979).

The Z-matrix, also known as an internal coordinate representation, describes the position of the atoms relative to other (previously defined) atoms in the molecule. Usually, Z-matrices are written in terms of bond lengths, angles and torsions which preserve the actual bonding characteristics and therefore also contain the molecular connectivity information. An illustrative Z-matrix for the gas-phase geometry of formamide (computed at the MP2/6-31G(d,p) level of theory, Figure 3.1) is given in Table 3.1 and described in more detail below.

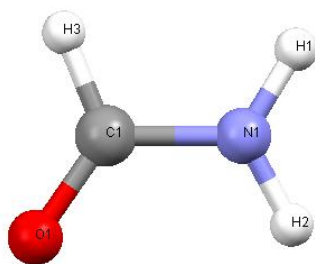


Figure 3.1 Molecular diagram and atom labels of formamide.

Table 3.1 Sample Z-matrix for the gas-phase geometry of formamide computed at MP2/6-31G(d,p) level of theory. Refer to Figure 3.1 for molecular diagram and atom labels.

#	Atom	Atom 2	Bond length (Å)	Atom 3	Bond angle (°)	Atom 4	Torsion angle (°)
1	C1						
2	N1	C1	1.362				
3	O1	C1	1.224	N1	124.74		
4	H1	N1	1.003	C1	121.37	O1	173.95
5	H2	N1	1.006	C1	118.47	O1	4.89
6	H3	C1	1.101	N1	112.17	H1	-6.95

The first two columns in the Z-matrix in Table 3.1 provide the respective numbers and labels of all the atoms in the molecule. The first row (atom C1) shows the reference atom relative to which all other atoms are defined. The third and fourth column define which atoms are directly bonded to the atom in column two and give the distance of this bond respectively. For instance, atom 2 (N1) is directly bonded to C1 with a bond length of 1.362 Å. The fifth and sixth columns provide a third reference atom and the bond angle between the three atoms respectively. For example, for the third atom in the Z-matrix (O1) the bond angle O1-C1-N1 is 124.74°. Finally, the last two columns provide the fourth reference atom and the torsion angle between the four atoms respectively. As an example of all of the above, the sixth atom in the Z-matrix, H3, is defined relative to atoms C1, N1 and H1 with bond length H3-C1 = 1.101 Å, bond angle H3-C1-N1 = 112.17° and torsion H3-C1-N1-H1 = -6.95°.

Overall, the Z-matrices are an efficient means of storing molecular geometry and can be automatically generated. A more detailed description of the Z-matrix generation procedure adopted in this work is provided in Appendix A and the conversion of Z-matrix variables to Cartesian coordinates is outlined in section 5.2.

3.2. Treatment of Molecular Flexibility

As described above, the geometry of a flexible molecule can be completely defined by its Z-matrix consisting of the set of all intramolecular degrees of freedom θ (torsion angles, bond angles and bond lengths). In molecular crystals, the intermolecular forces are significantly weaker than the energy of typical covalent interactions. Consequently, only a subset of the intramolecular degrees of freedom, θ , is expected to deviate significantly from their gas-phase values. These flexible degrees of freedom, θ^f , (such as torsions around single bonds), are often sufficient in capturing the effect of molecular flexibility and thus need to be explicitly modelled during lattice energy minimisation (Figure 3.2).

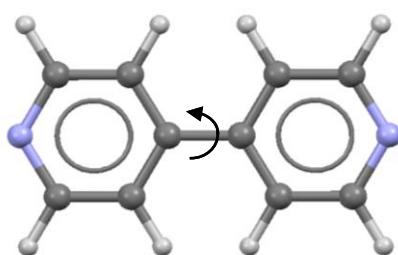


Figure 3.2 One flexible torsion, θ^f , shown as a black arrow captures most of the flexibility in 4,4'-bipyridine. The geometry of the aromatic rings can be assumed to be “rigid” for the purpose of lattice energy minimisation.

However, as the values of θ^f for a given molecule change significantly, the rest of the intramolecular degrees of freedom adjust so as to minimise the intramolecular energy. Hence, these remaining, more rigid degrees of freedom, θ^r , (such as torsions in aromatic systems, most bond angles and bond lengths), and the intramolecular energy, ΔE^{intra} , can be approximated as functions of the flexible degrees of freedom, θ^f , at the solution of a constrained isolated-molecule quantum mechanical geometry optimisation:

$$\Delta E^{intra}(\theta^f) = \min_{\theta^r} [E^{intra}(\theta^r; \theta^f)] - E^{vac} \quad \text{Eq. 3.1}$$

where E^{vac} is the global (or at least a local) minimum gas-phase molecular energy used as a correction and needs to be computed only once. Because the evaluation of ΔE^{intra} requires a minimisation with respect to the “rigid” degrees of freedom, θ^r , these are not

truly constant. The change $\boldsymbol{\theta}^r$ in response to changes in the flexible degrees of freedom, $\boldsymbol{\theta}^f$, can be seen explicitly in the following equation:

$$\boldsymbol{\theta}^r(\boldsymbol{\theta}^f) = \arg \min_{\boldsymbol{\theta}^r} [E^{\text{intra}}(\boldsymbol{\theta}^r; \boldsymbol{\theta}^f)] \quad \text{Eq. 3.2}$$

where “arg min” denotes the value of the optimisation variables at the solution of the minimisation problem shown in equation 3.1. These changes, however small, will have an effect on the intramolecular energy, molecular geometry and final structure reproduction (especially for large molecules) and therefore cannot be ignored by keeping $\boldsymbol{\theta}^r$ fixed at some nominal values, such as those in the gas-phase conformational minimum. The validity of this partitioning of intramolecular degrees of freedom will be examined in section 4.4.4 by contrasting the results obtained with different sets of $\boldsymbol{\theta}^f$. The main benefit of considering “rigid” degrees of freedom is a reduction in computational cost with little loss of accuracy. However, the approach presented is also valid when the vector $\boldsymbol{\theta}^r$ is empty, i.e. when all intramolecular variables are treated as flexible.

3.3. Intramolecular Energy LAM

The intramolecular energy for a given conformation in close proximity to a reference conformation, $\boldsymbol{\theta}_{ref} \equiv (\boldsymbol{\theta}_{ref}^f, \boldsymbol{\theta}_{ref}^r)$, can be estimated using a local approximate model (LAM) based on a quadratic Taylor expansion:

$$\begin{aligned}
\Delta E^{\text{intra}}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r) = & \Delta E^{\text{intra}}(\boldsymbol{\theta}_{\text{ref}}) + \left[\frac{\partial \Delta E^{\text{intra}}}{\partial \boldsymbol{\theta}^f} \right]_{\boldsymbol{\theta}_{\text{ref}}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) \\
& + \left[\frac{\partial \Delta E^{\text{intra}}}{\partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}}^T (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r) \\
& + \frac{1}{2} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f)^T \left[\frac{\partial^2 \Delta E^{\text{intra}}}{\partial \boldsymbol{\theta}^{f2}} \right]_{\boldsymbol{\theta}_{\text{ref}}} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) \\
& + (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r)^T \left[\frac{\partial^2 \Delta E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) \\
& + \frac{1}{2} (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r)^T \left[\frac{\partial^2 \Delta E^{\text{intra}}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{\text{ref}}} (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r)
\end{aligned} \tag{Eq. 3.3}$$

The above equation is valid for *any* reference point $\boldsymbol{\theta}_{\text{ref}} \equiv (\boldsymbol{\theta}_{\text{ref}}^f, \boldsymbol{\theta}_{\text{ref}}^r)$ in the coordinate space. Throughout this work, the subscript “*ref*” denotes the point around which the Taylor expansion is constructed. The values of $\boldsymbol{\theta}_{\text{ref}}^r$ at the reference point are obtained via an isolated-molecule quantum-mechanical constrained optimisation (equation 3.2) for the fixed values $\boldsymbol{\theta}^f = \boldsymbol{\theta}_{\text{ref}}^f$. This calculation also yields the minimum molecular deformation energy, $\Delta E^{\text{intra}}(\boldsymbol{\theta}_{\text{ref}})$, and the first- and second- order derivatives of ΔE^{intra} with respect to all intramolecular degrees of freedom, $\boldsymbol{\theta}_{\text{ref}}$, necessary to construct the LAM. Since $\boldsymbol{\theta}_{\text{ref}}^r$ is obtained by minimising ΔE^{intra} , it must satisfy the first-order optimality condition:

$$\left[\frac{\partial \Delta E^{\text{intra}}}{\partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}^r, \boldsymbol{\theta}_{\text{ref}}^f} = \mathbf{0} \tag{Eq. 3.4}$$

We require that equation 3.4 also apply to the intramolecular energy LAM. Hence, if $\boldsymbol{\theta}^f$ is changed by a small amount $\delta \boldsymbol{\theta}^f$ from the reference value, then $\boldsymbol{\theta}^r$ needs to change so that the intramolecular energy remains at a minimum. This can be enforced by ensuring that the first-order optimality conditions continue to be satisfied at $(\boldsymbol{\theta}^r + \delta \boldsymbol{\theta}^r, \boldsymbol{\theta}^f + \delta \boldsymbol{\theta}^f)$:

$$\left[\frac{\partial \Delta E^{\text{intra}}}{\partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}^r + \delta \boldsymbol{\theta}^r, \boldsymbol{\theta}_{\text{ref}}^f + \delta \boldsymbol{\theta}^f} = \mathbf{0} \tag{Eq. 3.5}$$

where $\delta\theta^r$ is the corresponding change in θ^r . Performing a first-order Taylor expansion, subtracting equation 3.4 from equation 3.5 and solving for $\delta\theta^r$ yields the approximate expression (refer to Appendix B for a more detailed derivation):

$$\delta\theta^r \approx - \left[\frac{\partial^2 \Delta E^{\text{intra}}}{\partial \theta^{r2}} \right]_{\theta_{\text{ref}}}^{-1} \left[\frac{\partial^2 \Delta E^{\text{intra}}}{\partial \theta^f \partial \theta^r} \right]_{\theta_{\text{ref}}}^T \delta\theta^f \quad \text{Eq. 3.6}$$

which then allows the approximation of θ^r via an explicit linear function of θ^{f1} :

$$\theta^r(\theta^f) = \theta_{\text{ref}}^r + \mathbf{A}(\theta_{\text{ref}})(\theta^f - \theta_{\text{ref}}^f) \quad \text{Eq. 3.7}$$

where the matrix $\mathbf{A}(\theta_{\text{ref}})$ is defined as:

$$\mathbf{A}(\theta_{\text{ref}}) \equiv \frac{\partial \theta^r}{\partial \theta^f} \equiv - \left[\frac{\partial^2 E^{\text{intra}}}{\partial \theta^{r2}} \right]_{\theta_{\text{ref}}}^{-1} \left[\frac{\partial^2 E^{\text{intra}}}{\partial \theta^f \partial \theta^r} \right]_{\theta_{\text{ref}}}^T \quad \text{Eq. 3.8}$$

By substituting equation 3.7 into equation 3.3 and also taking account equation 3.5, an estimate for the intramolecular energy as a quadratic function solely of the flexible degrees of freedom is obtained (refer to Appendix D for a detailed derivation):

$$\begin{aligned} \Delta E^{\text{intra}}(\theta^f) &= \Delta E^{\text{intra}}(\theta_{\text{ref}}^f) + \mathbf{b}(\theta_{\text{ref}})^T (\theta^f - \theta_{\text{ref}}^f) \\ &\quad + \frac{1}{2} (\theta^f - \theta_{\text{ref}}^f)^T \mathbf{C}(\theta_{\text{ref}}) (\theta^f - \theta_{\text{ref}}^f) \end{aligned} \quad \text{Eq. 3.9}$$

where the vector $\mathbf{b}(\theta_{\text{ref}})$ and matrix $\mathbf{C}(\theta_{\text{ref}})$ are defined as:

$$\mathbf{b}(\theta_{\text{ref}}) \equiv \left[\frac{\partial \Delta E^{\text{intra}}}{\partial \theta^f} \right]_{\theta_{\text{ref}}} \quad \text{Eq. 3.10}$$

¹ An alternative derivation of the approximation of the rigid degrees of freedom with the flexible degrees of freedom can be found in Appendix C.

$$\mathbf{C}(\boldsymbol{\theta}_{ref}) \equiv \left[\frac{\partial^2 \Delta E^{intra}}{\partial \boldsymbol{\theta}^{f2}} \right]_{\boldsymbol{\theta}_{ref}} - \left[\frac{\partial^2 \Delta E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}} \left[\frac{\partial^2 \Delta E^{intra}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{ref}}^{-1} \left[\frac{\partial^2 \Delta E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}}^T \quad \text{Eq. 3.11}$$

Equations 3.7 and 3.9 allow the explicit and fast calculation of the values of $\boldsymbol{\theta}^r$ and ΔE^{intra} for any given values of $\boldsymbol{\theta}^f$ without performing new quantum mechanical calculations; the values obtained are accurate in the proximity of a reference point $\boldsymbol{\theta}_{ref}$. If a molecule is optimised atomistically, i.e. all intramolecular degrees of freedom are treated as flexible, then equation 3.9 reduces to the standard quadratic Taylor expansion in UPACK (van Eijck et al. 2001).

3.4. Electrostatic Model LAM

For limited conformational changes, the conformational dependence of the intermolecular electrostatic model can be captured by rotating the multipole moments with their local environment (van Eijck et al. 2001, Karamertzanis et al. 2009). Once the distributed multipole moments (Ω) have been computed (Stone 2005) (up to the hexadecapole level) for a reference molecular conformation, $\boldsymbol{\theta}_{ref}$, each atom is assigned a local axis system using two directly connected atoms (or first and second bonded atoms for the case of terminal atoms). The calculated multipole moments are then converted to their Cartesian form and rotated to the local axis system of each atom. The locally expressed multipoles are kept constant for small conformational changes during lattice energy minimisation. The conformational variability of the electrostatic model is limited to the analytical rotation (using Cartesian tensors) of the local atomic multipoles to the molecular axis system of each newly generated conformation:

$$\begin{aligned} \Omega_{k_1 k_2 \dots k_n}^i(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r(\boldsymbol{\theta}^f)) \approx \\ \sum_{k_1'} \sum_{k_2'} \dots \sum_{k_n'} Rot_{k_1 k_1'}^i(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r(\boldsymbol{\theta}^f)) Rot_{k_2 k_2'}^i(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r(\boldsymbol{\theta}^f)) \times \\ \dots Rot_{k_n k_n'}^i(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r(\boldsymbol{\theta}^f)) \Omega_{k_1' k_2' \dots k_n'}^i(\boldsymbol{\theta}_{ref}^f, \boldsymbol{\theta}_{ref}^r(\boldsymbol{\theta}_{ref}^f)) \end{aligned} \quad \text{Eq. 3.12}$$

where the multipole moment, Ω , of rank n for atom i is calculated using the rotation matrix $Rot(\theta^f, \theta^r(\theta^f))$ that transforms the local axis system of each atom to the molecular axis system and $\theta^r(\theta^f)$ denotes the LAM-based estimate of the rigid degrees of freedom θ^r using equation 3.7. After each conformational rotation, the multipoles expressed in the molecular axis system are used to compute the intermolecular electrostatic contribution to the lattice energy.

The conformational transferability of multipole moments, and consequently the accuracy of equation 3.12, varies from one molecule to another. For certain functional groups, it cannot be assumed that the localised multipole moments remain constant even for small conformational changes (Polito et al. 2008). For instance, the pyramidalisation of the $-NH_2$ group has a direct influence on the position of the electron lone pair on the nitrogen atom that cannot be captured by the analytical rotation of the atomic multipole moments. In such cases, improved accuracy may be obtained by applying a linear correction to the result of equation 3.12 based on a first-order Taylor expansion:

$$\Omega(\theta^f) = \Omega(\theta^f, \theta^r(\theta^f)) + \left[\frac{\partial \Omega}{\partial \theta^f} \right]_{\theta_{ref}^f} (\theta^f - \theta_{ref}^f) \quad \text{Eq. 3.13}$$

In principle, the partial derivatives on the right hand side of the above equation could be computed from the results of the QM calculations. In practice, this is not straightforward with the currently available QM codes, and consequently the derivatives in the proposed approach are approximated using finite differences. This requires at least one (in the case of one-sided 1st order finite differences) additional QM charge density calculation for each flexible degree of freedom being perturbed.

It is important to note that the correction of equation 3.13 may not be necessary for all flexible degrees of freedom under consideration, as in many cases the computationally cheaper LAM of equation 3.12 will already lead to the required accuracy. More research is needed to establish the functional groups for which the application of 3.13 is necessary. Our preliminary analysis indicates that it is advisable to apply the correction in the case of torsional angles involving nitrogen atoms (e.g. $-NH_2$ group) or $-OH$ groups.

3.5. Range of LAM validity

The range of validity of the proposed LAMs for the estimation of the intramolecular energy and the rigid degrees of freedom (equations 3.9 and 3.7) has been tested against *ab initio* calculations using the GAUSSIAN suite of programs (Frisch et al. 2009). The tests were carried out for piracetam (2-oxo-pyrrolidine-acetamide, Figure 3.3), a molecule comprising 20 atoms (54 intramolecular degrees of freedom). For the purpose of illustration, the two most important torsional angles, N2-C6-C5-N1 and C6-C5-N1-C1, shown as blue arrows in Figure 3.3, have been considered as the only flexible degrees of freedom.

The *ab initio* intramolecular energy surface, ΔE^{intra} , (relative to the energy at the global conformation minimum) is shown in Figure 3.4 as a function of the two flexible degrees of freedom. The results shown were computed using a total of 64 points on an 8×8 grid. At each point, the two flexible degrees of freedom were fixed at the corresponding grid values, and the remaining rigid degrees of freedom were determined via a quantum mechanical, constrained, isolated-molecule geometry optimisation (equation 3.2).

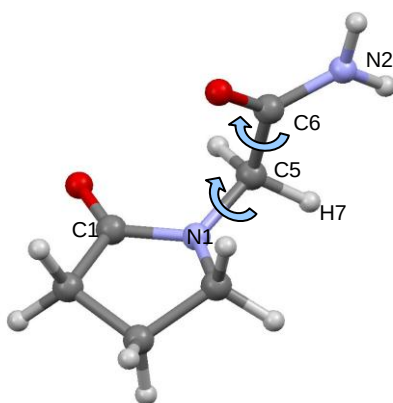


Figure 3.3 Molecular diagram and atom labelling for piracetam (2-oxo-pyrrolidine-acetamide).

As seen in Figure 3.4b, the intramolecular energy LAM approximates the QM surface with a maximum error of 0.15 kJ mol⁻¹ over a range of 5° around the reference point at 90.8° and 155.6° for C6-C5-N1-C1 and N2-C6-C5-N1 torsions respectively. This error is less than 3% of the 5 kJ mol⁻¹ intramolecular energy variation over the conformational region considered. The maximum error is reduced to 0.07 kJ mol⁻¹ within 3° of the reference point and 0.02 kJ mol⁻¹ within 2° of the reference point.

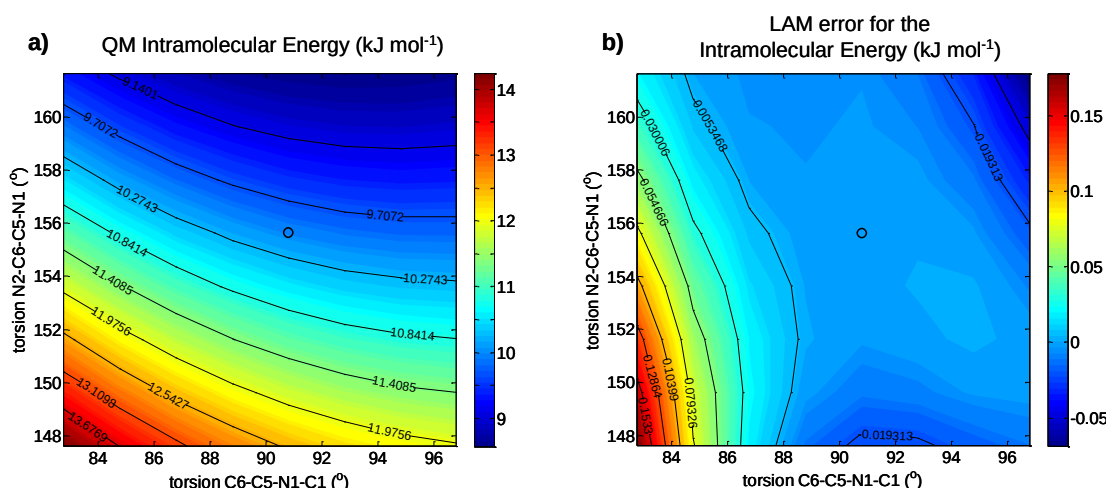


Figure 3.4 (a) QM intramolecular energy, $\Delta E^{\text{intra, QM}}$ at the HF/6-31G(d,p) level of theory and (b) LAM error for the intramolecular energy defined as $\Delta E^{\text{intra, QM}} - \Delta E^{\text{intra, LAM}}$ as a function of two “flexible” dihedral angles (blue arrows in Figure 3.3) for piracetam. Open circles correspond to the reference point for the LAM.

Similarly, Figure 3.5 shows that the LAM provides an excellent approximation for the dependence of the rigid degrees of freedom on the flexible torsions. The maximum errors for the rigid degrees of freedom are less than 0.10° for the torsional angle H7-C5-N1-C1 even though this torsion changes by up to 14.5° as the values of the C6-C5-N1-C1 and N2-C6-C5-N1 torsions are modified. Correspondingly, the maximum error is less than 0.06° for bond angle C5-N1-C1 and less than $0.0006 \text{ \AA}$ for bond length C1-N1. As expected, the absolute errors are seen to decrease for rigid degrees of freedom that are less sensitive to changes in the flexible degrees of freedom: bond angle C5-N1-C1 varies by 1.41° over the region considered and bond length C1-N1 by $0.0094 \text{ \AA}$.

Figure 3.6 is concerned with the accuracy of the multipole rotation LAM (equation 3.12) in modelling the electrostatic potential energy of xylitol (1,2,3,4,5-pentapentanol) as a selected torsional angle (H1-O1-C1-C2) deviates from the LAM’s reference point. The latter corresponds to the molecular conformation at the minimised experimental crystal structure (Kim and Jeffrey 1969) using 9 major torsional angles (all angles involving hydroxyl groups, and 4 selected backbone torsions, as shown in Table 4.2).

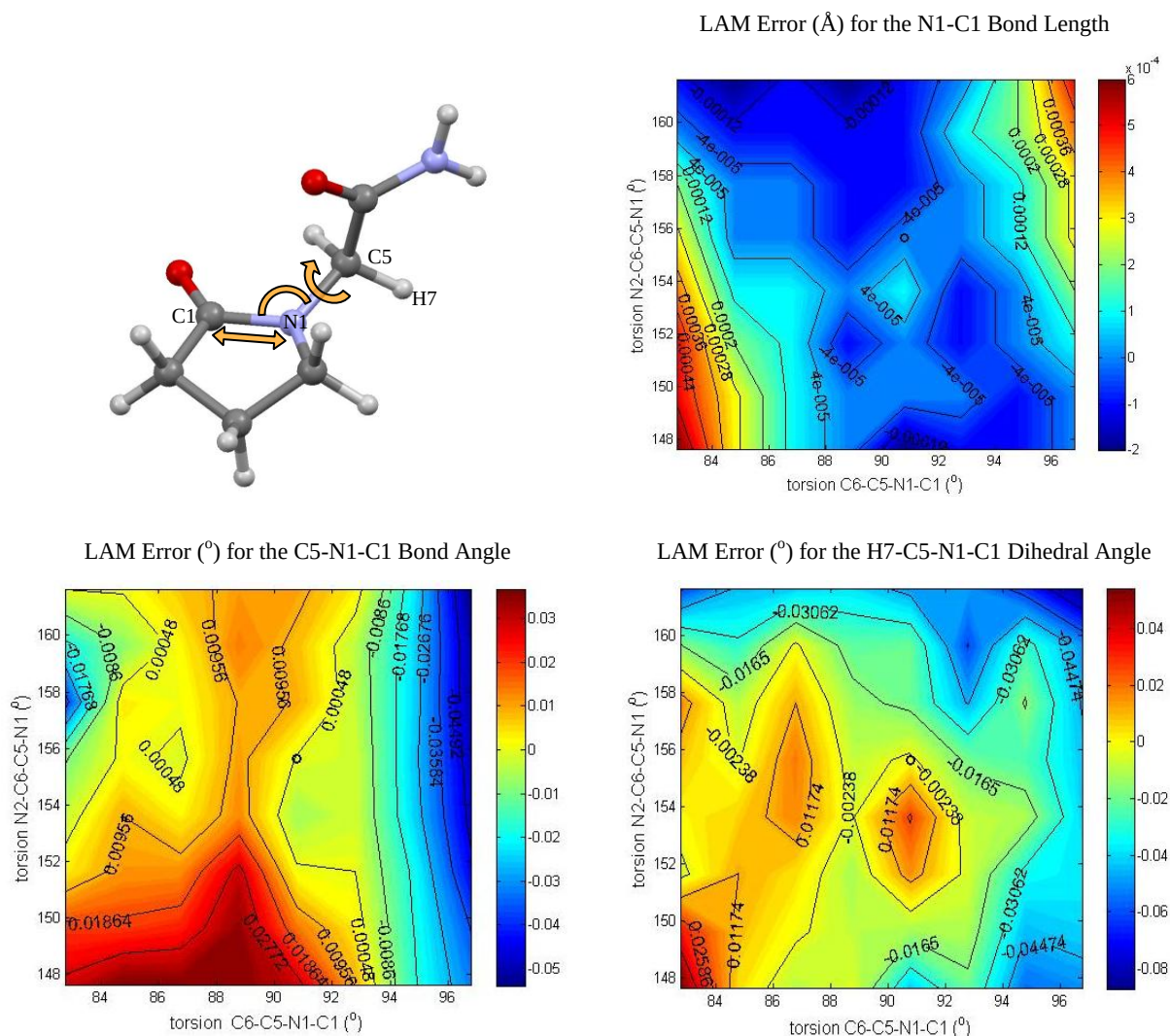


Figure 3.5 Difference between the quantum mechanical and estimated LAM values for selected “rigid” degrees of freedom (orange arrows top left) as a function of two “flexible” dihedral angles (blue arrows in Figure 3.3) for piracetam (2-oxo-pyrrolidine-acetamide). QM calculations performed at the HF/6-31G(d,p) level of theory. Open circles correspond to the reference point for the LAM.

The quantum mechanical electrostatic potential energy of the reference conformation is shown at the centre of Figure 3.6. For the perturbed conformations, the selected hydroxyl torsion (indicated by a blue arrow on the molecule in Figure 3.6) was varied by up to $\pm 10^\circ$ from its reference value while the other 8 flexible degrees of freedom were held constant at their reference values. For each perturbed molecular geometry, the rigid degrees of freedom were computed using a LAM (equation 3.7) constructed at the PBE/6-31G(d,p) level of theory. The atomic multipole moments and the electrostatic potential were consequently evaluated using both the LAM (equation 3.12) and explicit quantum mechanical isolated-molecule calculations at the PBE/6-31G(d,p) level of theory. The maximum error in the electrostatic potential energy increases the

further the conformation moves away from the reference molecular geometry, but does not exceed 0.03 eV over the entire range $\pm 10^\circ$ of the flexible torsion angle considered. This error is approximately 2% of the electrostatic potential energy range. The error is reduced to less than 0.02 eV when the LAM is used to model the electrostatic potential energy within $\pm 5^\circ$ of the reference geometry. Of course, what is important for the purposes of crystal structure prediction is not the electrostatic potential energy per se but the intermolecular electrostatic contributions to the lattice energy. In section 4.4.3, the accuracy of the multipole LAMs given by equations 3.12 and 3.13 will be considered in more details as applied to lattice energy minimisation.

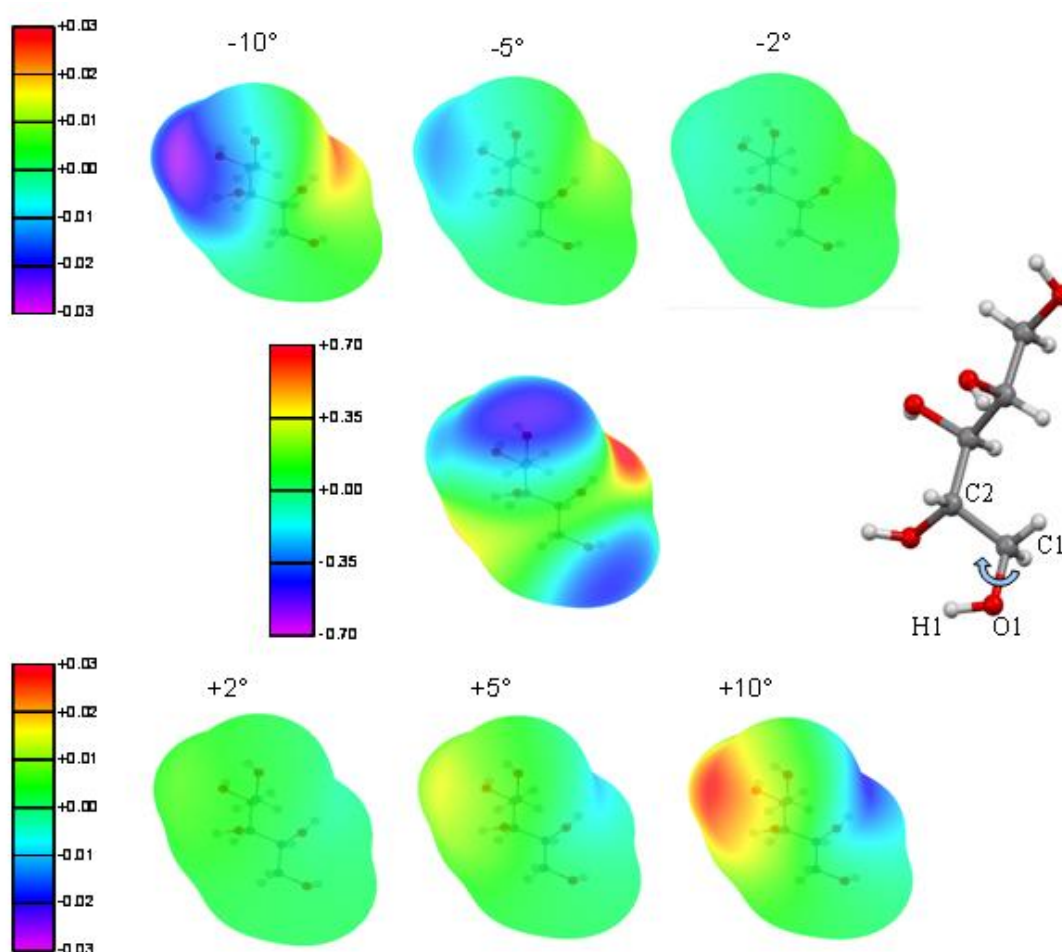


Figure 3.6 Error in the electrostatic potential energy (in eV; $1\text{eV} = 96.5\text{ kJ mol}^{-1}$) on the surface of xylitol (1,2,3,4,5-pentapentanol) at twice the van der Waals radii as a function of the H1-O1-C1-C2 torsional angle. LAM used to estimate the rigid degrees of freedom and rotate the atomic multipole moments with their local environment. The quantum mechanical electrostatic potential energy in eV for the reference molecular conformation is also shown for comparison (centre). All electrostatic potential energy surfaces were computed with atomic multipoles up to the hexadecapole level, at the PBE/6-31G(d,p) level of theory, using ORIENT (Stone et al. 2006). The van der Waals radius for hydroxyl hydrogen atoms was set to 1 au, the radii for other atoms were taken from Bondi (1964).

In conclusion, the error inherent in any LAM increases as one moves further away from the reference point around which the LAM was constructed. It is therefore necessary to reconstruct the LAM after a significant change in the flexible degrees of freedom. What constitutes a significant change is molecule dependent, but in our experience $\pm 5^\circ$ for torsional angles, $\pm 5^\circ$ for bond angles and 0.1 Å for bond lengths gives reliable results in the majority of systems. These ranges are comparable to the conformational changes observed during the local lattice energy minimisation of flexible molecules. Hence, it should be possible to perform such calculations with only a few explicit quantum mechanical optimisations, which provides the motivation for the algorithm for the local minimisation of the lattice energy presented in the next chapter.

3.6. Re-usability of QM calculations (LAM/QM databases)

For a given molecular species, a local approximate model as described above is derived completely from the results of an isolated-molecule QM calculation performed at a certain reference point θ_{ref} with a certain level of QM theory. These QM results are independent of the crystalline environment and the physical conditions (e.g. pressure) under which lattice energy minimisations is taking place. Consequently, they can be re-used in repeated calculations involving the same molecule. This is particularly useful in the case of *ab initio* crystal structure prediction studies where it is necessary to perform lattice energy minimisations starting from a large set (possibly hundreds or even thousands) of candidate crystal structures for a given molecule.

To take advantage of this re-usability, all QM-computed quantities (reference conformation, the intramolecular energy, first- and second- intramolecular energy derivatives with respect to all intramolecular degrees of freedom, the localised atomic multipole moments and their gradients with respect to the flexible degrees of freedom) used to construct a LAM at any point during a calculation are stored in a database. Torsional angles are allowed to take any value in the course of lattice energy minimisation, but are stored only in the $[-180^\circ, +180^\circ]$ range in the database. For molecules exhibiting chirality, an entry for the enantiomer of the molecule is also automatically generated from the same *ab initio* calculation, by inverting the values of

the torsional angles, intramolecular energy derivatives with respect to the torsional angles and the relevant components of the multipole moments.

Whenever it is necessary to create LAMs at a new point θ^f in the molecule's conformational space, the corresponding LAM/QM database is searched to identify whether any existing entry θ_{ref} in it can be used to create a LAM that would be valid at θ^f . This would be the case if the differences between the values of the elements of θ^f and the corresponding values in θ_{ref}^f were all within a given tolerance ε . If more than one database entry meets these validity criteria, the entry with the lowest root-mean-square deviation from θ^f is chosen. On the other hand, if no database entry satisfies the validity criteria, new QM calculations are performed to construct the required LAMs by solving the constrained minimisation problem defined by equation 3.1 and performing a charge density calculation. The QM results are then used to create a new entry in the database with a reference point $\theta_{ref} = (\theta_{ref}^f, \theta_{ref}^r)$, where θ^r is given by the solution of the constrained geometry optimisation in equation 3.2.

A LAM/QM database is specific to a particular molecule and level of QM theory. However, it can be re-used and, indeed, extended during more than one calculation involving this particular molecule. For example, a database created during the lattice energy minimisation of a given experimentally observed polymorph may be used on a later occasion for the minimisation of the same experimental structure under a different pressure, or of a different polymorph, or of a cocrystal that involves the given molecule together with a different one. For this reason, LAM/QM databases are stored as persistent computer files that can be used in any calculation relating to the corresponding molecule, potentially being extended during each such calculation to contain an increasing number of points, thereby becoming more and more useful as an increasing fraction of the molecule's conformational space is covered. An illustrative example of the performance gain that can be achieved due to the use of databases in an *ab initio* crystal structure prediction study is discussed in section 4.4.6.

Finally, it is worth pointing out that the information being stored in a LAM/QM database entry comprises the results of a QM calculation (i.e. the values of the intramolecular energy and its partial derivatives appearing in the right hand side of equation 3.3) and not the LAMs of equations 3.7 and 3.9 (e.g. the matrices $A(\theta_{ref})$ and $C(\theta_{ref})$, and the vector $b(\theta_{ref})$) derived from them. Thus, the LAM/QM database entries are independent of the specific way in which the conformational degrees of freedom θ^r

are partitioned between “flexible” θ^f and “rigid” θ^r variables. Consequently, this information may be used for different calculations pertaining to the same molecule under different degrees of flexibility, provided the first-order optimality condition in equation 3.4 still holds. In practice, this means that an existing database can be re-used by subsequent calculations considering the same degree or *higher* degree of flexibility. It is worth noting that the requirement for non-decreasing flexibility does not apply to the atomic multipole moments; these can be reused across different calculations involving any degree of flexibility as long as the definition of the local axis system for each atom in the molecule remains the same.

Chapter 4. Local minimisation of lattice energy – the CrystalOptimizer algorithm

The CrystalOptimizer algorithm (Kazantsev et al. 2010, Kazantsev et al. 2011a) described in this section is a local lattice energy minimisation scheme for crystal structures containing flexible molecules. It is designed to reduce the computational cost associated with quantum mechanical calculations without compromising accuracy. It is applicable to molecules of the size, complexity and flexibility typically encountered in pharmaceutical development. The main novelty of the approach is the use of local approximate models (LAMs) to represent the intramolecular energy and the conformationally-dependent charge density. These were already presented in Chapter 3. The formulation of the lattice energy minimisation problem and a detailed description of the CrystalOptimizer algorithm are provided in sections 4.1 and 4.3 respectively. The computational performance and the accuracy of the algorithm are critically assessed by its ability to reproduce the lattice geometry and conformational degrees of freedom for a set of experimentally determined crystal structures in sections 4.4.4 and 4.4.5. Finally, the applicability of CrystalOptimizer to the refinement of many hypothetical crystal structures in crystal structure prediction is discussed in section 4.4.6.

4.1. Formulation of the lattice energy minimisation problem

The lattice energy minimisation problem can be written as (Karamertzanis and Price 2006):

$$\min_{\mathbf{X}, \boldsymbol{\theta}} E^{\text{latt}} \equiv \min_{\mathbf{X}, \boldsymbol{\theta}} [\Delta E^{\text{intra}}(\boldsymbol{\theta}) + U^{\text{inter}}(\mathbf{X}; \boldsymbol{\theta}, \Omega(\boldsymbol{\theta}))] \quad \text{Eq. 4.1}$$

where ΔE^{intra} is the energy required to deform the molecule from its most stable gas-phase conformation and U^{inter} is the intermolecular energy describing the interactions between neighbouring molecules in the crystal. Stable crystal forms can therefore be identified by minimising the lattice energy with respect to the intramolecular degrees of freedom, $\boldsymbol{\theta}$, (bond lengths, bond angles and torsional angles) and lattice variables, $\mathbf{X}$,

which include the unit cell geometry and the position and orientation of all crystallographically independent molecules in the lattice. Here $\Omega(\boldsymbol{\theta})$ denotes the distributed multipole model (Stone 2005) used to represent the dominant electrostatic contributions to the intermolecular energy.

As has been explained in section 3.1, the dimensionality of the above optimisation problem can be reduced by the division of the intramolecular degrees of freedom, $\boldsymbol{\theta}$, into flexible, $\boldsymbol{\theta}^f$, and rigid, $\boldsymbol{\theta}^r$ degrees of freedom, leading to the modified minimisation problem:

$$\min_{\mathbf{X}, \boldsymbol{\theta}^f} E^{\text{latt}} \equiv \min_{\mathbf{X}, \boldsymbol{\theta}^f} [\Delta E^{\text{intra}}(\boldsymbol{\theta}^f) + U^{\text{inter}}(\mathbf{X}; \boldsymbol{\theta}^f, \boldsymbol{\theta}^r(\boldsymbol{\theta}^f), \Omega(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r(\boldsymbol{\theta}^f)))] \quad \text{Eq. 4.2}$$

where $\Delta E^{\text{intra}}(\boldsymbol{\theta}^f)$ and $\boldsymbol{\theta}^r(\boldsymbol{\theta}^f)$ are the solutions of the isolated-molecule constrained geometry optimisation defined by equations 3.1 and 3.2 respectively, and can be approximated accurately and efficiently via the LAMs shown in equations 3.9 and 3.7 respectively (refer to section 3.3 for the derivations). The multipoles $\Omega(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r(\boldsymbol{\theta}^f))$ can also be approximated via the LAMs defined by equations 3.12 and 3.13 defined in section 3.4.

For a given molecular conformation $\boldsymbol{\theta}$, the minimum intermolecular energy with respect to the lattice variables, $\mathbf{X}$, can be calculated with existing algorithms such as DMACRYS (Willock et al. 1995, Price et al. 2010) which also incorporates an anisotropic electrostatic model based on the distributed multipoles. Therefore, in order to make use of such codes, equation 4.2 is further reformulated as:

$$\min_{\boldsymbol{\theta}^f} E^{\text{latt}} = \min_{\boldsymbol{\theta}^f} [\Delta E^{\text{intra}}(\boldsymbol{\theta}^f) + \bar{U}^{\text{inter}}(\boldsymbol{\theta}^f)] \quad \text{Eq. 4.3}$$

where $\bar{U}^{\text{inter}}(\boldsymbol{\theta}^f)$ is given by the solution of another local minimisation problem in DMACRYS:

$$\bar{U}^{\text{inter}}(\boldsymbol{\theta}^f) \equiv \min_{\mathbf{X}} U^{\text{inter}}(\mathbf{X}; \boldsymbol{\theta}^f, \boldsymbol{\theta}^r(\boldsymbol{\theta}^f), \Omega(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r(\boldsymbol{\theta}^f))) \quad \text{Eq. 4.4}$$

Equations 4.3 and 4.4 define a bi-level optimisation problem. The inner minimisation (equation 4.4) determines the optimal crystal structure for *rigid* molecular entities, whose conformation is determined by the outer minimisation (equation 4.3) manipulating the flexible degrees of freedom θ^f . It should be noted that the DMAflex algorithm (Karamertzanis and Price 2006), presented in section 2.4.2, solves the same bi-level optimisation problem, also using DMACRYS for the solution of the inner minimisation problem (equation 4.4).

4.2. DMACRYS – inner minimisation algorithm

DMACRYS (Willock et al. 1995, Price et al. 2010) is a local lattice energy minimisation algorithm for rigid molecules used to solve the inner optimisation problem given by equation 4.4. It was developed with the aim to extend the distributed multipole approach to the organic solid state. The minimisation is performed using a modified Newton-Raphson procedure with or without space group constraints. In order to limit the number of independent variables, the molecules are modelled as rigid units rather than treating each atom independently. The electrostatic interactions are calculated using the distributed multipole model (up to hexadecapole level), which accurately represents the non-spherical features of the charge density essential for reproducing highly directional interactions such as hydrogen bonds. The slow converging charge-charge, charge-dipole and dipole-dipole interactions are calculated using the standard Ewald summation technique (Ewald 1921, Toukmaji and Board 1996). For higher multipole interactions a molecule-based cut-off is used for direct space summation. The repulsion-dispersion interactions can be modelled with a range of isotropic or anisotropic, empirical or *ab initio* atom-atom potentials with direct summation up to a predefined atom-based cut-off distance. Charge polarisation effects can also be included.

4.3. CrystalOptimizer – outer minimisation algorithm

CrystalOptimizer is a local lattice energy minimisation algorithm designed to solve the outer optimisation problem given by equation 4.3. It makes use of the LAMs

introduced in Chapter 3 in order to reduce the computational cost associated with quantum mechanical calculations. A flowchart of the CrystalOptimizer algorithm is shown in Figure 4.1 and the pseudo code is given in Figure 4.3.

In the initialisation step, the algorithm requires the user to specify the starting crystal structure. By representing the molecular geometry in the automatically generated Z-matrix form (that avoids near-linear bond angles), the proposed algorithm allows the user to select the extent of molecular flexibility to be considered during the minimisation. A more detailed description of the Z-matrix procedure is provided in Appendix A. This can range from a few selected torsional angles to a full atomistic minimisation.

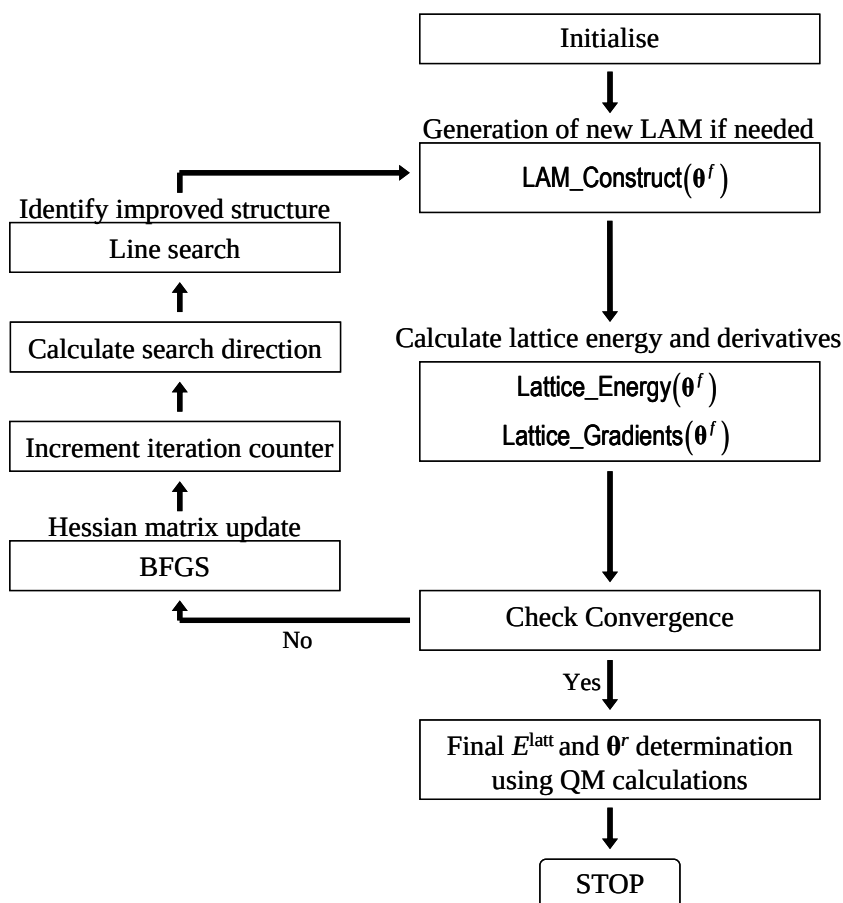


Figure 4.1 CrystalOptimizer local lattice energy minimisation algorithm flowchart. Refer to Figure 4.2 for the procedures for LAM construction (LAM_Construct), lattice energy (Lattice_Energy) and lattice energy derivative (Lattice_Gradients) calculations.

Before the optimisation is carried out, the user is also required to specify several model parameters such as optimisation convergence tolerances (Table 4.1), the choice of quantum mechanical methods and basis sets, and the tolerance vector ϵ that defines

the range of LAM validity for different types of flexible degrees of freedom (torsions, bond angles and bond lengths). If required, different tolerances may be specified for the LAMs relating, to intramolecular energy and multipoles. For each flexible degree of freedom, the user also specifies whether or not the linear update to the multiple moments (equation 3.13) is to be used. Input information relating to DMACRYS, such as the cut-off range for Ewald summation, the type of repulsion-dispersion potential and the values of parameters within it, must also be specified.

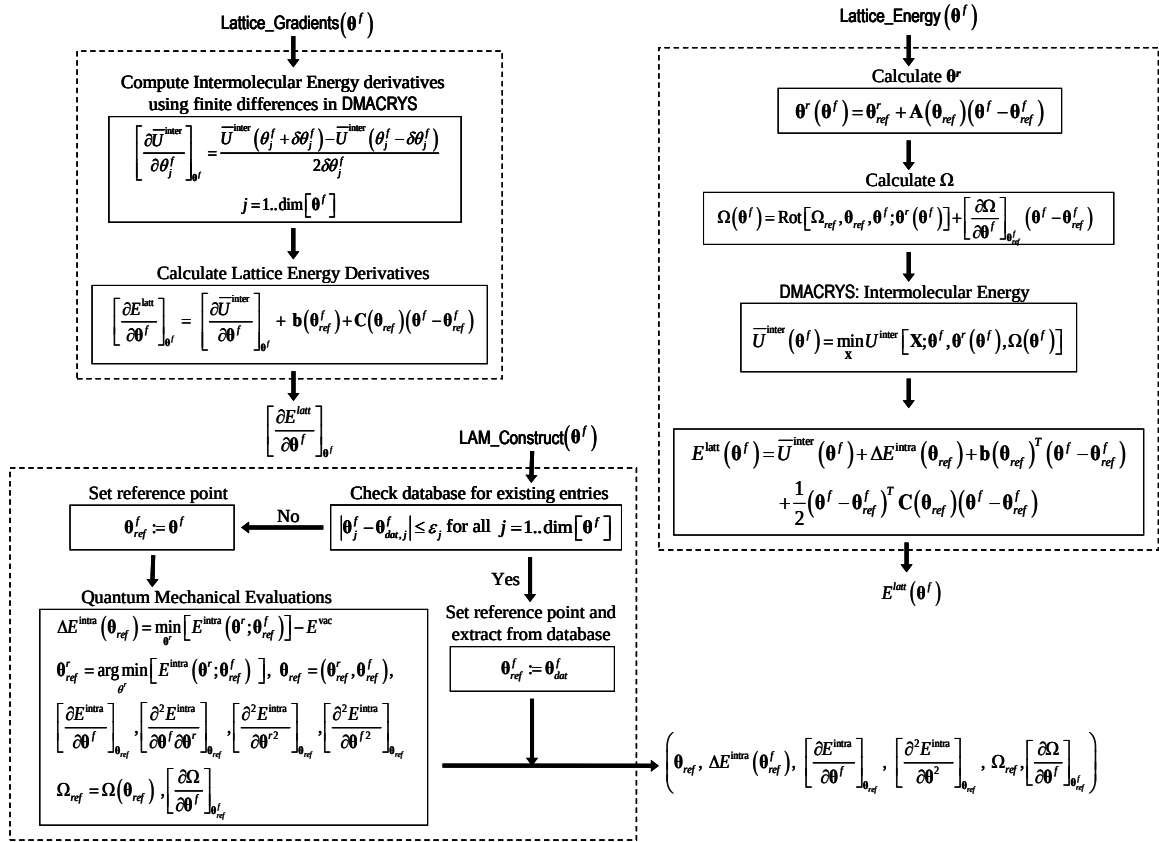


Figure 4.2 Procedures for lattice energy derivative (top left) and lattice energy (right) calculations and LAM construction (bottom left) in CrystalOptimizer. $dim[\theta^f]$ denotes the dimensionality of vector θ^f .

CrystalOptimizer uses a Quasi-Newton algorithm coupled with a line-search (Dennis and Schnabel 1983) to solve the outer minimisation problem (equation 4.3). This approach ensures rapid convergence even when there are many flexible degrees of freedom by using an approximation of the Hessian matrix of the second-order derivatives of the lattice energy with respect to the flexible degrees of freedom. At the start of the lattice energy minimisation, this Hessian approximation is normally

initialised to the unit matrix; at each subsequent iteration, it is updated via the Broyden-Fletcher-Goldfarb-Shanno method (BFGS) (Dennis and Schnabel 1983). This ensures that the Hessian approximation remains positive-definite, which guarantees the identification of a direction in which to change the flexible degrees of freedom θ^f , which results in a reduction of the lattice energy, thereby avoiding convergence to saddle points. The use of the BFGS approximation avoids the evaluation of the second-order derivatives of the lattice energy at every outer iteration, instead making use only of values of the lattice energy and its first-order gradients with respect to the flexible degrees of freedom.

As can be seen from the right hand side of equation 4.3, the computation of the lattice energy for given θ^f involves two components. The first one, $\Delta E^{\text{intra}}(\theta^f)$, is computed explicitly via the LAM of equation 3.9. The second component is computed by solving the inner minimisation problem (equation 4.4) using DMACRYS (Willock et al. 1995, Price et al. 2010), with the molecular conformation and distributed multipole moments fixed at the values determined via the LAMs of equations 3.7, 3.12 and 3.13.

Table 4.1 List of optimisation parameters in CrystalOptimizer.

	Parameter	Definition	Default Value
user-specified	k_{max}	max. number of outer iterations	0 (single-point lattice energy evaluation)
	ϵ	region of LAM validity	5.0° for torsions
			3.5° for bond angles
			0.05 Å for bond lengths
	ϵ_{grad}	optimality tolerance on gradients	1.14 kJ mol ⁻¹ rad ⁻¹ (0.02 kJ mol ⁻¹ deg ⁻¹) for torsional and bond angles
			1.14 kJ mol ⁻¹ Å ⁻¹ for bond lengths
algorithmic	i_{max}	max. # of iterations in line search	∞
	α_i	step length in line search at iteration i	updated dynamically
	ϵ_{ls}	minimum step length in line search	dynamic, close to machine precision
	ϵ_{dis}	optimality tolerance on displacements	4 x machine precision

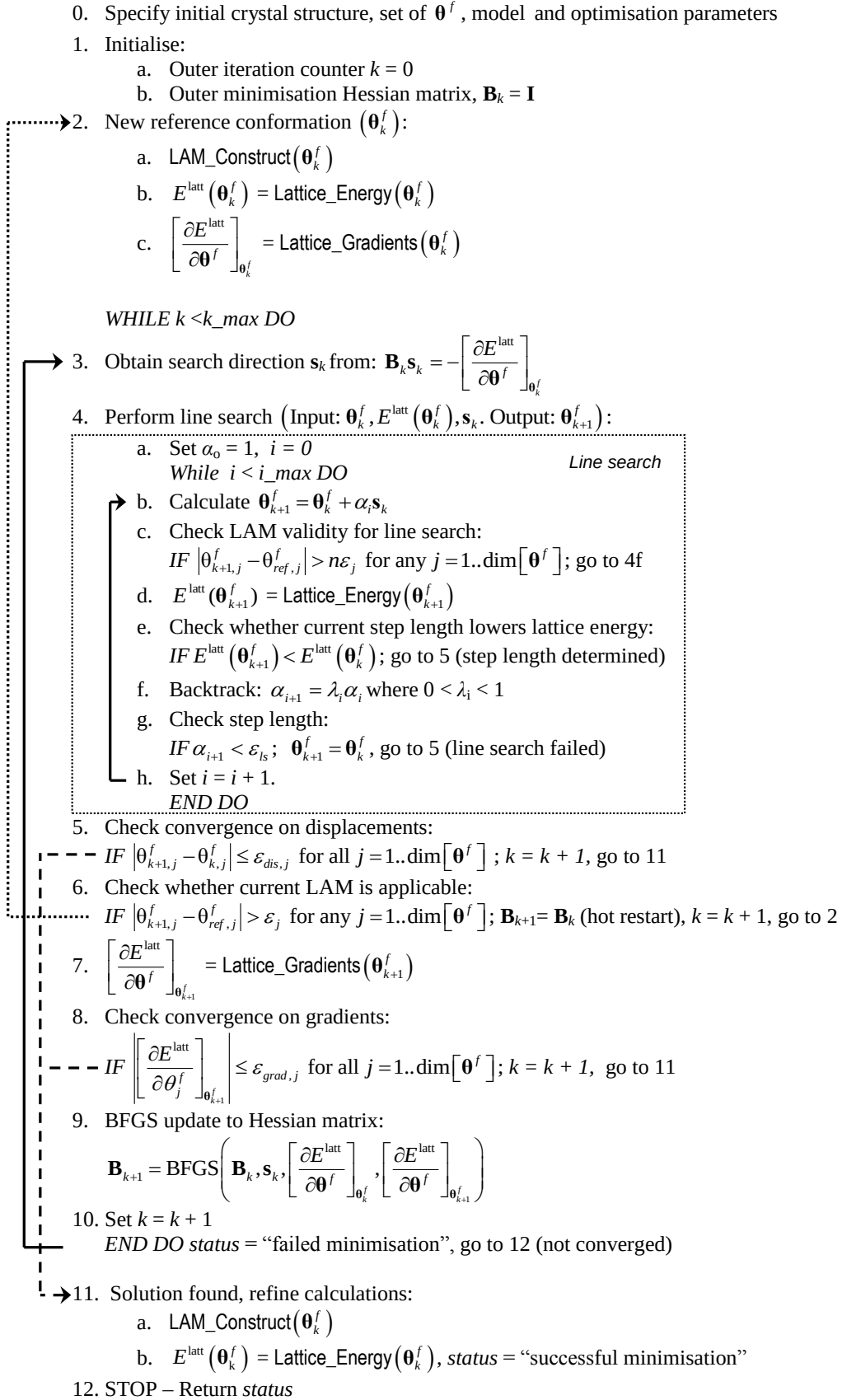


Figure 4.3 Pseudo code for the CrystalOptimizer local lattice energy minimisation algorithm. Refer to Table 4.1 for the list of optimisation parameters and Figure 4.2 for the procedures for LAM construction (LAM_Construct), lattice energy (Lattice_Energy) and lattice energy derivatives (Lattice_Gradients) calculation. $\text{dim}[\theta^f]$ denotes the dimensionality of vector θ^f .

In addition to the value of the lattice energy for given θ^f , the outer optimisation algorithm also requires values of its gradients with respect to θ^f . The gradients of its first component, $\Delta E^{\text{intra}}(\theta^f)$, can be computed in a straightforward manner by differentiating the LAM of equation 3.9:

$$\left. \frac{\partial \Delta E^{\text{intra}}}{\partial \theta^f} \right|_{\theta^f} = \mathbf{b}(\theta_{\text{ref}}) + \mathbf{C}(\theta_{\text{ref}})(\theta^f - \theta_{\text{ref}}^f) \quad \text{Eq. 4.5}$$

where the vector $\mathbf{b}(\theta_{\text{ref}})$ and matrix $\mathbf{C}(\theta_{\text{ref}})$ were defined earlier by equations 3.10 and 3.11 respectively.

The gradient of the second component, $\bar{U}^{\text{inter}}$, of the lattice energy with respect to θ^f cannot be obtained in a closed analytical form. In CrystalOptimizer, they are computed via a centred finite difference scheme. The gradient with respect to the k^{th} element θ_k^f of θ^f requires a positive and a negative perturbation of magnitude $\delta\theta_k^f$. For each perturbed value, the LAMs of equations 3.7, 3.12 and, where necessary, 3.13 are used to update the values of the rigid degrees of freedom θ^f and the multipoles Ω , before calling DMACRYS to determine $\bar{U}^{\text{inter}}$ via the inner minimisation described by equation 4.4. The required gradient is then obtained via the finite difference approximation:

$$\left. \frac{\partial \bar{U}^{\text{inter}}}{\partial \theta_k^f} \right|_{\theta^f} \approx \frac{\bar{U}_{k+}^{\text{inter}} - \bar{U}_{k-}^{\text{inter}}}{2\delta\theta_k^f} \quad \text{Eq. 4.6}$$

where $\bar{U}_{k+}^{\text{inter}}$ and $\bar{U}_{k-}^{\text{inter}}$ denote the values of $\bar{U}^{\text{inter}}$ returned by DMACRYS for the positive and negative perturbations respectively¹.

The required gradients $\partial E^{\text{latt}} / \partial \theta^f$ are then simply computed as the sum of expressions 4.5 and 4.6. The procedures for calculating the lattice energy and the

¹ The method of the finite differences (central, forward or backward) and magnitudes of the perturbations can be user defined. The default setting is to use central finite differences with 0.50°, 0.35° and 0.01 Å perturbations for torsions, bond angles and bond lengths respectively.

gradient of the lattice energy with respect to the flexible degrees of freedom are shown in Figure 4.2.

Based on the value of the lattice energy and the search direction calculated using the lattice energy gradient and the BFGS Hessian matrix approximation, the line-search procedure in the outer minimisation algorithm determines a new set of flexible degrees of freedom that results in a sufficiently large reduction of the lattice energy. This involves the evaluation of lattice energy at a sequence of points along the search direction. Points are evaluated for progressively reduced step sizes until a sufficient reduction in the lattice energy is found or the minimum step size is reached. Any point that does not lead to a decrease in the energy is immediately discarded from further consideration. As it is not generally worth performing any expensive QM calculations at such points, during the line search part of the algorithm we relax the LAM validity tolerances ε by a factor n which is typically in the range $1 < n \leq 2$.

After the line search identifies a new improved point in conformational space, θ^f , the LAM validity tolerances are reset to their original value ε , before evaluating the new lattice energy and its gradients with respect to θ^f in the manner detailed above. In some cases, this will result in the construction of new LAMs through QM calculations. The BFGS approximation is then used to provide a new estimate of the Hessian matrix and the algorithm proceeds to the next outer iteration by calculating a new search direction and performing a line search along it.

The optimisation terminates successfully when either the changes in all flexible degrees of freedom during the last step or all the lattice energy gradients with respect to θ^f are below specified tolerances. As a final step, CrystalOptimizer uses rigorous QM calculations to re-compute the lattice energy and molecular conformation at the values of θ^f determined by the optimisation, thereby eliminating any small errors that may have arisen from the use of LAMs. It should be noted however, that this final calculation does not eliminate any error caused by using LAMs in identifying the true minimum, as the gradients are not re-evaluated.

CrystalOptimizer returns a failure status if either the line search fails to identify a new set of flexible degrees of freedom that produce a sufficient reduction in the lattice energy along the search direction, or when a pre-defined maximum number of outer iterations is reached without either of the two convergence criteria having been satisfied.

4.4. Algorithm validation, Results and Discussion

The main question to be answered in validating the CrystalOptimizer algorithm is whether, for a given initial point, the use of the local approximate models leads to the same local energy minimum as when using explicit QM calculations at each minimisation iteration. Once this is established, the performance of the algorithm, and particularly the reduction in computational cost for a given number of flexible degrees of freedom can be investigated. The use of additional degrees of freedom can also be studied, and the impact of flexibility on crystal structure can be assessed.

4.4.1. Systems studied

The validation of the CrystalOptimizer algorithm was performed on the systems shown in Table 4.2. These were selected because of their scientific and/or practical importance and the fact that they possess sufficient conformational flexibility to present a significant challenge for lattice energy minimisation using current techniques.

Xylitol, a naturally occurring sugar alcohol, is a stereoisomer of 1,2,3,4,5-pentapentanol (pentose) for which only one crystal structure has been determined experimentally to date (Kim and Jeffrey 1969). No other organic systems containing xylitol were found in the Cambridge Crystallographic Database (CSD) (Allen 2002).

The second test system is glucose. Only the structure of one of its cyclical isomers, α -D-glucose (Brown and Levy 1979) is considered. A co-crystal of α -D-glucose with urea (Snyder and Rosenstein 1971) is also found in the CSD.

Piracetam (2-oxo-pyrrolidine-acetamide) is a pharmaceutical nootropic drug. There are five distinct polymorphs determined experimentally (Admiraal et al. 1982, Galdecki and Glowka 1983, Louer et al. 1995, Fabbiani et al. 2005, Fabbiani et al. 2007), two of which are observed only at high pressure. Two co-crystals of piracetam (Vishweshwar et al. 2005) are also reported in the CSD.

In order to examine the applicability of CrystalOptimizer to a range of systems of practical interest, three additional crystals were considered, namely the salt (R)-1-phenyl-2-(4-methylphenyl)ethylamonium-(S)-mandelate (Sakai et al. 2004) and the co-crystals of the pharmaceutically important steroid progesterone with resorcinol

(Dideberg et al. 1975) and pyrene (Friscic et al. 2010). These were selected primarily on the basis of posing significant computational challenges in terms of both the molecular size and flexibility.

Table 4.2 Systems considered for CrystalOptimizer refinement.

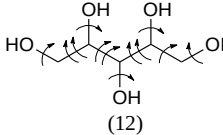
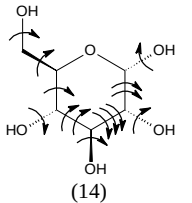
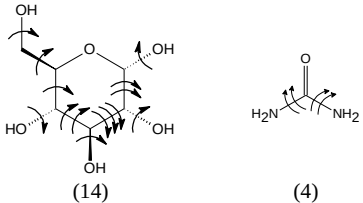
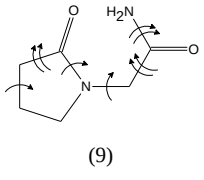
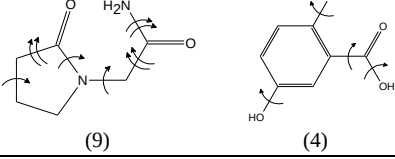
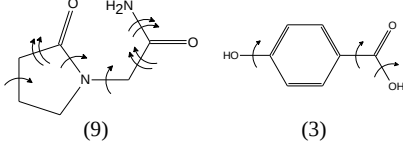
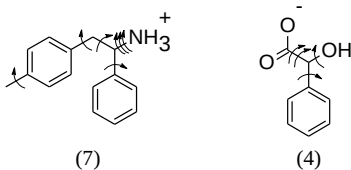
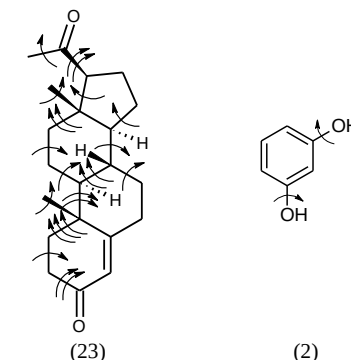
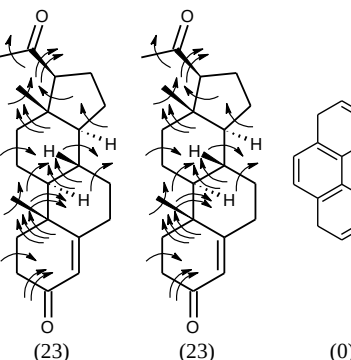
Name		Molecular Diagram (number of flexible degrees of freedom)	Space Group	z'
Xylitol		 (12)	$P2_12_12_1$	1
α -D-glucose		 (14)	$P2_12_12_1$	1
α -D-glucose Urea (1:1) co-crystal		 (14) (4)	$P2_12_12_1$	1
Piracetam	Form I	 (9)	$P2_1/n$	1
	Form II		$P\bar{1}$	1
	Form III		$P2_1/n$	1
	Form IV		$P2_1/c$	1
	Form V		$P\bar{1}$	1
Piracetam Gentisic acid (1:1) co-crystal		 (9) (4)	$C2/c$	1
Piracetam p-Hydroxybenzoic acid (1:1) co-crystal		 (9) (3)	$P2_1/n$	1

Table 4.2 continued.

Name	Molecular Diagram (number of flexible degrees of freedom)	Space Group	z'
(R)-1-phenyl-2-(4-methylphenyl)ethylammonium-(S)-mandelate salt	 (7) (4)	$P2_12_12_1$	1
Progesterone Resorcinol (1:1) co-crystal	 (23) (2)	$P2_12_12$	1
Progesterone Pyrene (2:1) co-crystal	 (23) (23) (0)	$P1$	1

4.4.2. Molecular modelling and computational considerations

The quantum mechanical molecular geometry optimisations and the charge density calculations were both evaluated at the PBE/6-31G(d,p) level of theory.

The LAM validity tolerances were set at 5° for torsional and bond angles and $0.02 \text{ \AA}$ for bond lengths. The linear update to the multipole moments (equation 3.13) was used for all flexible torsional angles that involved the nitrogen atom or the OH groups. Although the extent to which the LAMs are reliable is system-dependent, the above range of validity is expected to provide a sufficiently small error compared with the other approximations present in the computational model.

The repulsion-dispersion interactions were modelled with an empirical exp-6 potential (Coombes et al. 1996) and were summed in direct space up to a 30 Å cut-off.

CrystalOptimizer was used to perform local lattice energy minimisations starting from the experimental structures for each model system considered. The flexible degrees of freedom that were taken into consideration for each system are marked in the molecular diagrams shown in the second column of Table 4.2.

Most calculations were performed on a single Intel Xeon 5150 2.66 GHz processor using 1500 MB of memory. In the case of the salt and progesterone co-crystals, the minimisations were performed on 4 Intel Xeon 5150 2.66 GHz processors with 7 GB shared memory. The availability of multiple processors was exploited for performing the QM computations using the GAUSSIAN code (Frisch et al. 2009).

4.4.3. Use of linear updates for multipole LAMs

The range of validity of the LAMs for the distributed multipole model in terms of their ability to reproduce the electrostatic potential field surrounding an isolated molecule was examined in section 3.4. As noted there, what is more important is the accuracy with which these LAMs can approximate the intermolecular electrostatic contributions to the lattice energy.

This question is considered in more detail in this section by using xylitol, with the minimised experimental structure (using 9 flexible hydroxyl and backbone torsions) as the reference point. The H1-O1-C1-C2 hydroxyl angle is varied by up to $\pm 10^\circ$ from the reference point whilst maintaining the remaining flexible degrees of freedom at their reference values. DMACRYS is used to re-optimize the crystal lattice at each different conformation under consideration, and the intermolecular energy contribution to the lattice energy at the optimal structure is recorded.

Figure 4.4 shows results from three different sets of calculations performed for several values of the selected torsion angle over the range mentioned above.

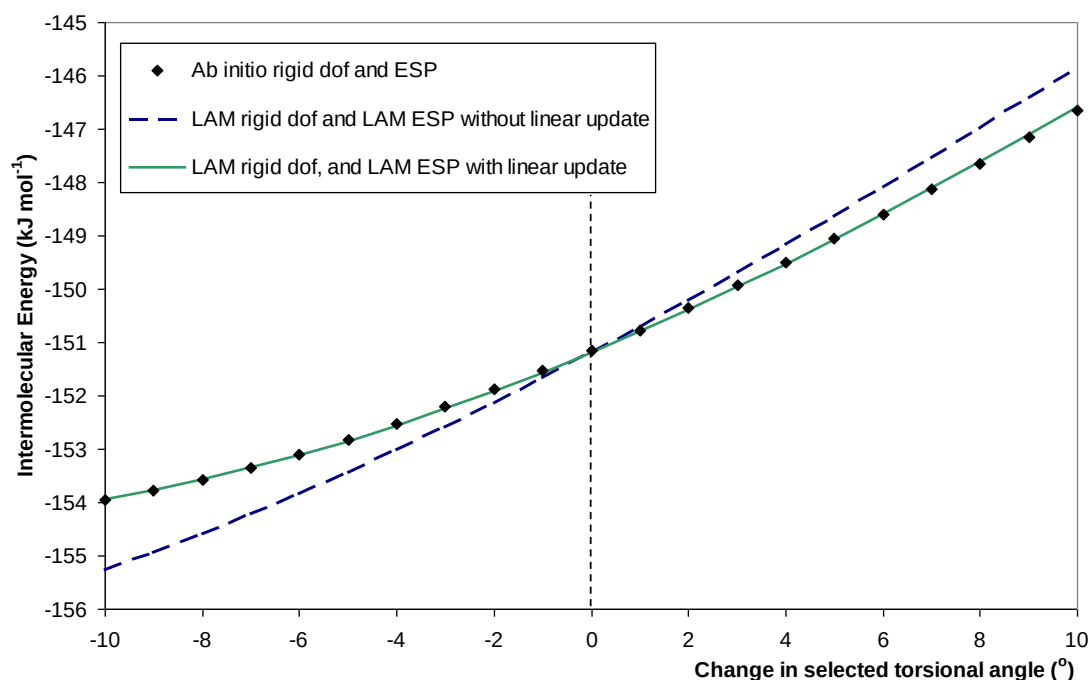


Figure 4.4 Intermolecular energy as a function of the H1-O1-C1-C2 torsional angle (blue arrow in Figure 3.6) for three different computational strategies: 1) molecular geometry and the electrostatic potential (ESP) obtained by full quantum mechanical calculations (black diamonds); 2) LAM used to update rigid degrees of freedom (dof) and multipole rotation (blue dashed line); 3) LAMs used to update the rigid degrees of freedom, multipole rotation and a linear update of the multipole moments (continuous green line).

First, for each value of the flexible hydroxyl angle, the rigid degrees of freedom were computed explicitly by solving the quantum mechanical constrained optimisation in equation 3.2. The electrostatic potential was also calculated quantum mechanically for each generated conformation. These molecular and electrostatic models were used to compute the benchmark intermolecular energy shown as black diamonds in Figure 4.4.

Next, the LAM of equation 3.7 was used to approximate the rigid degrees of freedom, and the multipoles were rotated according to equation 3.12. The results are shown as a blue dashed line in Figure 4.4. The maximum error in the intermolecular energy is 1.3 kJ mol^{-1} and occurs when the selected flexible torsion angle is decreased by 10° from its reference value. The absolute maximum errors for changes of up to $\pm 5^\circ$ and $\pm 2^\circ$ from the reference value are 0.6 kJ mol^{-1} and 0.3 kJ mol^{-1} respectively. In this scenario, the error arises due to the use of a LAM to obtain the rigid degrees of freedom (which affects both the repulsion-dispersion interactions and the molecular geometry for the calculation of the charge density) and the LAM used to rotate the multipole

model. A subsequent calculation¹ revealed that the use of the LAM-based rigid degrees of freedom does not significantly affect the electrostatic calculations. Hence, the error arises primarily due to the rotation of the multipole model.

Finally, the intermolecular energy is calculated using the LAMs for the estimation of the rigid degrees of freedom (equation 3.7), multipole rotation (equation 3.12) and a linear update to the multipole moments (equation 3.13). The results are shown as a continuous green line in Figure 4.4. The error in the intermolecular energy is within 0.06 kJ mol⁻¹ of the *ab initio* value (cf. case 1 above) for all values of the selected flexible degree of freedom.

The above results indicate that, for this particular system, the multipole rotation LAM of equation 3.12 leads to considerable errors in intermolecular energy. However, the subsequent application of the linear correction (equation 3.13) produces an accurate estimate of the electrostatic potential even for relatively large conformational changes.

4.4.4. Application to single component crystals

Firstly three model systems which involve single component crystals, namely xylitol, α -D-glucose and piracetam Form II are considered. In order to assess the computational performance and the effects of varying molecular flexibility on the performance of the CrystalOptimizer, these systems were studied using different optimisation settings and increasingly wider sets of flexible degrees of freedom.

Initially, only the hydroxyl (amide for piracetam Form II) and selected backbone dihedrals were treated as flexible degrees of freedom, while the rings in α -D-glucose and piracetam were assumed to be rigid. These simplifications result in 9, 6 and 4 flexible degrees of freedom for xylitol, α -D-glucose and piracetam respectively. Although these minimal sets of flexible torsion angles are not sufficient to capture the whole molecular flexibility, they have the advantage of being within the range of applicability of earlier algorithms for lattice energy. Of particular interest in this context is the DMAflex algorithm (Karamertzanis and Price 2006) which has a similar model of inter- and intra-molecular interactions as CrystalOptimizer, thereby allowing some validation of the results obtained with the proposed code, and a direct comparison of

¹ In this calculation, the LAM-based rigid degrees of freedom were used, but an explicit calculation QM calculation of the multipoles was performed for the resulting molecular conformation.

computational performance. Three different CrystalOptimizer runs were performed with the same number of degrees of freedom:

Case 1 CrystalOptimizer without LAMs or databases was tested. This requires quantum mechanical calculations at every iteration, as with DMAflex, and shows the effect of using a quasi-Newton algorithm instead of a simplex algorithm.

Case 2 All features of CrystalOptimizer are used. The databases are initially empty and are populated during these calculations. Therefore only the impact of the LAMs is assessed.

Case 3 All features of CrystalOptimizer are used, including the use of the LAM/QM database populated in Case 2.

Following this initial study, additional optimisations were carried out with CrystalOptimizer, by gradually increasing the degree of molecular flexibility under consideration in five further steps:

Case 4 All heavy-atom torsional angles (excluding H-C-X-X) treated as flexible, resulting in 12, 14 and 9 flexible degrees of freedom for xylitol, α -D-glucose and piracetam respectively.

Case 5 All hydroxyl (H-O-C) and amide (H-N-C) bond angles additionally treated as flexible, resulting in 17 flexible degrees of freedom for xylitol and 19 for α -D-glucose and piracetam.

Case 6 All heavy-atom bond angles (excluding H-C-X) additionally treated as flexible, resulting in 25 flexible degrees of freedom for xylitol, 29 for α -D-glucose and 21 for piracetam.

Case 7 The hydroxyl (O-H) and amide (N-H) bond lengths additionally treated as flexible, resulting in 30, 34 and 30 flexible degrees of freedom for xylitol, α -D-glucose and piracetam respectively.

Case 8 An atomistic representation corresponding to full molecular flexibility is considered, resulting in 60 flexible degrees of freedom for xylitol, 66 for α -D-glucose and 54 for piracetam.

Note that in the analysis of the results reported in all cases except Cases 2 and 3 and also in section 4.4.5, the LAM/QM database (cf. section 3.6) feature of CrystalOptimizer was disabled. An analysis of the impact of using these databases is presented in section 4.4.6.

The detailed results from the above studies are provided in Table 4.3, Table 4.4 and Table 4.5 for xylitol, α -D-glucose and piracetam Form II respectively. In all cases, agreement between the predicted crystal structures and the experimentally observed ones is assessed on the basis of the root-mean-square deviation of the molecular conformation and the 15-molecule coordination sphere (Chisholm and Motherwell 2005).

Overall, there is a gain in stability as the set of flexible degrees of freedom considered by CrystalOptimizer is widened, which indicates that the detailed modelling of molecular flexibility is important in order to capture the full extent of conformational distortions by the packing forces. For instance, the inclusion of the selected bond angles (Case 5) stabilises the structures of xylitol, α -D-glucose and piracetam by roughly 1 kJ mol^{-1} compared to the value obtained by considering the flexibility of only the main torsional angles (Cases 1-3). As the number of flexible degrees of freedom increases, small changes are observed in the lattice energy, but the balance between intermolecular and intramolecular energy tends to shift so that the molecule adopts a less stable conformation to reduce the energetic cost of intermolecular interactions. CrystalOptimizer makes it computationally feasible to consider a large number of degrees of freedom but it is up to the user to define the extent of flexibility required for each particular molecule or system.

More generally, explicit consideration of bond angles and stiff torsions, such as those in poly-aromatic ring systems, is expected to be especially significant for larger systems and under high pressure. On the other hand, it can be seen by comparing Cases 6-8 that the additional energy stabilisation arising from modelling the bond lengths is more limited (unless proton transfer takes place (Mohamed et al. 2009)). The energy required to perturb a bond length is usually very large, and consequently the modelled bond lengths do not change significantly during lattice energy minimisation. Thus, in order to reduce computational cost, the bond lengths can normally be assumed to be rigid with little loss of predictive accuracy

Table 4.3 Detailed lattice energy minimisation results for the experimental structure of Xylitol.

			DMAFlex ^a	CrystalOptimizer							
				No LAMs, no LAM databases	LAMs, no LAM databases	LAMs and LAM databases	More flexibility, LAMs but no LAM databases				
				Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8
No. of Flexible Degrees of Freedom ^b			9 main torsions	9 main torsions	9 main torsions	9 main torsions	12 all torsions ^c	17 all torsions ^c , OH angles	25 all torsions ^c , all angles ^d	30 all torsions ^c , all angles ^d , OH bonds	60 atomistic ^e
Lattice Energy (kJ mol ⁻¹)			-125.32	-125.14	-125.15	-125.16	-126.78	-127.04	-127.69	-127.79	-127.78
Intramolecular Energy (kJ mol ⁻¹)			23.31	24.10	23.55	23.43	23.59	23.18	24.53	24.84	24.47
Intermolecular Energy (kJ mol ⁻¹)			-148.62	-149.25	-147.70	-148.58	-150.37	-150.22	-152.22	-152.63	-152.25
rmsd ₁₅ ^f			0.20	0.19	0.19	0.19	0.19	0.19	0.23	0.23	0.23
Experimental											
Unit Cell	a (Å)	8.27	8.43	8.44	8.43	8.44	8.46	8.47	8.52	8.52	8.52
	b (Å)	8.90	8.95	8.95	8.95	8.95	8.92	8.91	8.84	8.84	8.85
	c (Å)	8.91	9.27	9.27	9.27	9.27	9.27	9.25	9.21	9.21	9.20
	α (°)	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0
	β (°)	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0
	γ (°)	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0
Density (g cm ⁻³)		1.542	1.445	1.444	1.445	1.443	1.448	1.446	1.457	1.457	1.457
CPU Time (hr:min) ^g			35:32	14:04	07:30	04:01	07:38	12:20	16:25	18:35	43:55
No. of outer iterations			127	10	13	13	13	14	15	15	23
No. of QM calculations	Molecular Optimisation	127	10	4	1	4	5	5	5	5	
	Molecular Hessian	-	10	3	0	3	4	4	4	4	
	Charge Density	127	60	19	1	19	25	25	25	25	
% CPU Time in	QM Molecular optimisation	53.83	4.59	4.79	1.07	5.00	3.33	2.32	2.02	0.50	
	QM Hessian calculation	-	23.63	12.95	0.00	13.35	11.04	8.30	7.24	2.94	
	QM Charge Density, DMA	41.01	49.30	29.02	2.82	29.05	23.55	17.73	15.61	6.50	
	DMACRYS	5.13	22.37	52.96	95.60	52.33	61.75	71.28	74.75	89.58	
	Other	0.03	0.11	0.28	0.51	0.27	0.33	0.37	0.38	0.48	

^a(Karamertzanis and Price 2006), ^b linear update to the multipole moments calculated for all torsions involving the OH group, ^c excluding dihedrals defined as H-C-X-X, ^d excluding H-C-X, ^e all intramolecular degrees of freedom, ^f root mean square deviation in 15-molecule coordination sphere compared with the experimental structure (excluding hydrogen atoms) (Chisholm and Motherwell 2005), ^g single Intel Xeon 5150 2.66 GHz processor using 1500 MB of memory.

Table 4.4 Detailed lattice energy minimisation results for the experimental structure of α -D-glucose.

			DMAFlex ^a	CrystalOptimizer							
				No LAMs, no LAM databases	LAMs, no LAM databases	LAMs and LAM databases	More flexibility, LAMs but no LAM databases				
				Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8
No. of Flexible Degrees of Freedom ^b			6 main torsions	6 main torsions	6 main torsions	6 main torsions	14 all torsions ^c	19 all torsions ^c , OH angles	29 all torsions ^c , all angles ^d	34 all torsions ^c , all angles ^d , OH bonds	66 atomistic ^e
Lattice Energy (kJ mol ⁻¹)			-146.08	-146.27	-146.41	-146.39	-147.44	-147.86	-148.59	-148.49	-148.62
Intramolecular Energy (kJ mol ⁻¹)			9.60	10.28	10.08	10.05	12.50	12.54	12.27	12.22	12.55
Intermolecular Energy (kJ mol ⁻¹)			-155.68	-156.55	-156.49	-156.44	-159.94	-160.40	-160.86	-160.71	-161.16
rmsd ₁₅ ^f			0.20	0.19	0.20	0.20	0.16	0.16	0.14	0.14	0.14
Experimental											
Unit Cell	a (Å)	10.37	10.60	10.65	10.66	10.66	10.59	10.59	10.53	10.53	10.52
	b (Å)	14.85	14.83	14.79	14.78	14.78	14.93	14.92	14.96	14.95	14.96
	c (Å)	4.98	5.06	4.99	4.97	4.97	4.93	4.93	4.95	4.96	4.95
	α (°)	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0
	β (°)	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0
	γ (°)	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0
Density (g cm ⁻³)		1.562	1.525	1.524	1.527	1.527	1.518	1.537	1.534	1.533	1.535
CPU Time (hr:min) ^g			44:13	22:46	09:20	02:17	10:27	13:22	19:04	21:26	41:32
No. of outer iterations			118	13	14	11	13	16	18	18	22
No. of QM calculations	Molecular Optimisation	118	13	5	1	4	4	4	4	4	4
	Molecular Hessian	-	13	4	0	3	3	3	3	3	3
	Charge Density	118	78	25	1	19	19	19	19	19	19
% CPU Time in	QM Molecular optimisation	55.47	5.14	8.33	2.71	5.88	4.73	2.44	2.16	0.59	
	QM Hessian calculation	-	30.89	21.69	0.00	14.94	11.20	8.23	7.33	3.57	
	QM Charge Density, DMA	41.14	52.80	41.53	6.66	28.20	21.91	15.52	13.80	7.02	
	DMACRYS	3.37	11.08	28.25	89.98	50.62	61.71	73.29	76.17	88.18	
	Other	0.02	0.09	0.20	0.65	0.36	0.45	0.52	0.54	0.64	

^a(Karamertzanis and Price 2006), ^b linear update to the multipole moments calculated for all torsions involving the OH group, ^c excluding dihedrals defined as H-C-X-X, ^d excluding H-C-X, ^e all intramolecular degrees of freedom, ^f root mean square deviation in 15-molecule coordination sphere compared with the experimental structure (excluding hydrogen atoms) (Chisholm and Motherwell 2005), ^g single Intel Xeon 5150 2.66 GHz processor using 1500 MB of memory.

Table 4.5 Detailed lattice energy minimisation results for the experimental structure of piracetam Form II.

			DMAFlex ^a	CrystalOptimizer							
				No LAMs, no LAM databases	LAMs, no LAM databases	LAMs and LAM databases	More flexibility, LAMs but no LAM databases				
				Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8
No. of Flexible Degrees of Freedom ^b			4 main torsions	4 main torsions	4 main torsions	4 main torsions	9 all torsions ^c	19 all torsions ^c , all angles ^d	21 all torsions ^c , all angles ^d , NH bonds	30 all torsions ^c , all angles ^d , all bonds ^e	54 atomistic ^f
Lattice Energy (kJ mol ⁻¹)			-93.49	-93.31	-93.31	-93.31	-93.94	-94.18	-94.16	-94.38	-94.33
Intramolecular Energy (kJ mol ⁻¹)			15.27	15.47	15.47	15.47	17.03	16.55	16.48	16.80	17.08
Intermolecular Energy (kJ mol ⁻¹)			-108.76	-108.78	-108.78	-108.78	-110.97	-110.73	-110.64	-111.18	-111.41
rmsd ₁₅ ^g			0.25	0.25	0.25	0.25	0.22	0.22	0.22	0.22	0.21
Experimental											
Unit Cell	a (Å)	6.40	6.41	6.41	6.41	6.41	6.42	6.43	6.43	6.43	6.42
	b (Å)	6.62	6.50	6.50	6.49	6.50	6.56	6.52	6.52	6.52	6.52
	c (Å)	8.56	8.78	8.75	8.77	8.75	8.64	8.71	8.71	8.69	8.70
	α (°)	79.9	82.9	83.0	83.0	83.0	82.2	82.7	82.7	82.6	82.6
	β (°)	102.4	100.5	100.2	100.2	100.2	99.8	100.3	100.3	100.2	100.7
	γ (°)	91.1	90.3	90.3	90.4	90.3	90.8	90.6	90.6	90.8	90.6
Density (g cm ⁻³)		1.355	1.325	1.325	1.325	1.325	1.329	1.326	1.325	1.327	1.329
CPU Time (hr:min) ^h			11:42	02:11	01:28	00:17	04:28	16:14	15:19	44:52	58:31
No. of outer iterations			55	2	2	2	9	32	26	63	54
No. of QM calculations	Molecular Optimisation	55	2	2	1	3	7	7	16	14	
	Molecular Hessian	-	2	1	0	2	6	6	15	13	
	Charge Density	55	10	6	1	17	49	49	121	131	
% CPU Time in	QM Molecular optimisation	57.45	18.18	27.61	15.48	9.46	3.67	3.74	2.17	1.05	
	QM Hessian calculation	-	27.04	19.79	0.00	19.88	10.27	10.74	9.30	6.17	
	QM Charge Density, DMA	39.97	48.15	43.14	36.54	40.08	31.29	33.05	28.13	23.29	
	DMACRYS	2.57	6.62	9.43	47.86	30.51	54.64	52.35	60.27	69.34	
	Other	0.01	0.01	0.03	0.12	0.07	0.13	0.12	0.13	0.15	

^a(Karamertzanis and Price 2006), ^b linear update to the multipole moments calculated for all torsions involving the N atom, ^c excluding dihedrals defined as H-C-X-X, ^d excluding H-C-X, ^e excluding bond lengths defined as H-C, ^f all intramolecular degrees of freedom, ^g root mean square deviation in 15-molecule coordination sphere compared with the experimental structure (excluding hydrogen atoms) (Chisholm and Motherwell 2005), ^h single Intel Xeon 5150 2.66 GHz processor using 1500 MB of memory.

Figure 4.5 demonstrates the accuracy of reproduction of the experimental conformation obtained by considering full molecular flexibility (atomistic representation). Even with the low-quality level of theory used for the QM evaluations, the reproduction of the crystal structures with *CrystalOptimizer* is acceptable for crystal structure prediction. As shown in Table 4.3, Table 4.4 and Table 4.5, the maximum error in the root-mean-square deviation of a 15-molecule coordination sphere in all three cases is below 0.3 Å which is less than the 0.4 to 0.5 Å criterion usually taken to indicate a successful prediction in crystal structure blind tests. In general, structure reproduction improves as a larger number of flexible degrees of freedom are considered. An exception to this trend is observed for xylitol (see the increase in rms_{15} in the last 6 columns of Table 4.4) but the overall rms_{15} variation for this molecule is too small to warrant further attention considering that thermal effects are not taken into account.

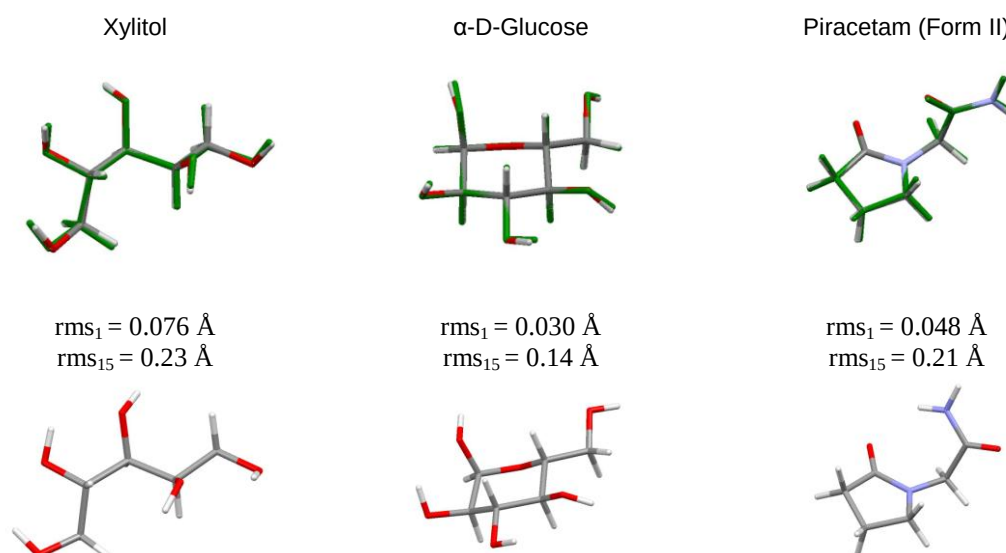


Figure 4.5 Top row: Overlay of the experimental (coloured by element) and minimised (green) structures of xylitol, α -D-glucose and piracetam when full molecular flexibility is allowed (atomistic representation). The root mean square deviation for one molecule (rms_1) and 15-molecule coordination sphere (rms_{15}) (Chisholm and Motherwell 2005) compared to the experimental structure are shown below. Bottom row: The gas-phase molecular conformations for the three molecules.

Comparisons of the DMAflex results to those of Cases 1 to 3, as presented in Table 4.3, Table 4.4 and Table 4.5, indicate that the results obtained from *CrystalOptimizer* and DMAflex in terms of energy and structure reproduction are within the errors of the underlying numerical methods, and in particular of the criteria used to determine what constitutes a “converged” solution of the optimisation calculations. In the case of

CrystalOptimizer, some convergence inaccuracies are introduced by the use of finite differences for the evaluation of the gradients of the intermolecular energy. For *DMAflex*, there are convergence issues arising from the use of the gradient-free simplex-based optimisation algorithm. This is especially apparent from Case 1 where all functions are evaluated in the same way in both algorithms.

CrystalOptimizer is significantly more computationally efficient than *DMAflex*. With all features implemented (Case 3), the CPU time is reduced by $\sim 89\%$ for xylitol, $\sim 95\%$ for glucose and $\sim 97\%$ for piracetam when considering the same degree of molecular flexibility. One reason for this is that *CrystalOptimizer*'s gradient-based minimisation scheme converges in an order of magnitude fewer outer iterations than the simplex algorithm in *DMAflex*, and scales better with the number of modelled flexible degrees of freedom, as can be observed by comparing the *DMAflex* results and Case 1 columns in Table 4.3, Table 4.4 and Table 4.5. This accounts for an average reduction in CPU time of 63%. A second reason for the much improved computational performance of *CrystalOptimizer* is that the use of LAMs means that only a small fraction of the outer iterations require full quantum mechanical calculations. It should be noted, however, that the types of QM calculations required by *DMAflex* and *CrystalOptimizer* are not the same. Each outer iteration in *DMAflex* simply requires one constrained geometry optimisation and one charge density calculation, the former usually being more computationally expensive. On the other hand, the construction of the LAMs in *CrystalOptimizer* requires a constrained geometry optimisation followed by a (usually much more expensive) analytical evaluation of the intramolecular Hessian matrix (i.e. the second-order partial derivative matrices appearing on the right hand side of equation 3.3), and a charge density calculation. Moreover, if the linear update to the multipole model LAM given by equation 3.13 is used for one or more flexible degrees of freedom, one or two additional charge density calculations for each of such flexible degrees of freedom are required, depending on the finite difference method used for approximating the partial derivatives on the right hand side of equation 3.13. Nevertheless, despite the additional cost of constructing the LAMs in *CrystalOptimizer*, the frequency with which this has to be done is small (Case 2, Table 4.3, Table 4.4 and Table 4.5), leading to a further reduction in computational cost of 46%. The poor scaling of the *DMAflex*'s simplex optimisation algorithm with the number of minimisation variables, coupled with the significant cost of the QM calculations, indicates that this approach is unlikely to be practically applicable to systems of even

moderate molecular size if they involve more than about 10 flexible degrees of freedom. It is likely that any other approach that relies on explicit QM calculations at every iteration would be subject to similar limitation.

Case 3 in Table 4.3, Table 4.4 and Table 4.5 indicates the computational cost if all the information required to construct the LAMs was already available in a LAM/QM database, which essentially removes the need to perform QM calculations during the minimisation. Note that one QM geometry optimisation and one charge density calculations are still necessary for the explicit evaluation of the *final* lattice energy to remove errors associated with the use of LAMs.

Based on Cases 2, 4-8 in Table 4.3, Table 4.4 and Table 4.5, the computational cost of the *CrystalOptimizer* algorithm seems to scale linearly with the number of flexible degrees of freedom, and is dominated by two elements, both of them associated with the use of finite difference approximations to partial derivatives. The first element arises from the calculation of the gradients of intermolecular energy with respect to the flexible degrees of freedom (cf. equation 4.6) by repeated calls to the DMACRYS package; for systems involving more than about 15 flexible degrees of freedom, this calculation accounts for more than 60% of the computational time. The second element is associated with the numerical calculation of multipole derivatives with respect to the specified flexible degrees of freedom (cf. equation 3.13) and can account for up to 50% of the computational time in cases involving less than 15 flexible degrees of freedom.

4.4.5. Application to multicomponent crystals

CrystalOptimizer is directly applicable to systems with more than one species in the asymmetric unit, including co-crystals and salts, such as those listed in Table 4.2. It can also take account of the effects of pressure on crystal structure, via a simple extension of the intermolecular energy (equation 4.4) to lattice enthalpy. Table 4.6 presents results for co-crystals involving α -D-glucose or piracetam, and for all 5 known forms of piracetam, two of which (forms IV and V) are studied under elevated pressures. Table 4.7 presents further results for a salt and two co-crystals of progesterone. Molecular diagrams of these systems can be found in Table 4.2.

As for the systems studied in section 4.4, agreement with the experimentally observed crystal structures is generally good, with the rms_{15} error being below 0.30 Å

for all structures other than the salt and the progesterone pyrene 2:1 co-crystal. All of the single-component polymorphs (including the high-pressure forms) of piracetam are reproduced well and are minimised in less than 8 CPU hours (Table 4.6). The computational cost for the co-crystals is larger (but still acceptable) due to the incorporation of more than one molecule in the asymmetric unit, for which QM calculations have to be performed independently, and a larger number of flexible degrees of freedom considered during the minimisation.

Table 4.6 Lattice energy minimisation of additional crystal structures containing α -D-glucose or piracetam in the asymmetric unit. Refer to Table 4.2 for molecular diagrams. The number of flexible degrees of freedom considered for each molecule during the minimisation is given in square brackets.

System		Lattice Energy (kJ mol ⁻¹) (Pressure) ^a	rmsd ₁₅ ^b (Å)	# outer iterations	# LAM updates ^c		CPU time (hr:min) ^d
α -D-glucose Urea (1:1) co-crystal		-221.62 (0.0 GPa)	0.11	13	2 [14]	3 [4]	19:08
Piracetam	Form I ^e	-91.56 (0.0 GPa)	0.21	5	1 [9]		02:58
	Form II	-93.94 (0.0 GPa)	0.22	9	2 [9]		04:28
	Form III	-93.97 (0.0 GPa)	0.13	13	3 [9]		07:28
	Form IV	-52.51 (0.4 GPa)	0.25	9	2 [9]		05:10
	Form V	291.58 (4.0 GPa)	0.30	12	3 [9]		06:04
Piracetam Gentisic acid (1:1) co-crystal		-188.45 (0.0 GPa)	0.26	37	2 [9]	1 [4]	13:32
Piracetam p-Hydroxybenzoic acid (1:1) co-crystal		-193.22 (0.0 GPa)	0.28	7	2 [9]	1 [3]	09:37

^a Two pressure polymorphs of piracetam, Forms IV and V, were minimised at the experimental pressure of 0.4 and 4.0 GPa respectively; ^b root mean square deviation in 15-molecule coordination sphere compared with the experimental structure (Chisholm and Motherwell 2005); ^c for each molecule in the asymmetric unit as shown in the system; ^d single Intel Xeon 5150 2.66 GHz processor using 1500 MB of memory; ^e disordered experimental form with atoms refined over two positions with occupancies 0.657:0.343. Only the most abundant conformer used for the minimisation

Despite the number of QM calculations needed to be performed during the course of lattice energy minimisation is kept to a minimum due to the use of LAMs, the computational cost associated with the minimisation of the progesterone-resorcinol crystal is large (Table 4.7). This is mainly due to the QM cost of updating the LAMs as each of the three terminal methyl groups in progesterone rotates by an average of 35°. In the case of the progesterone-pyrene co-crystal, the computational cost is large due to the computation of the numerical gradients for each of the 46 flexible torsions considered explicitly during the minimisation (Table 4.7). In practice, the rings in the progesterone molecule can be assumed to be relatively rigid. As a result, the

computational cost for these two systems can be significantly reduced by decreasing the number of torsions that have to be explicitly minimised.

It is worth noting that the reproduction of these larger experimental structures is very sensitive to the selected QM model. Due to the ionic nature of the salt, it is imperative to accurately capture and reproduce the electrostatic interactions in the crystal. If the quality of the charge density calculation is improved by using the MP2/6-31G(d,p) level of theory, the rms₁₅ deviation can be reduced from 1.34 Å to 0.84 Å. Similarly, if the hybrid B3LYP/6-31G(d,p) level of theory is used to represent the intramolecular interactions and the multipole model during lattice energy minimisation of the progesterone-pyrene (2:1) co-crystal, the error in rms₁₅ is reduced to 0.35 Å. The selection of the best available computational and flexibility model is therefore integral to the correct reproduction of experimental structures. This aspect is discussed in more detail in Chapter 6. Overall, there seems to be no algorithmic limitations of applying *CrystalOptimizer* to the systems of such size, flexibility and complexity.

Table 4.7 Lattice energy minimisation of selected crystal structures of varying size and complexity. Refer to Table 4.2 for molecular diagrams. The number of flexible degrees of freedom considered for each molecule during the minimisation is given in square brackets.

System	Lattice Energy (kJ mol ⁻¹)	rmsd ₁₅ ^a (Å)	# outer iterations	# LAM updates ^b			CPU time (hr:min) ^c
(R)-1-phenyl-2-(4-methylphenyl) ethylamonium-(S)-mandelate salt	-596.43	1.34	36	8 [7]	8 [4]		45:55
Progesterone Resorcinol (1:1) co-crystal	-221.92	0.15	61	13 [23]	4 [3]		179:11
Progesterone Pyrene (2:1) co-crystal	-458.81	0.72	32	3 [23]	3 [23]	1 [0]	130:53

^a root-mean-square deviation in 15-molecule coordination sphere compared with the experimental structure (Chisholm and Motherwell 2005); ^b for each molecule in the asymmetric unit as shown in the system; ^c 4 Intel Xeon 5150 2.66 GHz processors with 7 GB shared memory

4.4.6. Application to ab initio Crystal Structure Prediction

The main purpose for the development of an accurate local lattice energy minimisation algorithm for crystal structures containing flexible molecules is not the re-minimisation of the experimentally determined crystals, but the final refinement of a large set of hypothetical structures produced during a crystal structure prediction search.

It is in this context that the performance gains through the use of LAM/QM databases become more relevant (refer to section 3.6 for more details).

As an illustrative example, the application of *CrystalOptimizer* to the refinement of 60 distinct crystal structures for xylitol using 9 main hydroxyl and backbone torsions as the flexible degrees of freedom (cf. Case 2 in Table 4.3) is considered. The 60 structures are typical of those that would be considered in the context of *ab initio* crystal structure prediction; in this example case, they were generated using the *CrystalPredictor* code (Karamertzanis and Pantelides 2005) by treating xylitol as a rigid molecule.

Figure 4.6 shows three lines:

- The solid line shows the cumulative number of outer minimisation iterations against the number of crystal structures analysed. It can be seen that a total of 2412 iterations are needed, each effectively corresponding to a different molecular conformation. The computational cost would be prohibitively high if a new QM evaluation of intramolecular energy and charge density were performed at each such iteration. This effectively excludes the use of codes such as *DMAflex* for this kind of application.
- The dashed line shows the number of LAMs created if no LAM/QM databases are used. This was obtained by a modified form of the *CrystalOptimizer* code which, like the standard form, created and used LAMs; however, if an iteration involved a molecular conformation outside the range of validity of the most recently constructed LAM, the latter was simply discarded (i.e. without any information being recorded in the LAM/QM database for possible future use) and a new one created. The total number of LAMs that need to be constructed in this case is 633. Comparing this number with the total number of iterations, we conclude that a LAM is approximately re-used over a sequence of, on the average, 3.8 successive iterations.
- The dotted line shows the number of LAMs created if the standard *CrystalOptimizer* code (including LAM/QM databases) is used. In this case, the refinement of the first structure is started using an empty LAM/QM database, and the number of entries in the database potentially increases with the consideration of each new structure. The total number of LAMs

created is 408. The difference between this and the dashed line is a measure of the savings achieved via the use of a LAM/QM database.

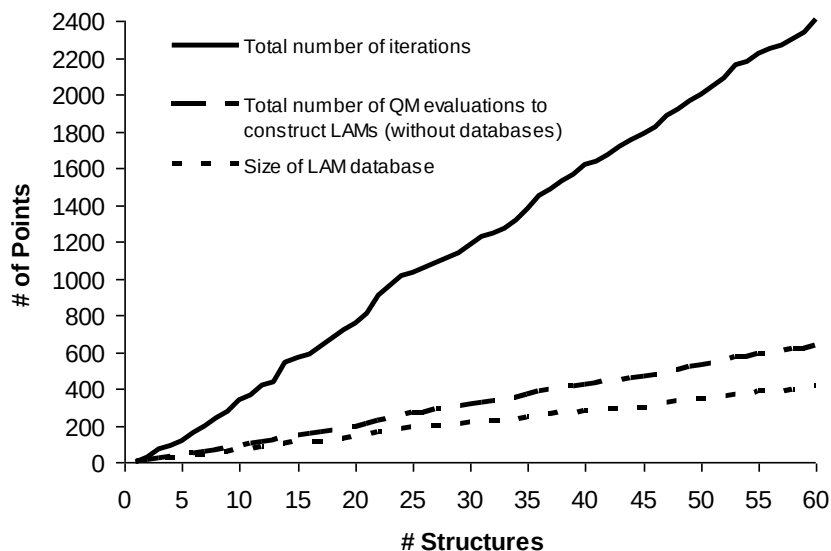


Figure 4.6 Total number of *CrystalOptimizer* minimisation iterations (solid), number of LAM updates (dashed) and the actual number of QM evaluation sets necessary to construct LAM/QM database (dotted) as a function of the number of crystal structures minimised.

The savings achieved via the use of a LAM/QM database would be higher if a larger number of crystal structures were refined: as the database is enlarged, there is an increasingly higher probability that a molecular conformation under consideration will be covered by one of the points that are already in the database, thereby obviating the need to perform a new set of QM calculations. This is particularly important since, in typical *ab initio* crystal structure predictions, it is not uncommon to refine many hundreds or even thousands of candidate structures. These structures are often far from a local minimum, so that additional optimisation iterations are needed for convergence. Thus, reducing the cost of each iteration is very important for a thorough search.

In summary, the use of LAM/QM databases is of significant benefit in the context of *ab initio* crystal structure prediction studies involving the refinement of large numbers of candidate crystal structures. It is also advantageous for the computational screening of compounds with different co-crystal formers and for crystal structure calculations under different physical conditions (e.g. a range of different pressures for identification of high-pressure polymorphs).

4.5. Conclusions

For the majority of organic crystal structures of practical interest, detailed modelling of molecular flexibility is essential for correct minimisation of the lattice energy. In order to obtain balanced models for the intra- and intermolecular energy contributions, it is necessary to use quantum mechanical molecular deformation energies and charge densities, making lattice energy minimisation computationally expensive.

An algorithm has been presented in this chapter that makes it possible to take molecular flexibility into account when identifying crystal structures that minimise the lattice energy. This algorithm, *CrystalOptimizer*, is a local optimisation method that overcomes the large computational cost of treating flexibility through the use of local approximate models (LAMs). This leads to a substantial reduction in the number of quantum mechanical calculations required. In addition, the algorithm is based on a Quasi-Newton scheme that ensures rapid convergence even with systems involving many flexible degrees of freedom.

The results presented for a variety of single-component crystals, cocrystals and a salt indicate that the implementation of local approximate models can significantly improve the computational efficiency without adversely affecting accuracy. In practice, during the course of an optimisation, most iterations rely on previously constructed LAMs (rather than new QM calculations) for the computation of the values of the rigid degrees of freedom, the intramolecular energy and the electrostatic multipole model. Moreover, the use of LAM/QM databases further reduces the computational cost associated with QM calculations allowing the accurate energy minimisations performed by *CrystalOptimizer* to be coupled directly within the crystal structure prediction techniques. The low-energy structures computed with *CrystalOptimizer* can also be used as input to the study of kinetic effects (Santiso and Trout 2011) and crystal growth (Snyder and Doherty 2009), which play an important role in crystal formation.

Overall, with the *CrystalOptimizer* code, the computational cost of QM calculations is no longer the limiting factor in accurate lattice energy minimisation, and this makes it possible to achieve even higher accuracy by using larger basis sets and/or post-HF levels of theory. The computational cost of the algorithm seems to scale linearly with the number of flexible degrees of freedom, and is dominated by the calculation of the

multipole derivatives and of the intermolecular energy derivatives with respect to the flexible degrees of freedom. Both currently require the use of finite difference approximations to partial derivatives. Removing these computational bottlenecks¹ will extend the practical applicability of *CrystalOptimizer* to even larger molecules.

On a more fundamental level, an important aspect of *CrystalOptimizer* is its ability to handle systems with more than one molecule or ion in the asymmetric unit, which makes it applicable to the prediction of crystal structures of cocrystals and salts. Our approach to date has been to treat each of these molecules or ions independently for the purposes of QM calculations, and then to rely on intermolecular descriptions of the interactions between them. Whether or not this is the most appropriate way of handling such systems, or related ones such as hydrates, requires further investigation.

¹ If the partial derivatives for the intermolecular energy with respect to the flexible degrees of freedom could be obtained analytically, then the bi-level optimisation algorithm in *CrystalOptimizer* could be potentially replaced with a more robust single-level minimisation.

Chapter 5. Global minimisation of lattice energy – extension to CrystalPredictor

A variety of computational methods and approaches for crystal structure generation have been reviewed in section 2.4.1. This chapter presents extensions to an existing search algorithm for flexible molecules, CrystalPredictor (Karamertzanis and Pantelides 2005, Karamertzanis and Pantelides 2007) to substantially increase its ability to handle molecules with significant flexibility, and to improve the reliability of its predictions. To start with, the original algorithm is described in detail and its main limitations illustrated in section 5.1.

Section 5.2 describes how these limitations can be overcome by adopting a more realistic representation of the molecular geometry (and flexibility) through the use of the molecular model and local approximate models presented in Chapter 3.

The precise formulation of the lattice energy minimisation problem and the modifications that need to be effected to the original algorithm are presented in section 5.3.

As an illustration of the proposed modifications, the algorithm is applied to the determination of the crystal structures of ROY, 5-methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile, in section 5.4. The accuracy of the algorithm is critically assessed by its ability to reproduce the ranking, lattice geometry and conformational degrees of freedom of the known polymorphs of this molecule (Yu et al. 2000, Chen et al. 2005).

5.1. Description and limitations of CrystalPredictor

CrystalPredictor (Karamertzanis and Pantelides 2005, Karamertzanis and Pantelides 2007) is a global energy minimisation algorithm that allows the identification of the crystal structures in the 64 most common space groups and with more than one molecule in the asymmetric unit (as in the case of salts, cocrystals and hydrates). The global search and minimisation approach, outlined in Figure 5.1, comprises of four distinct steps, which are described in more detail below.

5.1.1. Algorithm overview

In the **first step**, the global search procedure in the algorithm generates candidate crystal structures within a search space specified by the user in terms of the set of space groups to be considered and the number of molecules in the asymmetric unit. The decision variables that are sampled (within the user-specified parameter ranges) include the choice of the space group, the magnitudes of the lattice lengths and lattice angles, the position and orientation of the independent molecules in the unit cell, and, for flexible molecules, the values of the selected flexible torsional angles. The search method uses Sobol' deterministic low-discrepancy sequences (Sobol' 1967) to achieve the best coverage of this multidimensional search space. In this approach, N decision variables are generated simultaneously in an N -dimensional variable domain as binary fractions from a set of special direction numbers¹. A characteristic of the approach is that the projection of each and every Sobol' point onto each of the axes corresponds to a *distinct* combination of every decision variable. This is an advantage over the random number Monte Carlo techniques or grid methods used in other crystal structure generation methods where uniform and complete coverage of the domain of interest may not be guaranteed or is computationally prohibitive (L'Ecuyer 1988). Another advantage of the low-discrepancy sequences is that, unlike in a grid search, the final number of trial structures does not have to be decided before carrying out a search. Additional starting points can be always produced, taking into account all the previously generated crystals, if deemed necessary. The generated cells are checked for feasibility (density, lattice energy, steric hindrance, etc.) and used as a starting point for local minimisation of the lattice energy.

In the **second step** of the algorithm, the lattice energy of each candidate structure is optimised taking account of any symmetry constraints implied by its corresponding space group. Due to the properties of the structure generation sequence, it is theoretically² guaranteed that at some point during the global search, a point will be generated that is sufficiently close to each low-energy minimum to be within the local optimiser's region of attraction. For each local energy minimisation, a sequential

¹ Refer to the original literature for the details of the methodology and the generation of direction numbers (Sobol' 1967).

² In infinite time and provided the search space has been selected well.

quadratic programming algorithm¹ (Dennis and Schnabel 1983), is used guaranteeing convergence to a local minimum (rather than a saddle point). Furthermore, all lattice energy gradients are calculated analytically and the algorithm exhibits fast (super-linear) convergence in the region of the local minimum. The number of iterations, and consequently the computational cost, is further reduced by appropriate scaling of the decision variables.

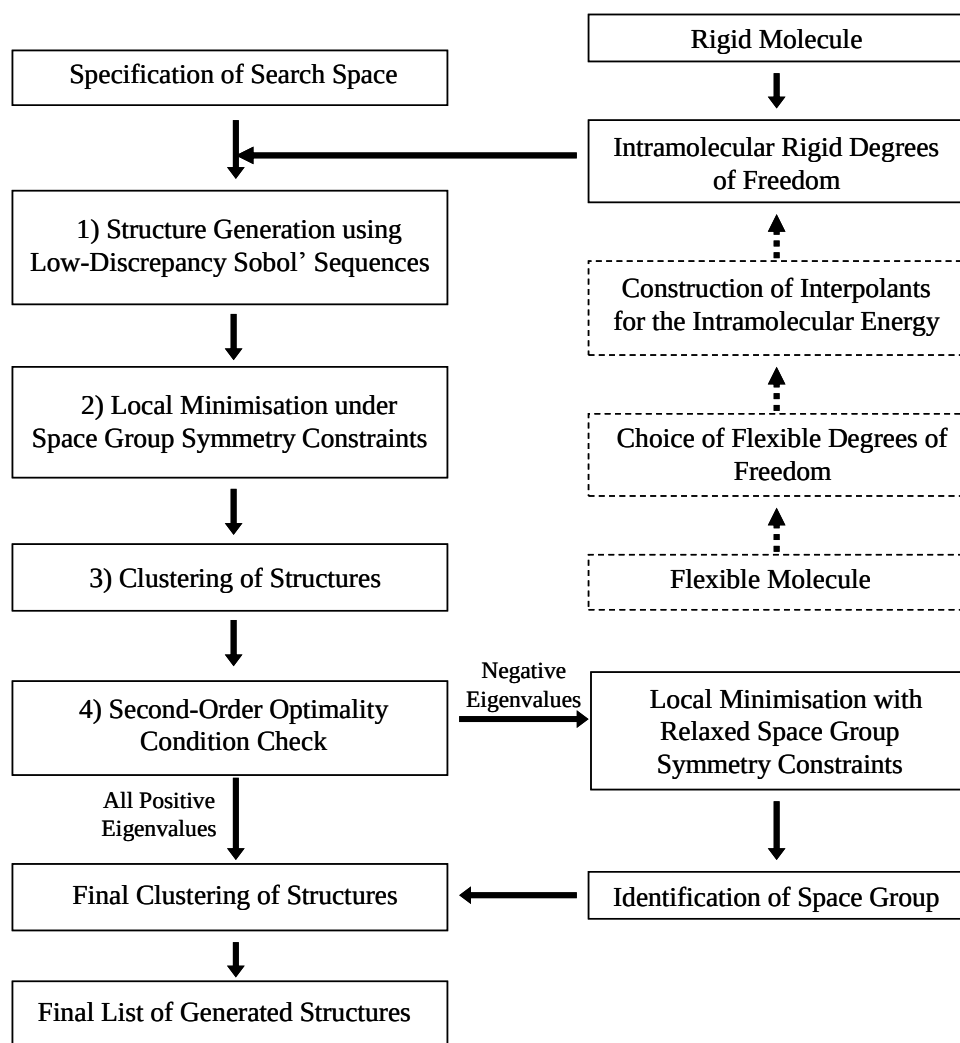


Figure 5.1 Flowchart of the CrystalPredictor search and minimisation algorithm.

During the calculation of the lattice energy, it is assumed that each molecule comprises a number of sites that can participate in, and contribute to, the repulsion-dispersion and electrostatic interactions, which are modelled using a set of distributed charges. The electrostatic interactions sites do not have to coincide with atomic

¹ Implemented via routine E04UFF of the Nag Library, see <http://www.nag.co.uk>

positions, consequently, the distributed charges may be located at ‘satellite’ positions in vicinity of certain atoms. The exact magnitudes and the positions of these charges are determined by fitting to molecular electrostatic field computed using quantum mechanical calculations (Karamertzanis and Pantelides 2004). Reciprocal and real Ewald sums (Ewald 1921, Toukmaji and Board 1996) are used to improve convergence arising from the modelling of electrostatic interactions. A user specified cut-off distance is employed to truncate infinite summations of pairwise interaction energies.

In the **third step** of the algorithm, the local minima identified are clustered according to their lattice energies, densities and interatomic distances to remove any duplicate structures. The clustering procedure may be tuned to achieve the desired extent of structure separation.

Local minimisations at step two are carried out subject to space group symmetry and, therefore, it is possible that some of the structures are local minima only with respect to the space group. Consequently, such points could be unstable saddle points on the lattice energy surface that have to be eliminated. As a result, in the **fourth step** of the algorithm, the second-order optimality conditions are checked by evaluating the Hessian matrix using centred finite differences. If its eigenvalues are non-positive (i.e. the second-order optimality conditions are not satisfied), a small perturbation to the structure is introduced, and it is subsequently minimised without the space group constraint. Following the minimisation, the space group of the resulting structure is identified (Spek 2003).

A parallel implementation of the algorithm, in which the generation and local minimisation of candidate structures are performed simultaneously using a network of processors, ensures an efficient sampling of the multidimensional search space.

5.1.2. Representation of molecular flexibility

In CrystalPredictor, flexible molecules are modelled as sets of rigid fragments of fixed geometry, connected by flexible torsion angles (Figure 5.2). The justification for this representation is that only a small set of flexible degrees of freedom (torsions) is often sufficient in capturing most of the effects of the molecular flexibility. However, whilst this assumption usually holds for smaller molecules (Mooij et al. 2000, Karamertzanis et al. 2009), it is, of course, molecule-dependent and may not be

applicable for larger pharmaceutical molecules. Moreover, as the values of the flexible degrees of freedom change from their gas-phase values, the rest of the intramolecular degrees of freedom adjust so as to minimise the intramolecular energy (cf. equation 3.2 in section 3.1). Hence, these remaining, more rigid intramolecular degrees of freedom, are not truly fixed as assumed in CrystalPredictor, but are indirectly affected by the intermolecular interactions. This introduces a degree of approximation, which leads to a miscalculation of the actual positions of the atoms in the molecule within the lattice, and therefore of the corresponding intermolecular interactions. Albeit individually small, these effects are additive and, thus, more pronounced for large and flexible molecules as shown in Figure 5.3 using one of the molecules studied in detail in Chapter 6. For this example, the molecular geometries obtained by the explicit QM constrained geometry optimisation and by rotating the fixed fragments from the gas-phase geometry are slightly different, despite the values of the five flexible torsions being the same (Figure 5.3a). The original CrystalPredictor algorithm is particularly sensitive to these effects, since the molecules are constructed successively starting from a base fragment which is usually terminal. This may introduce considerable deviations in atomic positions at the other end of the molecules (Figure 5.3b).

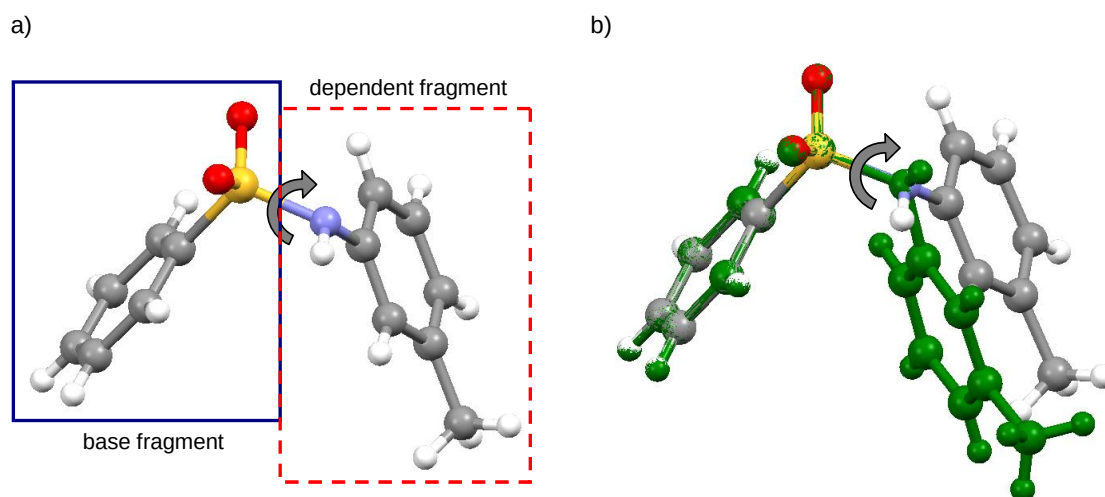


Figure 5.2 Description of flexible molecules used in the original CrystalPredictor algorithm. a) Fragment definition for a molecule with just one flexible torsion (grey arrow). Note that the sulphur atom (yellow) and the nitrogen atom (blue) are part of both fragments (a fact not indicated on the diagram in the interests of clarity). b) Overlay of the molecule geometry before (coloured by element) and after (green) a 60° change in the flexible torsion angle (grey arrow). The geometry of the individual fragments is kept fixed throughout.

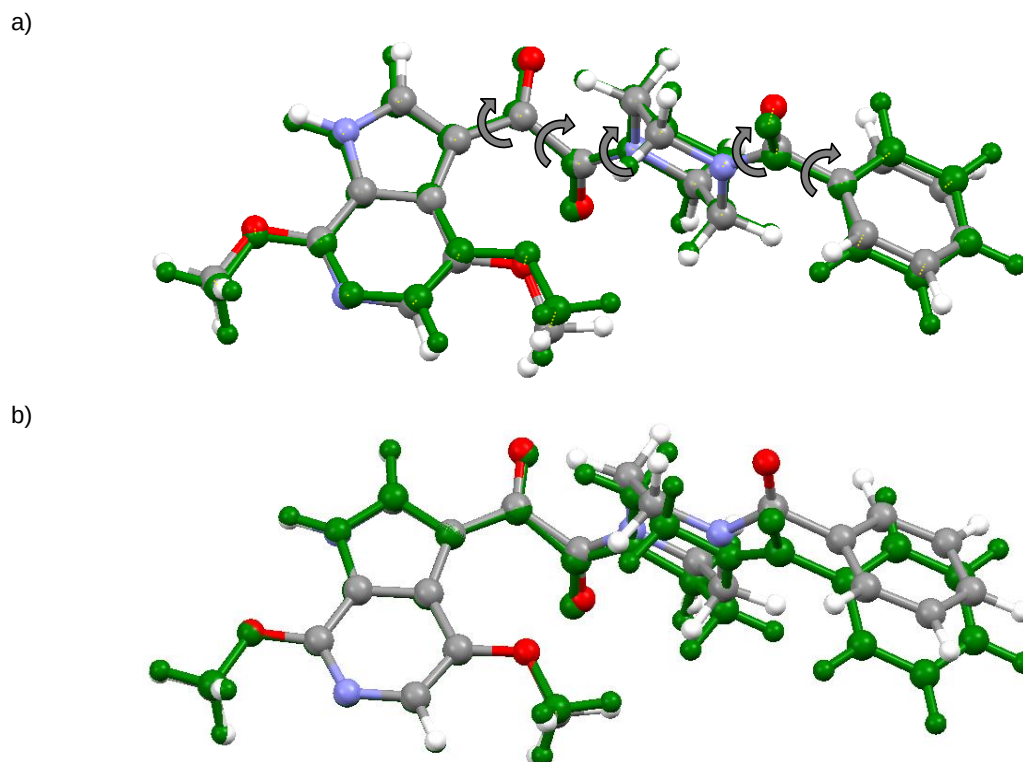


Figure 5.3 (a) Whole molecule and (b) the terminal conjugated ring (base fragment) overlays (excluding hydrogen atoms) of the molecular geometries obtained using constrained QM geometry optimisation at the PBE0/6-31G(d,p) level of theory (coloured by element) and that obtained by rotating the fragments of the molecule from its gas-phase geometry (also obtained at the PBE0/6-31G(d,p) level of theory). The values of the five flexible torsions (grey arrows) are the same for both molecular models in both cases. Detailed crystal structure prediction results for this molecule are provided in section 6.2.

As a result, the structure reproduction, the absolute energy and, consequently, the ranking of low-energy crystal structures produced by CrystalPredictor for many flexible molecules is not very reliable (Day et al. 2007, Karamertzanis et al. 2009). The errors are usually corrected during subsequent re-minimisation of the crystal lattice energy using more accurate (and expensive) models (Day and Cooper 2010, or for example using the algorithm described in Chapter 4). However, if these errors in the search stage were to be reduced, the ranking of plausible structures could be improved thereby reducing the number of candidates that needs to be subjected to further consideration. Overall, this could lead to significant savings in the computational cost of the crystal structure prediction methodology and increased likelihood of finding all experimentally relevant structures.

In order to take account of the effects of the flexible degrees of freedom on the molecular conformation, an alternative molecular geometry model is adopted in section 5.2. The necessary modifications to the lattice energy minimisation procedure in

CrystalPredictor are then presented in section 5.3. Finally, the relative benefits of the above implementations are discussed in section 5.4 by means of an illustrative example.

5.2. Geometric modelling of flexible molecules

This section presents an alternative description for the geometrical representation of molecules, which does not rely on the specification of fragments with fixed geometries as implemented in the original version of CrystalPredictor.

The positions of all atoms in a molecule can be determined from a knowledge of the position and orientation of three base reference atoms together with the values of all the intramolecular degrees of freedom (bond lengths, bond and torsional angles), which are contained within the Z-matrix (refer to section 3.1 for a more detailed description of the Z-matrices). All of the Z-matrix variables can be dynamically and efficiently updated as functions of the flexible torsions using the local approximate model of equation 3.7 described in Chapter 3. As a result, it is no longer necessary to treat the molecule as a set of fully rigid fragments, which allows a more realistic description of molecular conformations based directly on the results of QM calculations (i.e. eliminating the discrepancies with the QM molecular model illustrated in Figure 5.3).

To resolve the atomic Cartesian positions in the molecule from the Z-matrix variables, a common convention is adopted for the placement of three base atoms:

- The first atom is placed at the origin:

$$\mathbf{r}_A = (0 \ 0 \ 0)^T \quad \text{Eq. 5.1}$$

where $\mathbf{r}_i$ is the vector defining the position of atom i , $\mathbf{r}_i = (x_i \ y_i \ z_i)^T$

- The second atom is placed on the positive x-axis:

$$\mathbf{r}_B = (l_{AB} \ 0 \ 0)^T \quad \text{Eq. 5.2}$$

where l_{AB} is the bond length between atoms A and B contained in the Z-matrix.

- The third atom is positioned on the x-y plane with a positive y-coordinate. Note that the third atom can be bonded to either the first or the second atom. As a result, the x-coordinate of the third atom, x_C , for the generic case when both coordinates of atoms A and B are known, can be obtained from the dot product of two vectors:

$$\mathbf{r}_{BA} \cdot \mathbf{r}_{BC} = \|\mathbf{r}_{BA}\| \|\mathbf{r}_{BC}\| \cos \nu_{ABC} \quad \text{Eq. 5.3}$$

where $\mathbf{r}_{ii'} \equiv \mathbf{r}_{i'} - \mathbf{r}_i$ and ν_{ABC} is the bond angle between atoms A-B-C. The magnitudes of the two vectors ($\|\mathbf{r}_{BA}\|$ and $\|\mathbf{r}_{BC}\|$) correspond to the bond lengths l_{BA} and l_{BC} respectively, which are stored in the Z-matrix along with the bond angle between the first three atoms ν_{ABC} . By substituting these results into the above equation the following can be obtained:

$$(x_A - x_B \quad 0 \quad 0) \cdot \begin{pmatrix} x_C - x_B \\ y_C \\ 0 \end{pmatrix} = l_{BA} l_{BC} \cos \nu_{ABC} \quad \text{Eq. 5.4}$$

which after simplification yields the x-coordinate of the third atom¹:

$$x_C = x_B + \frac{l_{BA} l_{BC} \cos \nu_{ABC}}{x_A - x_B} \quad \text{Eq. 5.5}$$

The y-coordinate can then be calculated using Pythagoras' theorem:

$$y_C = \sqrt{l_{BC}^2 - (x_C - x_B)^2} \quad \text{Eq. 5.6}$$

¹ Note that the equation is valid for both bonding types – when the third atom is bonded to the second atom (1-2-3) and when the third atom is bonded to the first atom (3-1-2). In the former case (1-2-3), the quantity $x_A - x_B$ is identical to l_{BA} and the equation reduces to the standard trigonometric solution $x_C = x_B + l_{BC} \cos \theta_{ABC}$. In the latter case (3-1-2) the quantity $x_A - x_B$ equals $-l_{BA}$ indicating the previous trigonometric solution can not be used.

Once the coordinates of the first three reference atoms have been defined, the Cartesian position of any other atom in the molecule can easily be resolved. For a molecular fragment consisting of atoms A-B-C-D, the Cartesian position of atom D can be expressed as a function of the already resolved Cartesian positions, $\mathbf{r}_i$, $i \in \{A, B, C\}$, and the bond length l_{DC} , the bond angle ν_{BCD} and the torsion angle ξ_{ABCD} . All three of which are known parameters stored in the Z-matrix. The position of atom D can be expressed as (Karamertzanis 2004):

$$\mathbf{r}_D = \mathbf{r}_C + l_{DC} \mathbf{r}_w \quad \text{Eq. 5.7}$$

where the unit vector $\mathbf{r}_w$ gives the direction from atom C to atom D and is defined as:

$$\mathbf{r}_w = \sin \nu_{BCD} \cos \xi_{ABCD} \mathbf{r}_u + \sin \nu_{BCD} \sin \xi_{ABCD} \mathbf{r}_t - \cos \nu_{BCD} \frac{\mathbf{r}_{BC}}{\|\mathbf{r}_{BC}\|} \quad \text{Eq. 5.8}$$

with the unit vector $\mathbf{r}_t$ defined normal to the plane of atoms A, B, C:

$$\mathbf{r}_t = \frac{\mathbf{r}_{BA} \times \mathbf{r}_{CB}}{\|\mathbf{r}_{BA} \times \mathbf{r}_{CB}\|} \quad \text{Eq. 5.9}$$

and the unit vector $\mathbf{r}_u$ vertical to the plane defined by $\mathbf{r}_{CB}$ and $\mathbf{r}_t$:

$$\mathbf{r}_u = \mathbf{r}_t \times \frac{\mathbf{r}_{CB}}{\|\mathbf{r}_{CB}\|} \quad \text{Eq. 5.10}$$

Equations 5.1 to 5.10 define atomic positions¹ relative to the adopted reference frame for the specific molecule. In order to obtain the positions in the coordinate system within the *crystal*, the Cartesian coordinates have to be further expressed relative to the position and orientation of a reference fragment usually defined in terms of the first three atoms in the crystal. This takes the form:

¹ Or more generally, positions of interaction sites (which may or may not correspond to atomic positions).

$$\bar{\mathbf{r}}_i = \mathbf{r}_{ref} + \mathbf{A}^T \mathbf{r}_i \quad i = A, B, C, \dots, N \quad \text{Eq. 5.11}$$

Where , $\mathbf{r}_{ref}$ is the position of the first atom, N is the total number of atoms in the molecule and $\mathbf{A}^T$ is the rotation matrix given by the first three atoms:

$$\mathbf{A} = \left[\frac{\mathbf{r}_{AB}}{\|\mathbf{r}_{AB}\|}, \frac{\mathbf{r}_{AB} \times \mathbf{r}_{BC}}{\|\mathbf{r}_{AB} \times \mathbf{r}_{BC}\|} \times \frac{\mathbf{r}_{AB}}{\|\mathbf{r}_{AB}\|}, \frac{\mathbf{r}_{AB} \times \mathbf{r}_{BC}}{\|\mathbf{r}_{AB} \times \mathbf{r}_{BC}\|} \right]^T \quad \text{Eq. 5.12}$$

The rotation matrix can also be expressed using three Euler angles, defined in terms of a sequence of rotations of a set of Cartesian coordinate axes about the origin. The first rotation is through an angle ϕ about the z-axis, followed by a rotation θ around the new x-axis and a final rotation through an angle ψ about the new z-axis. The angles ϕ and ψ can vary in the range $[-\pi, \pi]$ and θ in the range $[0, \pi]$. The three matrices describing the individual rotations are given by:

$$\mathbf{P}_\phi = \begin{bmatrix} \cos \phi & \sin \phi & 0 \\ -\sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{Eq. 5.13}$$

$$\mathbf{P}_\theta = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & \sin \theta \\ 0 & -\sin \theta & \cos \theta \end{bmatrix} \quad \text{Eq. 5.14}$$

$$\mathbf{P}_\psi = \begin{bmatrix} \cos \psi & \sin \psi & 0 \\ -\sin \psi & \cos \psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{Eq. 5.15}$$

The full rotation matrix, $\mathbf{A}$, is the product of the individual rotation matrices:

$$\begin{aligned}
\mathbf{A} &= \mathbf{P}_\psi \mathbf{P}_\theta \mathbf{P}_\phi \\
&= \begin{bmatrix} \cos \phi \cos \psi - \sin \phi \cos \theta \sin \psi & \sin \phi \cos \psi + \cos \phi \cos \theta \sin \psi & \sin \theta \sin \psi \\ -\cos \phi \sin \psi - \sin \phi \cos \theta \cos \psi & -\sin \phi \sin \psi + \cos \phi \cos \theta \cos \psi & \sin \theta \cos \psi \\ \sin \phi \sin \theta & -\cos \phi \sin \theta & \cos \theta \end{bmatrix} \quad \text{Eq. 5.16}
\end{aligned}$$

The orientation of the molecule plays an integral part in the calculation of the Cartesian positions of all the atoms in the molecule (equation 5.11). As a result, Euler angles are explicit minimisation variables in CrystalPredictor. They constitute the minimum possible information than can describe uniquely the rotation of molecules (Frenkel and Smit 2002). However, from the computational point of view, trigonometric functions are more expensive to evaluate than simple polynomial expressions. In CrystalPredictor, the orientation of molecules is defined in an alternative, mathematically equivalent, way through the use of quaternions (Rapaport 2003, Karamertzanis 2004) described in more detail in Appendix E.

This section has focused on the description of a molecular geometry model for flexible molecules that is compatible with local approximate models described in Chapter 3. The incorporation of these modifications into the lattice energy minimisation problem is described in the next section.

5.3. Lattice energy minimisation of flexible molecules

5.3.1. Minimisation variables

The number and type of the optimisation variables in the lattice energy, E^{latt} , minimisation problem for flexible molecules remains unchanged from the original CrystalPredictor algorithm (Karamertzanis and Pantelides 2007). However, a few minor modifications are introduced into the precise definition of these variables:

$$\min_{\mathbf{x}, \boldsymbol{\theta}^f} E^{\text{latt}} \equiv \min_{\mathbf{x}, \boldsymbol{\theta}^f} \left[\Delta E^{\text{intra}}(\boldsymbol{\theta}^f) + U^{\text{inter}}(\mathbf{X}; \boldsymbol{\theta}^f, \boldsymbol{\theta}^r(\boldsymbol{\theta}^f), q(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r(\boldsymbol{\theta}^f))) \right] \quad \text{Eq. 5.17}$$

where ΔE^{intra} is the energy required to deform the molecule from its most stable gas-phase conformation and U^{inter} is the intermolecular energy describing the interactions between neighbouring molecules in the crystal. Because of the adoption of the new model for the molecular geometry (see section 5.2), the conformation of the *whole* molecule (θ) can now be updated. Therefore, as has been explained in sections 3.1 and 4.1, the dimensionality of the above optimisation problem can be managed by partitioning the intramolecular degrees of freedom, θ , into flexible, θ^f , and rigid, θ^r degrees of freedom. Consequently, $\theta^r(\theta^f)$ and $\Delta E^{\text{intra}}(\theta^f)$ can be approximated accurately and efficiently using the local approximate models shown in equations 3.7 and 3.9 respectively (refer to section 3.3 for the derivations). Here, $q(\theta^f, \theta^r(\theta^f))$ denotes the conformationally-dependent atomic (or distributed) charge model used to represent the electrostatic contributions to the intermolecular energy. For a given molecular conformation $\theta = (\theta^f, \theta^r(\theta^f))$, the intermolecular energy, U^{inter} , is computed with respect to the lattice variables, $\mathbf{X}^1$ defined as:

$$\mathbf{X} = (a, b, c, \alpha, \beta, \gamma, [\mathbf{r}_{\text{ref},i}, \phi_i, \theta_i, \psi_i \text{ for } i = 1..Z']) \quad \text{Eq. 5.18}$$

where a, b, c and α, β, γ are the three vectors and three angles, respectively, defining the unit cell. The Cartesian vector $\mathbf{r}_{\text{ref}}$ and the three Euler angles (ϕ, θ, ψ) define the position and orientation of every molecule in the asymmetric unit (Z'). Due to the use of the Z-matrix to define the molecular geometry, the notion of fully rigid fragments connected by flexible torsional angles, as implemented in the original CrystalPredictor algorithm, becomes redundant. As a result, the optimisation variables corresponding to the normalised position of the centres of mass of the base fragment in each independent molecule are replaced with the normalised position of the first atom in the Z-matrix, $\mathbf{r}_{\text{ref}}$, of each independent molecule. Similarly, the rotation of the molecules is defined by the Euler angles describing the orientations of the first three atoms in the Z-matrix (and not those of the base fragment, as is the case in the original CrystalPredictor algorithm).

Although perturbations in bond lengths and bond angles due to intermolecular interactions are also possible, the energy associated with such changes is high in

¹ Note that the dimensionality of this vector may be reduced due to space group constraints.

comparison to the weak intermolecular interactions found in the crystalline environment (Warshel et al. 1974, Harris and Lammertsma 1997, Hargittai and Levy 1999). Therefore, it is most likely that, at least within the degree of accuracy required by a global search algorithm, any major changes in molecular conformation due to the packing forces will be due to distortions in torsion angles. As a result, only a subset of the torsions present in the Z-matrix is allowed to be sampled explicitly (i.e. as defined by θ^f) during the search and lattice energy minimisation in CrystalPredictor. An extension to treating bond angles and bond lengths as explicit search variables could be implemented in a straightforward manner if deemed necessary.

5.3.2. Intramolecular energy and molecular conformation

In the original CrystalPredictor algorithm (Karamertzanis and Pantelides 2007), the intramolecular energy is evaluated using the restricted Hermite interpolation method (Stefanovic and Pantelides 2001, Karamertzanis 2004). The interpolant is a function only of a set of the selected conformational degrees of freedom. In order to achieve the desired accuracy with the minimal number of grid points (each one of which adds considerable computational cost as it requires QM calculations), the interpolation method exploits the values of the both the intramolecular energy and its partial derivatives with respect to the conformational degrees of freedom under consideration. Even so, the grid spacing should not exceed 20°-30° in order to maintain the required accuracy¹.

Although accurate in terms of the intramolecular energy, the above interpolation method does not provide any information on the dependence of the molecular conformation on the flexible torsional angles, which, as we have seen, may be significant in some circumstances. The molecular conformation along with the intramolecular energy can alternatively be obtained as explicit functions of the flexible degrees of freedom using local approximate models of the type presented in section 3.3 (equations 3.7 and 3.9 respectively). An additional benefit of the LAMs is that due to their linear or quadratic formulation, they are readily differentiable, thereby providing

¹ Consider a molecule with 4 torsional angles all of which may change by up to 180°. Even, with a grid spacing of 30°, 7 grid points will be evaluated for each torsion requiring 2401 (7^4) QM calculations, which is likely to be computationally prohibitive if anything other than basic level of theory is used for the calculations.

the partial derivatives required by the lattice energy minimisation algorithm in CrystalPredictor. In particular, the partial derivatives of the lattice energy, E^{latt} , with respect to the flexible torsional angles, θ^f , for each molecule in the asymmetric unit are computed as the sum of intra- and intermolecular contributions:

$$\frac{\partial E^{\text{latt}}}{\partial \theta^f} = \frac{\partial \Delta E^{\text{intra}}}{\partial \theta^f} + \sum_{i=1}^{N^s} \frac{\partial U^{\text{inter}}}{\partial \mathbf{r}_i} \frac{\partial \mathbf{r}_i}{\partial \theta^f} \quad \text{Eq. 5.19}$$

where $\mathbf{r}_i$ are the Cartesian coordinates of atom i and N^s is the number of atoms in the molecule. The first term in the expression is obtained by the analytical differentiation of the intramolecular energy LAM with respect to the flexible degrees of freedom shown in equation 4.5. The partial derivatives of the intermolecular energy with respect to the Cartesian coordinates of all atoms ($\partial U^{\text{inter}}/\partial \mathbf{r}_i$) are already determined in the original version of CrystalPredictor (Karamertzanis and Pantelides 2005, Karamertzanis and Pantelides 2007). The last partial derivative in the equation above can be further decomposed using the chain rule¹:

$$\frac{\partial \mathbf{r}_i}{\partial \theta^f} = \frac{\partial \mathbf{r}_i}{\partial \theta^r} \frac{\partial \theta^r}{\partial \theta^f} \quad \text{Eq. 5.20}$$

where $\partial \theta^r/\partial \theta^f$ represents the effect of changes in the flexible torsions on the molecular conformation and $\partial \mathbf{r}_i/\partial \theta^r$ describes how the Cartesian coordinates ($\mathbf{r}_i$) are affected by the changes in the rigid intramolecular degrees which must be calculated for bond lengths (l), bond angles (ν) and torsions (ξ):

$$\frac{\partial \mathbf{r}_i}{\partial \theta^r} \equiv \left(\frac{\partial \mathbf{r}_i}{\partial l}, \frac{\partial \mathbf{r}_i}{\partial \nu}, \frac{\partial \mathbf{r}_i}{\partial \xi} \right) \quad \text{Eq. 5.21}$$

For any molecular fragment consisting of bonded atoms A-B-C-D, the partial derivative of the Cartesian coordinates of atom D with respect to the intramolecular degrees of freedom contained in the Z-matrix (in this case bond length l_{DC} , the bond

¹ Note that in the original CrystalPredictor the $\partial \mathbf{r}_i/\partial \theta^f$ are obtained in a more direct manner since it is assumed that the geometry of the fragments is fixed.

angle ν_{BCD} and the torsion ξ_{ABCD}) can be determined via a vector formulation (Bradford Thompson 1967, Karamertzanis 2004). The partial derivative of the Cartesian coordinates with respect to the bond length is defined as:

$$\frac{\partial \mathbf{r}_D}{\partial l_{DC}} = \mathbf{r}_w = \frac{[\mathbf{r}_D - \mathbf{r}_C]}{\|\mathbf{r}_D - \mathbf{r}_C\|} \quad \text{Eq. 5.22}$$

where the unit vector $\mathbf{r}_w$ specifies the direction from atom C to atom D and has been previously defined in equation 5.8. The partial derivative with respect to the bond angle can be calculated using:

$$\frac{\partial \mathbf{r}_D}{\partial \nu_{DCB}} = -\mathbf{n}_{DBC} \times [\mathbf{r}_D - \mathbf{r}_C] \quad \text{Eq. 5.23}$$

where the unit vector $\mathbf{n}_{DBC}$ is perpendicular to the plane defined by atoms D-B-C:

$$\mathbf{n}_{DBC} = \frac{[\mathbf{r}_B - \mathbf{r}_C] \times [\mathbf{r}_D - \mathbf{r}_C]}{\|[\mathbf{r}_B - \mathbf{r}_C] \times [\mathbf{r}_D - \mathbf{r}_C]\|} \quad \text{Eq. 5.24}$$

Finally, the derivative with respect to the torsional angle is the vector product between a unit vector along the rotation axis and another vector from any position on that axis:

$$\frac{\partial \mathbf{r}_D}{\partial \xi_{ABCD}} = \frac{[\mathbf{r}_B - \mathbf{r}_C]}{\|\mathbf{r}_B - \mathbf{r}_C\|} \times [\mathbf{r}_D - \mathbf{r}_i] \quad \text{Eq. 5.25}$$

where $\mathbf{r}_i$ is the position of any atom lying on the axis of rotation (i.e. atoms B and C).¹

The last partial derivative term in equation 5.20 can be obtained by differentiating the LAM estimate of the rigid degrees of freedom, $\boldsymbol{\theta}^r$ (shown in equation 3.7):

¹ Note that changes in the intramolecular degrees of freedom in the beginning of the molecule will affect the positions (and gradients) all subsequent atoms. As a result, the derivatives are computed sequentially for all atoms and all intramolecular degrees of freedom and are summed up accordingly.

$$\frac{\partial \theta^r}{\partial \theta^f} = - \left[\frac{\partial^2 \Delta E^{\text{intra}}}{\partial \theta^{r^2}} \right]_{\theta_{\text{ref}}}^{-1} \left[\frac{\partial^2 \Delta E^{\text{intra}}}{\partial \theta^f \partial \theta^r} \right]_{\theta_{\text{ref}}}^T \quad \text{Eq. 5.26}$$

where the second-order partial derivatives of the intramolecular energy for a reference conformation are obtained analytically via a constrained QM minimisation of the molecule under consideration.

5.3.3. Generation of LAMs for CrystalPredictor

Performing the computationally expensive QM calculations “on-the-fly” (as done in CrystalOptimizer described in Chapter 4) as and when required during the search is impractical. This is mainly due to the difficulties with computational implementation in a parallelised environment. Since the local lattice energy minimisations in CrystalPredictor can be executed over a network of processors, it may be possible that different minimisations encounter similar conformational regions for which similar QM calculations have to be performed. As a result, it may be possible that the QM evaluations are duplicated, thereby wasting computational resources. A possible solution is to perform only one QM calculation, but this would leave the other processor(s) idle waiting for the result of this QM optimisation. An elegant solution would be to pause the execution of *all* CrystalPredictor minimisations as soon as a new QM calculation needs to be performed and redirect all the available computational resources to speed up the QM optimisation. However, the parallelised implementation of the GAUSSIAN package (Frisch et al. 2009) used for the QM calculations in this work requires *all* processors to have access to shared memory, which may not be the case on all computer clusters¹.

Instead, in our implementation of the modified CrystalPredictor algorithm, a grid of reference LAM points that completely covers the required conformational space is evaluated and stored in the LAM/QM *before* any structure generation is carried out. These are then used to estimate the intramolecular energy, the molecular conformation and the partial derivatives of the lattice energy as explicit functions of the flexible degrees of freedom during the course of the search and lattice energy minimisation. A

¹ Note that one of the strengths of CrystalPredictor is that it may be executed over a network of processors which do not have shared memory.

key computational benefit to this approach is that the LAM/QM databases can be re-used in the subsequent CrystalOptimizer refinement¹.

The generation of LAMs for a reference conformation requires the analytical evaluation of the intramolecular Hessian matrix (second-order derivatives of the intramolecular energy with respect to the intramolecular degrees of freedom). These calculations are more computationally demanding than the constrained geometry optimisations necessary to construct the Hermitian intramolecular energy grid used in the original CrystalPredictor. However, as we have seen in Chapter 4, LAMs are assumed to be valid within a hypercube of conformational space, which implies that one LAM can capture a conformational region that is represented by several grid points (see Figure 5.4 for a two-dimensional illustration). For example, in the case where the range of LAM validity (defined as distance from the reference conformation) is equal to the grid spacing (here 20°)², one LAM calculation can replace up to 9 grid evaluations when two flexible torsions are considered (Figure 5.4). Overall, only 9 (3×3) LAM points are required to capture the entire conformational region of the two torsions shown in Figure 5.4, compared to 49 (7×7) grid points required to construct the Hermite interpolant. This benefit is further improved when considering grids of higher flexibility. For instance, if considering three torsions in the previous example, 27 ($3 \times 3 \times 3$) LAM evaluations are sufficient to capture the entire conformational region instead of 343 ($7 \times 7 \times 7$) grid points. Therefore, despite the fact that more expensive QM calculations are necessary to construct the LAM/QM databases, fewer calculations typically have to be performed when compared to constructing the intramolecular energy grid in the original CrystalPredictor. The resulting computational cost is, thus, comparable, despite the fact that the use of LAMs allows a more accurate treatment of flexibility and hence a more reliable evaluation of the lattice energy. The computational advantage of LAMs becomes very significant as one moves to molecules with large number of flexible degrees of freedom, of the kind of interest in this work.

The local approximate models are used to *approximate* the intramolecular energy and the required partial derivatives. The intramolecular energy (equation 3.9) is a second-order function, whilst the gradient of the intramolecular energy with respect to

¹ Provided the refinement is performed with the same or higher level of flexibility. Refer to section 3.6 for more details.

² This is usually the case as the grid spacing is typically taken to be 20° . In our experience, the LAM validity of $\pm 20^\circ$ of the reference conformation for torsions provided sufficiently accurate results during the search stage.

the flexible torsion and the estimate of the rigid degrees of freedom are first-order (equations 4.5 and 3.7 respectively). This indicates that the value of the estimated intramolecular energy tends to be more accurate than the estimated values of both of its derivatives and of the rigid degrees of freedom, especially as one moves away from the reference point. The main point to be made is that the partial derivative of the rigid degrees of freedom with respect to the flexible degrees of freedom (equation 5.26) is assumed to be *constant* (zeroth-order) within the region of the LAM validity. As a result, the accuracy of this derivative is expected to be most adversely affected by moving away from the reference point at which it has been originally computed. Consequently, considerations about the accuracy of not just the intramolecular energy, but also of the relevant partial derivatives, should be taken in account when specifying the corresponding LAM validity region. In our experience, for torsional angles this validity range should not exceed 30° .

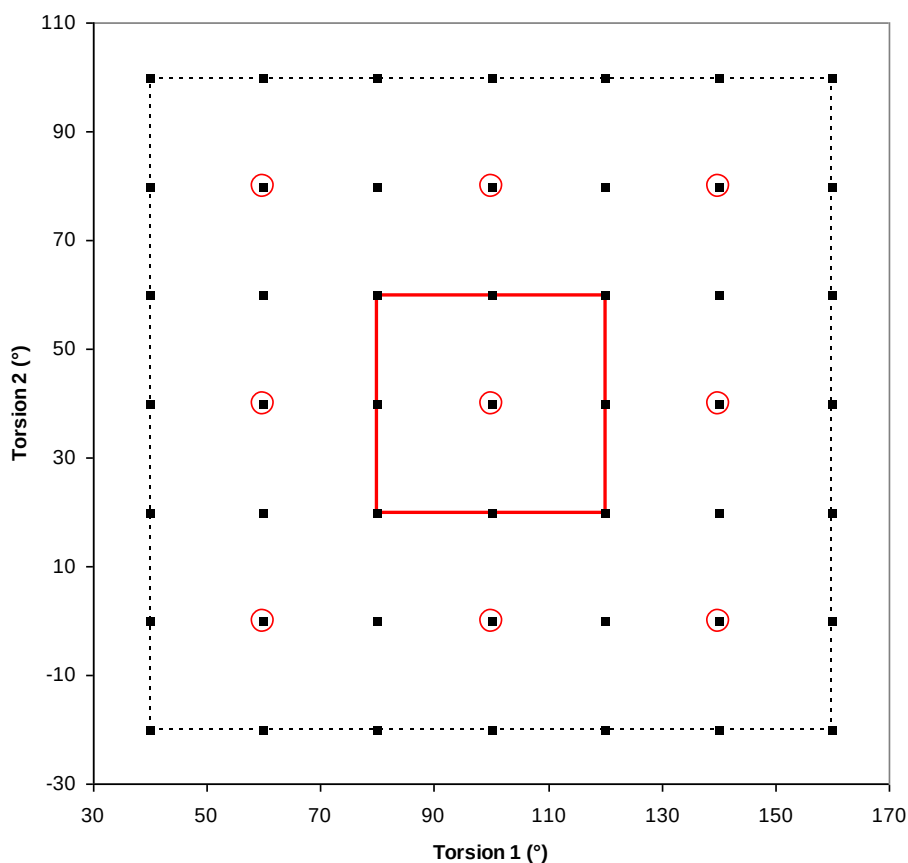


Figure 5.4 Location of Hermite interpolant grid points (black squares) and LAM points (open circles) necessary to capture the conformational space of two hypothetical torsions (defined by the boundaries of the dashed square). The conformational hypercube (defined by the boundaries of the red square) captured by the evaluation of one LAM at the reference conformation indicated by the open circle at the centre of the diagram is also shown. For this illustrative example, LAM validity (defined as the distance from the reference conformation) is taken to be equal to the grid spacing.

5.3.4. Conformationally-dependent electrostatic models

The description of the electrostatic interactions for flexible molecules in CrystalPredictor is based on atomic charges¹. As shown in section 3.4, the electrostatic field around the molecule is a function of molecular conformation. Therefore, it is unlikely that a single set of values for the charges can provide a sufficiently accurate representation of the electrostatic field for all the conformations that need to be sampled during the search (and lattice energy minimisation). Hence, the atomic charges also need to be considered as functions of the flexible torsional angles.

The QM calculations used to construct a set of local approximate models for the intramolecular energy and molecular geometry also provide a corresponding set of atomic charges. These discrete charges are then used without modification to represent the electrostatic interactions within the region of validity of LAM (Figure 5.4). Although this zeroth-order approximation leads to some loss of accuracy in the lattice energy, in our experience this is usually acceptable within the range of validity of the LAMs.

Alternatively, the atomic charges can be approximated as continuous functions of the torsional angles using the multi-dimensional Hermite interpolants of the type used in the original CrystalPredictor. Since the partial derivatives of the atomic charges with respect to the flexible torsional angles are not readily obtainable from the quantum mechanical calculations, the Hermite interpolants need to be obtained using only the values of the functions being interpolated. To compensate for the missing partial derivatives, additional QM electrostatic field points can be evaluated within each grid interval. However, this will increase the computational cost and, in our experience, only marginally improve the results in most cases.

¹ In the case of rigid molecules, satellite charges can be used for a better representation of the electrostatic field.

5.4. Illustrative example

The search algorithm needs to be computationally efficient in order to cover the multidimensional search space within practically feasible amounts of computation, but also sufficiently accurate to achieve reasonably good ranking and reproduction of the lowest energy structures. These requirements are imperative for the overall success of any given crystal structure prediction methodology. If the low-energy structures are not ranked sufficiently low by the algorithm, they may be omitted from subsequent refinement with more accurate local lattice energy minimisation models (such as CrystalOptimizer described in Chapter 4) and therefore missed, unless, of course, such refinement is applied to many more structures, which adds significantly to the computational burden. The impact of the search methodology (algorithms) and computational strategy on structure reproduction and rank is discussed in this section in the context of a model system.

5.4.1. Model system ROY

As an illustrative example we consider 5-methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile (Figure 5.5), also known as ROY for its red, orange and yellow polymorphs. There are ten known single-component polymorphs of ROY, and the structures of seven of these (Y, YT04, YN, R, ON, OP and ORP) have already been determined experimentally (Yu et al. 2000, Chen et al. 2005). All of the known ROY polymorphs contain an intramolecular N-H $\cdots$ O=N hydrogen bonds which essentially locks the position of the NO₂ group and prevents it from rotating significantly. The conformation of the molecule in the polymorphs differs significantly in the two torsions connecting the aromatic rings (Figure 5.5). The difference in these torsions has also been linked to the colour shift exhibited by the polymorphs (Yu 2002). A description of a detailed crystal structure prediction study for the seven polymorphs of this model system is in preparation (Vasileiadis et al. 2011). In the sections that follow, several different search strategies are presented and discussed, but the subsequent refinement stage using CrystalOptimizer is not considered.

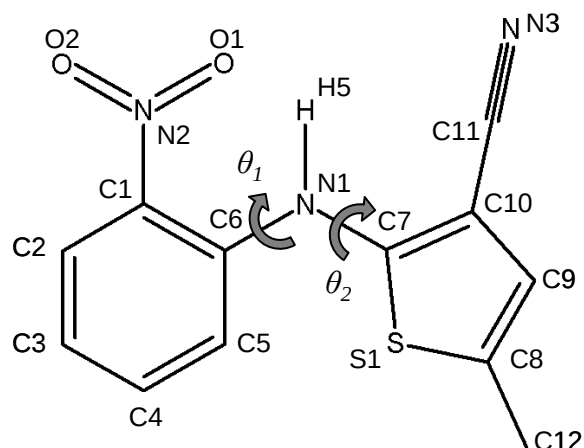


Figure 5.5 Molecular diagram, atom labels and the definition of the main two torsional angles in the molecule of ROY (5-methy-2-[(2-nitrophenyl) amino]-3-thiophenecarbonitrile).

5.4.2. Computational model and search strategies

Despite significant conformational differences, the intramolecular energies of all the ROY polymorphs are within 5 kJ mol⁻¹ of the global gas-phase minimum at the B3LYP/6-31G(d,p) level of theory (Figure 5.6). Overall, a large and continuous conformational region is energetically accessible by the molecule. As a result, the flexibility in the two main torsions has to be explicitly considered during the search. The calculations used to construct Figure 5.6 showed that the accessible range of torsion θ_1 (C5-C6-N1-C7) could be restricted to [-80.0, +80.0]. Similarly, the full rotation around torsion θ_2 (C6-N1-C7-C10) could be reduced by half to [0.0, 180.0] due to the symmetry of the molecule¹.

Three distinct cases, each employing a different computational strategy to capture the effects of molecular flexibility during the search were investigated. In Cases 1 and 2, the original CrystalPredictor is used, while the new version is used in Case 3. The three cases are described further below.

Case 1 The molecule was treated as a set of three rigid fragments connected by the two flexible torsions θ_1 and θ_2 . The molecular fragments were defined intuitively according to the atomic connectivity (Figure 5.7a). The fragment geometries were determined from the gas-phase minimum conformation computed at B3LYP/6-31G(d,p) level of theory and kept fixed during the search. The intermolecular

¹ Refer to Chapter 6 for a detailed description and explanation of conformational analyses.

electrostatic interactions were modelled using conformationally invariant atomic charges that were derived from the same gas-phase calculation. The intramolecular energy as a function of the two flexible torsions was computed on a 10×9 grid using 20° grid spacing for each torsion (top of Figure 5.6). At each grid point the deformation energy (and its derivatives with respect to the flexible torsions) was obtained at the solution of the isolated-molecule constrained geometry optimisation at B3LYP/6-31G(d,p) level of theory.

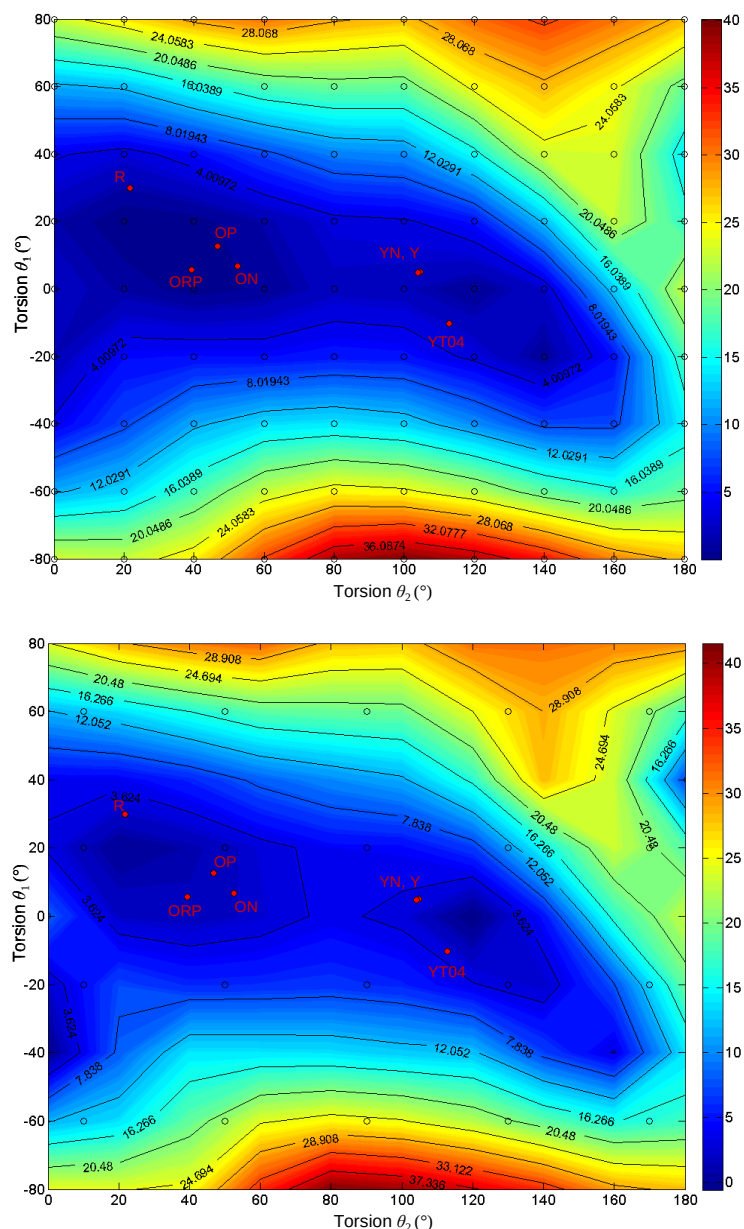


Figure 5.6 Intramolecular energy (kJ mol^{-1}) surfaces as a function of two main torsional angles of ROY (refer to Figure 5.5 for torsion definition) computed from a grid of QM calculations (top, Cases 1 and 2) and estimated using local approximate models (bottom, Case 3). Open circles indicate reference grid (top) or LAM (bottom) points at which explicit QM calculations (at the B3LYP/6-31G(d,p) level of theory) were performed. The location of the known polymorphs on the intramolecular energy surfaces is also shown as red dots.

Case 2 Identical to the above, except that the molecular fragments were defined in a different manner. The hydrogen atom of the N-H was grouped with the fragment containing the NO₂ group (Figure 5.7b) to maintain the geometry of the intramolecular hydrogen bond throughout the search. The intramolecular energy grid from Case 1 was re-used for the search. This is the computational strategy used as part of the crystal structure prediction methodology reported in (Vasileiadis et al. 2011).

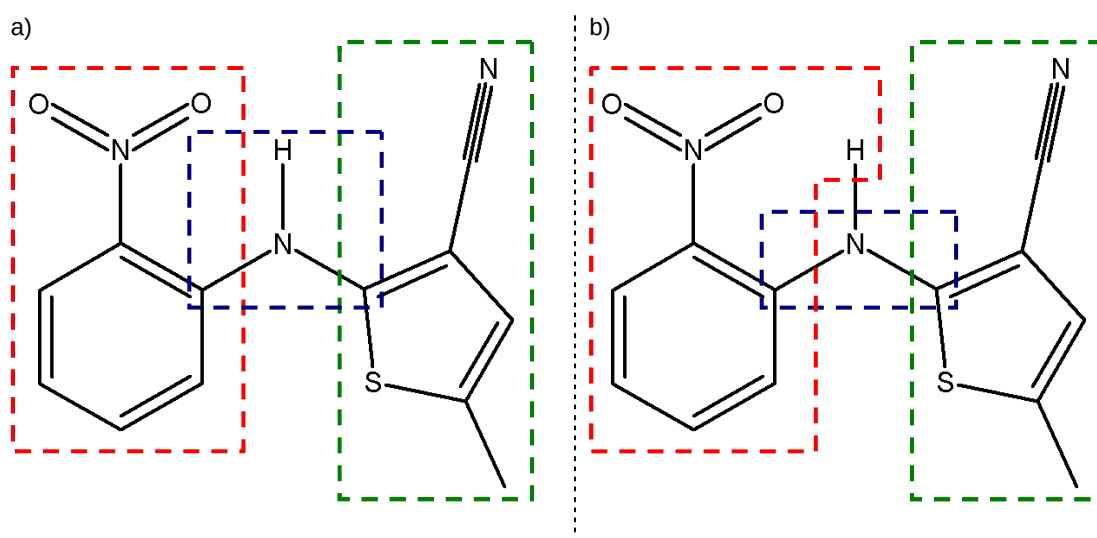


Figure 5.7 Two alternative definitions of rigid fragments in the molecule of ROY. a) The middle (blue) fragment defined as containing atoms C6, N1, C7 and H5; b) The middle (blue) fragment is defined without the H5 hydrogen (i.e. C6, N1, C7), whilst H5 is included in the definition of the red fragment.

Case 3 The modified version of CrystalPredictor was used for the search. No molecular fragments had to be defined as the calculations were performed on the whole molecule. The intramolecular energy and the molecular geometry were calculated as a function of the flexible torsion angles using local approximate models. A 5×4 grid of reference LAM points was constructed at the B3LYP/6-31G(d,p) level of theory assuming LAMs are valid for changes in torsional angles of up to 20° from the reference conformations (bottom of Figure 5.6). The electrostatic interactions were modelled using atomic charges derived from the conformation of the closest LAM reference point.

As has already been noted (cf. section 4.4.4), the QM calculations used to construct the LAMs are generally more expensive (due to the calculation of the intramolecular energy Hessian) than those necessary to create the Hermite interpolation grid for Cases

1 and 2. However, only 20 LAM points were required to completely cover the conformational range and quantitatively reproduce the intramolecular energy surface compared to 90 grid points necessary to construct the Hermite interpolation grid used in Cases 1 and 2 (Figure 5.6). As a result, the computational cost of the two models used for the intramolecular energy is comparable.

For the three cases described, the repulsion-dispersion intermolecular energy terms were derived from a standard empirical exp-6 potential (Coombes et al. 1996), except for the sulphur atoms. The intermolecular repulsion-dispersion parameters for the sulphur interactions were those used by Price et al. in the second blind test of crystal structure prediction (CSP 2001) (Motherwell et al. 2002). For each test case, 400 000 structure¹ were generated and minimised in the 64 most common space groups assuming one molecule in the asymmetric unit. The candidate structures were then clustered on the basis of lattice energy, density and inter-atomic distances to eliminate all duplicate structures.

In order to assess the effectiveness of the search in identifying the experimentally known structures, a set of “adjusted” experimental structures was generated by taking each experimental form and performing a local lattice energy minimisation using the same computational model as that used for the search in each test case. Each and every structure generated during the search was then compared against the adjusted experimental structure to check if a match has been achieved.

5.4.3. Results of crystal structure search

Almost all low-energy structures identified in the Case 1 search contain an intermolecular hydrogen bond (Figure 5.8b) in place of the intramolecular hydrogen bond present in all ROY polymorphs (Figure 5.8a). Theoretically, this would be possible if the intramolecular energy penalty were compensated by stronger intermolecular interactions. However, the intramolecular energy cost of breaking the hydrogen bond in ROY is over 20 kJ mol⁻¹ which is unlikely to be overcome by more favourable intermolecular interactions as suggested by the search results. In fact, these are spurious structures that arise as an artefact of fixing the geometry of the fragments

¹ This was deemed to be sufficient as the lowest energy structures were found multiple times and the rate of generation of new structures was considered to be negligible indicating completeness of the search.

in the molecule. As an illustrative example, consider only the red and blue fragments in Figure 5.7a. The geometry of these fragments is fixed and the molecular flexibility in the original CrystalPredictor is modelled by rotating the blue fragment along with the torsion θ_1 (assuming that the red fragment is the base fragment). With this formulation, the H5-N1-C6-C5 torsion (which is responsible for the orientation of the N-H bond) is directly correlated to the values of the torsion θ_1 . In fact, it is an alternative definition of torsion θ_1 . As the value of the torsion θ_1 is changed, the N-H rotates away from the NO₂ group by the same amount as torsion θ_1 , artificially breaking the hydrogen bond in the search structures. To make things even worse, this energy penalty is not captured by the estimated intramolecular energy, since the intramolecular energy grid was constructed using QM calculations in which the intramolecular hydrogen bond was maintained for *all* combinations of the flexible torsions. For these reasons, the experimental structures were not reproduced at all in the search Case 1.

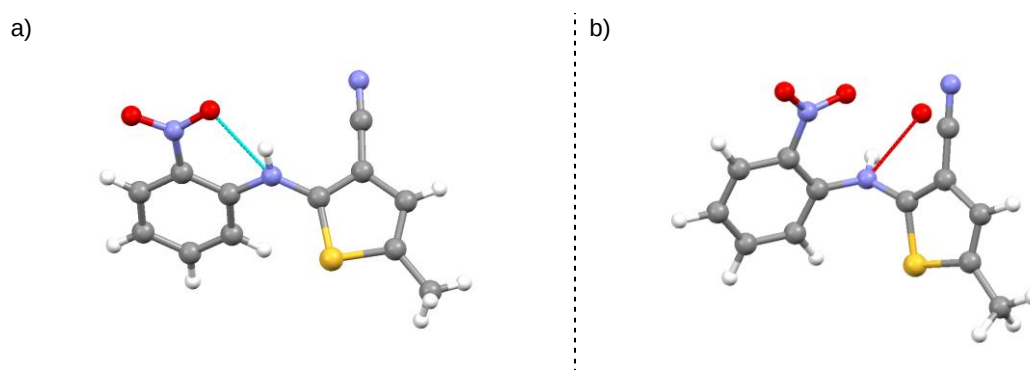


Figure 5.8 Geometries of low-energy structures obtained by search Case 1 a) expected intermolecular hydrogen bond in the structure of ROY and b) spurious intramolecular hydrogen bond (note that only one oxygen atom of the neighbouring ROY molecule is shown here).

The above limitation can be overcome by explicitly fixing the geometry of the intramolecular hydrogen bond when defining the fragments as was done in search Case 2 (Figure 5.7b) (Vasileiadis et al. 2011). This provides a more realistic model for the molecular geometry of ROY and the intramolecular energy could be estimated more reliably using the existing energy grid. As a result, it was possible to identify minima that are close to all of the experimental structures on the lattice energy surface (Figure 5.9a). Nevertheless, since the NH hydrogen was defined relative to the fragment containing the NO₂ group (Figure 5.7b), the N-H bond length was not reproduced accurately. Keeping the geometry of the fragments fixed throughout the search introduced further errors in the reproduction of molecular conformations and structures (Figure 5.10 and Table 5.1). Due to miscalculation of the atomic positions, along with

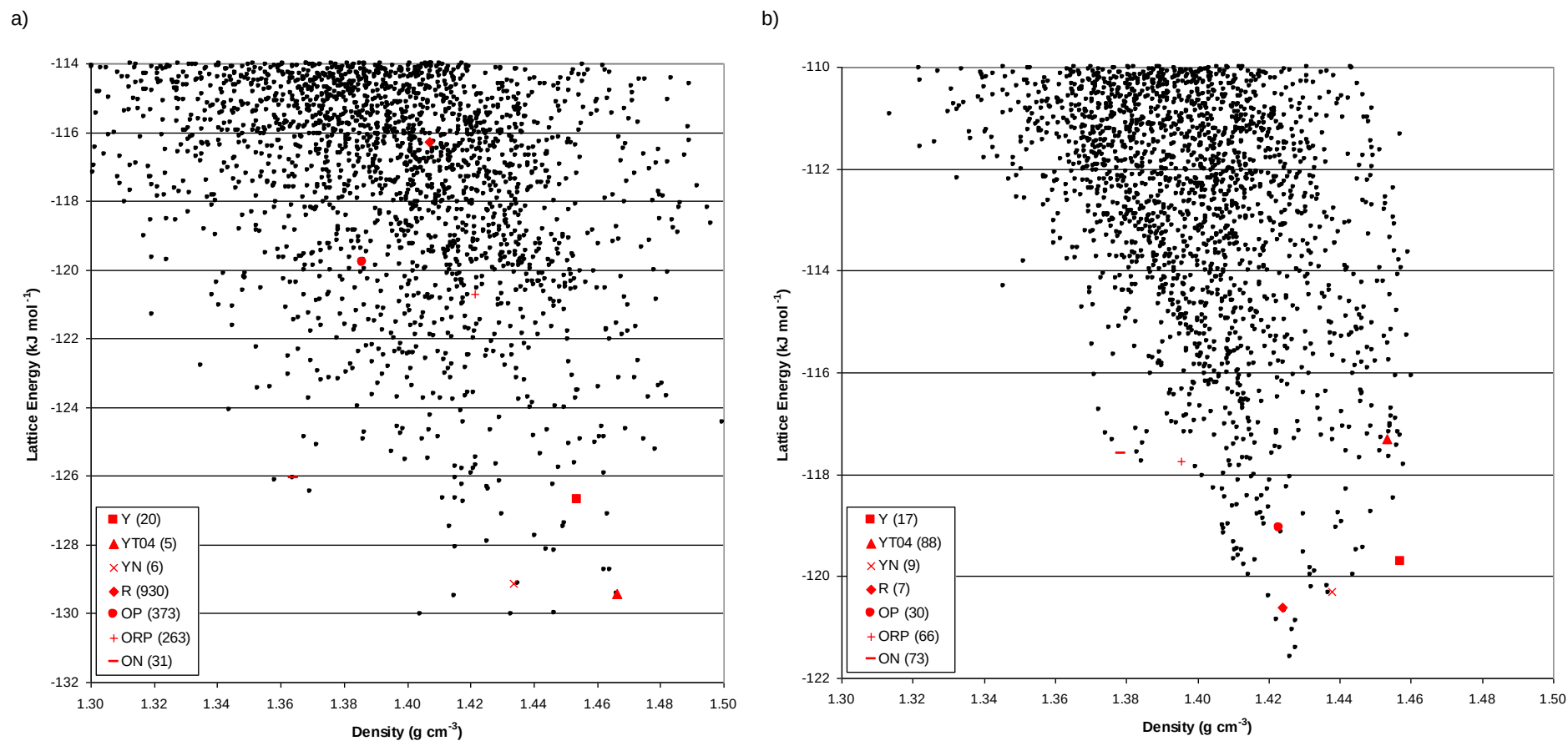


Figure 5.9 Lattice energy landscapes produced for search a) Case 2 and b) Case 3. The rank of the known experimental forms when minimised with the same computational model is shown in brackets.

the use of conformationally invariant charges, the absolute energy and, consequently, the ranking of the experimental forms on the lattice energy surface was not favourable (Figure 5.7a).

For instance, the R polymorph was predicted as the 930th structure above the global minimum (Figure 5.9a). Thus, to be able to identify this polymorph during a crystal structure prediction study, almost 1000 structures would need to be further refined with more accurate models for the lattice energy (Vasileiadis et al. 2011).

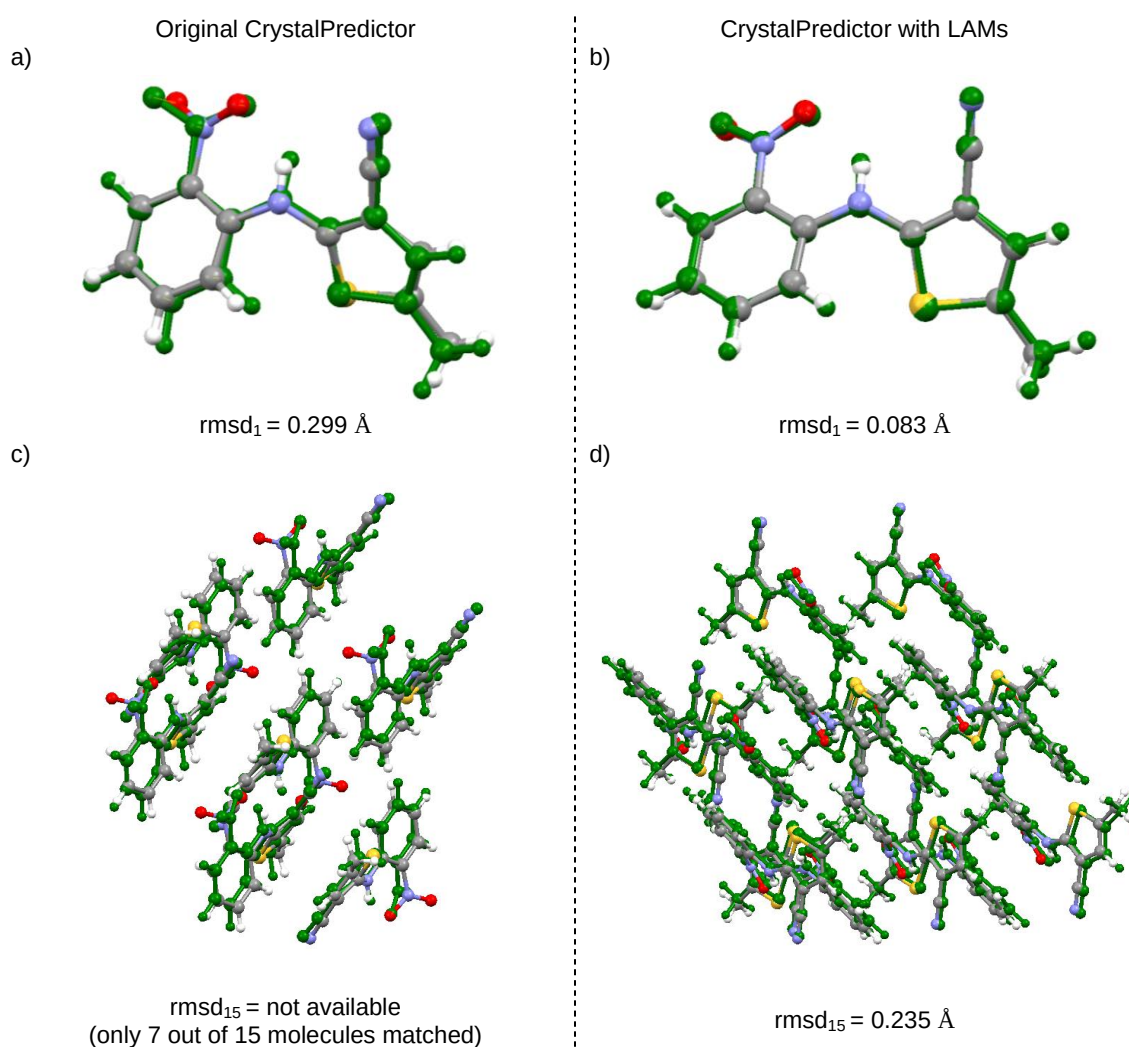


Figure 5.10 Overlay of the experimental R polymorph of ROY with the conformations (coloured by element) with the conformations obtained (green) after a) original CrystalPredictor (Case 2) and b) CrystalPredictor with LAMs. Similarly, overlay of the experimental structure of the R polymorph of ROY (coloured by element) and the predicted crystal structures (green) using c) Original Crystal Predictor and d) CrystalPredictor with LAMs. Root mean square deviations in 1- and 15-molecule coordination spheres are also shown.

In Case 3, the molecular geometry was updated (and the intramolecular energy calculated) as a function of the flexible torsional angles using local approximate models. This ensured that the dependence of the more rigid degrees of freedom (in particular the NH and NO₂ rotations defining the intramolecular hydrogen bond) in the flexible torsion angles was implicitly taken into account throughout the search stage. Furthermore, the electrostatic interactions were modelled using atomic charges derived for the closest LAM conformation, in essence capturing some of the conformational dependence of the electrostatic model. These more realistic models for the molecular conformation and the lattice energy allowed the identification of all experimentally known structures within 3.5 kJ mol⁻¹ of each other and amongst the 100 most stable structures in the search (Figure 5.9b). This is a significant improvement over Case 2, where the experimental structures differed by up to 13.2 kJ mol⁻¹ in stability and were scattered among the 1000 lowest energy structures. As a result, a much smaller number of the search structures produced in Case 3 have to be refined with more accurate and expensive models for the lattice energy, significantly reducing the computational cost of the crystal structure prediction methodology. Furthermore, even without any further refinement, the conformations and crystal structures of all experimental forms are reproduced to a degree of accuracy that would typically be considered acceptable for blind tests with a root mean square deviation in the 15-molecule coordination sphere is less than 0.266 Å for all polymorphs (Figure 5.10 and Table 5.1). This also indicates that the starting structures for the refinement algorithms are better than those produced in Case 2, suggesting the refinement algorithm can converge in fewer iterations as only small further changes to the crystal structures have to be made, further reducing the overall cost of the crystal structure prediction methodology.

5.5. Conclusions

An extension to *CrystalPredictor* (Karamertzanis and Pantelides 2005, Karamertzanis and Pantelides 2007), a crystal structure prediction algorithm for flexible molecules, has been presented in this chapter. In the original algorithm, the conformational flexibility was modelled by splitting molecules into fragments connected by flexible torsional angles. The main limitation of this approach was that the geometry of the fragments is assumed to be fixed. This introduces a degree of

approximation which leads to a miscalculation of the actual position of atoms in the molecule within the crystal lattice (and hence the energy).

Table 5.1 Structure reproduction and rank of ROY polymorphs obtained after search Case 2 and Case 3 respectively. Structures are shown in order of decreasing stability at a temperature of 40° C (Yu et al. 2000, Chen et al. 2005).

	Case 2			Case 3		
Polymorph	Rank	Lattice Energy (kJ mol ⁻¹)	rmsd ₁₅ [*] (Å)	Rank	Lattice Energy (kJ mol ⁻¹)	rmsd ₁₅ [*] (Å)
Y	20	-126.68	0.192	17	-119.70	0.173
YT04	5	-129.44	0.238	88	-117.29	0.186
YN	6	-129.12	0.357	9	-120.30	0.266
R	930	-116.29	(7/15) [‡]	7	-120.61	0.235
OP	373	-119.79	(6/15) [‡]	30	-119.05	0.165
ORP	263	-120.70	0.582	66	-117.74	0.210
ON	31	-126.04	0.725	73	-117.57	0.417

^{*} root mean square deviation in the 15-molecular coordination sphere compared to the experimental structure.

[‡] a match in the 15-molecule coordination sphere could not be established. Number of molecules in common out of the 15 considered is shown instead.

The key aspect of the proposed modification is to model the molecular geometry as an explicit function of the flexible torsional angles using the local approximate models presented (LAMs) in Chapter 3. The LAMs can also be used to estimate the intramolecular energy of the conformations and the relevant partial derivatives during the search and the subsequent lattice energy minimisation. The effects of the molecular conformation on the intermolecular electrostatic interactions are also taken into account as the atomic charges are calculated for every reference conformation used to construct the LAMs.

The limited experience gained with the ROY test case considered in this chapter indicates that the proposed approach is promising. By using more realistic models for the molecular conformation, the experimental structures are reproduced accurately and are predicted to be favourable low-energy minima on the lattice energy surface. This indicates that fewer structures have to be refined with subsequent, more accurate lattice energy minimisation schemes (such as the CrystalOptimizer presented in Chapter 4), resulting in a lower computational cost for the overall crystal structure prediction methodology. Additionally, the candidate structures provided by the modified search

algorithm are expected to be closer to the true minima, indicating that the refinement algorithm can converge in fewer iterations, thereby reducing the computational cost further.

Like any other type of interpolant, the *a priori* construction of the LAM grid can be computationally expensive especially as the intramolecular Hessian has to be evaluated at every grid point. Nevertheless, due to the geometrical properties of the LAM validity region, fewer LAM points are required to cover completely a given conformational region when compared to the Hermite interpolants (using equivalent grid spacing) implemented in the original *CrystalPredictor*. As a result, the computational costs of constructing LAM and Hermite grids are comparable. Another key computational benefit of the modified approach is that the LAM/QM databases constructed for the search can be re-used in subsequent re-minimisation of crystal structures with *CrystalOptimizer* using the same or *higher* degree of molecular flexibility.

Once the interpolants for the intramolecular energy are available, the cost of executing *CrystalPredictor* remains largely unchanged. The current version of the algorithm is slightly slower than the original due to the need to access LAM/QM databases that are stored as persistent computer files (i.e. not accessed from program memory). Nevertheless, lattice energy minimisations are still performed in the order of seconds for each structure considered.

The emphasis in this chapter has been placed on the accurate representation of the molecular geometry and the intramolecular energy as functions of selected torsional angles considered during the search. Atomic charges are calculated at grid points used to construct the LAM/QM database. These are then assumed to be conformationally invariant in the region of their respective LAM validity, or the charge magnitudes can be allowed to be functions of molecular conformations as implemented in the original *CrystalPredictor*. A useful extension to the algorithm would be, however, to allow the inclusion of conformationally-dependent satellite charges for a better reproduction of the electrostatic potential around the molecule.

Chapter 6. Crystal structure prediction of large flexible molecules

There have not been many attempts in the open literature to reproduce crystalline systems with molecular weights exceeding $\sim 200 \text{ g mol}^{-1}$, ~ 30 atoms and with more than just a handful of flexible torsional angles (Schmidt et al. 2010, Bond et al. 2011). In this chapter, the extent to which the algorithmic developments described in Chapter 4 and Chapter 5 can help to treat the larger and more flexible systems which are typical of modern pharmaceuticals with the accuracy of quantum mechanical models is investigated. A general crystal structure prediction methodology for such systems is outlined in section 6.1. This is followed by the detailed description of two examples – a drug candidate (now obsolete) provided by Bristol-Myers Squibb Company and the largest target in the latest blind test of crystal structure prediction. They are illustrated in section 6.2 and 6.3 respectively. Such systems are a significant challenge to current crystal structure prediction methodologies and can serve to highlight the areas requiring immediate attention and possible methodological innovation.

6.1. Crystal Structure Prediction methodology

The search for low-energy crystal structures is a complex multidimensional global optimisation problem. The decision variables are the space group, the unit cell dimensions, and the positions of all atoms in the unit cell. The size of the problem increases dramatically with molecular size and the number of molecules in the asymmetric unit (e.g. cocrystals, salts, and solvates). Furthermore, many hypothetical structures are very similar in stability (Price 2008b). Consequently, the correct ranking of structures requires an accurate representation of all the components of the lattice energy - the intermolecular energy contributions, U^{inter} , dominated by electrostatic interactions, and the energy cost associated with distorting the molecule from its gas-phase conformation, ΔE^{intra} . The accurate calculation of these two components requires elaborate energy models based on quantum mechanical (QM) calculations (as described

in Chapter 4). This is especially demanding for flexible molecules as the conformation adopted by the molecules also greatly affects the intermolecular interaction energy.

In order to keep the computational cost manageable, a multistage methodology is adopted in this work in which the multidimensional search space is firstly restricted to energetically meaningful regions and then efficiently sampled using increasingly accurate and complex models. This is a common approach for flexible molecules in the crystal structure prediction community (Day et al. 2007, Price 2008b, Karamertzanis et al. 2009), which has been successfully used in the blind test of crystal structure prediction (Lommerse et al. 2000, Motherwell et al. 2002, Day et al. 2005a, Day et al. 2009, Bardwell et al. 2011). This crystal structure prediction methodology can be typically split into four steps:

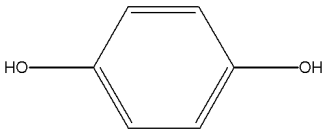
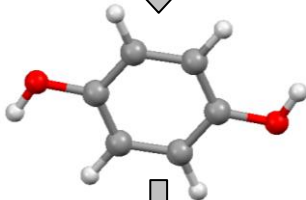
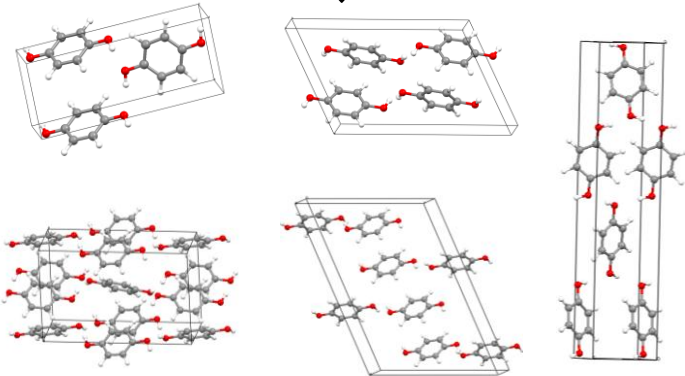
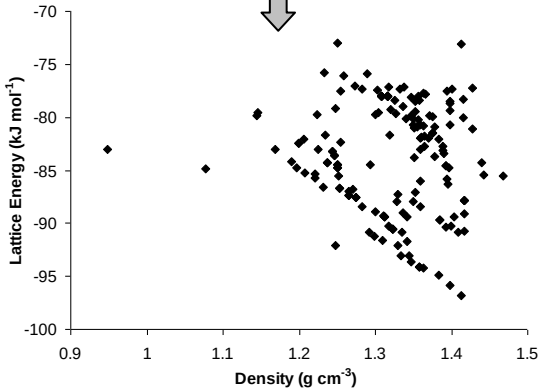
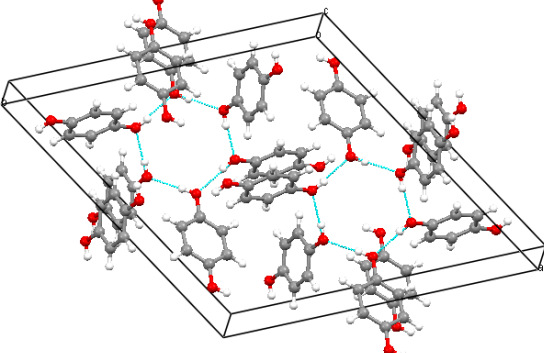
1. selection of computational models (both molecular and interaction potentials) and conformational analysis of the target molecule(s)
2. exhaustive search for all possible packing arrangements in the crystal
3. accurate lattice energy evaluation of the putative crystal structures, also termed “refinement”
4. selection of candidate structures based on additional considerations

An overview of each step in the multistage methodology is summarised in Table 6.1 and described in greater detail in the remainder of this section.

6.1.1. Step 1: Computational model selection and conformational analysis

Prior to performing any analysis, it is imperative to select a suitable computational model such that the intra- and intermolecular energy contributions are well-balanced and can realistically reproduce the deformations of the molecular structure caused by packing forces within the crystal. For the algorithms used in this work it is necessary to select the quantum mechanical level of theory for both the geometry optimisations and charge density calculations and the form of the empirical repulsion-dispersion potential. This can be achieved by taking into account any available experimental information and/or performing calculations on similar molecules found in the CSD. The computational model is usually selected by performing a conformational analysis, as described below.

Table 6.1 Crystal structure prediction methodology adopted in this work.

Input	Molecular connectivity diagram	
Step 1	Computational model selection and conformational analysis to obtain 3D molecular structure(s)	
Step 2	Crystal structure generation of many plausible close-packed molecular arrangements (CrystalPredictor)	
Step 3	Crystal structure refinement using more accurate models for the lattice energy (CrystalOptimizer)	
Step 4	Identification of relevant minima (polymorphs) on the lattice energy surface	

The molecular conformation within the crystal is a balance between the intra- and intermolecular forces. For some, more rigid molecules (such as fused aromatic rings) changes in conformation are not likely. However, for a majority of molecules of practical interest, molecular flexibility can become important. For these molecules the conformational preference could ideally be restricted to one or just a few *distinct* conformations (e.g. global and several local gas-phase minima geometries) that can be treated independently during a rigid-body generation step (Step 2 of Table 6.1). The molecular geometry can then be relaxed during final lattice energy minimisations (Step 3 of Table 6.1), but the molecule will normally remain close to the initial geometry. In practice, especially for larger molecules, a continuum of conformations is usually accessible. For instance, a terminal phenyl ring on an aliphatic chain (as in the case of molecule XX described in section 6.3) can rotate 360° with a maximum energy penalty of less than 2.5 kJ mol^{-1} . In such cases, the rigid body approach, where a very large number of distinct conformations have to be considered independently, will be impractical. As a result, molecular flexibility should be considered explicitly during the structure generation step (Step 2 of Table 6.1). The conformational analysis is then performed with the view of restricting (or simplifying) the extensive search space by identifying the expected ranges of variation of the flexible degrees of freedom.

There are two main considerations during the conformational analysis step: the number of stable (gas-phase) conformations of a particular molecule, and for a given conformation the effect of packing forces on the molecular geometry. Large changes in bond lengths and bond angles typically require much higher energies than can be readily supplied by the crystalline forces (refer to Chapter 4). As a result, these more rigid degrees of freedom are almost always ignored during the conformational analysis. The intramolecular degrees of freedom that are usually most significantly affected by the crystalline environment are torsions around single bonds. Such changes in geometry are often possible at a small cost in the conformational energy and must be considered explicitly. In this work, the exact energy cost for molecular distortions is calculated using isolated-molecule quantum mechanical (QM) scans. The intramolecular energy is computed by constraining the selected flexible torsions to specific values of interest, while allowing the remaining intramolecular degrees of freedom to relax to their gas-phase values. The intramolecular energy surface is then scanned by varying the values of the flexible degrees of freedom. Those conformational regions with intramolecular

energies exceeding $\sim 10 \text{ kJ mol}^{-1}$ are not expected to be observed in the experimental crystal structures and are usually discarded from further consideration.

Alternatively, studies of conformational preferences in small-molecule crystal structures (Brameld et al. 2008) have shown that similar fragments in different molecules often adopt similar conformations. As a result, a statistical search of the Cambridge Structural Database (CSD) for molecules or fragments of molecules that are similar can also highlight the range of likely conformations (Bruno et al. 2004). In this work, CSD analyses are used to complement and validate the QM intramolecular energy scans.

6.1.2. Step 2: Crystal structure generation

The main purpose of the search algorithm (Step 2 of Table 6.1) is to generate good initial starting crystal structures for subsequent lattice energy minimisation. This search is a complex multidimensional problem in which it is necessary to consider many optimisation variables such as the space group, the size and shape of the unit cell as well as the relevant molecular flexibility.

The search algorithm used throughout this work is CrystalPredictor (Karamertzanis and Pantelides 2005, Karamertzanis and Pantelides 2007)¹, which systematically and uniformly samples different space groups, torsion values, unit cell dimensions and position and orientation of both rigid and flexible molecules within the crystal. A more detailed description of the algorithm is provided in section 5.1 and its main characteristics are summarised in Table 6.2.

The search can be performed on either rigid molecules or the flexibility can be considered explicitly. From a practical point of view, rigid-molecule searches are simpler to implement than flexible searches, and preliminary results can be obtained almost immediately. The main assumption in this approach is that the error introduced in the lattice arrangement by imposing the rigid body constraint can be recovered by subsequent refinement of low energy structures where molecular flexibility is allowed during lattice energy minimisation (Step 3 of Table 6.1). However, even for moderately flexible molecules, many (usually tens of) different discrete conformations have to be

¹ Note that the developments described in chapter Chapter 5 were not used since they were implemented at a later stage.

considered explicitly, which may become computationally infeasible. Furthermore, for successful predictions, at least one of the selected conformations has to be sufficiently close to the experimental geometry since large conformational changes are unlikely during subsequent refinement (Step 3 of Table 6.1)¹ (refer to section 6.2 for a practical demonstration of this phenomenon). As a result, the initial conformational analysis (Step 1 of Table 6.1) is fundamental for the overall success of the prediction.

Table 6.2 Algorithm characteristics for crystal structure generation and crystal structure refinement steps of the proposed multistage methodology (Table 6.1).

	Step 2	Step 3
	Crystal Structure Generation	Crystal Structure Refinement
Algorithm	CrystalPredictor	CrystalOptimizer
Flexible degrees of freedom	fixed or optimised	optimised
Rigid degrees of freedom	fixed	computed as function of flexible degrees of freedom
Intramolecular energy	Hermite polynomial on grid determined via QM	“on-the-fly” QM / Local Approximate Models
Intermolecular electrostatic contribution	conformationally-invariant atomic charges	conformationally-dependent multipole moments
Intermolecular repulsion/dispersion contribution	isotropic atom-atom force fields	isotropic atom-atom force fields (also compatible with anisotropic potentials)
Number of structures considered	tens of thousands to millions	tens to hundreds
CPU time per minimisation	seconds	hours to days

Searches in which molecular flexibility is considered explicitly can be carried out in order to reduce the sensitivity of the success of the prediction to the initial choice of conformation(s). For such calculations, it is necessary to construct an intramolecular energy grid from relatively expensive QM calculations before any structure generation can take place. This usually becomes computationally infeasible for grids with more than 4 dimensions (torsions). Furthermore, because more variables have to be sampled by the algorithm than in a rigid-body search, more minimisations have to be performed to make sure the search is complete.

¹ This is particularly true for large and flexible molecules, where even small torsional changes can propagate through the molecule distorting the geometry in such a way that the experimental structure may not be recovered.

Overall the choice between a rigid and flexible search should be made on a case-by-case basis taking in account the precise flexibility and the size of the model system, as well as the computational cost associated with each computational strategy, both at the search and the subsequent refinement stage.

Since CrystalPredictor incorporates a local minimisation procedure, many generated structures converge to the same local minima, within the convergence tolerances. As a result all the structures are clustered on the basis of lattice energy, density and inter-atomic distances where all duplicate structures are eliminated on completion of the search. A set (usually several hundred) of independent low-energy structures is further refined with more accurate models for the lattice energy in Step 3 (Table 6.1). The remaining high energy structures are discarded from further consideration as it is assumed that the probability of these forms occurring experimentally is low.

6.1.3. Step 3: Crystal structure refinement

Despite the reliability of CrystalPredictor in identifying most low-energy minima (including the experimentally determined polymorphs) for crystals of both rigid and flexible molecules, a limitation with this technique is that, in order to be able to search efficiently what is a very large space of possible crystal structures, it has to rely on approximate estimates for the inter- and intramolecular energy contributions. For instance, intermolecular electrostatic contributions are represented using atomic (and sometimes satellite) charges derived from quantum mechanical calculations of the molecule's electrostatic field. Albeit computationally efficient, such representations are not as accurate as those based on multipole moments (Day et al. 2005b). Furthermore, in the case of flexible molecules, intramolecular energy is computed by interpolating the results of quantum mechanical calculations over pre-computed grids of the molecule's flexible torsion angles. The atomic charges may also be expressed as functions of these torsion angles in a similar fashion. Although the interpolants themselves are quite accurate, the distinction between "flexible" and "fixed" degrees of freedom in the molecule introduces a degree of approximation which leads to errors in the actual positions of the atoms in the molecule within the crystal lattice, and therefore the corresponding intermolecular interactions.

Because of the above reasons, the structure reproduction, the absolute energy and, consequently, the ranking of low-energy crystal structures identified by CrystalPredictor is not reliable. The ranking, therefore, has to be corrected by re-minimisation of the lattice energy using more accurate models. In this work, the refinement of crystal structures is performed using the CrystalOptimizer algorithm described in detail in Chapter 4 with its main characteristics outlined in Table 6.2. It is a local minimisation scheme that allows an arbitrary amount of molecular flexibility to be explicitly taken into account for systems of both single and multi-component crystals. Throughout the refinement, the computational cost is kept manageable through the use of local approximate models for the conformationally-dependent intramolecular energy and charge density, as discussed in Chapter 3. These are as accurate as explicit QM calculations but carry a much smaller computational burden (see Chapter 4 for more details). QM calculations are only performed for a limited number of conformations in the course of lattice energy minimisation and are stored in LAM/QM databases for future reuse (refer to Chapter 3 for their description).

At the end of the refinement, the minimised structures are clustered based on their root mean square deviation in the 15-molecule coordination sphere (Chisholm and Motherwell 2005) to remove any crystallographically similar minima.

6.1.4. Step 4: Identification of relevant minima on the lattice energy surface

There are numerous additional criteria based on statistical comparison to structural databases (such as the CSD), the calculation of the free energies, thermal stability, elastic properties and morphologies that can be used to further distinguish between the multitude of lattice energy minima typically predicted by the above methodology (refer to section 2.4.3 for more details). However, at present there are no definite conclusions on their merits and drawbacks.

In this work, the relevant minima are chosen solely on the basis of their lattice energy. The experimental structure is expected to appear within a narrow energy range of the global minimum on the lattice energy surface, where the range of polymorphism is typically taken to be 10 kJ mol⁻¹ (Bernstein 2002, Price 2008b).

6.2. Crystal structure prediction of a candidate active pharmaceutical ingredient

6.2.1. Introduction

The first application of the crystal structure prediction methodology described in this work arises from collaboration with Bristol-Myers Squibb (BMS), a global biopharmaceutical company. The target molecule, referred to as the BMS molecule, is an obsolete drug candidate with a molecular weight of 422.44 g mol⁻¹ and with molecular formula C₂₂H₂₂N₄O₅ (Figure 6.1). With 9 key torsions, it exhibits a high degree of conformational flexibility that has to be modelled explicitly throughout the crystal structure prediction framework. Two neat (single-component) forms have been determined experimentally. However, the structure of just one was initially disclosed and the main challenge was to identify the second, less stable, experimental form, as well as any additional hypothetical polymorphs.

After identifying feasible conformational regions of the molecule, a first attempt to predict the experimental structures was made through a series of rigid-body searches using a large number of distinct molecular conformations. However, once the structure of the second experimental form was revealed, it was not found within these search results. Since then, additional searches, in which a set of flexible torsions were treated explicitly as search and minimisation variables, were performed. These searches did manage to identify both experimental forms as favourable low-energy minima. Although no information on the second experimental structure was used in performing these searches, the results cannot be treated as successful *blind* predictions. Only the conformational analysis and the multiple rigid-body searches described in this section were performed when the structure of the second experimental structure was not known.

Throughout this section, the accuracy of each step in the adopted crystal structure prediction methodology is analysed in terms of the reproduction of the two experimental forms.

6.2.2. Step 1: Computational model selection and conformational analysis

A suitable computational model for the crystal structure prediction study was determined via the lattice energy minimisation of the first experimental structure by explicitly optimising the 9 flexible torsions shown in Figure 6.1. Given the size of the molecule (and atom types), the quantum mechanical level of theory for the geometry optimisations and charge density calculations had to be selected based on a tradeoff between computational cost and accuracy. It was determined that the experimental structure could be reproduced to an acceptable degree of accuracy with a root mean square deviation of the 15-molecule coordination sphere (rmsd_{15}) of 0.256 Å when using the PBE0/6-31G(d,p) level of theory for the intramolecular energy and charge density coupled with a standard isotropic exp-6 potential (Coombes et al. 1996) (Figure 6.2).

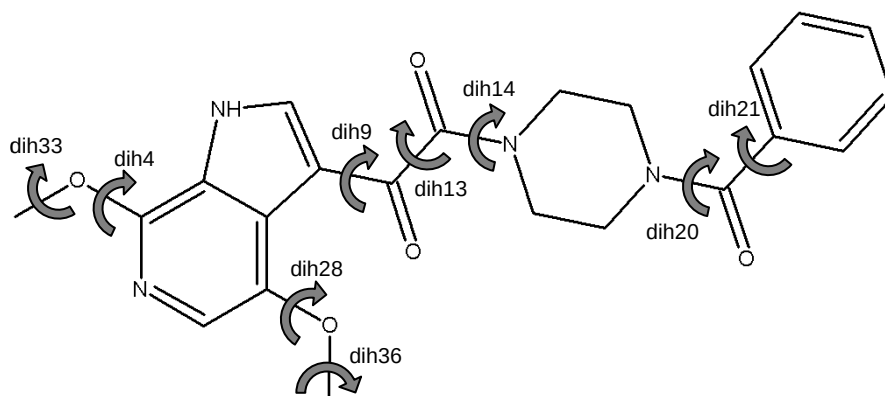


Figure 6.1 Molecular connectivity diagram of the BMS target molecule with the definitions of key flexible degrees of freedom (nine torsions).

A conformational analysis of the molecule was carried out to reduce the number of intramolecular degrees of freedom (torsional angles) that have to be considered when identifying low-energy molecular conformations. In this way, energy calculations showed that the terminal methyl groups showed limited flexibility and were expected to be pointing outwards from the main molecule. Similarly, the Cambridge Structural Database (CSD) was searched for the frequency of the boat/chair conformers occurring

for the $\text{N}_2\text{C}_4\text{H}_8$ ring. 218 crystal structures containing this particular fragment were identified, all of which were in the chair conformation except one structure which was a twisted boat. It was consequently decided to constrain the $\text{N}_2\text{C}_4\text{H}_8$ ring to the chair geometry. The remaining aromatic rings in the system were also assumed to be rigid. As a result, a detailed conformational analysis had to be performed for only 5 torsions (dih9, dih13, dih14, dih20 and dih21 on Figure 6.1).

The conformational analysis for the BMS molecule was performed using two different approaches, but with significant overlap. First, the intramolecular energy was calculated by explicit isolated-molecule quantum mechanical scans at the PBE0/6-31G(d,p) level of theory. In order to reduce the computational cost, the molecule was split into smaller fragments, assuming that the conformational changes in one fragment do not significantly affect the conformation (and energy) of the other fragments. Hydrogen atoms were added to these fragments accordingly to ensure complete connectivity. Model fragments used for QM scans are given in Figure 6.3a. Multiple flexible degrees of freedom in each fragment were scanned by varying them one at time, while keeping the other flexible degrees fixed at the selected reference values. For each perturbed molecular geometry, the intramolecular energy was calculated by allowing the remaining intramolecular degrees of freedom to relax to their gas-phase values.

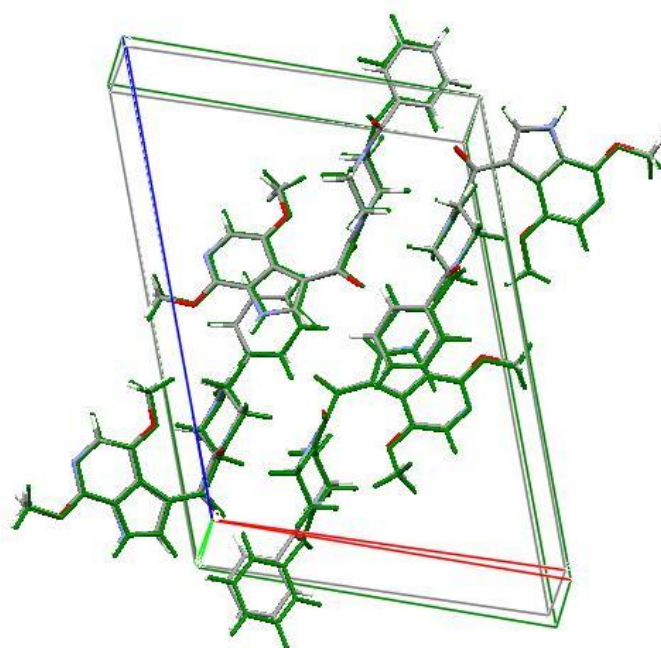


Figure 6.2 Overlay of experimental structure (solid colours) and the local minimum structure (green) obtained by lattice energy minimisation starting from the experimental form and allowing 9 flexible torsions to vary (Figure 6.1). Structure similarity, given by rmsd_{15} , is 0.256 Å.

Additionally, to support the QM energy calculations, the CSD was used to analyse the conformational preferences by comparing fragments of the BMS molecule (Figure 6.3b) with molecules of similar functionality whose crystal structures are present in the structural database. This was done efficiently using available tools such as Mogul or ConQuest (Bruno et al. 2002, Bruno et al. 2004). The relationship between the two types of conformational analyses is illustrated in Figure 6.4 and Figure 6.5. The DFT conformational energy and the CSD observations are generally in good agreement. At values for which the conformational energy penalty is high, few or no experimental occurrences are found in the CSD. This indicates that the crystals with the lowest intramolecular energies are favoured. This is particularly true for torsion dih20, where this information was used to restrict the feasible conformational space of this torsion (Figure 6.4) to just two low-energy regions ($[-80.0^\circ, +80.0^\circ]$ and $[+140.0^\circ, -140.0^\circ]$). In Figure 6.4, it also appears that the intramolecular energy calculation for an isolated fragment of the molecule is in a reasonable quantitative agreement with the intramolecular energy of that fragment based on calculations for the entire molecule. This validates the assumption about partitioning the molecule into fragments to ensure the QM scans remain computationally tractable.

The maximum intramolecular energy penalty for torsion dih21 (less than 20 kJ mol^{-1}) is much lower than for torsion dih20 (around 50 kJ mol^{-1}). As a result, the dih21 energy barrier is not significant enough to prevent some of the experimental crystals exhibiting the high-energy molecular geometries, although these are not as frequent as the more stable conformations (Figure 6.5). The energetic cut-off for the conformational analysis adopted in this work is 20 kJ mol^{-1} for any particular torsional angle. As a result, the complete rotation around the dih21 torsion is considered. However, in practice, this can be restricted to the $[0.0, +180.0]$ range due to the symmetry of the terminal phenyl ring.

A similar detailed conformational analysis was performed for the remaining three important flexible torsions in the BMS molecule. A summary of the accessible conformational space thus obtained is provided in Table 6.3.

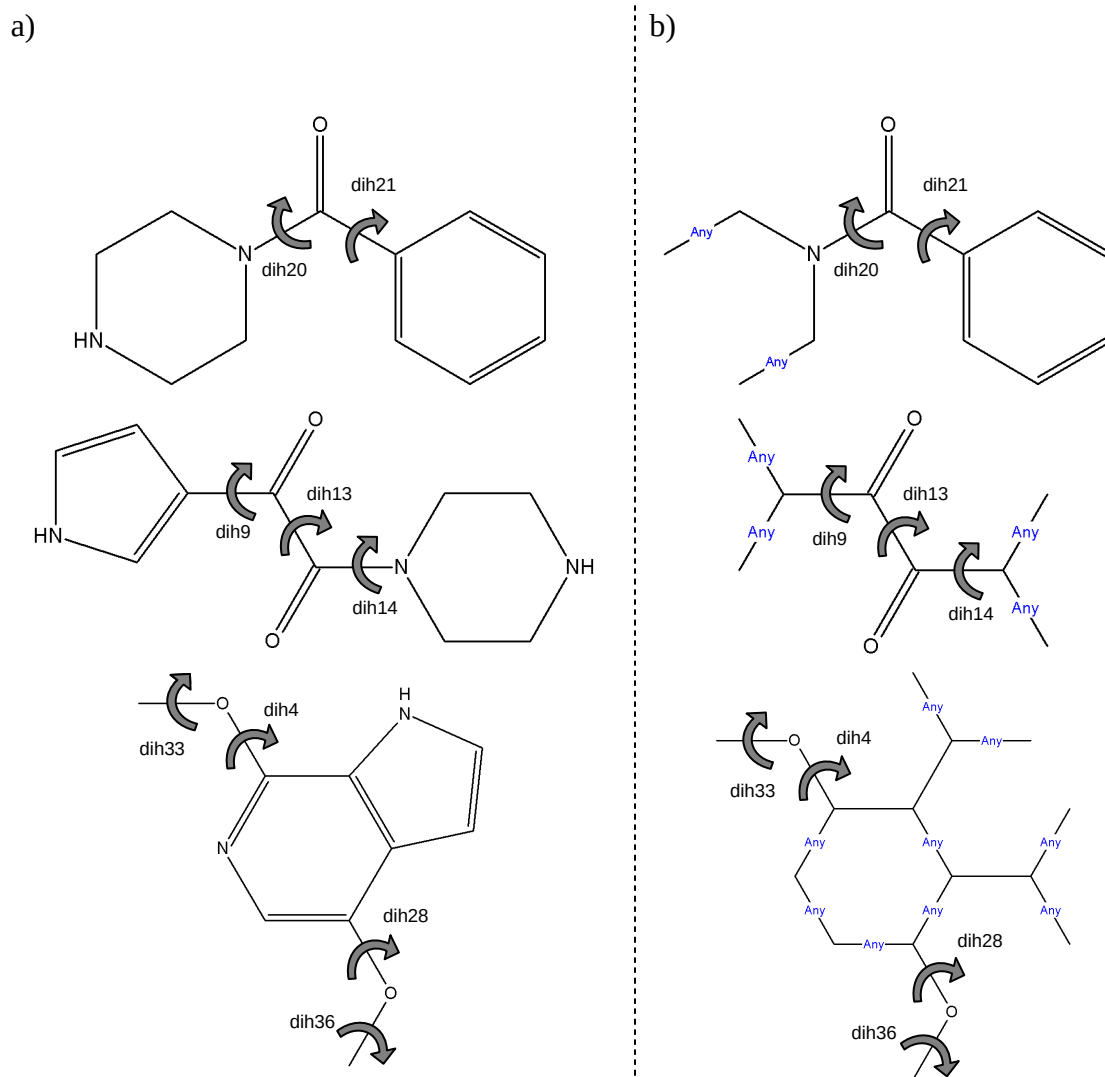


Figure 6.3 Three fragments of the BMS used for a) quantum mechanical scans and b) statistical search of the CSD during the conformational analysis.

The analysis of the BMS molecule confirms the observation made in other studies of conformational preferences in small-molecule crystal structures (Bruno et al. 2004) that similar fragments in different molecules often adopt similar conformations. Furthermore, where the conformational preferences are statistically significant, they are very rarely found to have comparatively high gas-phase energies (Allen et al. 1996). As a result, the CSD analysis and explicit QM calculations for the conformational analysis are complementary methods, but can also be used independently to identify the feasible ranges for the main torsional angles in the molecule. This information can then be used to restrict the conformational space that has to be considered during the crystal structure generation step.

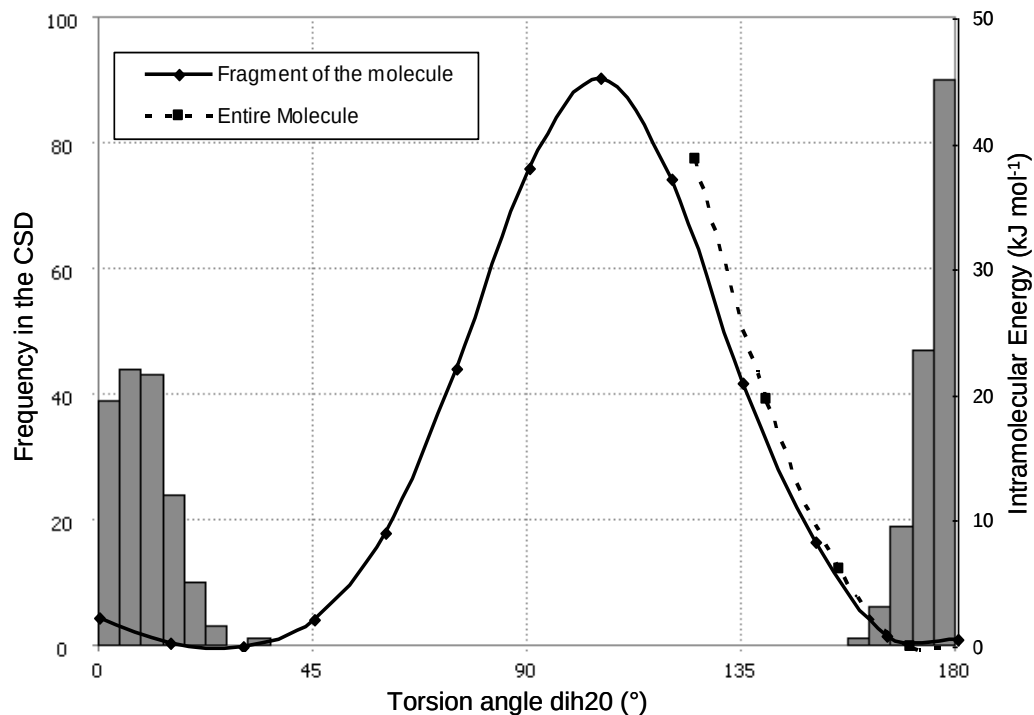


Figure 6.4 Comparison of the CSD MOGUL geometry analysis (bars, left axis) with the gas-phase intramolecular energy scan at the PBE0/6-31G(d,p) level of theory (lines, right axis) for the dih20 torsion in the BMS molecule (Figure 6.1). Results of the gas-phase calculations shown for a fragment of the BMS molecule defined at the top of Figure 6.3a (continuous black line) and the entire molecule (dashed line).

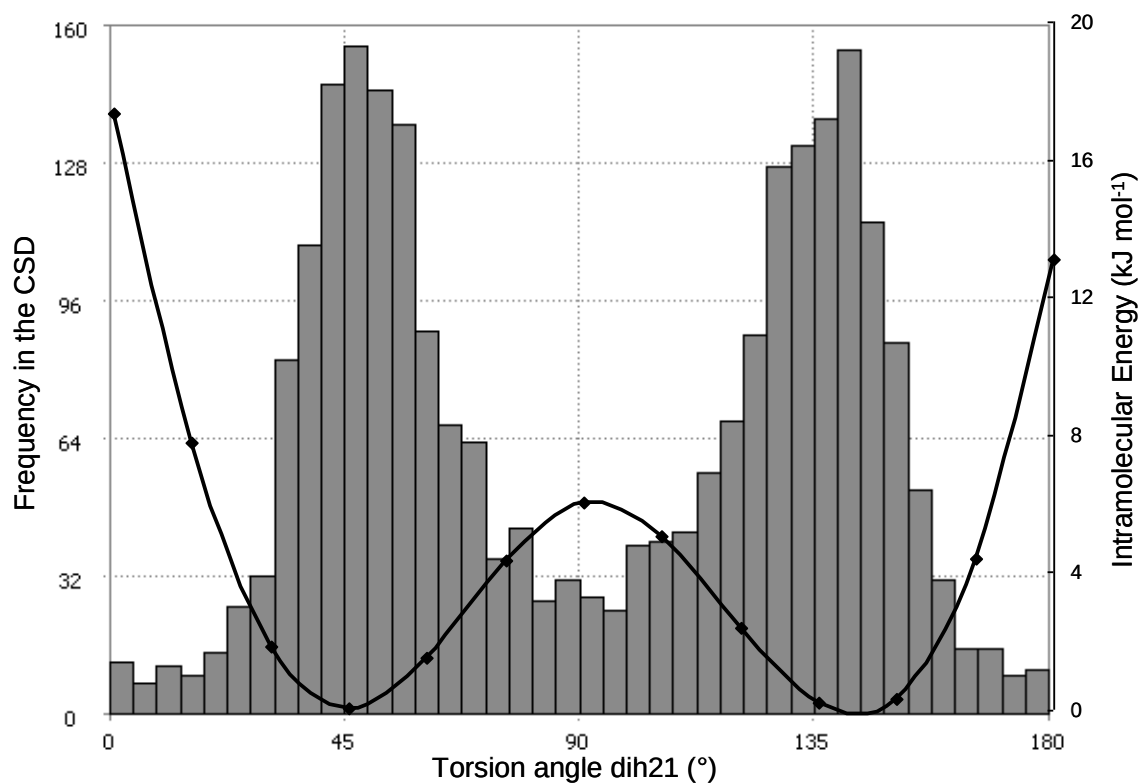


Figure 6.5 Comparison of the CSD MOGUL geometry analysis (bars, left axis) with the gas-phase intramolecular energy scan at the PBE0/6-31G(d,p) level of theory (black line, right axis) performed on a fragment of the BMS molecule defined at the top of Figure 6.3a for the dih21 torsion (Figure 6.1).

Table 6.3 Feasible ranges of torsion values for the BMS molecule. (Refer to Figure 6.1 for torsion definitions). The conformations of the experimental forms are also shown.

#	Torsion name	Feasible Torsion Range(s) (°)	Experimental Conformation 1 (°)	Experimental Conformation 2 (°)
1	dih9	[0.0, +40.0]	+12.8	+28.6
2	dih13	[-150.0, -110.0]	-114.5	-127.7
3	dih14	[+160.0, -160.0]; [-20.0, +20.0]	+9.4	+15.6
4	dih20	[+140.0, -140.0]; [-80.0, +80.0]	-178.1	-19.7
5	dih21	[0.0, +180.0]	+85.3	+126.0

6.2.3. Step 2: Crystal structure generation

To limit the size of the multivariable search space further, it was decided to generate crystal structures in the 13 most statistically significant crystallographic space groups ($P1$, $P1\bar{1}$, $P2_1$, $P2_1/c$, $P2_12_12$, $P2_12_12_1$, $Pna2_1$, $Pca2_1$, $Pbca$, $Pbcn$, $C2/c$, Cc and $C2$) with just one molecule in the asymmetric unit.

Two different strategies for accounting for conformational flexibility were successively attempted for the BMS molecule in the search stage. Initially, a set of rigid-body searches with distinct low-energy geometries was carried out. Molecular flexibility was explicitly allowed in the search in the second case. The details and the relative accuracy of both approaches are described and analysed in this section. The computational cost of the two strategies is compared further in section 6.2.5.

Case 1 – multiple rigid-body searches

In this case, the search for stable packing arrangements consisted of identifying low-energy molecular conformations and performing independent rigid-body searches on the selected conformers. It was assumed that the error introduced into the geometry using the rigid-body approach would be recovered by the subsequent refinement of the search structures in which molecular flexibility would be allowed during lattice energy minimisation.

Following the conformational analysis, 20 distinct molecular geometries were considered within 15 kJ mol^{-1} of the global gas-phase minimum. In order to compute the energies and the coordinates of all the atoms for each distinct geometry, the five torsional angles were constrained to the values specified in Table 6.4 and the remaining intramolecular degrees of freedom were allowed to relax to their gas-phase values. The conformers labelled as g_1 and g_2 correspond to global and local gas-phase minima for the BMS molecule respectively. These were determined by unconstrained geometry optimisations starting from the 18 distinct molecular geometries initially selected for the search.

The molecular geometries of the two experimental structures are also shown in Table 6.4. Note their untypically large intramolecular energies of around 10 kJ mol^{-1} . This indicates that within the crystalline environment the large intramolecular penalty is more than compensated by stronger intermolecular interactions suggesting that molecular flexibility is particularly important for this test case.

Two of the selected rigid molecular models for the search were very close to the experimental geometries (Table 6.4). The maximum deviation of the rigid conformers 2_8 and 2_9 are less than 20° for every main torsional angles considered during the conformational analysis when compared to the first and second experimental forms respectively. However, when the 20 independent rigid-body searches were carried out (generating 2,000,000 structures in total), and the lowest 90 energy structures were refined by allowing molecular flexibility during the more accurate lattice energy minimisations (using CrystalOptimizer in Step 3), neither of the experimental forms was reproduced¹. This indicates that the initial assumption that the error imposed by molecular rigidity in the search stage would be recovered during structure refinement did not hold in this particular case, highlighting the importance of explicit treatment of molecular flexibility even in the early stages of crystal structure prediction methodology.

¹ It is possible, however, that by refining a larger number of search structures using CrystalOptimizer would recover the experimental forms. However, a detailed analysis of the search structures showed that none were particularly close to the experimental forms (as determined by their rmsd_{15}).

Table 6.4 Molecular conformations considered for the preliminary rigid-body search. The molecular conformations of the two minimised experimental forms are also provided (Refer to Figure 6.1 for torsion definitions).

Conformation ID	Values for torsional angles (°)					Intramolecular Energy (kJ mol ⁻¹)
	dih9	dih13	dih14	dih20	dih21	
g_1	+27.3	-132.8	-166.4	+174.2	+143.7	0.00
g_2	+27.5	-138.0	+7.9	-29.2	+142.2	3.59
1_1	+20.0	-130.0	+180.0	+175.0	+140.0	1.83
1_2	+20.0	-130.0	+180.0	-175.0	+40.0	3.56
1_3	+20.0	-130.0	+180.0	+20.0	+50.0	2.49
1_4	+20.0	-130.0	+180.0	+20.0	+30.0	3.73
1_5	+20.0	-130.0	+180.0	-20.0	+150.0	6.03
1_6	+20.0	-130.0	+180.0	-175.0	+140.0	4.21
1_7	+20.0	-130.0	+180.0	0.0	+90.0	9.24
1_8	+20.0	-130.0	+180.0	-180.0	+90.0	9.05
1_9	+20.0	-130.0	+180.0	-20.0	+130.0	4.60
2_1	+20.0	-130.0	0.0	+175.0	+140.0	4.73
2_2	+20.0	-130.0	0.0	-175.0	+40.0	4.64
2_3	+20.0	-130.0	0.0	+20.0	+50.0	4.64
2_4	+20.0	-130.0	0.0	+20.0	+30.0	6.02
2_5	+20.0	-130.0	0.0	-20.0	+150.0	6.65
2_6	+20.0	-130.0	0.0	-175.0	+140.0	7.05
2_7	+20.0	-130.0	0.0	0.0	+90.0	10.80
2_8	+20.0	-130.0	0.0	-180.0	+90.0	11.02
2_9	+20.0	-130.0	0.0	-20.0	+130.0	4.71
experimental 1	+12.8	-114.5	+9.4	-178.1	+85.3	9.11
experimental 2	+28.6	-127.7	+15.6	-19.7	+126.0	10.39

Case 2 – explicit treatment of flexible torsions as search variables

Owing to the initial setback of not finding the experimental structures in the multiple rigid-body searches, the test system was recently revisited¹. Rather than treating each distinct molecular conformation independently, this time, full molecular flexibility was allowed during the search stage. Based on the original conformational analysis, the combination of all feasible ranges of the 5 flexible torsion angles (Table 6.3), resulted in four separate flexible CrystalPredictor searches (Table 6.5) that were carried out generating a total of 800,000 structures for $Z' = 1$ in the 13 most common space groups allowing five main torsional angles to change. A grid of intramolecular values for the flexible torsions was derived by performing QM energy scans on the entire molecule. To reduce the computational cost, the dimensionality of the grid was reduced by dividing the flexible torsions into two groups. The first group consisted of

¹ These calculations were performed after all of the experimental structures were disclosed. Consequently the results can no longer be considered “blind” as was intended by the original study.

torsions dih9, dih13 and dih14. The second group was made of torsions dih20 and dih21. The deformation energy for the first group was computed on two 3×3×3 grids and for the second torsion group on a 5×10 and 9×10 grid using a 20° grid spacing. At each grid point the deformation energy and its derivatives with respect to the flexible torsions are obtained at the solution of the isolated-molecule constrained geometry optimisation at the PBE0/6-31G(d,p) level of theory. The intramolecular energy of the BMS molecule was then approximated as the sum of the deformation energies of the two torsion groups, assuming that there are not significant interactions between the two parts. For the search in each conformational region, the intermolecular electrostatic interactions were modelled using the atomic charges that were derived from the PBE0/6-31G(d,p) electrostatic potential of the PBE0/6-31G(d,p) conformational minimum of the whole region. The same calculation provided the coordinates of the starting molecular geometry. All other intermolecular energy terms were derived from an empirical exp-6 potential (Coombes et al. 1996).

Table 6.5 Search definition and ranges of torsional angles considered explicitly during the (Refer to Figure 6.1 for torsion definitions).

Search ID	dih9 (°)	dih13 (°)	dih14 (°)	dih20 (°)	dih21 (°)
Search_1	[0.0, +40.0]	[-150.0, -110.0]	[+160.0, -160.0]	[140.0, -140.0]	[0.0, +180.0]
Search_2	[0.0, +40.0]	[-150.0, -110.0]	[+160.0, -160.0]	[-80.0, +80.0]	[0.0, +180.0]
Search_3	[0.0, +40.0]	[-150.0, -110.0]	[-20.0, +20.0]	[140.0, -140.0]	[0.0, +180.0]
Search_4	[0.0, +40.0]	[-150.0, -110.0]	[-20.0, +20.0]	[-80.0, +80.0]	[0.0, +180.0]

The clustered lattice energy surface for the four flexible searches is summarised in Figure 6.6. Each point on the graph represents a hypothetical crystal structure, some of which may be the experimental forms. There are a total of 196 structures within 20 kJ mol⁻¹ of the global minimum. The explicit treatment of flexibility during the search led to the identification of the experimental form 2 as the global minimum with an rmsd₁₅ of 0.442 Å. Experimental form 1 was also reproduced well during the search (rmsd₁₅ = 0.377 Å), but ranked 64th, 15.47 kJ mol⁻¹ above the global minimum. This is a significant improvement over the original rigid-body searches which failed to identify the experimental forms despite the initial conformations being relatively close to the experimental ones.

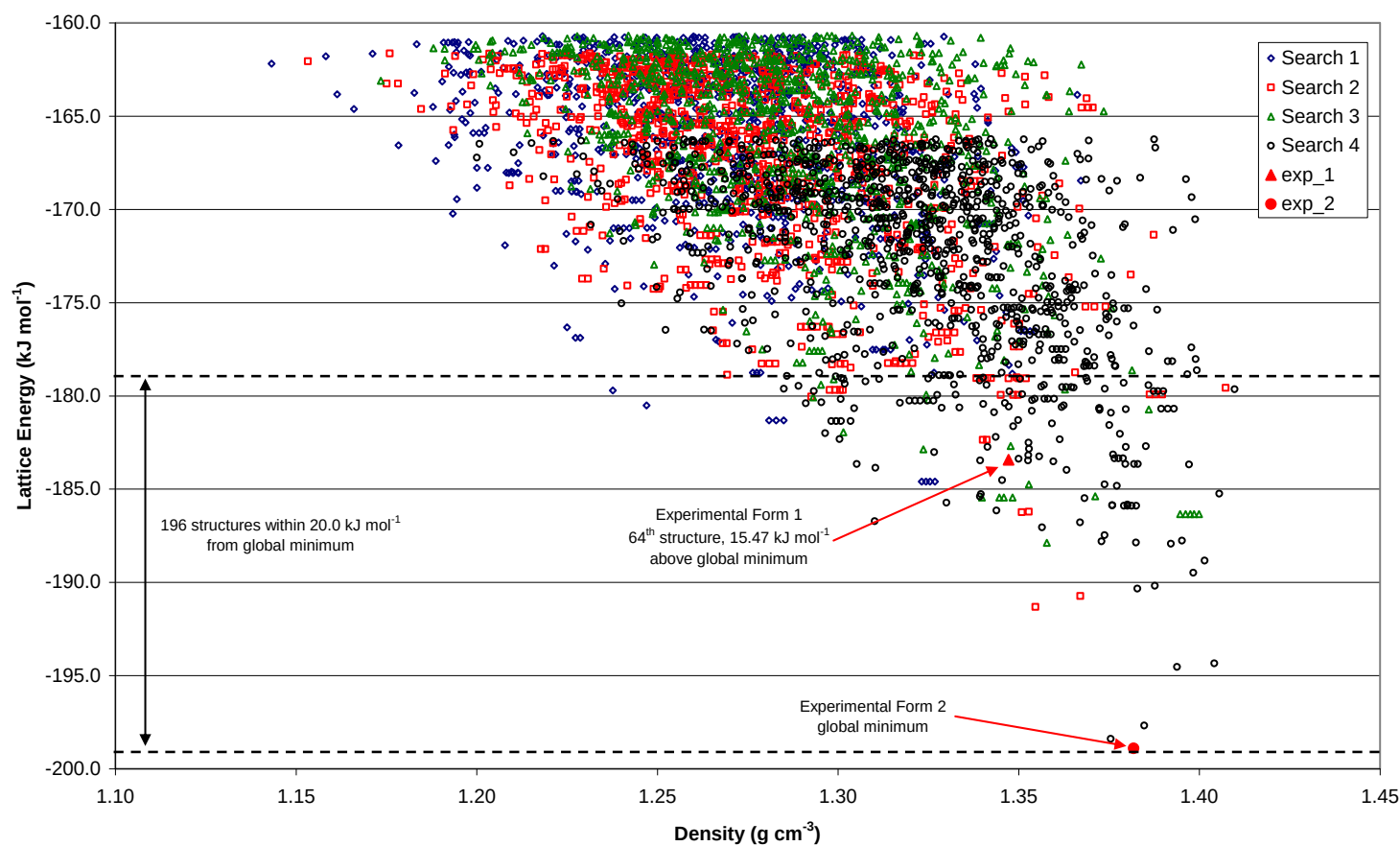


Figure 6.6 Lattice energy versus density plot for the generated crystal structures of the BMS molecule. Results shown for the four independent flexible searches outlined in Table 6.5. The most stable structure is that with the lowest lattice energy. The minima obtained when the experimental polymorphs are minimised with the same computational model are also shown.

6.2.4. Steps 3 & 4: Crystal structure refinement and relative lattice energies

As explained previously, the ability to search the very large space of possible crystal structures in Step 2 relies on simplified, relatively inexpensive models for the intermolecular and intramolecular energy contributions. As a result, all 196 structures within 20 kJ mol⁻¹ from the global minimum generated at the search stage (in which molecular flexibility was explicitly allowed) were refined with more realistic models for the lattice energy.

The molecular geometry and the lattice parameters were simultaneously re-optimised using CrystalOptimizer by allowing 9 flexible torsions to vary explicitly (Figure 6.1). A linear update to the multipole model (equation 3.13) was used for the four torsions containing the nitrogen atom. During lattice energy minimisation, the conformationally dependent intramolecular energy and charge density were modelled using LAMs computed from quantum mechanical calculations at the PBE0/6-31G(d,p) level of theory. The repulsion-dispersion interactions were computed using an empirical isotropic exp-6 potential (Coombes et al. 1996). The minimised structures were subsequently clustered based on their root mean square deviation of the 15-molecule coordination sphere, rmsd₁₅ (Chisholm and Motherwell 2005). Two structures were considered to be crystallographically similar if their rmsd₁₅ was below 0.50 Å.

The lattice energy landscape after the CrystalOptimizer refinement stage is depicted in Figure 6.7. Only 196 structures from the search stage were refined since the known experimental structures were known and already found within that list. The refinement of such a limited set of structures is usually not sufficient for blind predictions where structure re-ranking may be significant.

The two known experimental structures are both predicted as low-energy minima ranking 2nd and 8th, 4.89 kJ mol⁻¹ and 10.60 kJ mol⁻¹ above the global minimum respectively. The relative stability of the experimental structures has changed from the search stage. The second experimental form is now predicted to be 5.71 kJ mol⁻¹ less stable than the first experimental structure, which is in qualitative agreement with the experimental information.

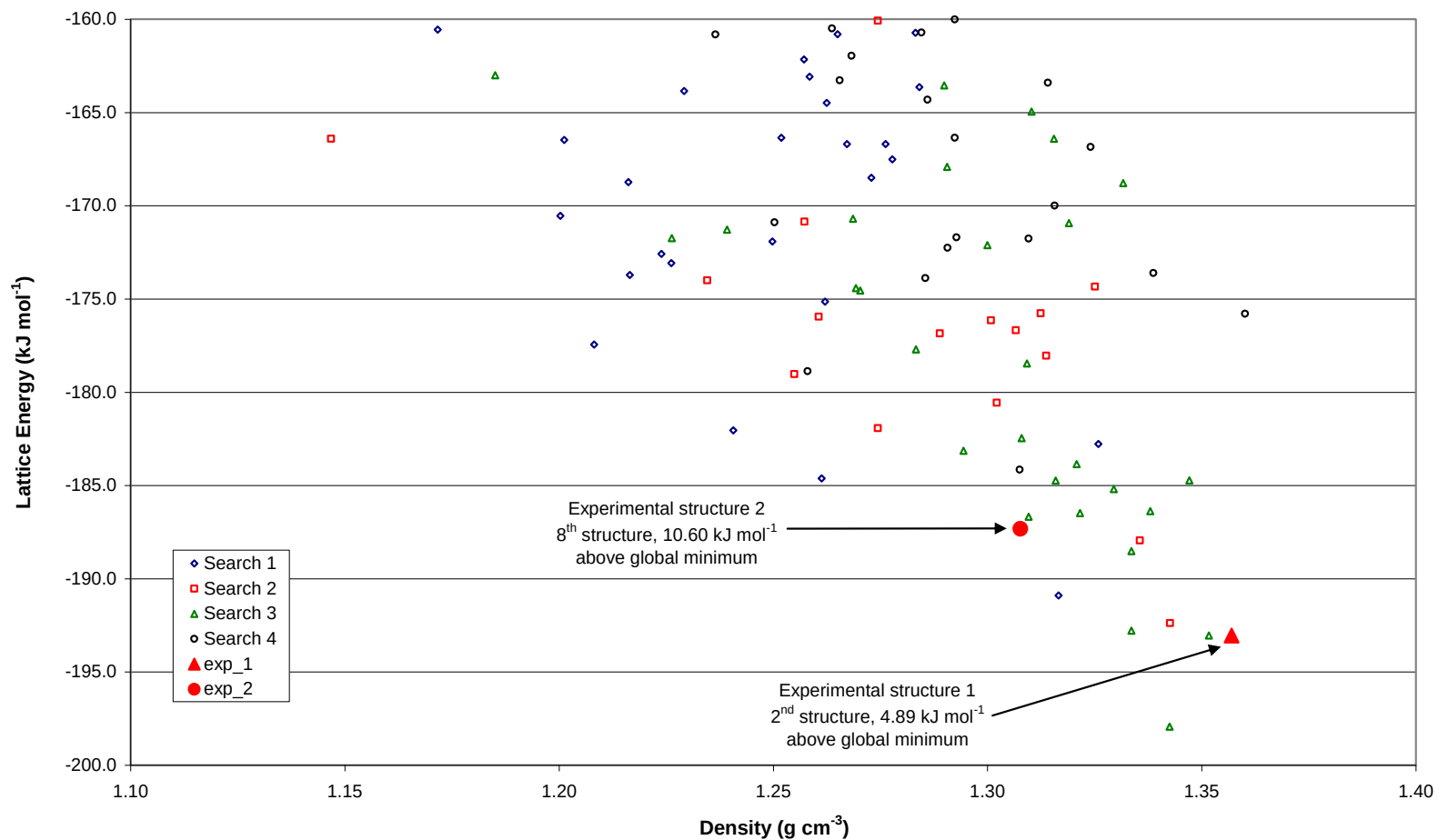


Figure 6.7 Final lattice energy landscape produced after refinement of 196 structures from the search using CrystalOptimizer algorithm and clustering. 126 structures remain as independent minima. The minima obtained when the experimental polymorphs are minimised with the same computational model are also shown.

Overlays of the predictions with the two experimental structures are given in Figure 6.8 along with the predicted and observed unit cell parameters in Table 6.6. The structures are reproduced accurately and would have been considered as successful predictions in a blind test (based on geometrical criteria only). There are also a number of more stable hypothetical structures with varying conformations and intermolecular contacts provided in Appendix F.

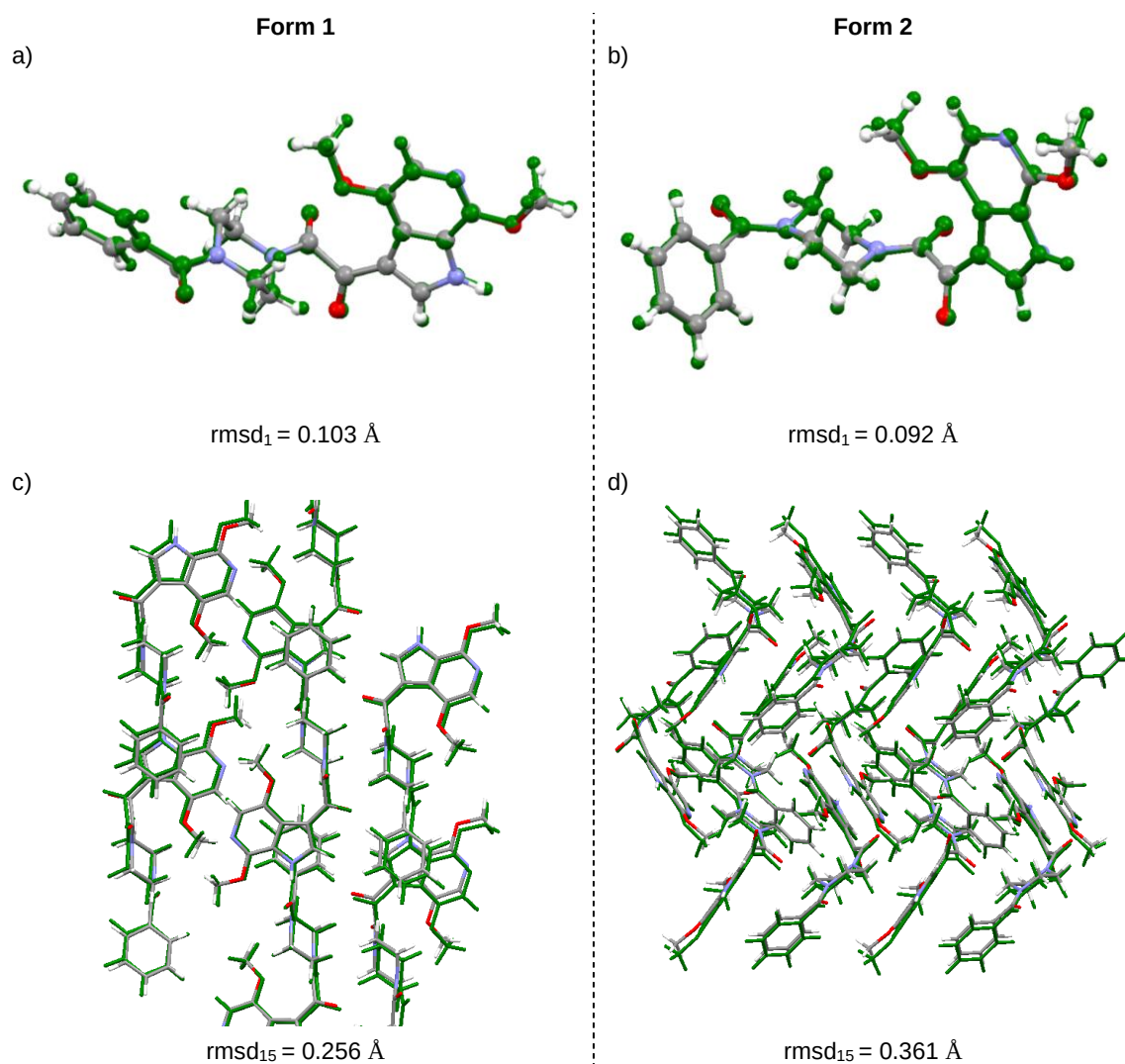


Figure 6.8 Overlay of the experimental conformations (coloured by element) with the final conformations obtained after CrystalOptimizer for experimental form 1 (a) and form 2 (b). Overlay of the experimental structures (coloured by element) and the predicted crystal structures (green) for experimental form 1 (c) and form 2 (d). Root mean square deviations in 1- and 15-molecule coordination spheres are also shown.

Table 6.6 Unit cell parameters and structure similarity between experimental and the predicted structures. For the space groups shown $\alpha = \gamma = 90^\circ$.

Structure	Density (g cm ⁻³)	Space Group	Unit Cell Dimensions				rmsd ₁₅ (Å)*
			a (Å)	b (Å)	c (Å)	β (°)	
Experimental Form 1	1.384	<i>P2₁/n</i>	14.51	7.28	20.08	107.1	-
Predicted Form 1	1.352	<i>P2₁/n</i>	14.74	7.16	20.75	108.6	0.256
Experimental Form 2	1.376	<i>P2₁/c</i>	13.22	11.30	13.78	97.9	-
Predicted Form 2	1.308	<i>P2₁/c</i>	13.26	11.33	14.43	98.5	0.361

* Root mean square deviation in atomic positions (excluding hydrogen atoms) of the 15-molecule coordination spheres compared to experimental structures.

6.2.5. Approximate computational cost

The conformational analysis and the candidate structure selection (steps 1 and 4 in the crystal structure prediction methodology outlined in section 6.1) are typically more demanding on the user rather than computer power. The former, involves a set of selected QM scans supported by a statistical search of the CSD can usually be performed within one week. In this work, the candidate structures are selected exclusively on the basis of their final lattice energy, which can be done almost instantly. A more detailed analysis of conformations and packing motifs of the low-energy structures will, of course, require further user input.

A summary of the approximate computational cost (on a single Intel Xeon 5150 2.66 GHz processor with 1500 MB of memory) for structure generation using the two different methods explored in this section and structure refinement with CrystalOptimizer is provided in Table 6.7.

From a practical point of view, rigid-molecule searches are simpler to implement than flexible searches, and preliminary results can be obtained almost immediately. However, due to the high degree of flexibility in the BMS molecule, 20 different conformations were initially considered and an extensive search involving 200,000 minimisations was completed for every conformation. As a result, the structure generation step required a total of 11500 hours for input file creation and search. However, in hindsight, 20 distinct geometries were not sufficient in identifying the experimental structures, and a larger number of conformations should have been considered which would have added proportionally to the computational cost.

In the flexible search approach, it was necessary to construct an intramolecular energy grid from relatively expensive QM calculations before any structure generation

could take place. Performing 200,000 minimisations in each of the 4 distinct flexibility regions was sufficient to capture most of the effects of the molecular flexibility exhibited by the BMS molecule within a total of 9,500 CPU hours for input file creation and structure generation.

The automated refinement using the CrystalOptimizer algorithm is computationally demanding due to the use of a large number of optimisation variables (molecular geometry and lattice parameters) and results of QM calculations during lattice energy minimisation. Nevertheless, the computational cost was kept manageable by using local approximate models (LAMs) and LAM/QM databases that provide QM accuracy but at a significantly reduced cost (refer to Chapter 4). This meant that the 196 structures could be refined within 20,000 CPU hours¹.

Table 6.7 Outline of the computational cost associated with structure generation and refinement (steps 2 and 3 in the crystal structure prediction methodology described in section 6.1) for the specific case of the BMS molecule.

Procedure	Number of structures considered	Computational Cost (CPU hours)*
Crystal Structure Generation (multiple rigid-body search) (Step 2)	2,000,000 structures generated	350 for obtaining starting geometries and charges 11500 for the search
Crystal Structure Generation (flexible search) (Step 2)	800,000 structures generated	4000 for the intramolecular grid generation 4500 for the search
Crystal Structure Refinement (Step 3)	196 structures considered 126 independent structures found	20,000

* Single Intel Xeon 5150 2.66 GHz processor, 1500 MB of memory

6.2.6. Conclusions

The BMS study was our first attempt at predicting crystal structures of large and flexible molecules. Hence, a key uncertainty at the beginning of the study was whether there would be any methodological and/or computational limits to the calculations. The adequate performance of the computational algorithms proved that such systems can now be routinely tackled.

The main challenge to the crystal structure prediction methodology was to account for the high degree of flexibility exhibited by the BMS molecule. Searching through the entire conformational space remains prohibitively expensive. Only a subset of the space,

¹ Note that only 196 structures were refined as the experimental forms were known to be in that list. A much larger set of structures would have been refined in a blind prediction adding to computational cost.

identified during the conformational analysis, can be considered in practice. At first, this was achieved by careful choice of a large number of low-energy rigid conformations and performing an independent search for every one of them. Despite several conformations being relatively close to the experimental structures, they were not correctly identified as stable minima on the lattice energy surface. The assumption that a $\pm 15^\circ$ error in the flexible torsional angles introduced at the search stage would be recovered during the refinement steps proved to be wrong. One of the solutions is to consider an even larger number of distinct rigid conformations during the search. However, this very quickly becomes computationally prohibitive for molecules of such size and flexibility.

A more tractable approach, as was shown by revisiting this test case, was to allow molecular flexibility during the search step¹. To do this, it was first necessary to create an intramolecular energy interpolation grid from relatively expensive DFT calculations before any structure generation could take place. Furthermore, because the intramolecular degrees of freedom had to be sampled, more crystal structures had to be generated and minimised by the search algorithm than in a typical rigid-body search. However, significantly fewer searches had to be performed (4 flexible searches versus 20 rigid-body searches). The relative computational cost of the two approaches turned out to be very similar. As a result, the more extensive flexible searches should be performed whenever possible. Had the flexible searches been implemented initially, the two experimental forms could have been predicted blind as low-energy structures.

The explicit treatment of full molecular flexibility using accurate conformationally-dependent models for the lattice energy based on QM calculations at the refinement stage led to an accurate reproduction of the unit cell parameters and conformations of both experimental forms.

Overall, the collaboration with Bristol-Myers Squibb proved to be very instructive in tackling the molecules of pharmaceutical importance and was a very helpful and timely practice before attempting to predict the structure of one of the molecules in the latest blind test of crystal structure prediction described in the next section.

¹ In the limit, a flexible search is equivalent to considering an infinite number of independent rigid conformations.

6.3. Target molecule XX in the Fifth Blind Test of Crystal Structure Prediction

6.3.1. Introduction

In the 2010 blind test of crystal structure prediction organised by the Cambridge Crystallographic Data Centre (CCDC) (Bardwell et al. 2011), the challenge was extended to include several more complex targets than were considered in the previous four blind tests (Lommerse et al. 2000, Motherwell et al. 2002, Day et al. 2005a, Day et al. 2009). This included a new category of crystal structures consisting of a flexible molecule with 50-60 atoms, 4-8 internal degrees of freedom, in any space group and with one or two independent molecules in the asymmetric unit. From the crystal structures provided in confidence to the CCDC, that of (benzyl-(4-(4-methyl-5-(p-tolylsulfonyl)-1,3-thiazol-2-yl)phenyl)carbamate was chosen for this category, noting that it was far more typical of modern pharmaceuticals than any other target previously considered. It became the 20th target in the series (and hence denoted molecule XX). In November 2009, the participants were given the molecular diagram (i.e. the covalent bonding) shown in Figure 6.9 and were informed that the crystal, with $Z' \leq 2$, was obtained by slow evaporation from an ethyl acetate solution. Each participant was required to submit three predictions of the crystal structure together with an extended list of low energy crystal structures on their crystal energy landscape by the 20th August 2010. Ten participants attempted predictions for this target. The approach adopted in this thesis and that of Day, Cruz-Cabeza and Galek were successful in predicting the observed crystal structure of blind test target XX as the global minimum (Kazantsev et al. 2011b). One other group (Neumann, Leusen, Kendrick and van de Streek) also located the observed crystal structure in its extended list of solutions (Bardwell et al. 2011). These results are summarised in Table 6.8. A more detailed description of the application of our approach to molecule XX is provided in this section.

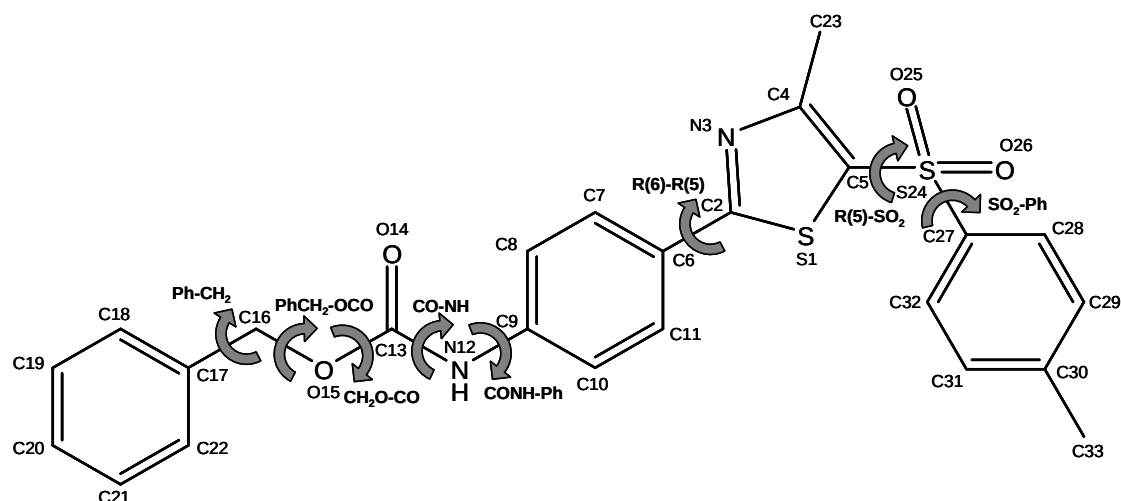


Figure 6.9 Molecular diagram of target molecule XX with the definitions of key torsion angles and atom labels.

Table 6.8 The unit cell parameters, rank and structure similarity between the experimental and predicted structures of target XX by three groups participating in the fifth blind test of crystal structure prediction (Bardwell et al. 2011). For the space group shown $\alpha = \gamma = 90^\circ$.

Method	Rank	Density (g cm ⁻³)	Space Group	Unit Cell Dimensions				rmsd ₁₅ (Å)*
				a (Å)	b (Å)	c (Å)	β (°)	
Experimental	-	1.410	<i>P2₁/n</i>	14.08	6.36	25.31	96.1	-
Kazantsev et al.	1	1.402	<i>P2₁/n</i>	14.26	6.32	25.36	97.3	0.178
Day et al.	1	1.375	<i>P2₁/n</i>	14.13	6.24	26.31	95.6	0.429
Neumann et al.	7	1.412	<i>P2₁/n</i>	14.16	6.30	25.37	95.4	0.113

* Root mean square deviation in atomic positions (excluding hydrogen atoms) of the 15-molecule coordination spheres compared to the experimental structure.

6.3.2. Step 1: Computational model selection and conformational analysis

Target molecule XX exhibits a large degree of conformational flexibility. As a result conformational analysis was performed to restrict the search space to energetically meaningful regions. This was done by quantum mechanical (B3LYP/6-31G(d,p)) isolated-molecule scans (Frisch et al. 2009), supported by an analysis of the Cambridge Crystallographic Database (CSD) on fragments of molecule XX for computational efficiency. The close relationship between the two types of conformational analysis is again (as previously discussed in section 6.2.2) illustrated in Figure 6.10 for torsion PhCH₂-OCO. The QM conformational energy and the CSD

observations are generally in good agreement. At values for which the conformational energy is high, few or no experimental occurrences are found in the CSD.

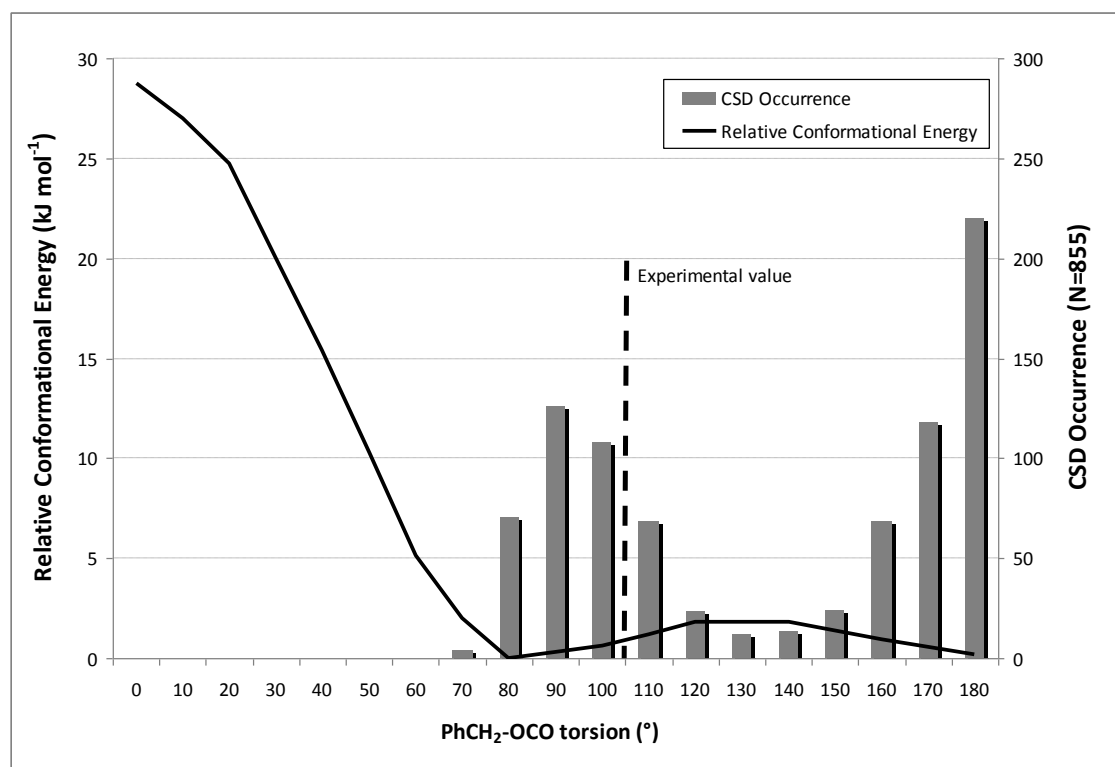


Figure 6.10 Comparison of the CSD MOGUL geometry analysis (bars, right axis) with the gas-phase intramolecular energy scan at the B3LYP/6-31G(d,p) level of theory (black line, left axis) for the torsion PhCH₂-OCO in molecule XX. N is the total number of similar fragments obtained from the CSD. The dashed line indicates the value as observed in the crystal structure of molecule XX. Note that the diagram is symmetric about 180°. (Assistance of A. J. Cruz-Cabeza in the preparation of this figure is gratefully acknowledged.)

The successful prediction of crystal structures of a wide range of small molecules using local torsional energy minima and rigid-molecule searches suggests the crystal conformation may be reasonably close to a gas phase conformation that corresponds to a low energy minimum. This assumption does not hold for flexible molecules such as molecule XX. A more appropriate assumption is that the crystal conformation usually has a gas-phase energy that is close to the minimum gas-phase energy (i.e. within a few kJ mol⁻¹). This does not preclude significant geometrical differences between the crystal conformation and the minimum energy gas phase conformation, when large differences in torsions angle incur a small energy penalty. Figure 6.11 shows that torsion SO₂-Ph, and especially torsion PhCH₂-OCO, in the crystal structure of molecule XX deviate significantly from the closest local minimum conformation from QM gas-

phase calculations. In fact, the torsion angle $\text{PhCH}_2\text{-OCO}$ can change by more than 100° with an energy penalty of less than 2 kJ mol^{-1} (Figure 6.10). Thus, it was appropriate to select a wide conformational range ($[+40.0^\circ, -50.0^\circ]$) for torsion $\text{PhCH}_2\text{-OCO}$ to ensure the experimental structure would not be missed. The experimental conformation takes the absolute value of 105.8° for this torsion, 25° from the closest local energy minimum and 16° from one of the maxima in the CSD statistical distribution (Figure 6.10). A similar analysis was carried out for all relevant torsions in the molecule by identifying all low energy conformational regions.

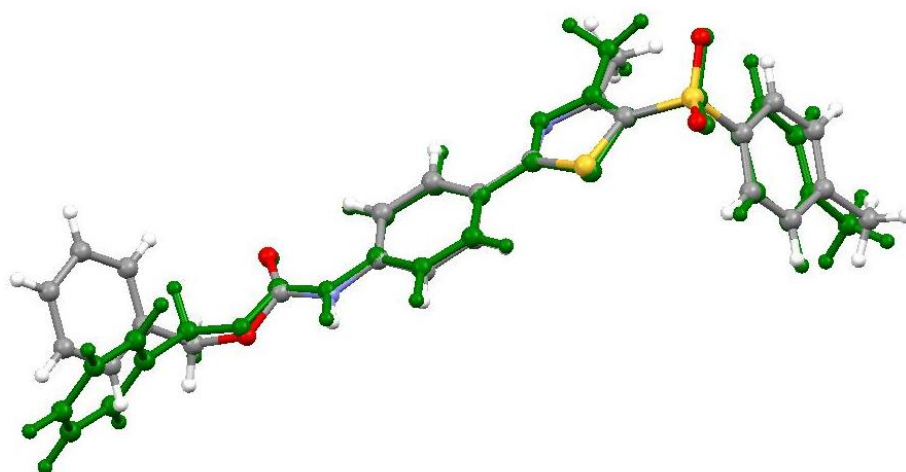


Figure 6.11 Overlay of the experimental conformation (coloured by element) with the closest local gas-phase minimum conformation obtained with the B3LYP/6-31G(d,p) level of theory (green).

The conformational analysis revealed that torsions $\text{CH}_2\text{O-CO}$ and CO-NH are relatively rigid i.e. very high energetic penalties associated with even small conformational changes. Consequently, they were assumed not to change in a continuum, but were enumerated instead. Note that the amide group (CO-NH) was allowed to adopt both the *trans* and the less energetically favourable ($\sim 10 \text{ kJ mol}^{-1}$) *cis* planar conformations. Furthermore, the conformational ranges of the two terminal phenyl ring rotations could be halved due to internal symmetry. The feasible ranges for the main torsional angles in the molecule are summarised in Table 6.9. The methyl group rotations were ignored at this stage as they are not expected to contribute significantly to the lattice energy.

Table 6.9 Energetically feasible torsion ranges for molecule XX. The torsion values in the experimental conformation are also given. Refer to Figure 6.9 for torsion definitions.

#	Torsion name	Torsion definition	Feasible Torsional Ranges or Values (°)	Experimental Conformation (°)
1	Ph-CH ₂	C22-C17-C16-O15	[0.0, +180.0]	+82.2
2	PhCH ₂ -OCO	C17-C16-O15-C13	[+40.0, -50.0]*	-105.8
3	CH ₂ O-CO	C16-O15-C13-O14	gas-phase value (close to 0.0)	+6.3
4	CO-NH	C9-N12-C13-O15	0.0 (<i>cis</i>) +180.0 (<i>trans</i>)	- +176.4
5	CONH-Ph	C8-C9-N12-C13	[-60.0, +60.0]	1.1
6	R(6)-R(5)	C7-C6-C2-S1	[-60.0, +60.0] [+120.0, -120.0]	-11.8 -
7	R(5)-SO ₂	S1-C5-S24-C27	[+50.0, +170.0] [-170.0, -50.0]	+104.7 -
8	SO ₂ -Ph	C28-C27-S24-C5	[+30.0, +150.0]	+107.0

* Note that this range spans 270° and corresponds to [+40.0, +310.0]

6.3.3. Step 2: Crystal structure generation

Before carrying out the crystal structure generation step, the dimensionality of the multivariable search space was reduced further. The CSD was analysed to determine the statistically expected number of molecules in the asymmetric unit and the space group of crystals containing molecules of similar size and functionality to molecule XX. The CSD was searched for all crystals containing one or two molecules in the asymmetric unit with more than 20 carbon atoms, but less than 100 atoms per molecule and not allowing for any metal atoms. A total of 11746 crystal structure matched the criteria. Almost 97% of the identified structures had just one molecule in the asymmetric unit and the vast majority were represented in just a handful of space groups (Figure 6.12).

Based on the above analyses, eight separate flexible CrystalPredictor searches were carried out for $Z' = 1$ structures in the 12 most common space groups ($P2_1/c$, $P2_12_12_1$, $P\bar{1}$, $P2_1$, $C2/c$, $Pbca$, $Pna2_1$, $C2$, $P1$, Cc , $Pca2_1$, $P2_12_12$)¹ in the crystal structure generation step (Table 6.10). All combinations of the torsion ranges/values identified in the conformational analysis (Table 6.9) were used to generate the search space. During the search, only 7 major torsional angles were allowed to change (Ph-CH₂, PhCH₂-OCO,

¹ Note that space groups $P2_1/a$ and $P2_1/n$ (Figure 6.12) are subgroups of the space group $P2_1/c$. Consequently, the search was only carried out for the latter space group.

CONH-Ph, R(6)-R(5), R(5)-SO₂, SO₂-Ph and CO-NH) with the amide group (CO-NH) in either the *trans* or *cis* planar conformation. To reduce the computational cost, interpolation was used in the evaluation of the intramolecular energy. For this purpose, two sub-molecules were derived from molecule XX that both include the central phenyl ring: one consisted of atoms 6 to 22 and the other of atoms 1 to 11 and 23 to 33 (defined in Figure 6.9), and a hydrogen atom was added to each of these fragments to ensure there were no free bonds. A grid of ΔE^{intra} values for the flexible torsions considered in the search was then derived by scanning three torsion angles on each sub-molecule.

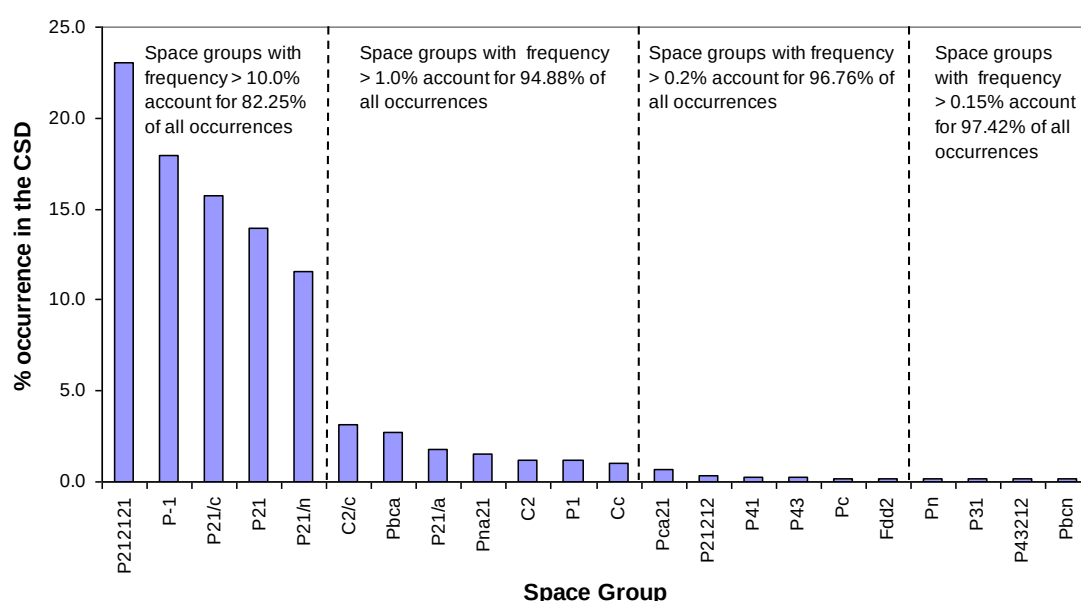


Figure 6.12 Space group frequency for $Z' = 1$ and $Z' = 2$ structures similar to molecule XX reported in the CSD.

The deformation energy for the first sub-molecule was computed on two $7 \times 10 \times 13$ grids (one each for *trans* and *cis* amide conformations) and for the second sub-molecule, on four $7 \times 7 \times 7$ grids, which corresponded to $20^\circ - 30^\circ$ grid spacing for each torsional angle. At each grid point the deformation energy was calculated (Schmidt et al. 1993) at the B3LYP/6-31G(d,p) level of theory with the flexible torsions fixed and the rest of the molecule optimised using the semi-empirical AM1 level of theory (Dewar et al. 1985). The intramolecular energy of molecule XX was then approximated as the sum of the deformation energy of the two sub-molecules, assuming that there are no significant interactions between these two parts. For the search in each conformational region, the intermolecular electrostatic interactions were modelled using the atomic charges that

were derived from the B3LYP/6-31G(d,p) electrostatic potential of the B3LYP/6-31G(d,p) conformational minimum of the whole molecule in this region. All other intermolecular energy terms, except for the sulphur atoms, were derived from a standard empirical exp-6 potential (Coombes et al. 1996). The intermolecular repulsion-dispersion parameters for the S...S interactions were those used by Price et al. in the second blind test of crystal structure prediction (CSP 2001) (Motherwell et al. 2002).

Table 6.10 Search enumeration and ranges of torsional angles considered explicitly during the (Refer to Figure 6.9 for torsion definitions).

#	Search name	Ph-CH ₂ (°)	PhCH ₂ - OCO (°)	CO-NH (°)	CONH-Ph (°)	R(6)-R(5) (°)	R(5)-SO ₂ (°)	SO ₂ -Ph (°)
1	Search_A1	[0.0, +180.0]	[+40.0, -50.0]	+180.0	[-60.0, +60.0]	[-60.0, +60.0]	[+50.0, +170.0]	[+30.0, +150.0]
2	Search_A2	[0.0, +180.0]	[+40.0, -50.0]	+180.0	[-60.0, +60.0]	[-60.0, +60.0]	[-170.0, -50.0]	[+30.0, +150.0]
3	Search_A3	[0.0, +180.0]	[+40.0, -50.0]	+180.0	[-60.0, +60.0]	[+120.0, -120.0]	[+50.0, +170.0]	[+30.0, +150.0]
4	Search_A4	[0.0, +180.0]	[+40.0, -50.0]	+180.0	[-60.0, +60.0]	[+120.0, -120.0]	[-170.0, -50.0]	[+30.0, +150.0]
5	Search_B1	[0.0, +180.0]	[+40.0, -50.0]	0.0	[-60.0, +60.0]	[-60.0, +60.0]	[+50.0, +170.0]	[+30.0, +150.0]
6	Search_B2	[0.0, +180.0]	[+40.0, -50.0]	0.0	[-60.0, +60.0]	[-60.0, +60.0]	[-170.0, -50.0]	[+30.0, +150.0]
7	Search_B3	[0.0, +180.0]	[+40.0, -50.0]	0.0	[-60.0, +60.0]	[+120.0, -120.0]	[+50.0, +170.0]	[+30.0, +150.0]
8	Search_B4	[0.0, +180.0]	[+40.0, -50.0]	0.0	[-60.0, +60.0]	[+120.0, -120.0]	[-170.0, -50.0]	[+30.0, +150.0]

For each of the eight separate flexible searches, 350,000 initial guesses were generated and minimised (2,800,000 in total). The search results after a preliminary clustering procedure are shown in Figure 6.13. The experimental structure was found with rank 223, 7.02 kJ mol⁻¹ above the global minimum.

In this flexible search method, the only assumptions about the molecular geometry were the choice of the 7 main torsion angles to be treated as flexible, and their specified ranges (as outlined in Table 6.9). The remaining set of intramolecular degrees of freedom was fixed at their gas-phase values. The explicit treatment of flexibility during the search led to the identification of a conformation similar to the experimental one (rmsd₁ = 0.167 Å, Figure 6.14) and of a crystal structure within rmsd₁₅ = 0.311 Å of the experimental structure.

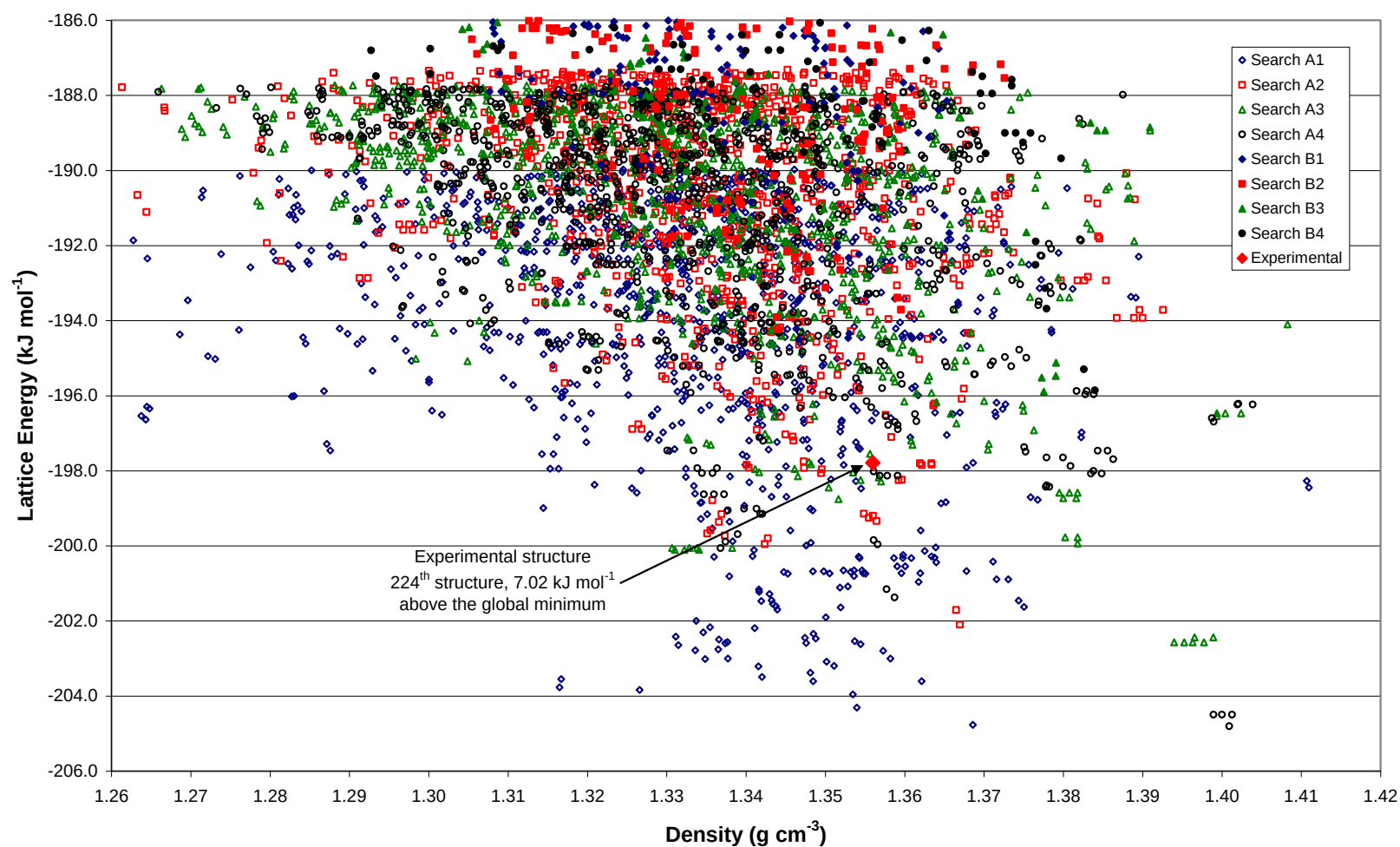


Figure 6.13 Lattice energy landscape generated by eight independent searches for molecule XX. Refer to Table 6.10 for search enumeration. The most stable structure is that with the lowest lattice energy. The minimum which most closely corresponds to the experimental structure is also shown.

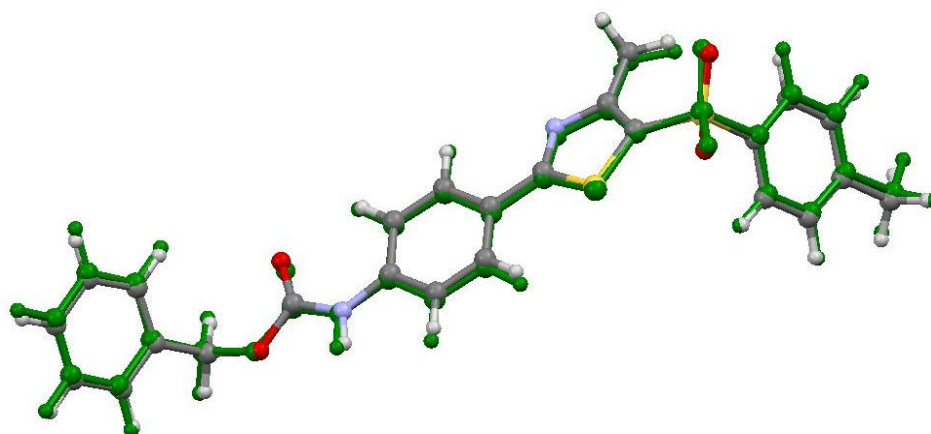


Figure 6.14 Overlay of the experimental conformation (coloured by element) with the closest conformation obtained from the flexible crystal structure generation step (green). Root mean square deviation in 1-molecule coordination sphere (rmsd_1) is 0.167 Å.

6.3.4. Steps 3 & 4: Crystal structure refinement and relative lattice energies

Even though the experimental structure was already accurately reproduced during the search, it ranked poorly (223rd) with respect to the other hypothetical structures. This is a direct consequence of the use of relatively inexpensive models for the intermolecular and intramolecular energy contributions during the search stage.

In Step 3, the molecular geometry and the lattice parameters of the 1500 most stable structures from the search were simultaneously re-optimised using the CrystalOptimizer algorithm with 19 flexible degrees of freedom (14 non-aromatic torsions and 5 selected chain bond angles, Figure 6.15). Local approximate models (LAMs) were constructed on-the-fly for the conformational variations of the intramolecular energy, molecular geometry and the distributed multipole moments at the PBE0/6-31G(d,p) level of theory and stored for reuse for similar conformations in subsequent lattice energy minimisations. The linear updates to the multipole model (multipole derivatives with respect to the flexible torsional angles) were calculated for all torsions involving a nitrogen atom. The minimised structures were then clustered based on their root mean square deviation in the 15-molecule coordination sphere, rmsd_{15} (Chisholm and Motherwell 2005). Two structures were considered to be crystallographically similar if

their rmsd_{15} was below 0.25 Å. Even if these are distinct minima mathematically, they are likely to interconvert to each other under thermal motion.

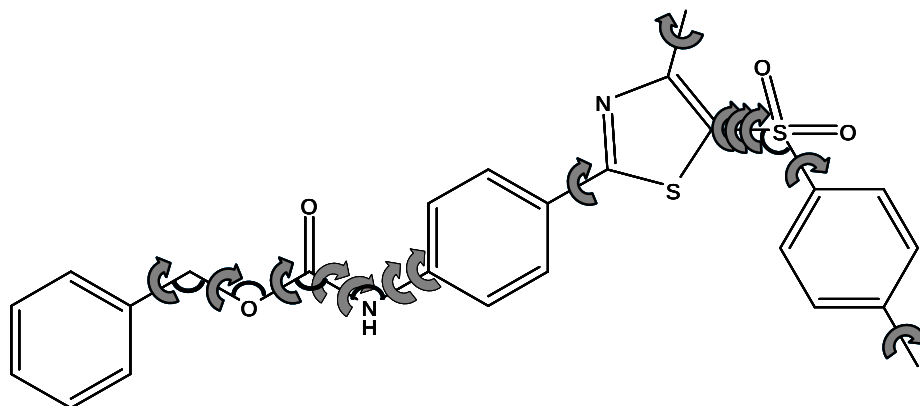


Figure 6.15 Molecular flexibility (14 non-aromatic torsions and 5 chain bond angles) considered explicitly during the CrystalOptimizer refinement stage for molecule XX.

The lattice energy landscape after the CrystalOptimizer refinement is shown in Figure 6.16. Only 553 out of 1500 structures considered for the refinement remained as independent structures. The structures selected for the blind test submission (Figure 6.17) were the two lowest in energy (with a 0.78 kJ mol⁻¹ gap). They differed significantly in that the second structure was less dense but had a conventional N-H...N hydrogen bond (N12...N3 2.9Å). Due to its lower density, the second structure can be potentially further stabilised by the entropic contributions. The third submission was the lowest energy structure with a *cis* amide conformation (11.43 kJ mol⁻¹ above the global minimum rank), in case this isomer of molecule XX had been synthesised. This conformation allows the formation of strong hydrogen bonded dimers. However, the intramolecular energy penalty that the molecule has to pay to adopt the *cis* conformation (~10 kJ mol⁻¹) significantly destabilises the crystal and cannot be counter-balanced by lower intermolecular energies.

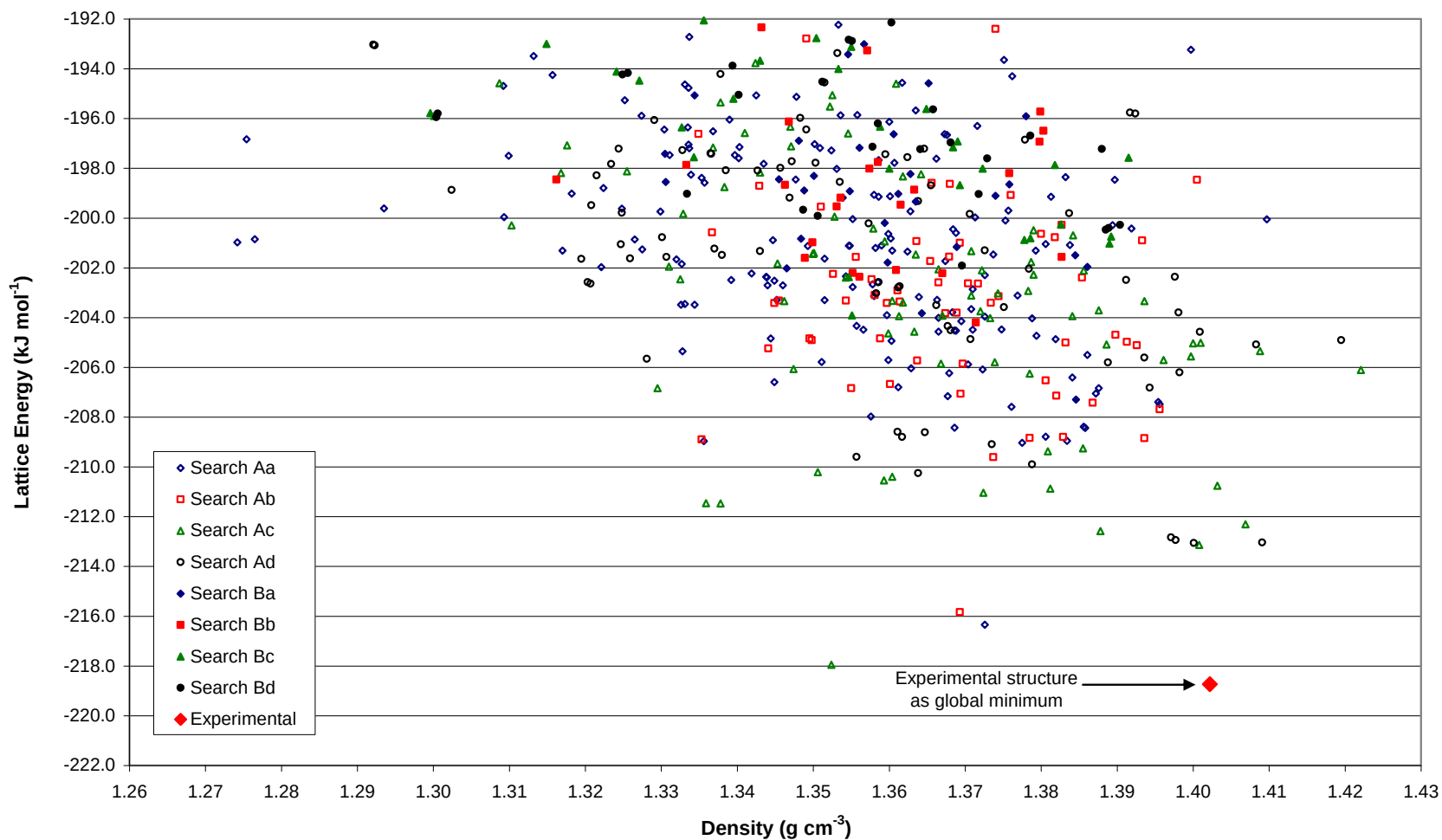


Figure 6.16 Final lattice energy landscape produced after the refinement of 1500 lowest energy structures from the search stage using the CrystalOptimizer algorithm and clustering. The minimum corresponding to the experimental structure is also shown. All 553 independent minima are shown.

# 1	A1_132	P2 ₁ /n	-218.73 kJ mol⁻¹ 1.402 g cm⁻³	# 2	A3_102	P2 ₁ /a	-217.95 kJ mol⁻¹ 1.352 g cm⁻³
Possible with S=O...S contact				Strong NH hydrogen bond to aromatic N			
# 3	B1_28	P2 ₁ /c	-207.30 kJ mol⁻¹ 1.385 g cm⁻³				
Strong NH...O=C hydrogen bond							

Figure 6.17 Crystal structures of the 3 candidates selected for the blind test submission. Structure identification number, space group, lattice energy, density and hydrogen bonding type are also shown.

The experimental structure was correctly predicted as the lowest energy structure (Figure 6.16 and Figure 6.17). The overlay of the prediction with the experimental form is given in Figure 6.18 and the predicted and observed unit cell parameters in Table 6.11. The remaining low-energy structures have very different conformations, packing motifs and hydrogen bonds (e.g. Figure 6.17).

Table 6.11 The unit cell parameters for the experimental and predicted structures. For the space group shown $\alpha = \gamma = 90^\circ$.

Structure	Density (g cm ⁻³)	Space Group	Unit Cell Dimensions			
			a (Å)	b (Å)	c (Å)	β (°)
Experimental	1.410	P2 ₁ /n	14.08	6.36	25.31	96.1
Predicted	1.402	P2 ₁ /n	14.26	6.32	25.36	97.3

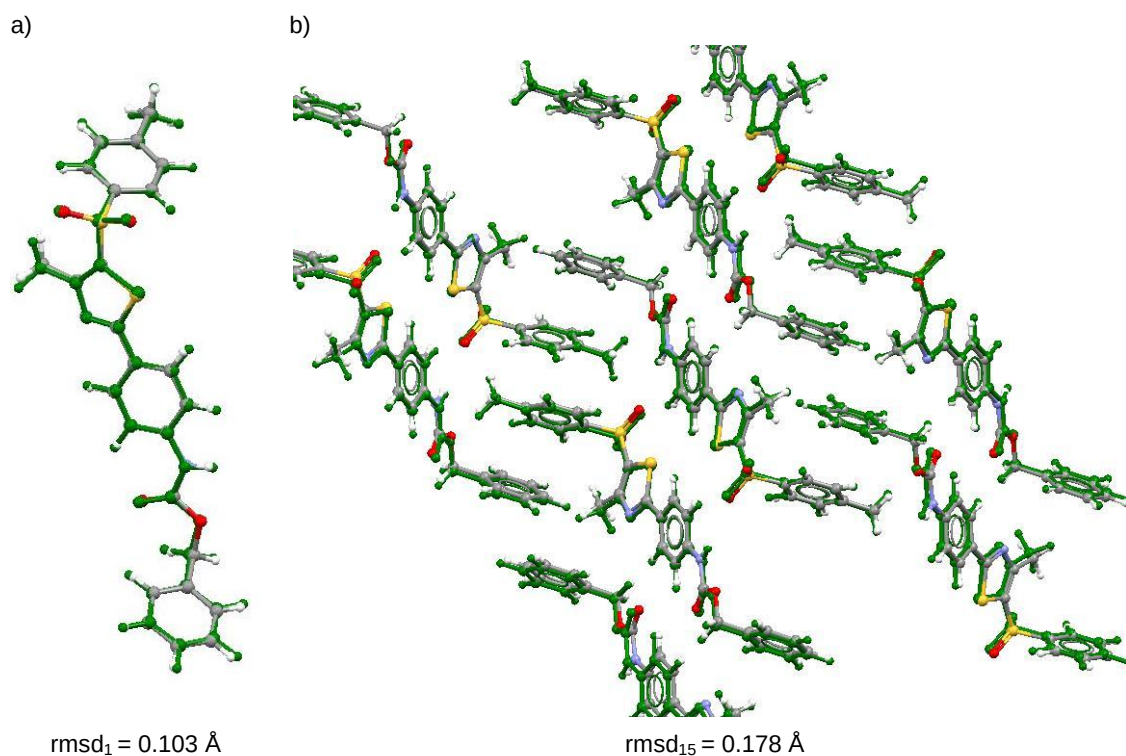


Figure 6.18 Overlay of the experimental conformation (a) and structure (b) (coloured by element) and the predicted lowest energy structure (green). The root mean square deviation in the 1- and 15-molecule coordination sphere is also shown.

A post-analysis of the blind test results for molecule XX was performed by each of the three research groups (Kazantsev et al., Day et al. and Neumann et al.) that correctly predicted the experimental structure as one of the submissions in the extended list of structures (Table 6.8). Each group re-calculated each other's low-energy structures using their most accurate models for the lattice energy (van Eijck 2005). This analysis revealed that, when evaluated with the other groups' energy models, our second submitted structure (Figure 6.17) was more stable than any of the structures submitted by the other two research groups, but was not present in their respective extended lists of structures (van de Streek 2011, Kazantsev et al. 2011b). This structure is less dense, but exhibits a strong conventional intermolecular hydrogen bond between the NH group and the aromatic N in the 5-membered ring. When re-minimised with our final computational model (CrystalOptimizer), no other structures were found to be more stable than our submitted global minimum (i.e. the experimental form).

6.3.5. Approximate computational cost

The crystal structure prediction methodology presented in this work is demanding on computational resources. The conformational analysis and the candidate structure selection (steps 1 and 4 in the crystal structure prediction methodology outlined in section 6.1) are typically more demanding on the user rather than computer power. The former, involving a statistical search of the CSD and selected QM intramolecular scans can usually be performed within one week. In this work, the candidate structures are selected mainly on the basis of their final lattice energy, which can be done almost instantly. A more detailed analysis of conformations and packing motifs of the low-energy structures will of course require further user input.

A summary of the approximate computational cost for structure generation and refinement (using a single Intel Xeon 5150 2.66 GHz processor using 1500 MB of memory) is provided in Table 6.12. In the flexible search approach adopted in this example, it was necessary to construct an intramolecular energy grid from relatively expensive QM calculations before any structure generation could take place. Performing 350,000 minimisations in each of the 8 distinct flexibility regions was sufficient to capture most of the effects of the molecular flexibility exhibited by molecule XX within a total of 18,000 CPU hours for input file creation and structure generation. The automated refinement using the CrystalOptimizer algorithm is computationally expensive (100,000 CPU hours for the refinement of 1500 structures) due to the use of a large number of optimisation variables (molecular geometry and lattice parameters) and results of QM calculations during lattice energy minimisation. Nevertheless, the computational cost was kept manageable by using local approximate models (LAMs) and LAM/QM databases that provide QM accuracy but at a significantly reduced cost (refer to Chapter 4). Results for a total of 3500 different conformations were stored and actively re-used in the LAM/QM databases. Also note that it was initially intended (and turned out to be more than sufficient) to refine a maximum of 1000 structures with CrystalOptimizer. However, due to the availability of computational resources prior to the submission deadline, an additional 500 structures from the search were further refined for added confidence. This was not necessary and would have reduced the overall cost of the computational methodology from a total of ~120,000 CPU hours to about ~90,000 CPU hours.

Table 6.12 Outline of the computational cost associated with structure generation and refinement (steps 2 and 3 in the crystal structure prediction methodology described in section 6.1) for the specific case of molecule XX in the fifth blind test of crystal structure prediction.

Procedure	Number of structures considered	Computational Cost (CPU hours)*
Crystal Structure Generation (Step 2)	2,800,000 structures generated 800,000 independent structures	2000 for the intramolecular grid generation 16,000 for search
Crystal Structure Refinement (Step 3)	1500 structures considered 553 independent structures found	100,000

* Single Intel Xeon 5150 2.66 GHz processor, 1500 MB of memory

6.3.6. Conclusions

The crystal structure of a highly flexible molecule, with a complexity typical of those currently being developed in the pharmaceutical industry, has been successfully predicted as the most stable structure in the fifth blind test of crystal structure prediction (CSP 2010).

The first challenge that had to be overcome was to effectively reduce the conformational space that is considered in crystal structure generation, without eliminating important conformations that could lead to low energy crystal structures. Simply identifying all low energy minima and assuming that these will be close to any solid state conformation proved to be clearly inadequate for molecule XX, although this strategy has been successful for some smaller molecules (Price 2004b). It is clear that all low energy conformational regions had to be considered. In this work, the flexible torsions were considered explicitly during the search. An alternative strategy based on multiple rigid-body conformations has also proven to be successful for this particular example (Bardwell et al. 2011, Kazantsev et al. 2011a).

Another key challenge was the computational cost of dealing with such a large and flexible molecule during the more accurate and demanding calculations of the final lattice energy. The development of the CrystalOptimizer algorithm allowed the explicit handling of 19 geometric variables (14 torsions and 5 bond angles) during lattice energy minimisation, which is a significantly higher degree of molecular flexibility than previously considered for any blind test targets. Furthermore, through the use of LAM/QM databases 1500 structures could be minimised at the QM accuracy at a high,

but nevertheless accessible computational cost. Previously, methods of similar accuracy were limited to the refinement of only a few tens of structures.

In treating molecule XX, the assumption that the target structure can be identified as the global minimum in the lattice energy, rather than the free energy surface, appears to be appropriate. The successful prediction of the crystal structure of molecule XX in the blind test indicates that current search methods and models for lattice energy may be capable of tackling this type of molecule to give worthwhile results, both in terms of the range of structures considered in the search and relative energies of the structures.

6.4. Conclusions

The successful identification of the experimental crystal structures of the BMS molecule and target XX presented in this chapter indicates that current search methods and models for lattice energy are capable of tackling this type of molecule to give worthwhile results, both in terms of the range of structures considered in the search and relative energies of the structures.

One of the main challenges was to account for the high degree of flexibility in the molecules during the search for crystal structures, as several torsion angles could vary considerably with only small variations in molecular energy. Searching through the entire conformational space remains prohibitively expensive for this molecular size, so that a subset of the conformational space was considered. In this work, the successful predictions were achieved by explicitly considering the torsion angles as variables in the search, but with limited ranges. The agreement in defining these regions from an analysis of existing crystal structures and from the use of quantum mechanical energy scans is encouraging.

Another key challenge was the computational cost of dealing with such large and flexible molecules during the more accurate and demanding calculations of the final lattice energy. The development of the CrystalOptimizer algorithm allowed the efficient, yet accurate, minimisations of many hundreds of structures with a significantly higher degree of flexibility than were previously considered for most molecules reported in the literature.

The hope that molecules with sufficient flexibility will find one mode of packing that is significantly more stable than any others (Day et al. 2005b), and therefore be readily predictable, has not been realised with both the BMS molecule and target XX. There are alternative structures with different conformations and intermolecular interactions that are well within the energy range of being possible polymorphs, let alone the likely errors in the models for the relative lattice energies. These structures all exhibit likely hydrogen bonds geometries and conformations derived from the CSD structures, as well as falling in a small density range. Hence, large flexible molecules, like most small molecules, provide a challenge to the development of sufficiently accurate and efficient models for the relative energies of crystal structures to be able to confidently predict the most thermodynamically stable form.

In pharmaceutical development, the calculation of the crystal energy landscape can complement solid form screening beyond confirming that the most thermodynamically stable form has been found (Lancaster et al. 2006, Price 2008b, Braun et al. 2011). Guiding the search for different types of crystal structures that appear to be feasible polymorphs should aid late stage polymorph screening (Arlin et al. 2011, Bond et al. 2011). Polymorphs that are formed by desolvation of metastable solvates are more likely to be kinetically trapped for large molecules, which are generally much less able than approximately spherical molecules to rearrange significantly late in the crystallisation process (Hulme et al. 2007). Isomorphic desolvates should be predictable, as relatively stable structures which contain voids do appear as local minima on the crystal energy landscapes of molecules which form inclusion compounds (Cruz-Cabeza et al. 2009).

An alternative application of computed crystal structures is in helping to characterise structures for which good single crystals cannot be grown, in conjunction with, for example, unindexable powder diffraction patterns (Tremayne et al. 2004, Cruz-Cabeza et al. 2010), terahertz spectra (Parrott et al. 2009) or solid-state NMR chemical shifts (Salager et al. 2010). Furthermore, the careful analysis of the crystal energy landscape can also point towards more complex behaviour. For instance, if the crystal energy landscape has related structures that are close in energy, this may suggest a tendency to certain forms of disorder that can complicate spectra (Li et al. 2010) and may hinder the growth of single crystals and the development of a robust production process (Copley et al. 2008).

Chapter 7. Conclusions and directions for future work

The incorporation of molecular flexibility throughout the course of crystal structure prediction has been the main focus of this work. A key aspect of the proposed methodology is the development of local approximate models (LAMs), based on Taylor series, described in Chapter 3. These allow the accurate, yet computationally efficient descriptions of the dependence of intramolecular energy, molecular geometry and conformationally-dependent charge density on molecular conformation. Overall, LAMs are as accurate as explicit quantum mechanical calculations within a limited range of variable changes, but carry a much smaller computational burden.

CrystalOptimizer, an algorithm for local minimisation of the lattice energy has been presented in Chapter 4. The main novelty of the algorithm lies in the use of dynamically constructed and updated LAMs which essentially make available the full accuracy of the quantum mechanical model at each and every iteration of the minimisation, while requiring the performance of only a small number of isolated-molecule quantum mechanical calculations. The results of the explicit QM calculations necessary to construct the local approximate models are stored in LAM/QM databases which are specific only to a particular molecule and level of QM theory. These databases can, therefore, be re-used (and extended) in the context of other minimisations involving the same molecule (e.g. a different polymorph, or under different physical conditions, or a cocrystal together with other molecules) with the same or higher degree of molecular flexibility. These developments have made possible the accurate treatment of single and multi-component crystals containing molecules with a relatively large number of atoms and significant flexibility in torsional and bond angles and even bond lengths, as demonstrated by the local minimisation of the lattice energy of a set of single-component crystals, cocrystals and a salt.

The computational cost of the CrystalOptimizer algorithm is dominated by the calculation of the partial derivatives of intermolecular energy (and in particular, distributed multipoles) with respect to the flexible degrees of freedom. This is currently done using finite difference approximations, which is both inefficient and potentially susceptible to numerical noise. Replacing these approximations by analytically calculated partial derivatives can reduce the computational cost, potentially, by a

further order of magnitude, and thus will extend the applicability of CrystalOptimizer to even larger molecules.

An extension to CrystalPredictor, a search algorithm for low-energy crystal structures involving flexible molecules, has been presented in Chapter 5. The conformational flexibility in the original approach was modelled by splitting molecules into fragments connected by flexible torsional angles. The main limitation of this approach is that the geometry of the fragments was assumed to be fixed. In the proposed modification, the molecular geometry is modelled as an explicit function of the flexible degrees of freedom using the LAMs presented in Chapter 3. As shown by applying the modified algorithm to the well-known ROY molecule, by using more realistic models for the molecular conformation, all experimentally-known structures are reproduced accurately and are predicted as favourable low-energy minima on the lattice energy surface. This indicates that fewer structures have to be refined with subsequent, more accurate, lattice energy minimisation algorithms, such as CrystalOptimizer, resulting in a lower computational cost of the overall crystal structure prediction methodology. Although the *a priori* construction of the necessary LAMs can be computationally expensive, the LAM/QM databases constructed for the search can be re-used in subsequent re-minimisation of crystal structures with CrystalOptimizer using the same or *higher* degree of molecular flexibility.

In the current implementation of CrystalPredictor, the intermolecular interactions are modelled using conformationally-dependent atomic charges. A useful extension would be to allow the inclusion of conformationally-dependent ‘satellite’ charges for a better reproduction of the electrostatic potential around the molecule.

The above algorithms combine to form an integral part of a reliable multi-stage crystal structure prediction methodology for flexible molecules, such as those typically encountered in the pharmaceutical industry (cf. Chapter 6). Based (almost) entirely on the knowledge of atomic composition and connectivity and the results of isolated-molecule quantum mechanical calculations and has been successfully applied to the prediction of a drug candidate provided by Bristol-Myers Squibb Company, and of the largest and most flexible target in the latest blind test of crystal structure prediction. The importance of taking proper account of molecular flexibility at all stages of crystal structure prediction is clearly demonstrated by these examples. However, the number of low-energy crystal structures predicted is still much larger than what is likely to appear

experimentally. It is, therefore, imperative to continue the development of reliable *a posteriori* analysis techniques of the type discussed in section 2.4.3 of this work.

The only empirical component of the crystal structure prediction methodology described in this work is the reliance on external information for the description of the van der Waals interactions. Even though this information is available for many molecules, it is not the case for more unusual atoms or functional groups. Furthermore, these empirically-derived potentials may not be fully transferable across different molecule classes or types. The drive towards non-empirical molecule-specific intermolecular potentials, described in section 2.2.2, is highly encouraged.

On a more fundamental level, an important aspect of the proposed crystal structure prediction methodology is its ability to handle systems with more than one molecule or ion in the asymmetric unit, which makes it applicable to the prediction of crystal structures of cocrystals, salts and solvates. Our approach to date has been to treat each of these molecules or ions independently for the purposes of quantum mechanical calculations, and then to rely on intermolecular descriptions of the interactions between them. Whether or not this is the most appropriate way of handling such systems, or related ones such as hydrates, requires further investigation. Models incorporating a continuum dielectric (Cooper et al. 2008) or based on dispersion-corrected periodic density functional theory (Neumann and Perrin 2005) seem to be more accurate in capturing the polarisability effects.

Although we are still a long way from understanding, let alone reliably predicting all solid forms of pharmaceutical molecules, the results presented in this work demonstrate a step change in the complexity of molecules for which a crystal lattice energy landscape can be calculated. Complementing experimental polymorph screening can have a profound effect on pharmaceutical development (Price 2009). However, further improvements in computational models and techniques are required to give more reliable crystal free energy landscapes. It is also necessary to consider kinetic and other factors (such as solvent effects), so that the field can move towards a predictive technology for the understanding and anticipation of polymorphism.

Bibliography

- [1] Aakeröy, C. B.; "Recent advances in crystal engineering", *CrystEngComm*, **2010**, 12, 1, 22-43.
- [2] Aakeröy, C. B.; "Crystal engineering: Strategies and architectures", *Acta Crystallographica Section B - Structural Science*, **1997**, 53, 4, 569-586.
- [3] Admiraal, G., Eikelenboom, J. C., and Vos, A.; "Structures of the triclinic and monoclinic modifications of (2-oxo-1-pyrrolidiny)acetamide", *Acta Crystallographica Section B - Structural Crystallography and Crystal Chemistry*, **1982**, 38, 10, 2600-2605.
- [4] Allen, F. H.; "The Cambridge Structural Database: a quarter of a million crystal structures and rising", *Acta Crystallographica Section B - Structural Science*, **2002**, 58, 3, 380-388.
- [5] Allen, F. H., Harris, S. E., and Taylor, R.; "Comparison of Conformer Distributions in the Crystalline State With Conformational Energies Calculated By Ab-Initio Techniques", *Journal of Computer-Aided Molecular Design*, **1996**, 10, 3, 247-254.
- [6] Allen, M. P. and Tildesley, D. J.; *Computer Simulations of Liquids*, Clarendon Press, Oxford, **1987**.
- [7] Andrews, L. C. and Bernstein, H. J.; "Lattices and Reduced Cells and Points in 6-Space and Selection of Bravais Lattice Type by Projections", *Acta Crystallographica Section A - Foundations of Crystallography*, **1988**, 44, 6, 1009-1018.
- [8] Anghel, A. T., Day, G. M., and Price, S. L.; "A study of the known and hypothetical crystal structures of pyridine: why are there four molecules in the asymmetric unit cell?", *CrystEngComm*, **2002**, 4, 62, 348-355.
- [9] Antoniadis, C. D., D'Oria, E., Karamertzanis, P. G., Tocher, D. A., Florence, A. J., Price, S. L., and Jones, A. G.; "A computationally inspired investigation of the solid forms of (R)-1-phenylethylammonium-(S)-2-phenylbutyrate", *Chirality*, **2010**, 22, 447-455.
- [10] Arikawa, T., Tajima, N., Tsuzuki, S., Tanabe, K., and Hirano, T.; "A Possible Crystal-Structure of 1,2-Dimethoxyethane - Prediction Based on A Lattice Variable Molecular-Dynamics", *Theochem-Journal of Molecular Structure*, **1995**, 339, 115-124.
- [11] Arlin, J. B., Price, L. S., Price, S. L., and Florence, A. F.; "A strategy for producing predicted polymorphs: catermeric carbamezapine form V", *Chemical Communications*, **2011**, in press.

- [12] Bardwell, D. A., Adjiman, C. S., Ammon, H. L., Arnautova, E. A., Bartashevich, E., Boerrigter, S. X. M., Braun, D. E., Cruz-Cabeza, A. J., Day, G. M., Della Valle, R. G., Desiraju, G. R., van Eijck, B. P., Facelli, J. C., Ferraro, M. B., Grillo, D., Habgood, M., Hofmann, D. W. M., Hofmann, F., Jose, J., Karamertzanis, P. G., Kazantsev, A. V., Kendrick, J., Kuleshova, L. N., Leusen, F. J. J., Maleev, A., Misquitta, A. J., Mohamed, S., Needs, R. J., Neumann, M. A., Nikylov, D., Orendt, A. M., Pal, R., Pantelides, C. C., Pickard, C. J., Price, L. S., Price, S. L., Scheraga, H. A., van de Streek, J., Thakur, T. S., Tiwari, S., Venuti, E., and Zhitkov, I.; "Towards crystal structure prediction of pharmaceutically relevant molecules - a report on the fifth blind test.", *Acta Crystallographica Section B - Structural Science*, **2011**, submitted.
- [13] Bauer, J., Spanton, S., Henry, R., Quick, J., Dziki, W., Porter, W., and Morris, J.; "Ritonavir: An extraordinary example of conformational polymorphism", *Pharmaceutical Research*, **2001**, 18, 6, 859-866.
- [14] Bazterra, V. E., Ferraro, M. B., and Facelli, J. C.; "Modified genetic algorithm to model crystal structures: III. Determination of crystal structures allowing simultaneous molecular geometry relaxation", *International Journal of Quantum Chemistry*, **2004a**, 96, 4, 312-320.
- [15] Bazterra, V. E., Ferraro, M. B., and Facelli, J. C.; "Modified genetic algorithm to model crystal structures. II. Determination of a polymorphic structure of benzene using enthalpy minimization", *Journal of Chemical Physics*, **2002b**, 116, 14, 5992-5995.
- [16] Bazterra, V. E., Ferraro, M. B., and Facelli, J. C.; "Modified genetic algorithm to model crystal structures. I. Benzene, naphthalene and anthracene", *Journal of Chemical Physics*, **2002a**, 116, 14, 5984-5991.
- [17] Bazterra, V. E., Ona, O., Caputo, M. C., Ferraro, M. B., Fuentealba, P., and Facelli, J. C.; "Modified genetic algorithms to model cluster structures in medium-size silicon clusters", *Physical Review A*, **2004b**, 69, 5.
- [18] Bennema, P., Meekes, H., Boerrigter, S. X. M., Cuppen, H. M., Deij, M. A., van Eupent, J., Verwer, P., and Vlieg, E.; "Crystal growth and morphology: New developments in an integrated Hartman-Perdok-connected net-roughening transition theory, supported by computer simulations", *Crystal Growth & Design*, **2004**, 4, 5, 905-913.
- [19] Bernstein, J.; *Polymorphism in Molecular Crystals*, Clarendon Press, Oxford, **2002**.
- [20] Beyer, T., Lewis, T., and Price, S. L.; "Which organic crystal structures are predictable by lattice energy minimisation?", *CrystEngComm*, **2001**, 3, 44, 178-212.
- [21] Bisker-Leib, V. and Doherty, M. F.; "Modeling the crystal shape of polar organic materials: Prediction of urea crystals grown from polar and nonpolar solvents", *Crystal Growth & Design*, **2001**, 1, 6, 455-461.

- [22] Bond, A. D., Solanko, K. A., van de Streek, J., and Neumann, M.; "Experimental verification of a subtle low-temperature phase transition suggested by DFT-D energy minimisation", *CrystEngComm*, **2011**, 13, 1768-1771.
- [23] Bondi, A.; "van der Waals Volumes and Radii", *Journal of Physical Chemistry*, **1964**, 68, 3, 441-451.
- [24] Bradford Thompson, H.; "Calculation of Cartesian Coordinates and Their Derivatives from Internal Molecular Coordinates", *Journal of Chemical Physics*, **1967**, 47, 9, 3407-3410.
- [25] Brameld, K. A., Kuhn, B., Reuter, D. C., and Stahl, M.; "Small molecule conformational preferences derived from crystal structure data. A medicinal chemistry focused analysis.", *Journal of Chemical Information and Modeling*, **2008**, 48, 1, 1-24.
- [26] Braun, D. E., Karamertzanis, P. G., Arlin, J. B., Florence, A. J., Kahlenberg, V., Tocher, D. A., Griesser, U. J., and Price, S. L.; "Solid-State Forms of beta-Resorcylic Acid: How Exhaustive Should a Polymorph Screen Be?", *Crystal Growth & Design*, **2011**, 11, 1, 210-220.
- [27] Brock, C. P. and Dunitz, J. D.; "Towards A Grammar of Crystal Packing", *Chemistry of Materials*, **1994**, 6, 8, 1118-1127.
- [28] Brodersen, S., Wilke, S., Leusen, F. J. J., and Engel, G.; "A study of different approaches to the electrostatic interaction in force field methods for organic crystals", *Physical Chemistry Chemical Physics*, **2003**, 5, 21, 4923-4931.
- [29] Brown, G. M. and Levy, H. A.; "alpha-D-Glucose: further refinement based on neutron-diffraction data", *Acta Crystallographica Section B - Structural Science*, **1979**, 35, 656-659.
- [30] Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J., and Taylor, R.; "New software for searching the Cambridge Structural Database and visualizing crystal structures", *Acta Crystallographica Section B - Structural Science*, **2002**, 58, 389-397.
- [31] Bruno, I. J., Cole, J. C., Kessler, M., Luo, J., Motherwell, W. D. S., Purkis, L. H., Smith, B. R., Taylor, R., Cooper, R. I., Harris, S. E., and Orpen, A. G.; "Retrieval of Crystallographically-Derived Molecular Geometry Information", *Journal of Chemical Information and Computer Sciences*, **2004**, 44, 6, 2133-2144.
- [32] Bruno, I. J., Cole, J. C., Lommerse, J. P. M., Rowland, R. S., Taylor, R., and Verdonk, M. L.; "IsoStar: A library of information about nonbonded interactions", *Journal of Computer-Aided Molecular Design*, **1997**, 11, 6, 525-537.
- [33] Buckingham, R. A.; "The classical equation of state of gaseous helium, neon and argon", *Proceedings of the Royal Society of London*, **1938**, A168, 264-283.

- [34] Byrd, E. F. C., Scuseria, G. E., and Chabalowski, C. F.; "An ab initio study of solid nitromethane, HMX, RDX, and CL20: Successes and failures of DFT", *Journal of Physical Chemistry B*, **2004**, 108, 35, 13100-13106.
- [35] Chan, H. C. S., Kendrick, J., and Leusen, F. J. J.; "Molecule VI, a Benchmark Crystal-Structure-Prediction Sulfonimide: Are Its Polymorphs Predictable?", *Angewandte Chemie-International Edition*, **2011**, 50, 2979-2981.
- [36] Chemburkar, S. R., Bauer, J., Deming, K., Spiwek, H., Patel, K., Morris, J., Henry, R., Spanton, S., Dziki, W., Porter, W., Quick, J., Bauer, P., Donaubaue, J., Narayanan, B. A., Soldani, M., Riley, D., and McFarland, K.; "Dealing with the impact of ritonavir polymorphs on the late stages of bulk drug process development", *Organic Process Research & Development*, **2000**, 4, 5, 413-417.
- [37] Chen, S., Guzei, I. A., and Yu, L.; "New Polymorphs of ROY and New Record for Coexisting Polymorphs of Solved Structures", *Journal of the American Chemical Society*, **2005**, 127, 27, 9881-9885.
- [38] Chisholm, J. A. and Motherwell, S.; "COMPACT: a program for identifying crystal structure similarity using distances", *Journal of Applied Crystallography*, **2005**, 38, 228-231.
- [39] Chisholm, J. A., Motherwell, S., Tulip, P. R., Parsons, S., and Clark, S. J.; "An ab initio study of observed and hypothetical polymorphs of glycine", *Crystal Growth & Design*, **2005**, 5, 4, 1437-1442.
- [40] Coombes, D. S., Price, S. L., Willock, D. J., and Leslie, M.; "Role of Electrostatic Interactions in Determining the Crystal Structures of Polar Organic Molecules. A Distributed Multipole Study", *Journal of Physical Chemistry*, **1996**, 100, 18, 7352-7360.
- [41] Cooper, T. G., Hejczyk, K. E., Jones, W., and 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, 10, 1795-1805.
- [42] Copley, R. C. B., Barnett, S. A., Karamertzanis, P. G., Harris, K. D. M., Kariuki, B. M., Xu, M., Nickels, E. A., Lancaster, R. W., and Price, S. L.; "Predictable disorder versus polymorphism in the rationalization of structural diversity: a multi-disciplinary study of eniluracil", *Crystal Growth & Design*, **2008**, submitted.
- [43] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A.; "A Second Generation Force-Field For the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules", *Journal of the American Chemical Society*, **1996**, 118, 9, 2309.
- [44] Cox, S. R., Hsu, L. Y., and Williams, D. E.; "Nonbonded Potential Function Models for Crystalline Oxohydrocarbons", *Acta Crystallographica Section A* -

- Crystal Physics, Diffraction, Theoretical and General Crystallography*, **1981**, 37, MAY, 293-301.
- [45] Cox, S. R. and Williams, D. E.; "Representation of the Molecular Electrostatic Potential By a Net Atomic Charge Model", *Journal of Computational Chemistry*, **1981**, 2, 3, 304-323.
- [46] Cruz-Cabeza, A. J., Day, G. M., and Jones, W.; "Towards Prediction of Stoichiometry in Crystalline Multicomponent Complexes", *Chemistry - A European Journal*, **2008**, 14, 8830-8836.
- [47] Cruz-Cabeza, A. J., Day, G. M., and Jones, W.; "Predicting Inclusion Behaviour and Framework Structures in Organic Crystals", *Chemistry - A European Journal*, **2009**, 15, 13033-13040.
- [48] Cruz-Cabeza, A. J., Day, G. M., Motherwell, W. D. S., and Jones, W.; "Prediction and observation of isostructurality induced by solvent incorporation in multicomponent crystals.", *Journal of the American Chemical Society*, **2006**, 128, 45, 14466-14467.
- [49] Cruz-Cabeza, A. J., Karki, S., Fabian, L., Friscic, T., Day, G. M., and Jones, W.; "Predicting stoichiometry and structure of solvates", *Chemical Communications*, **2010**, 46, 13, 2224-2226.
- [50] D'Oria, E., Karamertzanis, P. G., and Price, S. L.; "Spontaneous Resolution of Enantiomers by Crystallization: Insights from Computed Energy Landscapes", *Crystal Growth & Design*, **2010**, 10, 1749-1756.
- [51] Darden, T., York, D., and Pedersen, L.; "Particle Mesh Ewald - an $n \cdot \log(n)$ Method for Ewald Sums in Large Systems", *Journal of Chemical Physics*, **1993**, 98, 12, 10089-10092.
- [52] Davey, R. J.; "Crystallization: Separation at Molecular and Supramolecular Levels", in Separation Technology. The Next Ten Years, Garside, J., The Institution of Chemical Engineers: UK, **1994**, 73-96.
- [53] Day, G. M.; "Current approaches to predicting molecular organic crystal structures", *Crystallography Reviews*, **2011**, 17, 1, 3-52.
- [54] Day, G. M., Chisholm, J., Shan, N., Motherwell, W. D. S., and Jones, W.; "Assessment of lattice energy minimization for the prediction of molecular organic crystal structures", *Crystal Growth & Design*, **2004**, 4, 6, 1327-1340.
- [55] Day, G. M. and Cooper, T. G.; "Crystal packing predictions of the alpha-amino acids: methods assessment and structural observations", *CrystEngComm*, **2010**, 12, 8, 2443-2453.
- [56] Day, G. M., Cooper, T. G., Cruz-Cabeza, A. J., Hejczyk, K. E., Ammon, H. L., Boerrigter, S. X. M., Tan, J., Della Valle, R. G., Venuti, E., Jose, J., Gadre, S. R., Desiraju, G. R., Thakur, T. S., van Eijck, B. P., Facelli, J. C., Bazterra, V. E., Ferraro, M. B., Hofmann, D. W. M., Neumann, M., Leusen, F. J. J., Kendrick, J., Price, S. L., Misquitta, A. J., Karamertzanis, P. G., Welch, G. W. A., Scheraga,

- H. A., Arnautova, Y. A., Schmidt, M. U., van de Streek, J., Wolf, A., and Schweizer, B.; "Significant progress in predicting the crystal structures of small organic molecules - a report on the fourth blind test", *Acta Crystallographica Section B - Structural Science*, **2009**, 65, 107-125.
- [57] Day, G. M., Motherwell, W. D. S., Ammon, H. L., Boerrigter, S. X. M., Della Valle, R. G., Venuti, E., Dzyabchenko, A., Dunitz, J. D., Schweizer, B., van Eijck, B. P., Erk, P., Facelli, J. C., Bazterra, V. E., Ferraro, M. B., Hofmann, D. W. M., Leusen, F. J. J., Liang, C., Pantelides, C. C., Karamertzanis, P. G., Price, S. L., Lewis, T. C., Nowell, H., Torrisi, A., Scheraga, H. A., Arnautova, Y. A., Schmidt, M. U., and Verwer, P.; "A third blind test of crystal structure prediction", *Acta Crystallographica Section B - Structural Science*, **2005a**, 61, 5, 511-527.
- [58] Day, G. M., Motherwell, W. D. S., and Jones, W.; "A strategy for predicting the crystal structures of flexible molecules: the polymorphism of phenobarbital", *Physical Chemistry Chemical Physics*, **2007**, 9, 14, 1693-1704.
- [59] Day, G. M., Motherwell, W. D. S., and Jones, W.; "Beyond the isotropic atom model in crystal structure prediction of rigid molecules: Atomic multipoles versus point charges", *Crystal Growth & Design*, **2005b**, 5, 3, 1023-1033.
- [60] Day, G. M., Price, S. L., and Leslie, M.; "Atomistic calculations of phonon frequencies and thermodynamic quantities for crystals of rigid organic molecules", *Journal of Physical Chemistry B*, **2003**, 107, 39, 10919-10933.
- [61] Day, G. M., Price, S. L., and Leslie, M.; "Elastic constant calculations for molecular organic crystals", *Crystal Growth & Design*, **2001**, 1, 1, 13-26.
- [62] Della Valle, R. G., Venuti, E., Brillante, A., and Girlando, A.; "Inherent structures of crystalline pentacene", *Journal of Chemical Physics*, **2003**, 118, 2, 807-815.
- [63] Dennis, J. E. and Schnabel, R. B.; *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, London, **1983**.
- [64] Desiraju, G. R.; "Crystal engineering: A brief overview", *Journal of Chemical Sciences*, **2010**, 122, 5, 667-675.
- [65] Desiraju, G. R.; "Supramolecular Synthons in Crystal Engineering - a New Organic Synthesis", *Angewandte Chemie-International Edition in English*, **1995**, 34, 21, 2311-2327.
- [66] Dewar, M. J. S., Ziegler, E. G., Healy, E. F., and Stewart, J. J. P.; "AM1: a new general purpose quantum mechanical molecular model.", *Journal of the American Chemical Society*, **1985**, 107, 3902-3909.
- [67] Dideberg, O., Dupont, L., and Campsteijn, H.; "Structure cristalline et moléculaire du complexe 1:1 progestérone-résorcinol", *Acta Crystallographica Section B - Structural Science*, **1975**, 31, 3, 637-640.

- [68] Dunitz, J. D.; "Are crystal structures predictable?", *Chemical Communications*, **2003**, 5, 545-548.
- [69] Dunitz, J. D.; *X-ray analysis and the structure of organic molecules*, Cornell University Press, Ithaca, **1979**.
- [70] Dunitz, J. D., Filippini, G., and Gavezzotti, A.; "Molecular shape and crystal packing: A study of C₁₂H₁₂ isomers, real and imaginary", *Helvetica Chimica Acta*, **2000**, 83, 9, 2317-2335.
- [71] Dzyabchenko, A. V.; "Theoretical Structures of Crystalline Benzene - the Search for A Global Minimum of the Lattice Energy in 4 Space-Groups", *Journal of Structural Chemistry*, **1984**, 25, 3, 416-420.
- [72] Dzyabchenko, A. V.; "Method of Crystal-Structure Similarity Searching", *Acta Crystallographica Section B - Structural Science*, **1994**, 50, Pt4, 414-425.
- [73] Dzyabchenko, A. V., Agafonov, V., and Davydov, V. A.; "A theoretical study of the pressure-induced dimerization of C- 60 fullerene", *Journal of Physical Chemistry A*, **1999**, 103, 15, 2812-2820.
- [74] Espinosa, E., Lecomte, C., Ghermani, N. E., Devemy, J., Rohmer, M. M., Benard, M., and Molins, E.; "Hydrogen bonds: First quantitative agreement between electrostatic potential calculations from experimental X-(X+N) and theoretical ab initio SCF models", *Journal of the American Chemical Society*, **1996**, 118, 10, 2501-2502.
- [75] Ewald, P.; "Die Berechnung optischer und elektrostatischer Gitterpotentiale", *Annals of Physics*, **1921**, 369, 253-287.
- [76] Fabbiani, F. P. A., Allan, D. R., David, W. I. F., Davidson, A. J., Lennie, A. R., Parsons, S., Pulham, C. R., and Warren, J. E.; "High-pressure studies of pharmaceuticals: An exploration of the behavior of piracetam", *Crystal Growth & Design*, **2007**, 7, 6, 1115-1124.
- [77] Fabbiani, F. P. A., Allan, D. R., Parsons, S., and Pulham, C. R.; "An exploration of the polymorphism of piracetam using high pressure", *CrystEngComm*, **2005**, 7, 29, 179-186.
- [78] Fabbiani, F. P. A. and Pulham, C. R.; "High-pressure studies of pharmaceutical compounds and energetic materials", *Chemical Society Reviews*, **2006**, 35, 10, 932-942.
- [79] Fowler, P. W. and Stone, A. J.; "Induced Dipole-Moments of Van der waals Complexes", *Journal of Physical Chemistry*, **1987**, 91, 3, 509-511.
- [80] Frenkel, D. and Smit, B.; "Understanding molecular simulation: form algorithms to applications", Academic Press, San Diego, London, **2002**, 2nd edition.
- [81] Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. A.,

- Nakatsuji, H., Caricato, M., Li, X., Hratchian, H. P., Izmaylov, A. F., Bloino, J., Zheng, G., Sonnenberg, J. L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery, J. A., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Rega, N., Millam, J. M., Klene, M., Knox, J. E., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Martin, R. L., Morokuma, K., Zakrzewski, V. G., Voth, G. A., Salvador, P., Dannenberg, J. J., Dapprich, S., Daniels, A. D., Farkas, O., Foresman, J. B., Ortiz, J. V., Cioslowski, J., and Fox, D. J.; "Gaussian 09", Gaussian, Inc., Wallingford CT, **2009**.
- [82] Friscic, T., Lancaster, R. W., Laszlo, F., and Karamertzanis, P. G.; "Tunable recognition of the steroid a-face by adjacent p-electron density", *Proceedings of the National Academy of Science of the United States of America*, **2010**, 107, 30, 13216-13221.
- [83] Galdecki, Z. and Glowka, M. L.; "Crystal-Structure of Nootropic Agent, Piracetam - 2-Oxopyrrolidin-1-Ylacetamide", *Polish Journal of Chemistry*, **1983**, 57, 10-1, 1307-1312.
- [84] Galek, P. T. A, Fabian, L., Motherwell, W. D. S., Allen, F. H., and Feeder, N.; "Knowledge-based model of hydrogen-bonding propensity in organic crystals", *Acta Crystallographica Section B - Structural Science*, **2007**, B63, 768-782.
- [85] Galek, P. T. A., Allen, F. H., Fabian, L., and Feeder, N.; "Knowledge-based H-bond prediction to aid experimental polymorph screening", *CrystEngComm*, **2009**, 11, 12, 2634-2639.
- [86] Gavezzotti, A.; "Calculation of intermolecular interaction energies by direct numerical integration over electron densities. 2. An improved polarization model and the evaluation of dispersion and repulsion energies", *Journal of Physical Chemistry B*, **2003a**, 107, 10, 2344-2353.
- [87] Gavezzotti, A.; "Calculation of intermolecular interaction energies by direct numerical integration over electron densities. I. Electrostatic and polarization energies in molecular crystals", *Journal of Physical Chemistry B*, **2002a**, 106, 16, 4145-4154.
- [88] Gavezzotti, A.; "Quantitative ranking of crystal packing modes by systematic calculations on potential energies and vibrational amplitudes of molecular dimers", *Journal of Chemical Theory and Computation*, **2005b**, 1, 5, 834-840.
- [89] Gavezzotti, A.; *Theoretical Aspects and Computer Modeling of the Molecular Solid State*, John Wiley, Chichester, **1997**.
- [90] Gavezzotti, A.; "Generation of Possible Crystal-Structures from the Molecular-Structure for Low-Polarity Organic-Compounds", *Journal of the American Chemical Society*, **1991**, 113, 12, 4622-4629.

- [91] Gavezzotti, A.; "Ten years of experience in polymorph prediction: what next?", *CrystEngComm*, **2002b**, 4, 61, 343-347.
- [92] Gavezzotti, A.; "Are Crystal-Structures Predictable?", *Accounts of Chemical Research*, **1994**, 27, 10, 309-314.
- [93] Gavezzotti, A.; "Zip-Promet", University of Milan, Univeristy of Milan, **1999**.
- [94] Gavezzotti, A.; "Calculation of lattice energies of organic crystals: the PIXEL integration method in comparison with more traditional methods", *Zeitschrift fur Kristallographie*, **2005a**, 220, 5-6, 499-510.
- [95] Gavezzotti, A.; "Structure and intermolecular potentials in molecular crystals", *Modelling and Simulation in Materials Science and Engineering*, **2002c**, 10, 3, R1-R29.
- [96] Gavezzotti, A.; "Towards a realistic model for the quantitative evaluation of intermolecular potentials and for the rationalization of organic crystal structures. Part I. Philosophy", *CrystEngComm*, **2003b**, 5, 76, 429-438.
- [97] Gavezzotti, A.; "Towards a realistic model for the quantitative evaluation of intermolecular potentials and for the rationalization of organic crystal structures. Part II. Crystal energy landscapes", *CrystEngComm*, **2003c**, 5, 77, 439-446.
- [98] Gavezzotti, A. and Filippini, G.; "Polymorphic Forms of Organic-Crystals at Room Conditions - Thermodynamic and Structural Implications", *Journal of the American Chemical Society*, **1995**, 117, 49, 12299-12305.
- [99] Gdanitz, R. J.; "Prediction of Molecular-Crystal Structures by Monte-Carlo Simulated Annealing Without Reference to Diffraction Data", *Chemical Physics Letters*, **1992**, 190, 3-4, 391-396.
- [100] Ghermani, N. E., Bouhmaida, N., and Lecomte, C.; "Modelling electrostatic potential from experimentally determined charge densities. I. Spherical-atom approximation", *Acta Crystallographica Section A - Crystal Physics, Diffraction, Theoretical and General Crystallography*, **1993**, A49, 781-789.
- [101] Gil-Villegas, A., Galindo, A., Whitehead, P. J., Mills, S. J., and Jackson, G.; "Statistical associating fluid theory for chain molecules with attractive potentials of variable charge", *Journal of Chemical Physics*, **1997**, 106, 10, 4168-4186.
- [102] Glusker, J. P., Lewis, M., and Rossi, M.; *Crystal Structure Analysis for Chemists and Biologists*, VCH, Cambridge, **1994**.
- [103] Görbitz, C. H., Dalhus, B., and Day, G. M.; "Pseudoracemic amino acid complexes: blind predictions for flexible two-component crystals", *Physical Chemistry Chemical Physics*, **2010**, 12, 8466-8477.
- [104] Gourlay, M. D., Kendrick, J., and Leusen, F. J. J.; "Rationalization of racemate resolution: Predicting spontaneous resolution through crystal structure prediction", *Crystal Growth & Design*, **2007**, 7, 1, 56-63.

- [105] Greengard, L. and Rokhlin, V.; "A fast algorithm for particle simulations", *Journal of Computational Physics*, **1987**, 73, 2, 325-348.
- [106] Grimme, S.; "Accurate description of van der Waals complexes by density functional theory including empirical corrections", *Journal of Computational Chemistry*, **2004**, 25, 12, 1463-1473.
- [107] Groom, C. R. and Allen, F. H.; "Crystal structure information in drug discovery and development: current perspectives and new possibilities from the Cambridge Crystallographic Data Centre", *Future Medicinal Chemistry*, **2010**, 2, 6, 933-939.
- [108] Hargittai, I. and Levy, J. B.; "Accessible geometrical changes", *Structural Chemistry*, **1999**, 10, 5, 387-389.
- [109] Harris, N. J. and Lammertsma, K.; "Ab initio density functional computations of conformations and bond dissociation energies for hexahydro-1,3,5-trinitro-1,3,5-triazine", *Journal of American Chemical Society*, **1997**, 119, 28, 6583-6589.
- [110] Hartman, P. and Bennema, P.; "The Attachment Energy as a Habit Controlling Factor 1. Theoretical Considerations", *Journal of Crystal Growth*, **1980**, 49, 1, 145-156.
- [111] Hartman, P. and Perdok, W. G.; "On the Relations Between Structure and Morphology of Crystals. I", *Acta Crystallographica*, **1955**, 8, 1, 49-52.
- [112] Hirano, T., Tsuzuki, S., Tanabe, K., and Tajima, N.; "Totally ab initio prediction of the structures of CO₂ molecular crystal", *Chemistry Letters*, **1995**, 12, 1073-1074.
- [113] Hockney, R. and Eastwood, J.; *Computer simulation using particles*, McGraw-Hill, New York, **1981**.
- [114] Hofmann, D. W. M. and Apostolakis, J.; "Crystal structure prediction by data mining", *Journal of Molecular Structure*, **2003**, 647, 1-3, 17-39.
- [115] Hofmann, D. W. M. and Lengauer, T.; "Prediction of crystal structures of organic molecules", *Journal of Molecular Structure*, **1999**, 474, 13-23.
- [116] Hofmann, D. W. M. and Lengauer, T.; "Crystal structure prediction based on statistical potentials", *Journal of Molecular Modeling*, **1998**, 4, 3, 132-144.
- [117] Hofmann, D. W. M. and Lengauer, T.; "A discrete algorithm for crystal structure prediction of organic molecules", *Acta Crystallographica Section A - Foundations of Crystallography*, **1997**, 53, 225-235.
- [118] Hulme, A. T., Johnston, A., Florence, A. J., Fernandes, P., Shankland, K., Bedford, C. T., Welch, G. W. A., Sadiq, G., Haynes, D. A., Motherwell, W. D. S., Tocher, D. A., and Price, S. L.; "Search for a predicted hydrogen bonding motif - A multidisciplinary investigation into the polymorphism of 3-

- azabicyclo[3.3.1]nonane-2,4-dione", *Journal of the American Chemical Society*, **2007**, 129, 12, 3649-3657.
- [119] Hulme, A. T. and Price, S. L.; "Towards the prediction of organic hydrate crystal structures", *Journal of Chemical Theory and Computation*, **2007**, 3, 4, 1597-1608.
- [120] Karamertzanis, P. G.; "Prediction of crystal structure of molecular solids", Imperial College, London, **2004**.
- [121] Karamertzanis, P. G., Day, G. M., Welch, G. W. A., Kendrick, J., Leusen, F. J. J., Neumann, M. A., and Price, S. L.; "Modeling the interplay of inter- and intra-molecular hydrogen bonding in conformational polymorphs", *Journal of Chemical Physics*, **2008a**, 128, art-244708.
- [122] Karamertzanis, P. G., Kazantsev, A. V., Issa, N., Welch, G. W. A., Adjiman, C. S., Pantelides, C. C., and Price, S. L.; "Can the Formation of Pharmaceutical Cocrystals Be Computationally Predicted? II. Crystal Structure Prediction.", *Journal of Chemical Theory and Computation*, **2009**, 5, 5, 1432-1448.
- [123] Karamertzanis, P. G. and Pantelides, C. C.; "Ab initio crystal structure prediction. II. Flexible molecules", *Molecular Physics*, **2007**, 105, 2-3, 273-291.
- [124] Karamertzanis, P. G. and Pantelides, C. C.; "Ab initio crystal structure prediction - I. Rigid molecules", *Journal of Computational Chemistry*, **2005**, 26, 3, 304-324.
- [125] Karamertzanis, P. G. and Pantelides, C. C.; "Optimal site charge models for molecular electrostatic potentials", *Molecular Simulation*, **2004**, 30, 7, 413-436.
- [126] Karamertzanis, P. G. and Price, S. L.; "Energy Minimization of Crystal Structures Containing Flexible Molecules", *Journal of Chemical Theory and Computation*, **2006**, 2, 4, 1184-1199.
- [127] Karamertzanis, P. G., Raiteri, P., Parrinello, M., Leslie, M., and Price, S. L.; "The Thermal Stability of Lattice Energy Minima of 5-Fluorouracil: Metadynamics as an Aid to Polymorph Prediction", *Journal of Physical Chemistry B*, **2008b**, 112, 14, 4298-4308.
- [128] Karfunkel, H. R. and Gdanitz, R. J.; "Ab initio Prediction of Possible Crystal-Structures for General Organic-Molecules", *Journal of Computational Chemistry*, **1992**, 13, 10, 1171-1183.
- [129] Karfunkel, H. R., Leusen, F. J. J., and Gdanitz, R. J.; "The ab initio prediction of yet unknown molecular crystal structures by solving the crystal packing problem", *Journal of Computer-Aided Molecular Design*, **1993**, 1, 177-785.
- [130] Kazantsev, A. V., Karamertzanis, P. G., Adjiman, C. S., and Pantelides, C. C.; "Efficient handling of molecular flexibility in lattice energy minimization of organic crystals", *Journal of Chemical Theory and Computation*, **2011a**, 7, 1998-2016.

- [131] Kazantsev, A. V., Karamertzanis, P. G., Adjiman, C. S., Pantelides, C. C., Price, S. L., Galek, P. T. A., Day, G. M., and Cruz-Cabeza, A. J.; "Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction", *International Journal of Pharmaceutics*, **2011b**, in press.
- [132] Kazantsev, A. V., Karamertzanis, P. G., Pantelides, C. C., and Adjiman, C. S.; "An Efficient Algorithm for Lattice Energy Minimization of Organic Crystals Using Isolated-Molecule Quantum Mechanical Calculations", in *Process Systems Engineering: Volume 6: Molecular Systems Engineering*, Adjiman, C. S. and Galindo, A., Wiley-VCH, Hamburg, **2010**, 1-42.
- [133] Kim, H. S. and Jeffrey, G. A.; "The crystal structure of xylitol", *Acta Crystallographica Section B - Structural Science*, **1969**, 25, 2607-2613.
- [134] Kim, S., Orendt, A. M., Ferraro, M. B., and Facelli, J. C.; "Crystal Structure Prediction of Flexible Molecules Using Parallel Genetic Algorithms with a Standard Force Field", *Journal of Computational Chemistry*, **2009**, 30, 13, 1973-1985.
- [135] Kitaigorodski, A. I.; *Organic Chemical Crystallography*, Consultants Bureau, New York, **1961**.
- [136] Kitaigorodski, A. I.; *Molecular Crystals and Molecules*, Academic Press, New York, **1973**.
- [137] Klooster, W. T., Swaminathan, S., Nanni, R., and Craven, B. M.; "Electrostatic Properties of 1-Methyluracil from Diffraction Data", *Acta Crystallographica Section B - Structural Science*, **1992**, 48, 217-227.
- [138] Koch, U., Popelier, P. L. A., and Stone, A. J.; "Conformational Dependence of Atomic Multipole Moments", *Chemical Physics Letters*, **1995**, 238, 4-6, 253-260.
- [139] Kopsias, N. P. and Theodorou, D. N.; "Elementary structural transitions in the amorphous Lennard-Jones solid using multidimensional transition-state theory", *Journal of Chemical Physics*, **1998**, 109, 19, 8573-8582.
- [140] Krivy, I. and Gruber, B.; "Unified Algorithm for Determining Reduced (Niggli) Cell", *Acta Crystallographica Section A - Crystal Physics, Diffraction, Theoretical and General Crystallography*, **1976**, 32, MAR1, 297-298.
- [141] L'Ecuyer, P.; "Efficient and Portable Combined Random Number Generators", *Communications of the Acm*, **1988**, 31, 6, 742-&.
- [142] Lancaster, R. W., Karamertzanis, P. G., Hulme, A. T., Tocher, D. A., Covey, D. F., and Price, S. L.; "Racemic progesterone: predicted in silico and produced in the solid state", *Chemical Communications*, **2006**, 47, 4921-4923.
- [143] Lehmann, C. W.; "Crystal Structure Prediction - Dawn of a New Era", *Angewandte Chemie-International Edition*, **2011**, 50, 2-4.

- [144] Lennard-Jones, J. E.; "On the determination of molecular fields", *Proceedings of the Royal Society of London*, **1924**, A106, 463-477.
- [145] Li, R. Y., Zeitler, J. A., Tomerini, D., Parrott, E. P. J., Gladden, L. F., and Day, G. M.; "A study into the effect of subtle structural details and disorder on the terahertz spectrum of crystalline benzoic acid", *Physical Chemistry Chemical Physics*, **2010**, 12, 20, 5329-5340.
- [146] 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., and Williams, D. E.; "A test of crystal structure prediction of small organic molecules", *Acta Crystallographica Section B - Structural Science*, **2000**, 56, 697-714.
- [147] Louer, D., Louer, M., Dzyabchenko, A. V., Agafonov, V., and Ceolin, R.; "Structure of a Metastable Phase of Piracetam From X-Ray-Powder Diffraction Using Atom-Atom Potential Method", *Acta Crystallographica Section B - Structural Science*, **1995**, 51, Pt2, 182-187.
- [148] Lucia, A. and Gattupalli, R. R.; "Ab Initio Crystal Structure Determination Using Global Optimization", AICHE Annual Meeting, Session #545 - Multi-Scale Modeling, Philadelphia, PA, USA, **2008**.
- [149] Maddox, J.; "Crystals from first principles", *Nature*, **1988**, 335, 6187, 201.
- [150] Martonak, R., Laio, A., Bernasconi, M., Ceriani, C., Raiteri, P., Zipoli, F., and Parrinello, M.; "Simulation of structural phase transitions by metadynamics", *Zeitschrift fur Kristallographie*, **2005**, 220, 5-6, 489-498.
- [151] Martonak, R., Oganov, A. R., and Glass, C. W.; "Crystal structure prediction and simulations of structural transformations: metadynamics and evolutionary algorithms", *Phase Transitions*, **2007**, 80, 4-5, 277-298.
- [152] Mayo, S. L., Olafson, B. D., and Goddard, W. A.; "Dreiding - A Generic Force-Field for Molecular Simulations", *Journal of Physical Chemistry*, **1990**, 94, 26, 8897-8909.
- [153] McCrone, W. C.; "Polymorphism", in *Physics and Chemistry of the Organic Solid State*, vol. II Fox, D., Labes, M., and Weissberger, M. M., Interscience, New York, **1965**, 726-727.
- [154] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H.; "Equation of State Calculations by Fast Computing Machines", *Journal of Chemical Physics*, **1953**, 21, 6, 1087-1092.
- [155] Misquitta, A. J., Welch, G. W. A., Stone, A. J., and Price, S. L.; "A first principles solution of the crystal structure of C₆Br₂ClFH₂", *Chemical Physics Letters*, **2008**, 456, 1-3, 105-109.
- [156] Mitchell, J. B. O. and Price, S. L.; "A systematic nonempirical method of deriving model intermolecular potentials for organic molecules: Application to amides", *Journal of Physical Chemistry A*, **2000**, 104, 46, 10958-10971.

- [157] Mohamed, S., Tocher, D. A., Vickers, M., Karamertzanis, P. G., and Price, S. L.; "Salt or Cocrystal? A New Series of Crystal Structures Formed from Simple Pyridines and Carboxylic Acids", *Crystal Growth & Design*, **2009**, 9, 6, 2881-2889.
- [158] Mooij, W. T. M. and Leusen, F. J. J.; "Multipoles versus charges in the 1999 crystal structure prediction test", *Physical Chemistry Chemical Physics*, **2001**, 3, 22, 5063-5066.
- [159] Mooij, W. T. M., van Duijneveldt, F. B., van Duijneveldt-van de Rijdt, J. G. C. M., and van Eijck, B. P.; "Transferable ab initio intermolecular potentials. 1. Derivation from methanol dimer and trimer calculations", *Journal of Physical Chemistry A*, **1999a**, 103, 48, 9872-9882.
- [160] Mooij, W. T. M., van Eijck, B. P., and Kroon, J.; "Ab initio crystal structure predictions for flexible hydrogen- bonded molecules", *Journal of the American Chemical Society*, **2000**, 122, 14, 3500-3505.
- [161] Mooij, W. T. M., van Eijck, B. P., and Kroon, J.; "Transferable ab initio intermolecular potentials. 2. Validation and application to crystal structure prediction", *Journal of Physical Chemistry A*, **1999b**, 103, 48, 9883-9890.
- [162] Motherwell, W. D. S.; "Crystal structure prediction and the Cambridge Structural Database", *Molecular Crystals and Liquid Crystals*, **2001**, 356, 559-567.
- [163] Motherwell, W. D. S., Ammon, H. L., Dunitz, J. D., Dzyabchenko, A., Erk, P., Gavezzotti, A., Hofmann, D. W. M., Leusen, F. J. J., Lommerse, J. P. M., Mooij, W. T. M., Price, S. L., Scheraga, H., Schweizer, B., Schmidt, M. U., van Eijck, B. P., Verwer, P., and Williams, D. E.; "Crystal structure prediction of small organic molecules: a second blind test", *Acta Crystallographica Section B - Structural Science*, **2002**, 58, 647-661.
- [164] Mulliken, R. S.; "Electronic Population Analysis on Lcao-Mo Molecular Wave Functions .1", *Journal of Chemical Physics*, **1955**, 23, 10, 1833-1840.
- [165] Nelder, J. A. and Mead, R.; "A SIMPLEX-method for function minimisation", *Computer Journal*, **1965**, 7, 4, 308-313.
- [166] Neumann, M. A.; "Tailor-made force fields for crystal-structure prediction.", *Journal of Physical Chemistry B*, **2008**, 112, 32, 9810-9829.
- [167] Neumann, M. A.; "GRACE", *Avant-garde Materials Simulation*, **2007**.
- [168] Neumann, M. A., Leusen, F. J. J., and Kendrick, J.; "A Major Advance in Crystal Structure Prediction", *Angewandte Chemie-International Edition*, **2008**, 47, 13, 2427-2430.
- [169] Neumann, M. A. and Perrin, M. A.; "Energy ranking of molecular crystals using density functional theory calculations and an empirical van der Waals correction", *Journal of Physical Chemistry B*, **2005**, 109, 32, 15531-15541.

- [170] Nobeli, I. and Price, S. L.; "A Non-Empirical Intermolecular Potential for Oxalic Acid Crystal Structures", *Journal of Physical Chemistry A*, **1999**, 103, 32, 6448-6457.
- [171] Nyburg, S. C. and Faerman, C. H.; "A Revision of van der Waals Atomic Radii For Molecular Crystals - N, O, F, S, Cl, Se, Br and I Bonded to Carbon", *Acta Crystallographica Section B - Structural Science*, **1985**, 41, AUG, 274-279.
- [172] Nyvlt, J.; "The Ostwald Rule of Stages", *Crystal Research and Technology*, **1995**, 30, 4, 443-449.
- [173] Parrott, E. P. J., Zeitler, J. A., Friscic, T., Pepper, M., Jones, W., Day, G. M., and Gladden, L. F.; "Testing the Sensitivity of Terahertz Spectroscopy to Changes in Molecular and Supramolecular Structure: A Study of Structurally Similar Cocystals", *Crystal Growth & Design*, **2009**, 9, 3, 1452-1460.
- [174] Pertsin, A. J. and Kitaigorodski, A. I.; *The Atom-Atom Potential Method. Applications to Organic Molecular Solids*, Springer-Verlag, Berlin, **1987**.
- [175] Pillardy, J., Arnautova, Y. A., Czaplewski, C., Gibson, K. D., and Scheraga, H. A.; "Conformation-family Monte Carlo: A new method for crystal structure prediction", *Proceedings of the National Academy of Sciences of the United States of America*, **2001**, 98, 22, 12351-12356.
- [176] Pillardy, J., Czaplewski, C., Wedemeyer, W. J., and Scheraga, H. A.; "Conformation-Family Monte Carlo (CFMC): An Efficient Computational Method for identifying the Low-energy States of a Macromolecule", *Helvetica Chimica Acta*, **2000**, 83, 2214-2230.
- [177] Polito, M., D'Oria, E., Maini, L., Karamertzanis, P. G., Grepioni, F., Braga, D., and Price, S. L.; "The Crystal Structures of Chloro and Methyl Ortho-Benzoic Acids and Their Co-crystal: Rationalizing Similarities and Differences", *CrystEngComm*, **2008**, 10, 1848-1854.
- [178] Potoff, J. J. and Bernard-Brunel, D. A.; "Mie potentials for Phase Equilibria Calculations: Application to Alkanes and Perfluoroalkanes", *Journal of Physical Chemistry B*, **2009**, 113, 14725-14731.
- [179] Price, S. L.; "From crystal structure prediction to polymorph prediction: interpreting the crystal energy landscape", *Physical Chemistry Chemical Physics*, **2008b**, 10, 15, 1996-2009.
- [180] Price, S. L.; "Computational prediction of organic crystal structures and polymorphism", *International Reviews in Physical Chemistry*, **2008a**, 27, 3, 541-568.
- [181] Price, S. L.; "Computed Crystal Energy Landscapes for Understanding and Predicting Organic Crystal Structures and Polymorphism", *Accounts of Chemical Research*, **2009**, 42, 1, 117-126.
- [182] Price, S. L.; "Towards More Accurate Model Intermolecular Potentials for Organic Molecules", in *Reviews in Computational Chemistry* 14, vol. 14

- Lipkowitz, K. B. and Boyd, D. B., John Wiley & Sons, New York, **2000**, 225-289.
- [183] Price, S. L.; "Quantifying intermolecular interactions and their use in computational crystal structure prediction", *CrystEngComm*, **2004a**, 6, 344-353.
- [184] Price, S. L.; "The computational prediction of pharmaceutical crystal structures and polymorphism", *Advanced Drug Delivery Reviews*, **2004b**, 56, 3, 301-319.
- [185] Price, S. L., Leslie, M., Welch, G. W. A., Habgood, M, Price, L. S., Karamertzanis, P. G., and Day, G. M.; "Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials", *Physical Chemistry Chemical Physics*, **2010**, 12, 8478-8490.
- [186] Price, S. L. and Stone, A. J.; "Electrostatic Models For Polypeptides: Can We Assume Transferability?", *Journal of the Chemical Society-Faraday Transactions*, **1992**, 88, 13, 1755-1763.
- [187] Raiteri, P., Martonak, R., and Parrinello, M.; "Exploring Polymorphism: The Case of Benzene", *Angewandte Chemie-International Edition*, **2005**, 44, 3769-3773.
- [188] Rapaport, D.; *The art of molecular dynamics simulation*, 2nd edn, Cambridge University Press, Cambridge, **2003**.
- [189] Reed, T. M. and Gubbins, K. E.; "Applied statistical mechanics: thermodynamic and transport properties of fluids", McGraw-Hill, Tokyo, **1973**.
- [190] Rovira, C. and Novoa, J. J.; "A first-principles computation of the low-energy polymorphic forms of the acetic acid crystal. A test of the atom-atom force field predictions", *Journal of Physical Chemistry B*, **2001**, 105, 9, 1710-1719.
- [191] Sakai, K., Sakurai, K., Nohira, H., Tanaka, R., and Hirayama, N.; "Practical resolution of 1-phenyl-2-(4-methylphenyl)ethylamine using a single resolving agent controlled by the dielectric constant of the solvent", *Tetrahedron: Asymmetry*, **2004**, 15, 3495-3500.
- [192] Salager, E., Day, G. M., Stein, R. S., Pickard, C. J., Elena, B., and Emsley, L.; "Powder Crystallography by Combined Crystal Structure Prediction and High-Resolution H-1 Solid-State NMR Spectroscopy", *Journal of the American Chemical Society*, **2010**, 132, 8, 2564-2566.
- [193] Santiso, E. E. and Trout, B. L.; "A general set of order parameters for molecular crystals", *Journal of Chemical Physics*, **2011**, 134, 6, 064109.
- [194] Sarma, J. A. R. P. and Desiraju, G. R.; "The supramolecular synthon approach to crystal structure prediction", *Crystal Growth & Design*, **2002**, 2, 2, 93-100.
- [195] Schmidt, M. U. and Englert, U.; "Prediction of crystal structures", *Journal of the Chemical Society-Dalton Transactions*, **1996**, 10, 2077-2082.

- [196] Schmidt, M. U., Paulus, E. F., and Day, G. M.; "Experimental and predicted crystal structures of Pigment Red 168 and other dihalogenated anthanthrones", *Acta Crystallographica Section B*, **2010**, 66, 515-526.
- [197] Schmidt, M. W., Baldrige, K. K., Boatz, J. A., Elbert, S. T., Gordon, M. S., Jensen, J. H., Koseki, S., Matsunaga, N., Nguyen, K. A., Su, S., Windus, T. L., Dupuis, M., and Montgomery, J. A.; "General Atomic and Molecular Electronic Structure Systems", *Journal of Computational Chemistry*, **1993**, 14, 1347-1363.
- [198] Shimada, J., Kaneko, H., and Takada, T.; "Performance of fast multipole methods for calculating electrostatic interactions in biomacromolecular simulations", *Journal of Computational Chemistry*, **1994**, 15, 1, 28-43.
- [199] Sigfridsson, E. and Ryde, U.; "Comparison of Methods for Deriving Atomic Charges from the Electrostatic Potential and Moments", *Journal of Computational Chemistry*, **1998**, 19, 4, 377-395.
- [200] Snyder, R. C. and Doherty, M. F.; "Predicting crystal growth by spiral motion", *Proceedings of the Royal Society of London, Series A*, **2009**, 465, 2104, 1145-1171.
- [201] Snyder, R. L. and Rosenstein, R. D.; "The crystal and molecular structure of the 1:1 hydrogen bond complex between [alpha]-D-glucose and urea", *Acta Crystallographica Section B - Structural Crystallography and Crystal Chemistry*, **1971**, 27, 10, 1969-1975.
- [202] Sobol', I. M.; "On the distribution of points in a cube and the approximate evaluation of integrals", *Computational Mathematics and Mathematical Physics*, **1967**, 7, 86-112.
- [203] Spek, A. L.; "PLATON, A Multipurpose Crystallographic Tool", Utrecht University, Utrecht, The Netherlands, **2003**.
- [204] Stefanovic, J. and Pantelides, C. C.; "Molecular dynamics as a mathematical mapping. III. Efficient evaluation of the differentiable force functions and their derivatives", *Molecular Simulation*, **2001**, 26, 5, 323-352.
- [205] Steiner, T.; "Frequency of Z' values in organic and organometallic crystal structures", *Acta Crystallographica Section B - Structural Science*, **2000**, 56, 673-676.
- [206] Stone, A. J.; "Distributed multipole analysis: Stability for large basis sets", *Journal of Chemical Theory and Computation*, **2005**, 1, 6, 1128-1132.
- [207] Stone, A. J.; *The Theory of Intermolecular Forces*, 1st edn, Clarendon Press, Oxford, **1996**.
- [208] Stone, A. J. and Alderton, M.; "Distributed Multipole Analysis - Methods and Applications", *Molecular Physics*, **1985**, 56, 5, 1047-1064.
- [209] Stone, A. J., Dullweber, A., Engkvist, O., Fraschini, E., Hodges, M. P., Meredith, A. W., Nutt, D. R., Popelier, P. L. A., and Wales, D. J.; "ORIENT: a

- program for studying interactions between molecules", University of Cambridge, **2006**.
- [210] Stone, A. J. and Misquitta, A. J.; "Atom-atom potentials from ab initio calculations", *International Reviews in Physical Chemistry*, **2007**, 26, 1, 193-222.
- [211] Stout, G. H. and Jensen, L. H.; *X-Ray Structure Determination. A Practical Guide*, 2nd Edition edn, John Wiley and Son, **1989**.
- [212] Tajima, N., Tanaka, T., Arikawa, T., Sakurai, T., Teramae, S., and Hirano, T.; "A Heuristic Molecular-Dynamics Approach For the Prediction of a Molecular-Crystal Structure", *Bulletin of the Chemical Society of Japan*, **1995**, 68, 2, 519-527.
- [213] Tajima, N., Tsuzuki, S., Tanabe, K., Aoki, K., and Hirano, T.; "First principles prediction of crystal structures of CO₂", *Electronic Journal of Theoretical Chemistry*, **1997**, 2, 139-148.
- [214] Tomasi, J., Mennucci, B., and Cammi, R.; "Quantum Mechanical Continuum Solvation Models", *Chemical Reviews*, **2005**, 105, 2999-3093.
- [215] Toukmaji, A. Y. and Board, J. A.; "Ewald summation techniques in perspective: A survey", *Computer Physics Communications*, **1996**, 95, 2-3, 73-92.
- [216] Tremayne, M., Grice, L., Pyatt, J. C., Seaton, C. C., Kariuki, B. M., Tsui, H. H. Y., Price, S. L., and Cherryman, J. C.; "Characterization of complicated new polymorphs of chlorothalonil by X-ray diffraction and computer crystal structure prediction", *Journal of the American Chemical Society*, **2004**, 126, 22, 7071-7081.
- [217] Trimarchi, G. and Zunger, A.; "Global space-group optimization problem: Finding the stablest crystal structure without constraints", *Physical Review B*, **2007**, 75, 10, 104113.
- [218] Tsui, H. H. Y. and Price, S. L.; "A non-empirical method of determining atom-atom repulsion parameters: application to the crystal structure prediction of an oxyboryl derivative", *CrystEngComm*, **1999**, 1, 1, 7.
- [219] van de Streek, J.; "Some post-analysis for compound XX", *personal communication*, **2011**.
- [220] van Eijck, B. P.; "Comparing hypothetical structures generated in the third Cambridge blind test of crystal structure prediction", *Acta Crystallographica Section B - Structural Science*, **2005**, 61, 528-535.
- [221] van Eijck, B. P.; "Ab initio crystal structure predictions for flexible hydrogen-bonded molecules. Part III. Effect of lattice vibrations", *Journal of Computational Chemistry*, **2001**, 22, 8, 816-826.
- [222] van Eijck, B. P.; "Crystal structure predictions for disordered halobenzenes", *Physical Chemistry Chemical Physics*, **2002**, 4, 19, 4789-4794.

- [223] van Eijck, B. P. and Kroon, J.; "UPACK program package for crystal structure prediction: Force fields and crystal structure generation for small carbohydrate molecules", *Journal of Computational Chemistry*, **1999**, 20, 8, 799-812.
- [224] van Eijck, B. P. and Kroon, J.; "Structure predictions allowing more than one molecule in the asymmetric unit (vol 56, pg 535, 2000)", *Acta Crystallographica Section B - Structural Science*, **2000**, 56, 745.
- [225] van Eijck, B. P. and Kroon, J.; "Fast clustering of equivalent structures in crystal structure prediction", *Journal of Computational Chemistry*, **1997**, 18, 8, 1036-1042.
- [226] van Eijck, B. P., Mooij, W. T. M., and Kroon, J.; "Attempted Prediction of the Crystal-Structures of 6 Monosaccharides", *Acta Crystallographica Section B - Structural Science*, **1995**, 51, 99-103.
- [227] van Eijck, B. P., Mooij, W. T. M., and Kroon, J.; "Ab initio crystal structure predictions for flexible hydrogen- bonded molecules. Part II. Accurate energy minimization", *Journal of Computational Chemistry*, **2001**, 22, 8, 805-815.
- [228] van Mourik, T., Karamertzanis, P. G., and Price, S. L.; "Molecular conformations and relative stabilities can be as demanding of the electronic structure method as intermolecular calculations", *Journal of Physical Chemistry A*, **2006**, 110, 1, 8-12.
- [229] Vasileiadis, M., Kazantsev, A. V., Adjiman, C. S., and Pantelides, C. C.; "Ab Initio Prediction of ROY polymorphs", *Crystal Growth & Design*, **2011**, in preparation.
- [230] Verwer, P. and Leusen, F. J. J.; "Computer Simulation to Predict Possible Polymorphs", in *Reviews in Computational Chemistry Volume 12*, Lipkowitz, K. B. and Boyd, D. B., John Wiley and Sons Inc., New York, **1998**, 327-365.
- [231] Vishweshwar, P., McMahon, J. A., Peterson, M. L., Hickey, M. B., Shattock, T. R., and Zaworotko, M. J.; "Crystal engineering of pharmaceutical co-crystals from polymorphic active pharmaceutical ingredients", *Chemical Communications*, **2005**, 36, 4601-4603.
- [232] Warshel, A., Huler, E., Rabinovich, D., and Shakked, Z.; "Examination of Intramolecular Potential Surfaces of Flexible Conjugated Molecules by Calculation of Crystal Structures. Equilibrium Geometries of Chalcones and Diphenyloctatetraene in the Crystal and Gaseous State", *Journal of Molecular Structure*, **1974**, 23, 2, 175-191.
- [233] Wawak, R. J., Pillardy, J., Liwo, A., Gibson, K. D., and Scheraga, H. A.; "Diffusion equation and distance scaling methods of global optimization: Applications to crystal structure prediction", *Journal of Physical Chemistry A*, **1998**, 102, 17, 2904-2918.
- [234] Weissbuch, I., Popovitzbiro, R., Lahav, M., and Leiserowitz, L.; "Understanding and Control of Nucleation, Growth, Habit, Dissolution and Structure of 2-Dimensional and 3-Dimensional Crystals Using Tailor-Made

- Auxiliaries", *Acta Crystallographica Section B - Structural Science*, **1995**, 51, 2, 115-148.
- [235] Welch, G. W. A., Karamertzanis, P. G., Misquitta, A. J., Stone, A. J., and Price, S. L.; "Is the induction energy important for modeling organic crystals?", *Journal of Chemical Theory and Computation*, **2008**, 4, 3, 522-532.
- [236] Williams, D. E.; "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, 11, 1154-1166.
- [237] Williams, D. E.; "Nonbonded Interatomic Potential-Energy Functions and Prediction of Crystal-Structures", *Acta Crystallographica Section A - Foundations of Crystallography*, **1984**, 40, C95.
- [238] Williams, D. E. and Cox, S. R.; "Nonbonded Potentials For Azahydrocarbons: the Importance of the Coulombic Interaction", *Acta Crystallographica Section B - Structural Science*, **1984**, 40, AUG, 404-417.
- [239] Williams, D. E. and Weller, R. R.; "Lone-Pair Electronic Effects On the Calculated Ab Initio Scf-Mo Electric-Potential and the Crystal-Structures of Azabenzenes", *Journal of the American Chemical Society*, **1983**, 105, 13, 4143-4148.
- [240] Willock, D. J., Price, S. L., Leslie, M., and Catlow, C. R. A.; "The Relaxation of Molecular Crystal Structures Using a Distributed Multipole Electrostatic Model", *Journal of Computational Chemistry*, **1995**, 16, 5, 628-647.
- [241] Wilson, A. J. C.; "Space-Groups Rare for Organic Structures .3. Symmorphisms and Inherent Molecular Symmetry", *Acta Crystallographica Section A - Foundations of Crystallography*, **1993**, 49, 795-806.
- [242] Yu, L.; "Survival of the fittest polymorph: how faster nucleator can lose to fast grower", *CrystEngComm*, **2007**, 9, 847-851.
- [243] Yu, L.; "Color Changes Caused by Conformational Polymorphism: Optical-Crystallography, Single-Crystal Spectroscopy, and Computational Chemistry", *Journal of Physical Chemistry A*, **2002**, 106, 3, 544-550.
- [244] Yu, L., Stephenson, G. A., Mitchell, C. A., Bunnell, C. A., Snorek, S. V., Bowyer, J. J., Borchardt, T. B., Stowell, J. G., and Byrn, S. R.; "Thermochemistry and Conformational Polymorphism of a Hexamorphic Crystal System", *Journal of the American Chemical Society*, **2000**, 122, 4, 585-591.

Appendix A

Proposed Z-matrix generation procedure

The Z-matrix generation procedure adopted in this work is outlined using the example of ethyl ethanoate (Figure A.1).

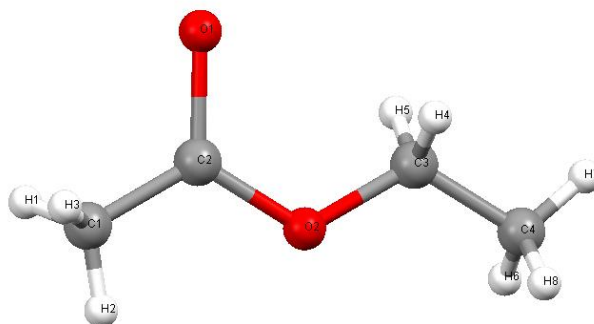


Figure A.1 Molecular diagram and atom labelling of ethyl ethanoate.

The connectivity of atoms is first established by providing the Cartesian coordinates of a sample conformation (Table A.1) and standard bond length data. The first atom in the Z-matrix is assigned to the atom which has the most atoms bonded to it. If there are two or more atoms with the same number of connections, the atom that came first in the input is chosen. Hence, the order of atoms in the Z-matrix (Table A.2) does not necessarily correspond to the order of atoms specified in the input (Table A.1). For instance, in the case of ethyl ethanoate, both atoms C1 and C4 are connected to 4 other atoms (Figure A.1), however atom C1 is chosen as the first atom in the Z-matrix as it comes first in the user-specified input (Table A.1). Next, atoms in decreasing order of connectivity are considered, but preference is given to atoms which are directly bonded to already defined atoms¹. The procedure is repeated until all atoms have been defined or further progress cannot be made.

In a general case, all atoms are usually defined with respect to atoms that are all bonded together. For example, in the case of ethyl ethanoate, atom C4 is fully defined relative to atoms C3, O2, and C2 (Table A.2), which form a connected aliphatic chain (Figure A.1). However, if the angle C4-C3-O2 happened to be linear (180°), the torsion C4-C3-O2-C2 would not be defined. As a result, the proposed Z-matrix generation

¹ To avoid any ambiguity for the atoms that have the same connectivity, the atoms are chosen according to the order of the atoms in the input (Table A.1).

procedure avoids linear angles by defining an improper bond angle and torsion, by switching the order of the last to atoms. For the previous hypothetical example, the improper bond angle would be defined relative to atoms C4-C3-C2, and the improper torsion described relative to C4-C3-C2-O2.

Table A.1 Gas-phase minimum conformation, computed at the MP2/6-31G(d,p) level of theory, used to generate the Z-matrix for ethyl ethanoate (refer to Figure A.1 for atom labels).

#	Atom label	Cartesian coordinates (Å)		
		x	y	z
1	C1	0.0000	0.0000	0.0000
2	C2	0.0000	0.0000	1.5045
3	C3	1.2726	0.0000	1.9791
4	C4	1.3744	-0.0005	3.4221
5	O1	2.8473	-0.0003	3.7543
6	O2	-0.9873	-0.0001	2.2197
7	H1	-1.0238	0.0013	-0.3584
8	H2	0.5269	0.8782	-0.3673
9	H3	0.5245	-0.8797	-0.3672
10	H4	0.8640	-0.8806	3.8125
11	H5	0.8637	0.8793	3.8131
12	H6	3.3312	0.8833	3.3441
13	H7	3.3316	-0.8834	3.3435
14	H8	2.9833	-0.0006	4.8349

Table A.2 Generated Z-matrix for the gas-phase geometry of ethyl ethanoate (Table A.1). Refer to Figure A.1 for molecular diagram and atom labels.

#	Atom	Atom 2	Bond length (Å)	Atom 3	Bond angle (°)	Atom 4	Torsion angle (°)
1	C1						
2	C2	C1	1.505				
6	O2	C2	1.358	C1	110.45		
3	C3	O2	1.447	C2	114.48	C1	-179.98
4	C4	C3	1.510	O2	106.74	C2	-179.99
5	O1	C2	1.219	O2	123.63	C2	0.01
7	H1	C1	1.085	C2	109.29	O1	-179.93
8	H2	C1	1.088	C2	109.73	O1	-59.04
9	H3	C1	1.088	C2	109.72	O1	59.19
10	H4	C3	1.090	O2	108.94	C2	58.62
11	H5	C3	1.090	O2	108.95	C2	-58.61
12	H6	C4	1.088	C3	110.55	O2	60.15
13	H7	C4	1.088	C3	110.56	O2	-60.15
14	H8	C4	1.089	C3	109.88	O2	180.00

During the derivation of the Z-matrix all internal coordinates (bond lengths, angles and torsions) are calculated from the Cartesian coordinates of atoms provided for the initial conformation (Table A.1). The description of the mathematical formulation used to calculate the required parameters is provided in detail below.

Bond lengths

A bond length (l) is the distance between two atoms directly bonded together. Mathematically, it is the magnitude of the interatomic vector between atoms A and B and can be calculated by the application of the Pythagoras' theorem:

$$l_{BA} = \|\mathbf{r}_{BA}\| = \|\mathbf{r}_A - \mathbf{r}_B\| = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2} \quad \text{Eq. A.1}$$

where $\mathbf{r}_A$ and $\mathbf{r}_B$ are the Cartesian coordinate vectors of atoms A and B respectively.

Bond angles

A bond angle, ν_{ABC} , is defined as the angle between three atoms A-B-C (Figure A.2).

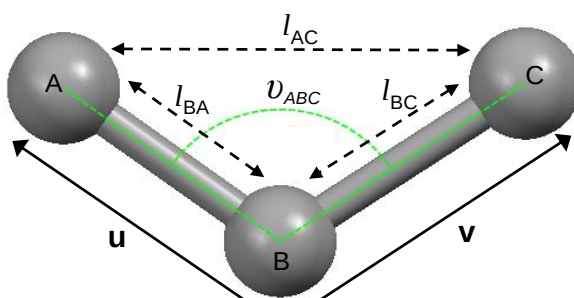


Figure A.2 Bond angle defined with respect to three atoms A-B-C.

The bond angle, ν_{ABC} , can be calculated using the cosine rule:

$$\cos \nu_{ABC} = \frac{l_{BA}^2 + l_{BC}^2 - l_{AC}^2}{2l_{BA}l_{BC}} \quad \text{Eq. A.2}$$

where the bond distances l_{BA} , l_{BC} and l_{AC} (shown in Figure A.2) are calculated using equation D.1.

Alternatively, and computationally more efficiently, the bond angles can be calculated from the dot product of the two vectors:

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| \cos \phi, \text{ where } \phi = \nu_{ABC} \text{ or } \xi_{ABCD} \quad \text{Eq. A.3}$$

where vectors $\mathbf{u}$ and $\mathbf{v}$ are defined using the Cartesian coordinates of atoms A, B and C (Figure A.2):

$$\mathbf{u} = \mathbf{r}_{BA} = \mathbf{r}_A - \mathbf{r}_B \quad \text{Eq. A.4}$$

$$\mathbf{v} = \mathbf{r}_{BC} = \mathbf{r}_C - \mathbf{r}_B \quad \text{Eq. A.5}$$

the magnitudes of vectors $\mathbf{u}$ and $\mathbf{v}$ are the corresponding bond lengths and are calculated using equation A.1.

Dihedral angles

A torsion angle measures the conformational twist about a series of four atoms, A-B-C-D. The atoms do not necessarily have to be bonded together. The torsion angle is defined as the angle between the directions A-B and C-D as viewed down BC (Glusker et al. 1994). Mathematically, a torsional angle is the angle between the normal directions of the planes through ABC and BCD.

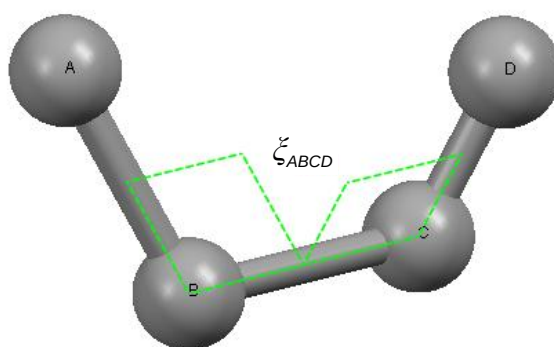


Figure A.3 Torsion angle defined with respect to four atoms A-B-C-D.

By convention, a clockwise rotation corresponds to a positive torsion value and torsion D-C-B-A is identical in magnitude and sign to A-B-C-D (Dunitz 1979). Furthermore, the torsion angles for equivalent sets of atoms in a pair of enantiomers have equal magnitudes but opposite signs. The torsion angles are not defined when one or both bond angles are 0° or 180°. As a result, the Z-matrix generation procedure avoids near-linear bond angles by redefining the Z-matrix with respect to different atoms.

In order to calculate the torsional angle, the normal directions of the planes through ABC and BCD have to be computed first. The normal directions can be obtained using the vector cross product:

$$\mathbf{n}_{ABC} = \mathbf{r}_{BA} \times \mathbf{r}_{BC} \quad \text{Eq. A.6}$$

$$\mathbf{n}_{DCB} = \mathbf{r}_{DC} \times \mathbf{r}_{CB} \quad \text{Eq. A.7}$$

where interatomic vectors $\mathbf{r}_{BA}, \mathbf{r}_{BC}, \mathbf{r}_{CB}$ and $\mathbf{r}_{DC}$ are calculated from the Cartesian coordinates using equation A.1. The torsional angle, ζ , is then obtained through the dot product of the normal directions defined in equation A.3 where:

$$\mathbf{u} = \mathbf{n}_{ABC} \quad \text{Eq. A.8}$$

$$\mathbf{v} = \mathbf{n}_{DCB} \quad \text{Eq. A.9}$$

Alternatively the dihedral angle, ξ_{ABCD} , can be calculated using an equivalent definition of the plane from three non-colinear vectors, $\mathbf{r}_{AB}, \mathbf{r}_{BC}$ and $\mathbf{r}_{CD}$, and the two-argument function atan2 , which returns the angle between the positive x-axis of a plane and the point provided by the coordinates (x, y) (Glusker et al. 1994):

$$\xi_{ABCD} = \text{atan2}(y, x) = 2 \arctan \frac{y}{\sqrt{x^2 + y^2} + x} \quad \text{Eq. A.10}$$

where the coordinates x and y are obtained from:

$$x = [\mathbf{r}_{AB} \times \mathbf{r}_{BC}] \cdot [\mathbf{r}_{BC} \times \mathbf{r}_{CD}] \quad \text{Eq. A.11}$$

$$y = \|\mathbf{r}_{BC}\| \mathbf{r}_{AB} \cdot [\mathbf{r}_{BC} \times \mathbf{r}_{CD}] \quad \text{Eq. A.12}$$

Appendix B

Derivation of the approximation of the “rigid” degrees of freedom as an explicit linear function of the flexible degrees of freedom

Assume an equation involving two sets of vector variables $\mathbf{x}$ and $\mathbf{y}$, such that it is non-singular with respect to the variable $\mathbf{x}$:

$$f(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0} \quad \text{Eq. B.1}$$

The function can be approximated with the first order Taylor expansion:

$$f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}_0, \mathbf{y}_0) + \left. \frac{\partial f}{\partial \mathbf{x}} \right|_{\mathbf{x}_0, \mathbf{y}_0} (\mathbf{x} - \mathbf{x}_0) + \left. \frac{\partial f}{\partial \mathbf{y}} \right|_{\mathbf{x}_0, \mathbf{y}_0} (\mathbf{y} - \mathbf{y}_0) \quad \text{Eq. B.2}$$

Consider changing the initial variables by a small amount, $\delta \mathbf{x}$ and $\delta \mathbf{y}$, so that the equation B.1 is still satisfied, i.e.:

$$f(\mathbf{x}_0 + \delta \mathbf{x}, \mathbf{y}_0 + \delta \mathbf{y}) = \mathbf{0} \quad \text{Eq. B.3}$$

Using the Taylor expansion in equation B.2, the above can be expanded up to the first order terms to obtain:

$$\begin{aligned} f(\mathbf{x}_0 + \delta \mathbf{x}, \mathbf{y}_0 + \delta \mathbf{y}) &= f(\mathbf{x}_0, \mathbf{y}_0) + \left. \frac{\partial f}{\partial \mathbf{x}} \right|_{\mathbf{x}_0, \mathbf{y}_0} ([\mathbf{x}_0 + \delta \mathbf{x}] - \mathbf{x}_0) \\ &\quad + \left. \frac{\partial f}{\partial \mathbf{y}} \right|_{\mathbf{x}_0, \mathbf{y}_0} ([\mathbf{y}_0 + \delta \mathbf{y}] - \mathbf{y}_0) \\ &= 0 \end{aligned} \quad \text{Eq. B.4}$$

Because of equation B.1, the first term on the right hand side of the above equation is zero. The following equation can be obtained by simplifying further:

$$\left. \frac{\partial f}{\partial \mathbf{x}} \right|_{\mathbf{x}_0, \mathbf{y}_0} (\delta \mathbf{x}) + \left. \frac{\partial f}{\partial \mathbf{y}} \right|_{\mathbf{x}_0, \mathbf{y}_0} (\delta \mathbf{y}) = 0 \quad \text{Eq. B.5}$$

Solving for $\delta \mathbf{x}$, we obtain:

$$\delta \mathbf{x} = - \left[\left. \frac{\partial f}{\partial \mathbf{x}} \right|_{\mathbf{x}_0, \mathbf{y}_0} \right]^{-1} \left[\left. \frac{\partial f}{\partial \mathbf{y}} \right|_{\mathbf{x}_0, \mathbf{y}_0} \right] \delta \mathbf{y} \quad \text{Eq. B.6}$$

Noting that $\frac{\delta \mathbf{x}}{\delta \mathbf{y}}$ is the definition of the partial derivative of variable $\mathbf{x}$ with respect to variable $\mathbf{y}$, the above equation can be reformulated as:

$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}} = - \left[\left. \frac{\partial f}{\partial \mathbf{x}} \right|_{\mathbf{x}_0, \mathbf{y}_0} \right]^{-1} \left[\left. \frac{\partial f}{\partial \mathbf{y}} \right|_{\mathbf{x}_0, \mathbf{y}_0} \right] \quad \text{Eq. B.7}$$

Applying the derivation of the general equation to our particular case where:

$$\mathbf{x} = \boldsymbol{\theta}^r \quad \text{Eq. B.8}$$

$$\mathbf{y} = \boldsymbol{\theta}^f \quad \text{Eq. B.9}$$

And f is the first-order optimality condition defined by equation 3.4, which must hold in all cases:

$$f = \left[\frac{\partial E^{\text{intra}}}{\partial \boldsymbol{\theta}^r} \right] = 0 \quad \text{Eq. B.10}$$

Note that the above matrix is non-singular since it is the Hessian of the objective function E^{intra} with respect to the optimisation variables $\boldsymbol{\theta}^r$ which is supposed to be positive definite at the minimum, satisfying the condition for $\mathbf{x}_0$ in equation B.1.

Furthermore, by noting that $\boldsymbol{\theta}_{\text{ref}} \equiv (\boldsymbol{\theta}_{\text{ref}}^f, \boldsymbol{\theta}_{\text{ref}}^r)$:

$$\left. \frac{\partial f}{\partial \mathbf{x}} \right|_{\mathbf{x}_0, \mathbf{y}_0} = \left. \frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{r2}} \right|_{\boldsymbol{\theta}_{ref}} \quad \text{Eq. B.11}$$

$$\left. \frac{\partial f}{\partial \mathbf{y}} \right|_{\mathbf{x}_0, \mathbf{y}_0} = \left. \frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right|_{\boldsymbol{\theta}_{ref}}^T \quad \text{Eq. B.12}$$

We can define how the rigid degrees of freedom change with the flexible degrees of freedom using equation B.7:

$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}} \equiv \frac{\partial \boldsymbol{\theta}^r}{\partial \boldsymbol{\theta}^f} = - \left[\left. \frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{r2}} \right|_{\boldsymbol{\theta}_{ref}} \right]^{-1} \left[\left. \frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right|_{\boldsymbol{\theta}_{ref}} \right]^T \quad \text{Eq. B.13}$$

The rigid degrees of freedom can then be approximated by an explicit linear function of $\boldsymbol{\theta}^f$ via a first order Taylor expansion:

$$\boldsymbol{\theta}^r(\boldsymbol{\theta}^f) = \boldsymbol{\theta}_{ref}^r + \frac{\partial \boldsymbol{\theta}^r}{\partial \boldsymbol{\theta}^f} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \quad \text{Eq. B.14}$$

Substituting equation B.13 into equation B.14 we finally obtain the local approximate model for the rigid degrees of freedom as a function of the flexible degrees of freedom:

$$\boldsymbol{\theta}^r(\boldsymbol{\theta}^f) = \boldsymbol{\theta}_{ref}^r - \left[\left. \frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{r2}} \right|_{\boldsymbol{\theta}_{ref}} \right]^{-1} \left[\left. \frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right|_{\boldsymbol{\theta}_{ref}} \right]^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \quad \text{Eq. B.15}$$

Appendix C

An alternative derivation of the approximation of the “rigid” degrees of freedom as an explicit linear function of the flexible degrees of freedom

The intramolecular energy for a given conformation in close proximity to a reference conformation, $\boldsymbol{\theta}_{ref} \equiv (\boldsymbol{\theta}_{ref}^f, \boldsymbol{\theta}_{ref}^r)$, can be estimated using a local approximate model (LAM) based on a quadratic Taylor expansion:

$$\begin{aligned}
 \Delta E^{intra}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r) = & \Delta E^{intra}(\boldsymbol{\theta}_{ref}) + \left[\frac{\partial E^{intra}}{\partial \boldsymbol{\theta}^f} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \\
 & + \left[\frac{\partial E^{intra}}{\partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{ref}^r) \\
 & + \frac{1}{2} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f)^T \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{f2}} \right]_{\boldsymbol{\theta}_{ref}} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \\
 & + (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{ref}^r)^T \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \\
 & + \frac{1}{2} (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{ref}^r)^T \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{ref}} (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{ref}^r)
 \end{aligned} \tag{Eq. C.1}$$

The above equation is valid for *any* reference point $\boldsymbol{\theta}_{ref} \equiv (\boldsymbol{\theta}_{ref}^f, \boldsymbol{\theta}_{ref}^r)$ in the coordinate space. The values of $\boldsymbol{\theta}_{ref}^r$ at the reference point are obtained via an isolated-molecule quantum-mechanical constrained optimisation for the fixed values $\boldsymbol{\theta}^f = \boldsymbol{\theta}_{ref}^f$. This calculation also yields the minimum molecular deformation energy, $\Delta E^{intra}(\boldsymbol{\theta}_{ref})$, and the first- and second- order derivatives of ΔE^{intra} with respect to all intramolecular degrees of freedom, $\boldsymbol{\theta}_{ref}$, necessary to construct the LAM. Since $\boldsymbol{\theta}_{ref}^r$ is obtained by minimising ΔE^{intra} , it must satisfy the first-order optimality condition:

$$\left[\frac{\partial E^{intra}}{\partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}^r, \boldsymbol{\theta}_{ref}^f} = \mathbf{0} \tag{Eq. C.2}$$

We require that equation C.2 also apply to the intramolecular energy LAM (equation C.1). Hence, differentiating equation C.1 with respect to the rigid degrees of freedom and setting the left-hand side equal to zero yields:

$$\begin{aligned}
 \mathbf{0} = & \frac{\partial}{\partial \boldsymbol{\theta}^r} \left\{ \Delta E^{\text{intra}}(\boldsymbol{\theta}_{\text{ref}}^f) \right\} + \frac{\partial}{\partial \boldsymbol{\theta}^r} \left\{ \left[\frac{\partial E^{\text{intra}}}{\partial \boldsymbol{\theta}^f} \right]_{\boldsymbol{\theta}_{\text{ref}}^f}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) \right\} \\
 & + \frac{\partial}{\partial \boldsymbol{\theta}^r} \left\{ \left[\frac{\partial E^{\text{intra}}}{\partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}^r}^T (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r) \right\} \\
 & + \frac{\partial}{\partial \boldsymbol{\theta}^r} \left\{ \frac{1}{2} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f)^T \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{f2}} \right]_{\boldsymbol{\theta}_{\text{ref}}^f} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) \right\} \\
 & + \frac{\partial}{\partial \boldsymbol{\theta}^r} \left\{ (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r)^T \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) \right\} \\
 & + \frac{\partial}{\partial \boldsymbol{\theta}^r} \left\{ \frac{1}{2} (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r)^T \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{\text{ref}}^r} (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r) \right\}
 \end{aligned} \tag{Eq. C.3}$$

Only the last two terms remain non-zero, which after simplification become:

$$\left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) + \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{\text{ref}}^r} (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r) = \mathbf{0} \tag{Eq. C.4}$$

Assuming the matrix is of full rank, the rigid degrees of freedom can thus be expressed explicitly as a function of the flexible degrees of freedom by rearranging the above equation:

$$\boldsymbol{\theta}^r(\boldsymbol{\theta}^f) = \boldsymbol{\theta}_{\text{ref}}^r - \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{\text{ref}}^r}^{-1} \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) \tag{Eq. C.5}$$

Appendix D

Derivation of the intramolecular energy LAM as an explicit function of the flexible degrees of freedom

Given the quadratic Taylor expansion of the intramolecular energy in close proximity of a reference conformation, $\boldsymbol{\theta}_{ref} \equiv (\boldsymbol{\theta}_{ref}^f, \boldsymbol{\theta}_{ref}^r)$:

$$\begin{aligned}
 \Delta E^{intra}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r) = & \Delta E^{intra}(\boldsymbol{\theta}_{ref}) + \left[\frac{\partial E^{intra}}{\partial \boldsymbol{\theta}^f} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \\
 & + \left[\frac{\partial E^{intra}}{\partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{ref}^r) \\
 & + \frac{1}{2} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f)^T \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{f2}} \right]_{\boldsymbol{\theta}_{ref}} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \\
 & + (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{ref}^r)^T \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \\
 & + \frac{1}{2} (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{ref}^r)^T \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{ref}} (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{ref}^r)
 \end{aligned} \tag{Eq. D.1}$$

Substituting the first-order optimality condition, which the intramolecular energy LAM needs to satisfy:

$$\left[\frac{\partial E^{intra}}{\partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}^r, \boldsymbol{\theta}_{ref}^f} = \mathbf{0} \tag{Eq. D.2}$$

The third term in equation D.1 vanishes:

$$\begin{aligned}
\Delta E^{\text{intra}}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r) = & \Delta E^{\text{intra}}(\boldsymbol{\theta}_{\text{ref}}) + \left[\frac{\partial E^{\text{intra}}}{\partial \boldsymbol{\theta}^f} \right]_{\boldsymbol{\theta}_{\text{ref}}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) \\
& + \frac{1}{2} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f)^T \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{f2}} \right]_{\boldsymbol{\theta}_{\text{ref}}} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) \\
& + (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r)^T \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) \\
& + \frac{1}{2} (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r)^T \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{\text{ref}}} (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r)
\end{aligned} \tag{Eq. D.3}$$

Given the (slightly rearranged) approximation of the rigid degrees of freedom as an explicit linear function of the flexible degrees of freedom:

$$(\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r) = - \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{\text{ref}}}^{-1} \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) \tag{Eq. D.4}$$

In order to remove the dependency of the equation D.3 on the rigid degrees of freedom, $\boldsymbol{\theta}^r$, equation D.4 can be substituted into the last two terms of equation D.3. Each term will be taken in turn.

Starting with the penultimate term in equation D.3:

$$(\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r)^T \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) \tag{Eq. D.5}$$

This can be rewritten as:

$$(\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f)^T \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}} (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{\text{ref}}^r) \tag{Eq. D.6}$$

Substituting equation C.4 into the above yields:

$$(\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f)^T \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}} \left(- \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{ref}}^{-1} \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \right) \quad \text{Eq. D.7}$$

This can be simplified as:

$$-(\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f)^T \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}} \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{ref}}^{-1} \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \quad \text{Eq. D.8}$$

Now, taking the last term in equation D.3:

$$\frac{1}{2} (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{ref}^r)^T \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{ref}} (\boldsymbol{\theta}^r - \boldsymbol{\theta}_{ref}^r) \quad \text{Eq. D.9}$$

Substituting equation D.4 into the above yields:

$$\begin{aligned} & \frac{1}{2} \left(- \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{ref}}^{-1} \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \right)^T \times \\ & \quad \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{ref}} \times \\ & \quad \left(- \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{ref}}^{-1} \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \right) \end{aligned} \quad \text{Eq. D.10}$$

The negative signs cancel out and one of the matrices is multiplied with its inverse to obtain:

$$\frac{1}{2} \left(\left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{ref}}^{-1} \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \right)^T \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \quad \text{Eq. D.11}$$

By noting that the Hessian matrices (and their inverse matrices) are of full rank and symmetric meaning that $(A)^T=A$, and that the matrix algebra dictates that $(AB)^T = B^T A^T$ and $(A^T)^T=A$, the above equation can be simplified further to:

$$\frac{1}{2}(\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}} \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{ref}}^{-1} \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \quad \text{Eq. D.12}$$

Now substituting the two terms (equations D.8 and D.12) back into the Taylor expansion of equation D.3, we obtain:

$$\begin{aligned} \Delta E^{intra}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r) &= \Delta E^{intra}(\boldsymbol{\theta}_{ref}) + \left[\frac{\partial E^{intra}}{\partial \boldsymbol{\theta}^f} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \\ &+ \frac{1}{2}(\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f)^T \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{f2}} \right]_{\boldsymbol{\theta}_{ref}} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \\ &- (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f)^T \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}} \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{ref}}^{-1} \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \\ &+ \frac{1}{2}(\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}} \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{ref}}^{-1} \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \end{aligned} \quad \text{Eq. D.13}$$

The last two terms can be added together to obtain:

$$\begin{aligned} \Delta E^{intra}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r) &= \Delta E^{intra}(\boldsymbol{\theta}_{ref}) + \left[\frac{\partial E^{intra}}{\partial \boldsymbol{\theta}^f} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \\ &+ \frac{1}{2}(\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f)^T \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{f2}} \right]_{\boldsymbol{\theta}_{ref}} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \\ &- \frac{1}{2}(\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}} \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{ref}}^{-1} \left[\frac{\partial^2 E^{intra}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{ref}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{ref}^f) \end{aligned} \quad \text{Eq. D.14}$$

Finally, the last two terms can be further factorised to obtain the final form of the intramolecular energy LAM:

$$\begin{aligned}
\Delta E^{\text{intra}}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^r) = & \Delta E^{\text{intra}}(\boldsymbol{\theta}_{\text{ref}}) + \left[\frac{\partial E^{\text{intra}}}{\partial \boldsymbol{\theta}^f} \right]_{\boldsymbol{\theta}_{\text{ref}}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) \\
& + \frac{1}{2} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f)^T \times \\
& \left[\left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{f2}} \right]_{\boldsymbol{\theta}_{\text{ref}}} - \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}} \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{\text{ref}}}^{-1} \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}}^T \right] \\
& \times (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f)
\end{aligned} \tag{Eq. D.15}$$

The above can be rewritten as:

$$\begin{aligned}
\Delta E^{\text{intra}}(\boldsymbol{\theta}^f) = & \Delta E^{\text{intra}}(\boldsymbol{\theta}_{\text{ref}}^f) + \mathbf{b}(\boldsymbol{\theta}_{\text{ref}})^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) \\
& + \frac{1}{2} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f)^T \mathbf{C}(\boldsymbol{\theta}_{\text{ref}}) (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f)
\end{aligned} \tag{Eq. D.16}$$

where the vector $\mathbf{b}(\boldsymbol{\theta}_{\text{ref}})$ and matrix $\mathbf{C}(\boldsymbol{\theta}_{\text{ref}})$ are defined as:

$$\mathbf{b}(\boldsymbol{\theta}_{\text{ref}}) \equiv \left[\frac{\partial E^{\text{intra}}}{\partial \boldsymbol{\theta}^f} \right]_{\boldsymbol{\theta}_{\text{ref}}} \tag{Eq. D.17}$$

$$\mathbf{C}(\boldsymbol{\theta}_{\text{ref}}) \equiv \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{f2}} \right]_{\boldsymbol{\theta}_{\text{ref}}} - \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}} \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{r2}} \right]_{\boldsymbol{\theta}_{\text{ref}}}^{-1} \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^f \partial \boldsymbol{\theta}^r} \right]_{\boldsymbol{\theta}_{\text{ref}}}^T \tag{Eq. D.18}$$

It should be noted that if the vector $\boldsymbol{\theta}^r$ is empty, i.e. the molecule is treated atomistically, the second term on the right hand side of equation D.18 disappears and the intramolecular energy LAM in equation D.16 reduces to the standard second-order Taylor expansion around $\boldsymbol{\theta}^f$:

$$\begin{aligned}
\Delta E^{\text{intra}}(\boldsymbol{\theta}^f) = & \Delta E^{\text{intra}}(\boldsymbol{\theta}_{\text{ref}}^f) + \left[\frac{\partial E^{\text{intra}}}{\partial \boldsymbol{\theta}^f} \right]_{\boldsymbol{\theta}_{\text{ref}}}^T (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f) + \\
& \frac{1}{2} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f)^T \left[\frac{\partial^2 E^{\text{intra}}}{\partial \boldsymbol{\theta}^{f2}} \right]_{\boldsymbol{\theta}_{\text{ref}}} (\boldsymbol{\theta}^f - \boldsymbol{\theta}_{\text{ref}}^f)
\end{aligned} \tag{Eq. D.19}$$

Appendix E

Description of molecular rotation through quaternions

Three Euler angles (ϕ , θ and ψ) contain the minimum required information that can uniquely describe the rotation of molecules (Frenkel and Smit 2002). However, from the computational point of view, trigonometric functions involving these variables are more expensive to calculate than simple polynomial expressions. Hence, an alternative definition of the orientation of molecules is often adopted in molecular simulations through the use of quaternions (Rapaport 2003). A quaternion, $\mathbf{Q}$, is a set of four scalar quantities:

$$\mathbf{Q} = [q_0, q_1, q_2, q_3]^T \quad \text{Eq. E.1}$$

which satisfy the constraint $\sum_{i=0}^3 q_i^2 = 1$. The quaternions can be expressed as explicit functions of the three Euler angles:

$$q_0 = \cos\left(\frac{\theta}{2}\right) \cos\left(\frac{\phi + \psi}{2}\right) \quad \text{Eq. E.2}$$

$$q_1 = \sin\left(\frac{\theta}{2}\right) \cos\left(\frac{\phi - \psi}{2}\right) \quad \text{Eq. E.3}$$

$$q_2 = \sin\left(\frac{\theta}{2}\right) \sin\left(\frac{\phi - \psi}{2}\right) \quad \text{Eq. E.4}$$

$$q_3 = \cos\left(\frac{\theta}{2}\right) \sin\left(\frac{\phi + \psi}{2}\right) \quad \text{Eq. E.5}$$

The rotation matrix can then be expressed in terms of the quaternions:

$$\mathbf{A} = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1 q_2 + q_0 q_3) & 2(q_1 q_3 - q_0 q_2) \\ 2(q_1 q_2 - q_0 q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2 q_3 + q_0 q_1) \\ 2(q_1 q_3 + q_0 q_2) & 2(q_2 q_3 - q_0 q_1) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix} \quad \text{Eq. E.6}$$

The quaternions of the molecular conformation (q_0, q_1, q_2, q_3), and hence the Euler angles, can be derived from the elements of the above rotation matrix (Karamertzanis 2004). The above formulation provides an efficient means of calculating the initial Euler angles from a given molecular conformation and the rotation matrix in the course of lattice energy minimisation where Euler angles are explicit minimisation variables.

Appendix F

Several low-energy hypothetical structures for the BMS molecule

Table F.1 Crystal structures and molecular conformations of several plausible single-component crystal forms of the BMS molecule. Rank, lattice energy, density, space group and unit cell parameters are also provided for every structure.

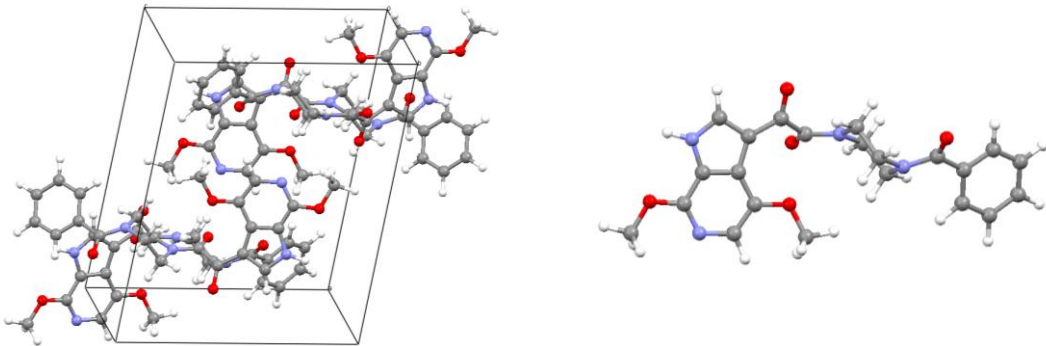
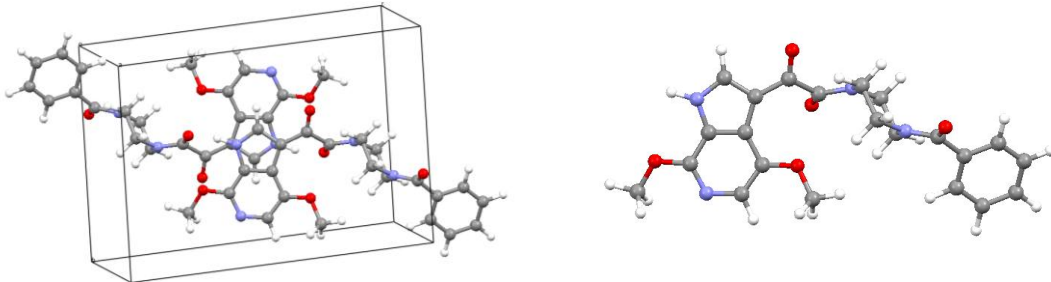
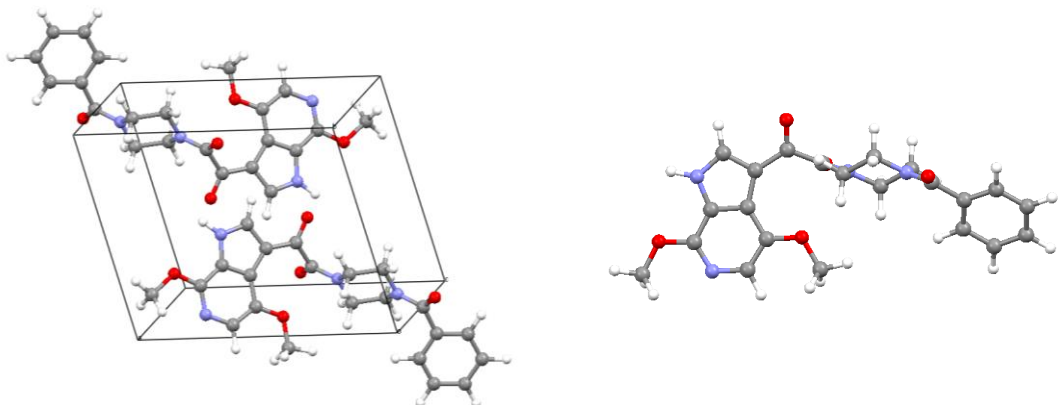
and cell parameters are also provided for every structure.					
1 (global minimum)	-197.93 kJ mol ⁻¹	1.3425 g cm ⁻³	<i>P2₁/n</i>		
					
a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)
14.78	11.54	12.42	90.0	80.6	90.0
3	-192.79 kJ mol ⁻¹	1.3336 g cm ⁻³	<i>P1̄</i>		
					
a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)
7.36	10.99	16.24	92.5	57.3	76.6

Table F.1 continued.

4	-192.38 kJ mol ⁻¹	1.3426 g cm ⁻³	$P\bar{1}$		
					
a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)
12.77	9.53	9.01	82.9	78.9	77.1

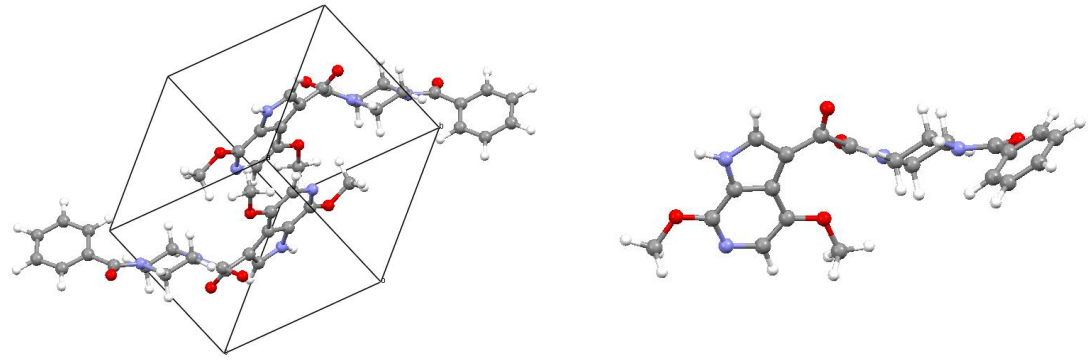
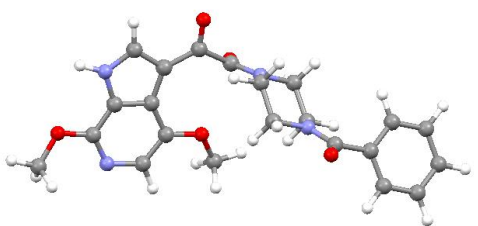
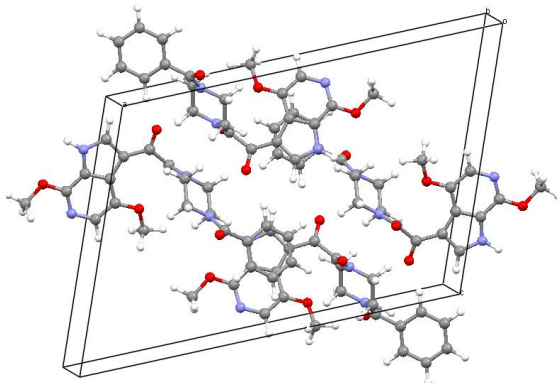
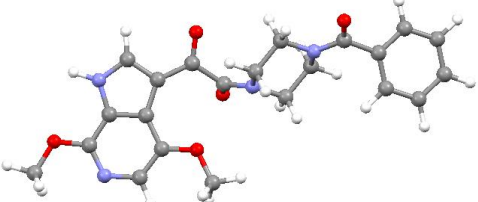
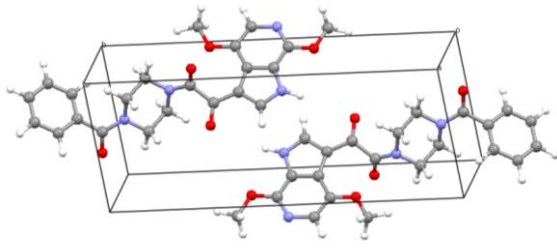
5	-190.90 kJ mol ⁻¹	1.3166 g cm ⁻³	$P\bar{1}$		
					
a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)
7.78	12.31	12.17	72.3	79.0	75.3

Table F.1 continued.

6	-188.52 mol ⁻¹		1.3336 g cm ⁻³		P2 ₁ /n
					
a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)
21.18	7.17	14.82	90.0	69.2	90.0

7	-187.95 kJ mol ⁻¹		1.3077 g cm ⁻³		P1̄
					
a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)
8.63	18.50	6.85	99.5	82.3	81.0