
IMPERIAL

Department of Physics
Matter Community

Physically Informed Machine Learning for Functional Materials

Author:

Keerati Keeratikarn

November 21, 2025

Supervisors:

Dr Jarvist Moore Frost

Department of Chemistry
Imperial College London
80-92 Wood Lane
London W12 7TA

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy
of Imperial College London

Keeratikarn, Keerati:

Physically Informed Machine Learning for Functional Materials

PhD Thesis, Imperial College London

London, 2025.

Statement of Originality

The thesis titled "Physically Informed Machine Learning for Functional Materials" submitted to Imperial College London for the degree of Doctor of Philosophy is my original work and has not been submitted for any other degree or qualification at any other institution. I, Keerati PK Keeratikarn, hereby declare this assertion.

This thesis is the result of my independently conducted research and analysis, and I confirm that all sources and references have been appropriately acknowledged. I have maintained the academic integrity and ethical standards that Imperial College London expects, and all contributions from other sources have been appropriately acknowledged.

I accept that this work may be subject to plagiarism and AI detection software and hereby verify that the content of this thesis is devoid of any form of academic dishonesty.

Copyright Statement

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Licence (CC BY-NC-ND 4.0).

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

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

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

Abstract

This thesis investigates the incorporation of physically informed machine learning techniques—particularly Gaussian Process (GP) regression [1] and Atomic Cluster Expansion (ACE) [2]—to effectively study and model the complex physical behaviours of functional materials [3; 4]. Understanding the electronic, thermal, and optoelectronic properties and applications of these materials requires accurate atomistic modelling. The atomistic modelling continues to pose computational difficulties due to the sophisticated anharmonic effects in crystalline inorganic systems and the complex electronic coupling dependencies in organic semiconductors.

The first section presents a derivative Gaussian Process framework [3; 5; 6; 7; 8], (GPFC.JL [9; 10]) that effectively evaluates harmonic and cubic anharmonic lattice force constants directly from data obtained through density functional theory (DFT) [11; 12; 13; 14]. This GP-based approach significantly improves computational efficiency, surpassing conventional finite-displacement techniques in predicting phonon dispersion, lifetimes, and thermal conductivity. Our model assesses multiple descriptor strategies, encompassing Cartesian and phonon-based representations, emphasising their influence on accuracy and computational efficiency.

The second section enhances the Atomic Cluster Expansion methodology [2] to precisely predict charge transfer integrals essential for charge transport in organic semiconductor materials. ACE models (ACEPOTENTIAL.JL [15]), employing systematically developed symmetry-adapted descriptors [16], effectively capture

the electronic coupling's dependence on molecular orientation and separation in various molecular dimers, including naphthalene, thiophene, ethylene, and their fluorinated derivatives. Our results underscore the significance of descriptor selection and illustrate enhanced model generalisation across chemical systems. Our research collectively establishes a unified methodology that integrates rigorous physical insights with sophisticated machine learning techniques. This integration greatly enhances the computational resources for predicting material properties, facilitating efficient exploration and precise optimisation of innovative functional materials. The thesis concludes by delineating future research avenues, encompassing computational improvements, sophisticated descriptor formulation, and applications to non-equilibrium and real-time dynamic processes.

Acknowledgements

I would like to extend my most heartfelt appreciation to all those who have provided me with assistance during my doctoral studies and thesis.

Initially, I am profoundly appreciative of my supervisors, Dr. Jarvist Moore Frost. I am grateful for your unwavering support, encouragement, and invaluable guidance during this research. Your expert advice and insightful feedback are greatly appreciated. You have played a critical role in my development as an academic and in my motivation as a scientist.

I am grateful to the Ministry of Education, Thailand, for providing me with a fully funded scholarship that has spanned my entire academic career, the Development and Promotion of Science and Technology Talent project (DPST).

I am grateful to Prof. Christoph Ortner, Dr. William Chuck Witt, and Cheuk Hin Ho (Jerry) for their unwavering support. They consistently offer assistance, provide guidance on how to apply their coding program to my research, and provide a comprehensive understanding.

I am grateful to the Imperial Physics and Chemistry Departments for providing the requisite resources and a supportive environment for my research. In the same vein, I am extremely grateful for the administrative and technical assistance provided by the Matter Group.

I am unable to express my gratitude to the Frost group members, both past and present, for their unwavering support and encouragement throughout the duration of this PhD. I am particularly grateful to Bradley Andrew Alan Martin,

Hanbo Yang, Lucius Liu, and Jack Coker for accompanying me on this academic journey, particularly during the COVID-19 pandemic. I will always cherish the time we have shared, not only for academic-related activities but also for Friday game nights, particularly during the initial phase of my PhD.. You all have been so supportive, helpful, and have made me chuckle so much. I am grateful to Yande, Ingvars, Ryo, and the new Frost Group cohort for contributing to the second half of my PhD. I always wish for Yande's residence when we shared hotpot.

I am deeply grateful to my family and friends for their unwavering support, patience, and understanding throughout the difficult periods of this PhD. Their unwavering affection and encouragement provided me with the determination and fortitude to persist.

I extend thanks to the Southamton community for consistently hosting Thai meal parties, and I also extend my gratitude to the DOTA2 team for providing an unparalleled level of relaxation and enjoyment.

I would like to express my deepest gratitude to my exceptional partner, Jasmine (Ja). Over the past four and a half years, I am deeply grateful for your unwavering and sincere love and support. I am unable to do so without your presence. I have a deep affection for you.

Lastly, I would like to express my gratitude to my mom, Nus, and my dad, Sung. Every step of the way, you have both been physically and remotely present, consistently believed in me, and consistently provided me with support. I have a deep affection for both of you. I am grateful to both of you for your assistance.

Contents

Abstract

List of Publications

List of Figures

List of Tables

1 Introductions, Motivations and Outlines

1.1	Introductions and Motivations	1
1.1.1	Gaussian Process Force Constants for Lattice Dynamics . .	2
1.1.2	Machine Learning for Charge Transfer Integrals	4
1.2	Thesis Outline	5

2 Theoretical background

2.1	Phonons and (an)harmonic potentials	8
2.1.1	Lattice Force Constants	8
2.1.2	Phonon Calculations	12
2.2	Density Functional Theory	20
2.2.1	Theoretical Foundations	21
2.2.2	Computational methodologies	22

2.3	Gaussian Processes	26
2.3.1	Conventional Model	26
2.3.2	Reviews on Gaussian Processes for Molecules and Materials	31
2.3.3	Derivative Model	32
2.4	Inductive Bias: Equivariant Descriptors	35
2.4.1	Reviews on Atomistic Descriptors	37
2.4.2	GAP & SOAP	39
2.5	Atomic Cluster Expansion	48
2.5.1	Canonical Cluster Expansion	48
2.5.2	Self-interaction Cluster Expansion	50
2.5.3	Many-body Product Basis	51
2.5.4	Rotation Invariance	53
2.5.5	Multi-element System & Learnable Radial Basis	57
2.6	Charge Transfer Integrals	60
2.6.1	Tight-binding Hamiltonian	60
2.6.2	Transfer Integrals	62
2.6.3	Computational Approaches	66
2.6.4	Machine Learning-Based Approaches	70
3	Anharmonic Phonons with Gaussian Processes	
3.1	GPFC.JL with Cartesian descriptors	74
3.1.1	Higher Derivative GP Model	74
3.1.2	Dataset Preparation & Training	80
3.1.3	Anharmonic Phonons from GPFCs	84
3.2	GPFC.JL with Phonon descriptors	89

3.2.1	Phonon descriptors	89
3.2.2	Phonon-basis GPFC	91
3.2.3	Recovering Real Space FCs	93
3.3	Atomic Cluster Expansion Model	95
3.3.1	ACE descriptors	96
3.3.2	ACE fitting	101
3.3.3	Phonons from Linear ACE data	102
3.3.4	Active learning with Linear ACE data	103
3.4	Conclusion and Future work	104
 4 ACE-Transfer Integrals		
4.1	Transfer Integrals in Organic Semiconductors	108
4.2	ACE Model Incorporation	111
4.3	Naphthalene Dimer	113
4.4	Ethylene and Thiophene Dimers	123
4.5	Conclusions and Future Work	132
 5 Conclusion and Future directions		
 A Detailed Derivations and Results		
A.1	Differentiable GP Kernels	137
A.1.1	Arbitrary Order Derivatives	137
A.1.2	Complex Gaussian Process	145
A.2	GPFC Results	151
A.2.1	Phonon Dispersion and Density of States	151
A.2.2	Phonon Lifetime and Lattice Thermal Conductivity	152

A.2.3	Acoustic Sumrule	153
A.2.4	HiPHIVE	156
A.3	Fine-tuning ACE Transfer Integrals	158
A.3.1	Interaction-order	158
A.3.2	totaldegree	160
A.3.3	rcut	162
A.3.4	q parameter	165
A.3.5	p parameter	168

B Bibliography

List of Publications

- Keerati Keeratikarn, Jarvist Moore Frost *Anharmonic phonons with Gaussian processes*, arXiv:2405.05113, 2024. (Submitted to IOP Science: Machine Learning: Science and Technology)
- Keerati Keeratikarn, Jarvist Moore Frost *Machine learning intermolecular transfer integrals with compact atomic cluster representations*, arXiv:2511.06551, 2025. (Submitted to Physical Review B)

List of Figures

2.1	The infinite one-dimensional diatomic chain consists of an infinite number of unit cells. In the l^{th} unit cell, two atoms with masses m_1 (yellow) and m_2 (green) are labeled by $\kappa = 1$ and $\kappa = 2$, respectively.	8
2.2	Schematic representation of harmonic and anharmonic force interactions in a lattice.	10
2.3	Phonon dispersion relation for a diatomic chain.	13
2.4	Diagrams show invariant and equivariant symmetry properties of functions f and f^* respectively.	35
2.5	Examples of the invariance and equivariance for the prediction of a molecule's properties.	36
2.6	This figure shows the 2- and 3-body order clusters for a 4-body system ($J = 4$).	49
2.7	These charts, presented by Ortner in GAP & (M)ACE 2024 workshop, show the allowed j_1, j_2 (yellow circles) for the restricted sums in a <i>Canonical</i> 3-body's cluster (left chart) and the unrestricted sums in an <i>Self-interaction</i> 3-body's cluster (right chart).	50
3.1	Figure illustrates the tensor contraction in Equation (3.7), where the derivative GP model is trained on a single data point at descriptor $\mathbf{x}$.	78

3.2	Julia implementation of a third-order derivative of a kernel function, computed by applying forward-mode automatic differentiation (AD) via <code>ForwardDiff.jacobian</code> on the Hessian obtained from reverse-mode AD via <code>Zygote.hessian</code>	79
3.3	Learning curves in (a) and (b) showing the RMSE of predicted GPFC2 and GPFC3 in Cartesian descriptors compared to FC2 and FC3 values, respectively, obtained via the finite displacement method (FDM) using PHONOPY.	84
3.4	Panels (a) and (b) show the phonon band structures and density of states for d- Si and PbTe , respectively.	85
3.5	These figures show the predicted phonon lifetimes at 300 K of d- Si and PbTe	86
3.6	This figure illustrates the lattice thermal conductivities of d- Si and PbTe over the temperature range from 300 K to 1200 K.	87
3.7	Figure illustrates the tensor contraction in Equation (3.7) using phonon-mode descriptors.	90
3.8	Learning curves in (a) and (b) illustrate the direct comparison of GPFC2 and GPFC3 in phonon descriptors to FC2 and FC3 from FDM, respectively, using PHONO(3)PY.	92
3.9	Construction of a 3-body order ACE descriptor for a silicon structure using a total degree constraint of 5, an Agnesi radial transform, and a smooth polynomial envelope. The code prints the resulting size of basis functions.	97
3.10	Radial distribution function (RDF) and radial basis functions (RBF) for the d- Si training set.	98

3.11	Construction of a 3-body order ACE descriptor for a PbTe -structure using a total degree constraint of 5, an Agnesi radial transform, and a smooth polynomial envelope.	99
3.12	Radial distribution function (RDF) and radial basis functions (RBF) for the PbTe training set.	100
3.13	Panels (a) and (b) show the ACE-fitted potentials for an input d- Si and PbTe training data point (blue) alongside the corresponding target energies (yellow).	101
3.14	Panels (a) and (b) show the phonon band structures and density of states for d- Si and PbTe , respectively.	102
4.1	Coupling between frontier molecular orbitals (FMOs) of two conjugated organic molecules. The top and bottom FMOs represent the coupling between the lowest unoccupied molecular orbitals (LUMO–LUMO) and the coupling between the highest occupied molecular orbitals (HOMO–HOMO), respectively.	108
4.2	Schematic illustration of charge transport in organic semiconductors. The active layer consists of an intermixed donor/acceptor morphology between anode and cathode electrodes. A zoomed-in view shows the electronic coupling between two π -conjugated molecules (Mol <i>a</i> and Mol <i>b</i>), characterised by their relative centre-of-mass displacement (x_{ab}, y_{ab}, z_{ab}) and relative rotational orientations ($\theta_x, \theta_y, \theta_z$). The transfer integral J_{ab} , governing the charge transfer rate between molecules, is a function of these relative coordinates.	109

4.3	(a), (b), and (c) show the surface plots of the transfer integral $J(\theta, \phi)$ of a naphthalene dimer, the squared transfer integral $J^2(\theta, \phi)$, and the logarithm of the squared transfer integral.	114
4.4	(a), (b), and (c) show the heatmaps of the naphthalene dimer's transfer integral, the squared transfer integral, and the logarithm of the squared transfer integral.	115
4.5	Illustration of two types of molecular representations used for constructing ACE descriptors.	116
4.6	Analysis of radial distribution functions (RDFs) and radial basis functions (RBFs) used in the ACE descriptor construction for the coarse-grained (Bead) representation.	117
4.7	Construction of a 3-body order ACE descriptor for a Bead structure using a total degree constraint of 10, an Agnesi radial transform, and a smooth polynomial envelope.	118
4.8	Analysis of naphthalene dimer RDFs and RBFs used in the ACE descriptor construction for the Heavy-atom representation.	119
4.9	The plots comparing predicted and reference energies for ACE models trained on 1350 structures out of a total of 4,753 dimer configurations. (a) Model trained using the <i>Bead</i> representation. (b) Model trained using the <i>Heavy-atom</i> representation.	120
4.10	Hyperparameter sensitivity analysis of ACE models fitted to transfer integrals using Bead (solid) and Heavy-atom representations (dashed).	121
4.11	Growth of ACE basis size as a function of key hyperparameters for 2- (solid) and 4- (dashdot) chemical species.	122

4.12	Illustration of three types of molecular representations used for constructing ACE descriptors for the ethylene and thiophene dimer (Mol <i>a</i> and Mol <i>b</i>).	123
4.13	(a), (b), and (c) show the heatmaps of the ethylene dimer's transfer integral, the squared transfer integral, and the logarithm of the squared transfer integral. (d), (e), and (f) illustrate the corresponding heatmap.	124
4.14	(a), (b), and (c) show the heatmaps of the thiophene dimer's transfer integral, the squared transfer integral, and the logarithm of the squared transfer integral. (d), (e), and (f) illustrate the corresponding heatmap.	125
4.15	Ethylene dimer's RDFs and RBFs used in the ACE descriptor construction for the Heavy-atom representation.	126
4.16	Thiophene dimer RDFs and RBFs used in the ACE descriptor construction for the Heavy-atom representation.	127
4.17	The extension of the construction of the ACE basis to support 4-chemical element structures.	128
4.18	The plots comparing predicted and reference transfer integrals for ACE models trained on 2,000 structures out of a total of 8,000 dimer configurations.	129
4.19	Hyperparameter sensitivity analysis of ACE models fitted to transfer integrals using Heavy-atom representations for fitting the ethylene's transfer integrals.	130

4.20	Hyperparameter sensitivity analysis of ACE models fitted to transfer integrals using Heavy-atom representations (green) for fitting the ethylene's transfer integrals.	131
A.1	Construction of the derivative kernel (covariance) matrix used in our GP model.	140
A.2	Complete construction of the derivative kernel matrix used in our GP model.	142
A.3	Figure illustrates the tensor contraction in Equation (A.37), where the derivative GP model is trained on a single data point at descriptor $\mathbf{z}$. This serve as an extension to the tensor contraction shown in Fig. 3.1 and 3.7.	149
A.4	Panels (a), (b) and (c) show the phonon band structures and density of states for NaCl , GaAs and CdTe , respectively.	152
A.5	Figures (a), (b), and (c) present the predicted phonon lifetimes at 300 K for NaCl , GaAs , and CdTe , respectively.	153
A.6	Figures (a), (b), and (c) illustrate the lattice thermal conductivities of these materials across the temperature range of 300 K to 1200 K.	154
A.7	Learning curves showing violations of the acoustic sum rule for GP-predicted force constants (GPFCs) across various materials. (a) Translational symmetry violations in GPFC2, quantified by the total magnitude of the residuals in Eq. (A.38), as a function of the number of training points. (b) Translational symmetry violations in GPFC3, computed from the residuals in Eq. (A.39), using phonon-based descriptors.	155

A.8	Figure shows the phonon band structures and density of states for d-Si.	157
A.9	Panel (a), (b), (c) and (d) show the model trained using the <i>Bead</i> representation parameterising with order = 2, 3, 4 and 5, respectively.	159
A.10	Panel (a), (b), (c) and (d) show the model trained using the <i>Heavy-atom</i> representation parameterising with order = 2, 3, 4 and 5, respectively.	160
A.11	Panel (a), (b), (c) and (d) show the model trained using the <i>Bead</i> representation parameterising with totaldegree = 5, 15, 20 and 30, respectively.	161
A.12	Panel (a), (b), (c) and (d) show the model trained using the <i>Heavy-atom</i> representation parameterising with totaldegree = 5, 15, 20 and 30, respectively.	162
A.13	Panel (a), (b), (c) and (d) show the model trained using the <i>Bead</i> representation parameterising with rcut = 13.0 Å, 13.3 Å, 13.6 Å, and 14.2 Å, respectively.	163
A.14	Radial distribution functions (RDFs) and radial basis functions (RBFs) used in the ACE descriptor construction for the bead and Heavy-atom representation.	164
A.15	Panel (a), (b), (c) and (d) show the model trained using the <i>Heavy-atom</i> representation parameterising with rcut = 13.0 Å, 13.3 Å, 13.6 Å, and 14.2 Å, respectively.	165

A.16 Radial distribution functions (RDFs) and radial basis functions (RBFs) used in the ACE descriptor construction for the bead and Heavy-atom representation.	166
A.17 Panel (a), (b), (c) and (d) show the model trained using the <i>Bead</i> representation parameterising with $q = 39, 41, 43$, and 45 , respectively.	167
A.18 Panel (a), (b), (c) and (d) show the model trained using the <i>Heavy-atom</i> representation parameterising with $q = 39, 41, 43$, and 45 , respectively.	168
A.19 Radial distribution functions (RDFs) and radial basis functions (RBFs) used in the ACE descriptor construction for the bead and Heavy-atom representation.	169
A.20 Panel (a), (b), (c) and (d) show the model trained using the <i>Bead</i> representation parameterising with $p = 17, 19, 21$, and 23 , respectively.	170
A.21 Panel (a), (b), (c) and (d) show the model trained using the <i>Heavy-atom</i> representation parameterising with $p = 17, 19, 21$, and 23 , respectively.	171

List of Tables

2.1	Comparison of atomic descriptors based on invariance properties and smoothness.	40
2.2	Comparison of atomic descriptors based on uniqueness, complete- ness, compactness, and scalability.	41

Introductions, Motivations and Outlines

1.1 Introductions and Motivations

The ability to predict thermal conductivity and electron transport properties is essential for designing next-generation functional materials, including efficient thermoelectrics, improved semiconductors, and sustainable energy systems. These properties influence performance in applications including energy conversion, storage, miniaturised electronics, and optoelectronic devices. Traditional experimental and computational techniques encounter significant challenges involving scalability and accuracy, underscoring the necessity for innovative physically informed modelling frameworks.

Functional materials, characterised by their critical roles in electronic, thermal, and optoelectronic devices, often present significant challenges for atomistic modelling due to their intrinsic complexity. Accurately characterising the vibrational properties of crystalline inorganic materials requires the consideration of subtle anharmonic effects, which are computationally demanding when employing conventional methods. Meanwhile, organic semiconductors complicate modelling endeavours because electronic transport properties are highly sensitive to molecular orientation and intermolecular interactions. Conventional techniques which include density functional theory (DFT) and semi-empirical methods are effective, but they may be-

come computationally impractical or insufficiently accurate for large-scale or complex molecular systems.

These computational challenges require novel approaches that integrate physical knowledge with machine learning, facilitating scalable and precise prediction of transport parameters across many categories of functional materials. This thesis fulfils this requirement by developing and implementing physically informed machine learning frameworks, such as Gaussian process regression and atomic cluster expansion based linear regression, that can improve the study of anharmonic phonons and charge transfer integrals in complex materials.

Physically informed machine learning (ML) has arisen as an effective approach to address these challenges. This thesis examines two principal methodologies within this framework: Gaussian Process (GP) regression and Atomic Cluster Expansion (ACE). Gaussian processes are Bayesian machine learning methods that can model intricate, nonlinear functional relationships from data, offer inherent uncertainty quantification, and effortlessly integrate derivative observations like atomic forces [1; 17; 18]. In contrast, ACE systematically develops rotationally invariant and symmetry-compliant representations of atomic environments [2; 16]. Collectively, these methodologies form a comprehensive framework for simulating the complex behaviours of functional materials.

1.1.1 Gaussian Process Force Constants for Lattice Dynamics

The mechanical and thermal properties of crystalline materials are fundamentally determined by their potential energy surfaces (PESs). Conventional methods for modelling vibrational properties, like the finite-displacement method (FDM), generally depend on a Taylor expansion of the potential energy surface (PES) near equilibrium positions. This method effectively captures harmonic interactions in simple systems; however, its computational complexity increases prohibitively with larger system sizes

and higher-order anharmonic interactions. The precise calculation of third-order (anharmonic) force constants, crucial for forecasting finite phonon lifetimes and realistic thermal conductivities, continues to be computationally prohibitive.

Gaussian process regression provides an effective solution owing to its intrinsic differentiability and probabilistic structure. Gaussian processes facilitate concurrent training on energies, atomic forces, and higher-order derivatives of the potential energy surface, thereby enabling the efficient extraction of both harmonic and anharmonic force constants via analytical and automatic differentiation [5; 19]. Prior GP applications have effectively modelled molecular potential energy surfaces for diverse small systems, including H_3O^+ , N_4 , and protonated imidazole dimers, in addition to bulk crystalline materials like elemental silicon [20; 21; 22; 23]. These studies utilised advanced kernel engineering and dimensionality-reduction techniques to address the scalability challenges associated with GP models.

Recent research has shown that Gaussian Process models trained on energies and forces calculated via density functional theory (DFT) can accurately predict both second- and third-order force constants. [9]. The Gaussian Process Force Constant (GPFC) methodology decreases the requisite number of ab initio calculations by an order of magnitude relative to traditional finite difference methods, particularly when utilising phonon-coordinate descriptors. We present a comprehensive framework for higher-derivative Gaussian process models, employing both Cartesian and phonon-based descriptors. Our models facilitate efficient and precise predictions of phonon dispersions, lifetimes, and lattice thermal conductivities. Therefore, they offer a feasible computational alternative for investigating complex, anharmonic lattice dynamics in crystalline materials.

1.1.2 Machine Learning for Charge Transfer Integrals

Meanwhile, charge transport in organic and molecular materials is fundamentally reliant on intermolecular electronic couplings, referred to as charge transfer integrals. These transfer integrals determine the hopping rates of charge carriers and demonstrate significant sensitivity to molecular orientation and interatomic distance. The precise computation of these integrals generally necessitates resource-intensive wavefunction-based quantum chemical techniques, rendering it impractical for systems marked by significant structural disorder.

Machine learning methodologies provide effective solutions for predicting transfer integrals based on restricted quantum chemical information. Initial applications utilised Gaussian processes alongside geometric descriptors to effectively forecast integrals in crystalline molecular systems, including pentacene [24]. Recent developments have employed advanced descriptors like Coulomb matrices and deep learning techniques, including neural networks, to improve accuracy and generalisability in amorphous and chemically varied systems [25]. These prior methodologies often depend on heuristically selected descriptors, constraining their transferability and general applicability.

In this thesis, I employ a linear model with Atomic Cluster Expansion (ACE) descriptors as a systematic and invariant approach—respecting translational, rotational, and permutational symmetries—to characterise molecular environments, thereby addressing these limitations. Training machine learning models to predict the transfer integrals using ACE descriptors effectively captures fundamental physical trends while maintaining numerical stability. We employ this methodology on a range of chemically diverse molecular dimers, such as naphthalene, thiophene, and ethylene, showcasing exceptional predictive accuracy and robustness in both coarse-grained and atomistic representations.

Together, these two research strands offer a unified methodology for modelling vibrational and electronic phenomena in functional materials. This thesis enhances computational methods by integrating quantum mechanical data, physically informed descriptors, and differentiable machine learning models, thereby facilitating the systematic exploration and rational design of next-generation functional materials beyond conventional constraints.

1.2 Thesis Outline

Outline of the thesis

The thesis consists of three primary chapters, each examining distinct facets of employing physically informed machine learning to comprehend and forecast the behaviour of functional materials.

In Chapter 2, I lay the theoretical groundwork essential for our ensuing discussions. The chapter commences with an introduction to lattice dynamics, encompassing both harmonic and anharmonic phonon theories, and elucidates the application of methods such as density functional theory (DFT) and finite-difference techniques for the computation of lattice force constants. I subsequently present a comprehensive introduction to Gaussian Processes (GP), emphasising their enhanced derivative capabilities for learning from atomic forces and energies. Additionally, we examine inductive biases in machine learning models, highlighting equivariant descriptors—representations that transform consistently under symmetry operations such as translations, permutations, and rotations—specifically the Atomic Cluster Expansion (ACE) descriptors (see Section 2.4). After that, I finish the chapter with an analysis of the theory and importance of charge transfer integrals in organic semiconductors. In Chapter 3, we establish and implement our higher-derivative Gaussian Process framework (GPFC.jl), intended to effectively acquire harmonic and cubic anharmonic force con-

starts directly from *ab initio* data. We illustrate the proficiency of GP models in precisely forecasting thermal conductivity, phonon lifetimes, and dispersion relations. Subsequently, we provide a comprehensive comparison of descriptor options—specifically phonon-based and Cartesian descriptors—and examine how the incorporation of ACE descriptors could further improve our model’s accuracy and efficiency. Furthermore, we examine prospective methodological enhancements and extensive active learning strategies.

Chapter 4 provides the extension of our ACE methodology to accurately and efficiently predict charge transfer integrals in molecular dimers typically present in organic semiconductor materials. We evaluate both coarse-grained and atom-resolved ACE representations by examining various molecular dimers, including naphthalene, thiophene, ethylene, and fluorinated derivatives. Our analysis specifically emphasises the development of descriptors that enable our models to generalise effectively across various chemical systems, explicitly addressing the sensitivity of transfer integrals to molecular orientation and distance.

In Chapter 5, I summarise the key contributions and insights obtained from our research. I provide the comprehensive implications of incorporating physical insights with machine learning models and propose directions for future research. These future directions consist of the development of more sophisticated descriptors, the explicit incorporation of electronic correlation effects, and the enhancement of our models to explore real-time dynamics and non-equilibrium processes.

2

Theoretical background

This chapter reviews the fundamental physics theory and machine learning concepts underpinning this thesis. The chapter begins with an overview of lattice vibrations, covering topics from the calculation of force constants to thermal properties, with a focus on **Anharmonic Phonons**. Section 2.2 provides a concise introduction to **Density Functional Theory (DFT)**, the electronic structure theory typically used for calculating force constants. In section 2.3, the **Gaussian Process (GP)** model is introduced as a surrogate tool for predicting lattice force constants based on DFT results. This section also highlights how incorporating physics-informed constraints into the GP model can enhance its predictions. Section 2.4, **Inductive Bias: Equivariant Representations**, explores the integration of physical principles into machine learning models. Among the widely used equivariant representations, the $O(3)$ representation, as used in e.g. **Atomic Cluster Expansions (ACE)**, is described in detail in section 2.5. The utility of the ACE representation extends beyond lattice force constant predictions, as these are also a useful inductive bias for modelling transfer integrals in molecular dimers. Finally, section 2.6 provides a brief introduction to electronic **Transfer Integrals** as used to model electronic properties of organic semiconductors, concluding the chapter.

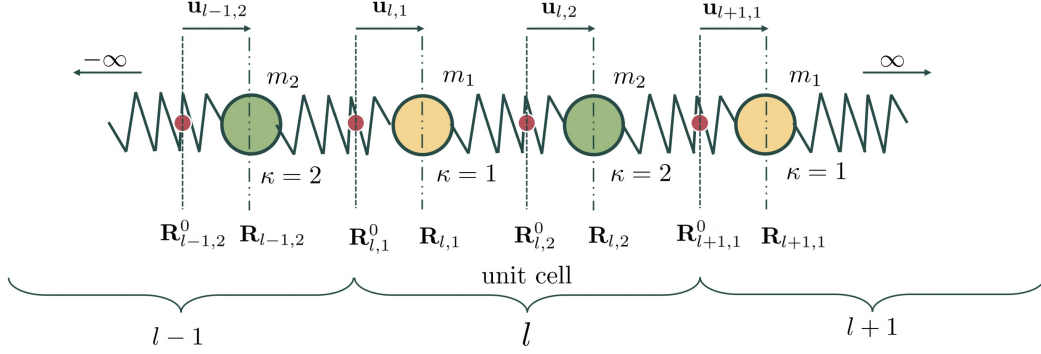


Figure 2.1: The infinite one-dimensional diatomic chain consists of an infinite number of unit cells. In the l^{th} unit cell, two atoms with masses m_1 (yellow) and m_2 (green) are labeled by $\kappa = 1$ and $\kappa = 2$, respectively.

2.1 Phonons and (an)harmonic potentials

This section introduces the foundational concepts of lattice vibrations in crystalline solids, focusing on both harmonic and anharmonic descriptions. It is divided into two subsections: Subsection 2.1.1, which discusses the calculation of lattice force constants using finite difference methods, and Subsection 2.1.2, which delves into the theory of phonons and lattice thermal conductivity derived from these force constants.

2.1.1 Lattice Force Constants

In crystalline materials, atoms are assumed to vibrate around their equilibrium positions, $\mathbf{R}_{l\kappa}^0$, with displacements $\mathbf{u}_{l\kappa} = \mathbf{R}_{l\kappa} - \mathbf{R}_{l\kappa}^0$, where κ and l denote the κ^{th} atom in the l^{th} unit cell. As an example, Fig. 2.1 illustrates an infinite diatomic chain. To derive the lattice vibrational properties, we begin by examining the Hamiltonian of the system, expressed as:

$$\mathcal{H} = \sum_{l\kappa} m_{\kappa} \left(\frac{d\mathbf{u}_{l\kappa}}{dt} \right)^2 + \Phi, \quad (2.1)$$

where Φ represents the lattice potential energy. The potential term can be expanded using a Taylor series around the equilibrium positions $\mathbf{R}_{l\kappa}^0$. Following [11; 12; 13; 14], this expansion is expressed as:

$$\begin{aligned}\Phi = & \Phi^{(0)} + \sum_{\alpha l\kappa} \Phi_{\alpha l\kappa}^{(1)} u_{l\kappa}^{\alpha} \\ & + \frac{1}{2} \sum_{\alpha l\kappa} \sum_{\beta l'\kappa'} \Phi_{\alpha l\kappa, \beta l'\kappa'}^{(2)} u_{l\kappa}^{\alpha} \cdot u_{l'\kappa'}^{\beta} \\ & + \frac{1}{3!} \sum_{\alpha l\kappa} \sum_{\beta l'\kappa'} \sum_{\gamma l''\kappa''} \Phi_{\alpha l\kappa, \beta l'\kappa', \gamma l''\kappa''}^{(3)} u_{l\kappa}^{\alpha} \cdot u_{l'\kappa'}^{\beta} \cdot u_{l''\kappa''}^{\gamma} \\ & + \dots,\end{aligned}\tag{2.2}$$

where α , β , and γ represent the Cartesian coordinate directions and $\Phi^{(n)}$ is the n^{th} -order force constants. The zeroth-order term corresponds to the potential energy at equilibrium. The first-order force constant of the potential energy vanishes at equilibrium positions because $\Phi_{\alpha l\kappa}^{(1)} = \left. \frac{\partial \Phi}{\partial u_{l\kappa}^{\alpha}} \right|_{\mathbf{u}=0}$ represents the force, which is always zero at the equilibrium positions of the system. The harmonic (second-order) force constants, which describe the vibrational properties within the harmonic approximation, can be derived in a similar manner to $\Phi^{(1)}$, as:

$$\Phi_{\alpha l\kappa, \beta l'\kappa'}^{(2)} = \left. \frac{\partial^2 \Phi}{\partial u_{l\kappa}^{\alpha} \partial u_{l'\kappa'}^{\beta}} \right|_{\mathbf{u}=0}.\tag{2.3}$$

Terms beyond the harmonic approximation represent anharmonic contributions. The cubic anharmonic (third-order) force constant, essential for collecting vibrational properties, is given by:

$$\Phi_{\alpha l\kappa, \beta l'\kappa', \gamma l''\kappa''}^{(3)} = \left. \frac{\partial^3 \Phi}{\partial u_{l\kappa}^{\alpha} \partial u_{l'\kappa'}^{\beta} \partial u_{l''\kappa''}^{\gamma}} \right|_{\mathbf{u}=0}.\tag{2.4}$$

Equations (2.3) and (2.4) can be calculated using the Finite Difference Method (FDM). In this approach, the harmonic force constant (correlating atom A and B) is determined by displacing atom B by $u^{\beta}(B)$ and calculating

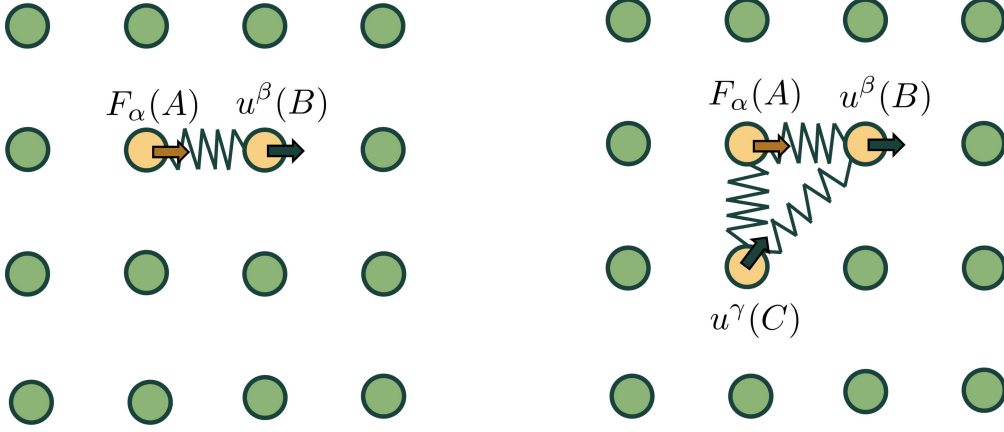


Figure 2.2: Schematic representation of harmonic and anharmonic force interactions in a lattice.

the corresponding force $F_\alpha(A)$ acting on atom A , as illustrated in the left panel of Figure 2.2. Generally, this involves generating a displaced crystal structure and using it as input for electronic structure calculations, such as Density Functional Theory (DFT), which will be discussed in Section 2.2. The resulting forces acting on atoms in the structure are then obtained via the Hellman–Feynman theorem from these calculations. The harmonic force constant correlation between A and B is evaluated as follows:

$$\Phi_{\alpha\beta}^{(2)}(A, B) = -\frac{F_\alpha(A)}{u^\beta(B)}. \quad (2.5)$$

At equilibrium the net force on each atom vanishes ($F_\alpha(A) = 0$), so the reference force at the undisplaced geometry cancels. Only the perturbed force appears in Eq. (2.5), which explains the absence of an explicit “difference” term. In practice, a centred finite difference scheme is usually employed—displacing atom B by $+u^\beta(B)$ and $-u^\beta(B)$ —to symmetrise the calculation about the equilibrium position. This cancels odd-order truncation errors and yields a more accurate $\mathcal{O}(h^2)$ estimate of the force constants compared with forward or backward differences.

This centred finite difference practically requires moving each atom a small displacement separately in x, y, z , resulting in $3N$ independent displacements. For each displacement, the forces on all $3N$ atomic degrees of freedom must be evaluated, yielding a full $3N \times 3N$ force constant matrix. Using central finite differences, this requires $6N$ total force evaluations (two per displacement), leading to $9N^2$ distinct elements in the force constant matrix [26]. (Typically, these forces are obtained from an electronic structure method, such as density functional theory.)

For the cubic anharmonic force constants, a similar procedure is employed. This requires the simultaneous displacement of two atoms (e.g., atoms B and C), while the resulting force acting on a third atom (e.g., atom A) is calculated, as depicted in the right panel of Figure 2.2.

These calculations follow the same electronic structure calculation procedure as for the harmonic force constants, enabling the numerical determination of the cubic anharmonic force constants, expressed as:

$$\Phi_{\alpha\beta\gamma}^{(3)}(A, B, C) = -\frac{F_{\alpha}(A)}{u^{\beta}(B)u^{\gamma}(C)}. \quad (2.6)$$

The process of evaluating them through central finite differences requires the displacement of all pairs of atoms (B, C) in all Cartesian directions (β, γ), and the subsequent recording of the change in force on atom A . Consequently, $9N^2$ distinct displacement pairings are generated, each necessitating four force evaluations, resulting in a total of $36N^2$ required force calculations. The third-order force constant tensor that results is significantly more expensive than the harmonic case due to the presence of $27N^3$ elements.

The computational cost arises from the extensive number of force evaluations (CPU time) required to generate the tensor and the considerable memory required for storing all those $27N^3$ independent elements. Thus, calculations of third-order force constants increase combinatorially with

system size and are far more resource-intensive than their harmonic equivalents.

2.1.2 Phonon Calculations

Phonons are quantised periodic vibrations that describe the collective motion of atoms in a crystalline material. Their study is essential for understanding vibrational, thermal, and electrical transport properties in solids. Phonon calculations begin with the construction of the dynamical matrix using harmonic force constants and solving the equations of motion for the lattice vibrations.

In the harmonic approximation, the potential energy of the lattice, Eq. (2.2), is expanded to the second order around the equilibrium positions. The equation of motion for the displacement of atoms is given by:

$$m_{\kappa} \frac{\partial^2 u_{l\kappa}^{\alpha}(t)}{\partial t^2} = - \sum_{\beta l' \kappa'} \Phi_{\alpha l \kappa, \beta l' \kappa'}^{(2)} u_{l' \kappa'}^{\beta}, \quad (2.7)$$

where m_{κ} is the mass of the κ^{th} atom in the l^{th} unit cell.

Due to the periodic boundary conditions of a crystal structure, the atomic displacements can be expressed as plane waves [27; 28]:

$$u_{l\kappa}^{\alpha}(t) = \sum_j \epsilon_{\kappa, j}^{\alpha}(\mathbf{q}) \mathbf{e}^{-i(\mathbf{q} \cdot \mathbf{R}_{\kappa l}^0 - \omega_j(\mathbf{q}) t)}. \quad (2.8)$$

In this expression, $\omega_j(\mathbf{q})$ is the phonon frequency associated with the j^{th} phonon mode characterised by the wave vector $\mathbf{q}$. The plane wave amplitude $\epsilon_{\kappa, j}^{\alpha}(\mathbf{q})$ represents the phonon polarisation in the α^{th} Cartesian direction. Substituting this expression into equation (2.7) leads to the dynamical matrix $\mathbf{D}$ eigenvalue problem:

$$\omega_j^2(\mathbf{q}) \epsilon_{\kappa, j}^{\alpha}(\mathbf{q}) = \sum_{\beta \kappa'} D_{\alpha \beta, \kappa \kappa'}(\mathbf{q}) \epsilon_{\kappa', j}^{\beta}(\mathbf{q}), \quad (2.9)$$

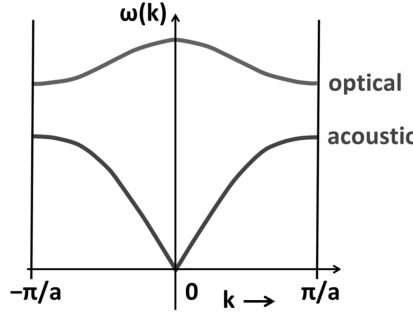


Figure 2.3: Phonon dispersion relation for a diatomic chain.

where $D_{\alpha\beta, \kappa\kappa'}(\mathbf{q})$ is the discrete Fourier transform of the harmonic force constants¹, defined as:

$$D_{\alpha\beta, \kappa\kappa'}(\mathbf{q}) = \sum_{l'} \frac{1}{\sqrt{m_{(\kappa)} m_{(\kappa')}}} \Phi_{\alpha l \kappa, \beta l' \kappa'}^{(2)} \mathbf{e}^{i\mathbf{q} \cdot (\mathbf{R}_{\kappa' l'}^0 - \mathbf{R}_{\kappa l}^0)}. \quad (2.10)$$

Solving Eq. (2.9) results in the phonon frequencies $\omega_j(\mathbf{q})$ and phonon polarisation vectors $\epsilon_{\kappa, j}^\alpha(\mathbf{q})$ as eigenvalues and eigenvectors respectively.

The fluctuation of phonon frequencies in respect to the wave vector $\mathbf{q}$ within the Brillouin zone is depicted by the phonon dispersion relations $\omega_j(\mathbf{q})$. These connections are essential to lattice dynamics and provide a thorough understanding of the vibrational modes present within the crystal. In crystals, the number of phonon branches corresponds to the total degrees of freedom per unit cell. For a unit cell containing N atoms, there are $3N$ branches: 3 acoustic branches, where phonon frequencies linearly approach zero as $\mathbf{q} \rightarrow 0$ and $3N - 3$ optical branches, where phonon frequencies remain finite as $\mathbf{q} \rightarrow 0$. Acoustic branches are the collective, in-phase vibrations of atoms that propagate through the crystal as sound waves. On the other hand, optical branches are the result of out-of-phase, relative motions between atoms within the unit cell and typically occur at higher frequencies.

¹ The translational symmetry, i.e. setting $l = 0$, can be used to simplify Eq. (2.10) more symmetric, which appears in [29].

I also provide the phonon dispersion relations of the infinite one-dimensional diatomic chain, shown in figure 2.3.

Once acquiring the phonon dispersion relations, the phonon density of states (DoS) can be computed to quantify the potential vibrational modes distributed across the range of frequencies. This distribution is crucial for determining thermodynamic properties such as heat capacity and entropy, as well as for analysing vibrational contributions to thermal transport. The phonon DoS [11] is defined as:

$$g(\omega) = \frac{1}{N_l} \sum_{\mathbf{q}j} \delta(\omega - \omega_j(\mathbf{q})), \quad (2.11)$$

where N_l is the number of unit cells in the lattice, and $\delta(\omega - \omega_j(\mathbf{q}))$ is the Dirac delta function, which ensures contributions only at the frequencies corresponding to the phonon modes.

The description of vibrational modes in the lattice, defined by acoustic and optical branches, can be improved by including quantum mechanics. This enables a quantised treatment of lattice vibrations by adding phonons as discrete energy-carrying quasiparticles. Such a framework is required for accurately collecting thermal characteristics at low temperatures and modelling phonon scattering processes that influence thermal conductivity.

In a quantum mechanical framework, the atomic displacement $u_{l\kappa}^\alpha$ can be treated as a quantum operator and expressed as a discrete Fourier transform of the phonon normal mode displacements, $u_{\kappa,j}(\mathbf{q})$ associated with the wave vector $\mathbf{q}$ and the normal mode of band index j :

$$u_{l\kappa}^\alpha = \frac{1}{\sqrt{N_l}} \sum_{\mathbf{q}j} u_{\kappa,j}(\mathbf{q}) \mathbf{e}^{-i\mathbf{q} \cdot \mathbf{R}_{\kappa l}^0} \epsilon_{\kappa,j}^\alpha(\mathbf{q}), \quad (2.12)$$

where the summation over j runs over all phonon modes. The normal mode displacement $u_{\kappa,j}(\mathbf{q})$ represent the quantum-mechanical position operator and is given by:

$$u_{\kappa,j}(\mathbf{q}) = \sqrt{\frac{\hbar}{2m_{\kappa}\omega_j(\mathbf{q})}} \left[\hat{a}_j^{\dagger}(-\mathbf{q}) + \hat{a}_j(\mathbf{q}) \right], \quad (2.13)$$

where $\hbar$ is the reduced Planck constant, $\hat{a}_j^{\dagger}(-\mathbf{q})$ is the phonon creation operator and $\hat{a}_j(\mathbf{q})$ is the phonon annihilation operator. This formulation incorporates the quantum nature of lattice vibrations, where phonon creation and annihilation operators govern the quantised energy exchange in the lattice.

Under the harmonic approximation, the lattice Hamiltonian can be expressed in terms of these operators. Starting from the Hamiltonian in equation (2.1), the total Hamiltonian is rewritten as the sum of the kinetic energy ($\mathcal{T}$) and the harmonic potential energy Φ_{Harm} :

$$\begin{aligned} \mathcal{H}_{\text{Harm}} &= \mathcal{T} + \Phi_{\text{Harm}} \\ &= \sum_{\mathbf{q}j} \hbar\omega_j(\mathbf{q}) \left[\frac{1}{2} + \hat{a}_j^{\dagger}(-\mathbf{q})\hat{a}_j(\mathbf{q}) \right], \end{aligned} \quad (2.14)$$

which is the Hamiltonian of quantum simple harmonic oscillation. The second term, $\hat{a}_j^{\dagger}\hat{a}_j$, in equation (2.14) is the phonon number operator, which corresponds to the phonon occupation number.

From the principles of statistical mechanics, phonons, as bosonic quasi-particles, obey the Bose-Einstein distribution, as their occupation number quantifies the thermal population of each vibrational mode in equilibrium. The phonon occupation number can, therefore, be expressed as:

$$n_j(\mathbf{q}) = \frac{1}{e^{(\hbar\omega_j(\mathbf{q})/k_{\text{B}}T)} - 1}, \quad (2.15)$$

where $n_j(\mathbf{q})$ gives the average number of phonons in mode j at wave vector $\mathbf{q}$, k_{B} is the Boltzmann constant, and T is the temperature. In the canonical

ensemble system, the Hamiltonian in equation (2.14) can be solved to obtain the energy of phonons as:

$$E_{\text{Harm}} = \sum_{\mathbf{q}j} \hbar \omega_j(\mathbf{q}) \left(\frac{1}{2} + n_j(\mathbf{q}) \right). \quad (2.16)$$

This phonon energy can be used to calculate the isochoric heat capacity C_V , which is essential to our research since it serves as a foundation for calculating lattice thermal conductivity [11]. By employing the thermodynamic relation of the heat capacity [30], we, therefore, obtain

$$\begin{aligned} C_V(T) &= \left(\frac{\partial E_{\text{Harm}}(T)}{\partial T} \right)_V \\ &= \sum_{\mathbf{q}j} C_{V,j}(\mathbf{q}, T) \\ &= \sum_{\mathbf{q}j} k_B \left(\frac{\hbar \omega_j(\mathbf{q})}{k_B T} \right)^2 \frac{e^{(\hbar \omega_j(\mathbf{q}) / k_B T)}}{\left[e^{(\hbar \omega_j(\mathbf{q}) / k_B T)} - 1 \right]^2}. \end{aligned} \quad (2.17)$$

This equation describes how phonons contribute to the heat capacity of the system, with contributions from all phonon modes indexed by $\mathbf{q}$ and j .²

Here, we can see the harmonic approximation simplifies the lattice potential by preserving solely quadratic (second-order) terms in atomic displacements. Consequently, atomic vibrations can be represented as a collection of isolated, non-interacting phonon modes with constant frequencies that remain unaffected by temperature variations. In this context, temperature-dependent thermal parameters, including isochoric heat capacity, can be precisely computed utilising Bose-Einstein statistics for phonon occupancy. However, the lattice thermal conductivity cannot be computed realistically under this approximation: because phonon-phonon interactions are absent, no intrinsic phonon scattering occurs, and the thermal conductivity

² In addition to phonons, electrons also contribute to the heat capacity, particularly in metals. This electronic contribution is typically linear in temperature at low T , but is negligible in insulators and semiconductors where electronic excitations are thermally suppressed. [31]

formally diverges and becomes infinite. To model thermal resistance and predict finite conductivity, one must include cubic (third-order) anharmonic terms, which give rise to three-phonon scattering processes.

The cubic anharmonic crystal potential, which introduces phonon-phonon interactions, can be expressed as a sum of three-phonon collisions:

$$\begin{aligned}\Phi_{\text{Anh}} = & \sum_{j_1, j_2, j_3} \sum_{\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3} \Phi_{j_1, j_2, j_3}(\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3) \\ & \times \left(\hat{a}_{j_1}^\dagger(-\mathbf{q}_1) + \hat{a}_{j_1}(\mathbf{q}_1) \right) \\ & \times \left(\hat{a}_{j_2}^\dagger(-\mathbf{q}_2) + \hat{a}_{j_2}(\mathbf{q}_2) \right) \\ & \times \left(\hat{a}_{j_3}^\dagger(-\mathbf{q}_3) + \hat{a}_{j_3}(\mathbf{q}_3) \right),\end{aligned}\tag{2.18}$$

where $\Phi_{j_1, j_2, j_3}(\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3)$ denotes the strength of interaction among the three phonons with wave vector $(\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3)$ in the normal mode (j_1, j_2, j_3) . The explicit form of this interaction strength can be expressed through a discrete Fourier transform of the cubic anharmonic force constant in Cartesian coordinates, $\Phi_{\alpha l_1 \kappa_1, \beta l_2 \kappa_2, \gamma l_3 \kappa_3}^{(3)}$, as:

$$\begin{aligned}\Phi_{j_1, j_2, j_3}(\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3) = & \frac{1}{6\sqrt{N_l}} \sum_{\alpha \beta \gamma} \sum_{\kappa_1 \kappa_2 \kappa_3} \epsilon_{\kappa_1, j_1}^\alpha(\mathbf{q}_1) \epsilon_{\kappa_2, j_2}^\beta(\mathbf{q}_2) \epsilon_{\kappa_3, j_3}^\gamma(\mathbf{q}_3) \\ & \times \sqrt{\frac{\hbar}{2m_{\kappa_1}\omega_{j_1}(\mathbf{q}_1)}} \sqrt{\frac{\hbar}{2m_{\kappa_2}\omega_{j_2}(\mathbf{q}_2)}} \sqrt{\frac{\hbar}{2m_{\kappa_3}\omega_{j_3}(\mathbf{q}_3)}} \\ & \times \sum_{l_1, l_2, l_3} \Phi_{\alpha l_1 \kappa_1, \beta l_2 \kappa_2, \gamma l_3 \kappa_3}^{(3)} \cdot \mathbf{e}^{i\mathbf{q}_2 \cdot (\mathbf{R}_{l_2 \kappa_2}^0 - \mathbf{R}_{l_1 \kappa_1}^0)} \\ & \cdot \mathbf{e}^{i\mathbf{q}_3 \cdot (\mathbf{R}_{l_3 \kappa_3}^0 - \mathbf{R}_{l_1 \kappa_1}^0)} \cdot \mathbf{e}^{i(\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3) \cdot \mathbf{R}_{l_1 \kappa_1}^0} \\ & \times \delta^*(\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3),\end{aligned}\tag{2.19}$$

where $\delta^*(\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3)$ enforces momentum conservation and is defined as:

$$\delta^*(\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3) = \begin{cases} 1, & \text{if } \mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3 \text{ is a reciprocal lattice vector,} \\ 0, & \text{otherwise.} \end{cases} \quad (2.20)$$

Many-body perturbation theory (MBPT)³ can be used to extract the anharmonic self-energy of phonons. The phonon self-energy is generally a complex function comprising a real part, $\Delta(\omega)$, and the imaginary part $\Gamma(\omega)$ corresponding to a phonon of frequency ω [33]. The real part, $\Delta(\omega)$, is associated with the phonon frequency renormalisation. The imaginary part, $\Gamma(\omega)$, represents an energy loss term associated with phonon scattering processes and is inversely related to the phonon lifetime.

To calculate the realistically consistent thermal conductivity of a crystal necessitates consideration of the imaginary component, which, within the context of second-order perturbation theory, is expressed as:

$$\begin{aligned} \Gamma_j(\omega; \mathbf{q}) = \frac{18\pi}{\hbar^2} \sum_{j_2, j_3} \sum_{\mathbf{q}_2, \mathbf{q}_3} |\Phi_{j, j_2, j_3}(-\mathbf{q}, \mathbf{q}_2, \mathbf{q}_3)|^2 \\ \left[(n_{j_2}(\mathbf{q}_2) + n_{j_3}(\mathbf{q}_3) + 1) \cdot \delta(\omega - \omega_{j_2}(\mathbf{q}_2) - \omega_{j_3}(\mathbf{q}_3)) \right. \\ + (n_{j_2}(\mathbf{q}_2) - n_{j_3}(\mathbf{q}_3)) \cdot \delta(\omega - \omega_{j_2}(\mathbf{q}_2)) \\ \left. - (n_{j_2}(\mathbf{q}_2) - n_{j_3}(\mathbf{q}_3)) \cdot \delta(\omega - \omega_{j_2}(\mathbf{q}_2) - \omega_{j_3}(\mathbf{q}_3)) \right] \end{aligned} \quad (2.21)$$

where $n_j(\mathbf{q})$ is the phonon occupation number defined in Eq. (2.15). $2\Gamma_j(\omega; \mathbf{q})$ represents the phonon line-width for the phonon mode $(\mathbf{q}, j)$. By

³ Many-body perturbation theory (MBPT) provides highly accurate calculations of phonon anharmonicity in crystalline materials with weak anharmonic interactions. However, for materials exhibiting strong anharmonicity, such as perovskites, thermoelectric materials, and superconducting hydrides, alternative methods like first-order self-consistent phonon theory are often more suitable [32].

taking the inverse of the phonon line-width, the corresponding phonon lifetime can be determined,

$$\tau_j(\mathbf{q}) = \frac{1}{2\Gamma_j(\omega; \mathbf{q})}. \quad (2.22)$$

From the linearised phonon Boltzmann transport equation (LBTE) under the single-mode relaxation-time (SMRT) approximation [34], the phonon lifetime $\tau_j(\mathbf{q})$ is approximated to be the single-mode relaxation time $\tau_j(\mathbf{q})^{\text{SMRT}}$. This approximation assumes that each phonon mode relaxes independently of all others, simplifying the complex interactions in the phonon system. This allows us to efficiently calculate the thermal conductivity by neglecting mode-coupling effects that may be substantial in strongly anharmonic systems. Thus combining the mode-resolved specific heat capacity (Equation (2.17)) and the phonon lifetime (Equation (2.22)), the lattice thermal conductivity is expressed as

$$\kappa = \frac{1}{N_l \cdot \Omega_{\text{cell}}} \sum_{\mathbf{q}, j} \left[\mathbf{v}_j(\mathbf{q}) \otimes \mathbf{v}_j(\mathbf{q}) \right] \cdot C_{V, j}(\mathbf{q}, T) \cdot \tau_j(\mathbf{q})^{\text{SMRT}}, \quad (2.23)$$

where Ω_{cell} is the volume of each unit cell, and N_l denotes the total number of atoms per a unit cell. The phonon group velocities $\mathbf{v}_j(\mathbf{q})$ can be given by the slope of the phonon dispersion relation,

$$\mathbf{v}_j(\mathbf{q}) = \frac{\partial \omega_j(\mathbf{q})}{\partial \mathbf{q}}. \quad (2.24)$$

Alternatively, the components of the group velocity, $v_j^\alpha(\mathbf{q})$, can be directly calculated by differentiating the eigenvalue in Equation (2.9) with respect to the phonon wave vector [12]:

$$v_j^\alpha(\mathbf{q}) = \frac{1}{2\omega_j(\mathbf{q})} \sum_{\kappa\beta} \sum_{\kappa'\gamma} \epsilon_{\kappa, j}^\beta(\mathbf{q}) \frac{\partial D_{\beta\gamma, \kappa\kappa'}(\mathbf{q})}{\partial q_\alpha} \epsilon_{\kappa', j}^\gamma(\mathbf{q}). \quad (2.25)$$

I have discussed the fundamental relations as used in PHONOPY and PHONO3PY, which form the basis for understanding the functionalities

and computational procedures underlying these widely used tools for phonon and thermal transport calculations.

While the force-constant approach described above is the most widely used route to obtain phonon spectra from first principles, phonons can also be computed from finite-temperature molecular dynamics simulations. In such cases, we can extract vibrational information by analysing the velocity autocorrelation function or the dynamical structure factor obtained from a sufficiently long MD trajectory (classical or *ab initio* MD). These time-correlation methods yield phonon densities of states and, when combined with appropriate projections, phonon dispersions, providing an alternative to the direct finite-difference or density-functional perturbation theory approaches.

2.2 Density Functional Theory

Harmonic and cubic anharmonic force constants are the key parameters which describe the phonon properties of materials. The calculation of these force constants via the finite difference method requires a set of forces acting on atoms in the system. One way to calculate these forces is by using electronic structure calculations, such as Density Functional Theory (DFT).

Density Functional Theory (DFT) is a powerful computational method widely used in quantum chemistry and condensed matter physics to study the electronic structure of atoms, molecules, and solids. By approximating the many-body Schrödinger equation with an $O(N^3)$ -based method, DFT offers a computationally efficient framework to investigate fundamental properties such as electronic energies, molecular geometries, and spectroscopic features. This chapter provides a detailed explanation of the theoretical foundations and computational methodologies associated with DFT.

2.2.1 Theoretical Foundations

The first point of quantum mechanical calculations is the time-independent Schrödinger equation for a system of interacting electrons,

$$\hat{H}\Psi = E\Psi,$$

where $\hat{H}$ is the Hamiltonian operator, Ψ is the many-body wavefunction, and E is the total energy of the system. The Hamiltonian includes the kinetic energy of the electrons, the electron-nucleus attraction, and the electron-electron Coulomb repulsion. Solving this equation directly is computationally infeasible for systems with many electrons due to the exponential growth of the wavefunction's complexity $O(N!)$.

Establishing the Hohenberg-Kohn (H-K) theorems is a key step in the development of density functional theory (DFT). These theorems prove that the ground-state energy of a many-electron system is uniquely determined by its electron density. Furthermore, the ground-state electron density minimises a total energy functional of the electron density, which includes contributions from the kinetic energy, the external potential, the Hartree energy, and the exchange-correlation energy.

Building on this H-K foundation, the Kohn-Sham (KS) approach simplifies the many-electron problem by introducing a system of non-interacting electrons that reproduces the ground-state electron density of the interacting system. The KS equations, which govern this non-interacting system, are expressed as:

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V_{\text{eff}}(\mathbf{r})\right)\psi_i(\mathbf{r}) = \epsilon_i\psi_i(\mathbf{r}), \quad (2.26)$$

where $\psi_i(\mathbf{r})$ are the Kohn-Sham orbitals, ϵ_i are their associated eigenvalues, and $V_{\text{eff}}(\mathbf{r})$ is the effective potential, given by:

$$V_{\text{eff}}(\mathbf{r}) = V_{\text{ext}}(\mathbf{r}) + V_{\text{Hartree}}(\mathbf{r}) + V_{\text{xc}}(\mathbf{r}), \quad (2.27)$$

with $V_{\text{ext}}(\mathbf{r})$, $V_{\text{Hartree}}(\mathbf{r})$, and $V_{\text{xc}}(\mathbf{r})$ representing the known external, exact Hartree, and approximate exchange–correlation potentials, respectively.

In typical condensed-phase systems the external potential $V_{\text{ext}}(\mathbf{r})$ due to the nuclei dominates, as it contains the strong Coulomb attraction at each nuclear site. The Hartree term $V_{\text{Hartree}}(\mathbf{r})$ represents the classical electrostatic repulsion of the electron density. It is generally an order of magnitude smaller but still comparable to V_{ext} in regions of high electron density. The exchange–correlation potential $V_{\text{xc}}(\mathbf{r})$ is the smallest component. Although crucial for accuracy, it typically contributes only a few to tens of percent of the total effective potential depending on the material and the functional used. These relative magnitudes motivate why V_{ext} sets the overall scale of the problem, while V_{Hartree} and V_{xc} provide successively finer corrections to the electronic structure.

2.2.2 Computational methodologies

The first step in a DFT calculation is to define the atomic positions and the system geometry. For molecular systems, atomic arrangements are specified using Cartesian or internal coordinates. For solid-state systems, additional lattice parameters and unit cell configurations are required to define the periodic structure of the material.

The electronic wavefunctions in a DFT calculation are typically represented using a finite basis set, which is chosen based on the nature of the system being studied. Common types of basis sets include:

- 1 **Gaussian-Type Orbitals (GTOs):** Frequently employed in molecular systems because of their computational efficacy in depicting localised wavefunctions. However, GTO basis sets sometimes exhibit overcompleteness, affecting the extrapolation to the complete (basis-free) limit.
Plane Waves (PW): Commonly used in periodic systems, as they effectively reflect delocalised electronic states in crystals. Although plane

waves may be inefficient for systems with high ionic density due to the requirement for a substantial energy cutoff, their intrinsic orthogonality allows systematic convergence to the complete basis set limit.

The choice of basis set significantly impacts the computational cost and accuracy of the calculation, making it an important consideration during problem setup.

The accuracy of a DFT calculation is also strongly influenced by the selection of the exchange-correlation (XC) functional, which approximates the many-body effects of electron exchange and correlation. Common categories of XC functionals are

- 1 **Local Density Approximation (LDA):** Relies solely on the local electron density, parametrised using quantum Monte Carlo results for the homogeneous electron gas (HEG), and is most accurate for systems with slowly varying electron density, such as metals.
- 2 **Generalised Gradient Approximation (GGA):** Improves upon LDA by incorporating the gradient of the electron density, allowing for greater accuracy in systems with sharply varying electron densities, such as insulators and semiconductor. Popular examples include the Perdew-Burke-Ernzerhof (PBE) functional.
- 3 **Meta-GGA Functionals:** Represent the next rung on “Jacob’s ladder” of DFT approximations by including, in addition to the density and its gradient, information about the kinetic-energy density or the Laplacian of the density. Examples include SCAN (Strongly Constrained and Appropriately Normed) [35] and TPSS (Tao–Perdew–Staroverov–Scuseria) [36] functionals. These functionals can improve predictions of structural and thermochemical properties compared with standard GGAs, although the improvement is not universal.
- 4 **Hybrid Functionals:** Combine DFT correlation with a fraction of exact Hartree–Fock exchange, providing a more accurate treatment of elec-

tron interactions for many systems. Examples include B3LYP, widely used in molecular chemistry, and HSE06, commonly applied in solid-state calculations.

Climbing Jacob’s ladder [36] often increases the formal sophistication of the functionality and often enhances accuracy for specific system categories; but, it does not ensure consistently superior outcomes. For example, a well-parameterised GGA can outperform a meta-GGA for specific properties, and hybrids can sometimes degrade predictions for metals even though they improve molecular energetics. Thus, the choice of functional should be made on a property- and system-specific basis, taking into account both accuracy requirements and computational cost.

After selecting the exchange-correlation (XC) functional, the Kohn-Sham equations are solved using the self-consistent field (SCF) procedure, an iterative algorithm designed to achieve convergence of both the electron density and the total energy. The SCF procedure involves the following steps:

- 1 Initial Guess: An initial guess for the electron density is generated, which may be based on atomic orbitals, a uniform density, or results from previous calculations.
- 2 Potential Update: the current electron density is used to calculate the effective potential. This potential includes contributions from the external potential, Hartree potential, and exchange-correlation potential.
- 3 Solve Kohn-Sham Equations: The Kohn-Sham equations are solved to obtain the updated Kohn-Sham orbitals (ψ_i) and their associated eigenvalues (ϵ_i).
- 4 Convergence Check: The updated electron density and total energy are compared to the values from the previous iteration. If the changes are

below predefined thresholds, the calculation is considered converged. Otherwise, the process is repeated using the updated density.

The SCF procedure ensures the self-consistency of the electronic structure, which is essential for achieving accurate and reliable results in DFT calculations. It is essential to obtain well-converged energy and forces from DFT calculations for our research, as we seek to predict machine-learned force constants based on this reference data. We give the SCF convergence parameters in the techniques part of Chapter 3 to furnish an appropriate dataset for training. [35] Although DFT provides a precise framework for calculating electronic structures and associated properties, its computational expense becomes excessive for large systems or when extensive configuration space sampling is required, such as in the generation of training data for machine learning models. To come across this limitation, machine learning, particularly Gaussian processes (GPs), has emerged as an efficient surrogate model, as an approximate model, for predicting energy, forces, and related observables with significantly reduced computational costs. In the subsequent section, we outline the fundamental notions underpinning Gaussian process regression.

2.3 Gaussian Processes

We present the mathematical foundations of the conventional Gaussian process (GP) model and its derivative model in this section on GPs. After that, a survey of relevant research will be given, emphasising its uses and importance in the domains of computational chemistry and physics.

2.3.1 Conventional Model

The concept of a (multivariate) Gaussian distribution can be extended using a *Bayesian* framework, leading to the formulation of a Gaussian process (GP) model [1]. Mathematically, Gaussian processes are defined as a collection of random variables, any finite subset of which follows a (multivariate) Gaussian distribution. A random vector (descriptor), $\mathbf{x}$, can be indexed by time, space, or other inputs; this set of descriptors constitutes the **input space**. In our case, the input space corresponds to Atomic Cluster Expansion descriptors of a given configuration, while the **target values** are the physical quantities to be predicted (such as energies, forces).

A GP is fully characterised by its mean function and covariance function. These functions are defined for a real function $f(\mathbf{x})$ as:

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})], \\ k(\mathbf{x}, \mathbf{x}^*) &= \mathbf{Cov}(f(\mathbf{x}), f(\mathbf{x}^*)) \\ &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}^*) - m(\mathbf{x}^*))], \end{aligned} \tag{2.28}$$

where $m(\mathbf{x})$ represents the expected value (mean) of the function at descriptor $\mathbf{x}$, and $k(\mathbf{x}, \mathbf{x}^*)$ defines the covariance between the function evaluated at $\mathbf{x}$ and $\mathbf{x}^*$. The Gaussian process, fully defined by its mean function and covariance function, can be interpreted as a **prior distribution** over the space of possible functions,

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}^*)), \tag{2.29}$$

where $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}^*)$ are defined in equation (2.28), describing the relationship between function values at different points in the input space. This means that before seeing any data we have a prior belief about the function's behaviour (encoded by m and k). Given observed data $\mathcal{D} = \{\mathbf{x}_i, y_i\}$, Bayes' theorem updates this belief to a **posterior distribution**:

$$p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f) p(f)}{p(\mathcal{D})}, \quad (2.30)$$

where:

- $p(f)$ is the **prior**, our assumption about the function before seeing the data (here, the GP defined by m and k);
- $p(\mathcal{D}|f)$ is the **likelihood**, which measures how probable the observed targets are for a given function;
- $p(f|\mathcal{D})$ is the **posterior**, our updated belief about the function after observing the data.

Because GPs maintain full probability distributions over functions and explicitly update them via Bayes' theorem, they are referred to as **Bayesian** models.

The mean function is often assumed to be zero, i.e. $m(\mathbf{x}) = 0$, simplifying in many Gaussian process applications. The covariance function $k(\mathbf{x}, \mathbf{x}^*)$ is commonly referred to as the GP "*kernel*". This *kernel* is a pairwise function defined on two d -dimensional descriptors, $\mathbf{x}$ and $\mathbf{x}^*$, and specifies the correlation between the function values at these points. In essence, it maps two d -dimensional vectors to a single scalar value,

$$k(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R},$$

The simplest kernel used in a GP model is the *dot-product* (or *linear*) kernel, which is defined as:

$$k(\mathbf{x}, \mathbf{x}^*) = \mathbf{x} \cdot \mathbf{x}^*. \quad (2.31)$$

This kernel reduces Gaussian process regression to linear regression. A more generalized version of the *linear* kernel is the *polynomial* kernel, expressed as:

$$k(\mathbf{x}, \mathbf{x}^*) = (\mathbf{x} \cdot \mathbf{x}^*)^\zeta. \quad (2.32)$$

where ζ represents the degree of the polynomial. In addition to these, various other kernels have been developed, including the *Matérn*₃₂, *Matérn*₅₂, and *periodic* kernels. Detailed definitions of these kernels can be found in Rasmussen's book [1].

In addition to the linear and polynomial kernels, one of the most widely used kernels in Gaussian process models is the *squared exponential* kernel, also known as the *radial basis function (RBF)* or simply the *Gaussian* kernel, defined as:

$$k(\mathbf{x}, \mathbf{x}^*) = \exp \left[-\frac{1}{2} d(\mathbf{x}, \mathbf{x}^*)^2 \right], \quad (2.33)$$

where $d(\mathbf{x}, \mathbf{x}^*) = \|\mathbf{x} - \mathbf{x}^*\|_2$ is the Euclidean metric.

This kernel is stationary and isotropic, depending only on the distance between points, and is infinitely differentiable, implying a very strong prior preference for smooth. These properties make the RBF kernel a default choice for modelling smooth physical processes. However, it can perform poorly for rough or discontinuous data-generating functions, or for non-stationary behaviour where the characteristic lengthscale varies across the input space; in such cases kernels with weaker smoothness assumptions (e.g. Matérn) or locally adaptive/composite kernels may be more appropriate.

Building on the concept of Gaussian processes (GPs) discussed above, we now explore how this framework can be applied in surrogate modelling physics, such as regression tasks. A common application of GP models in this domain is the analysis and prediction of potential energy surfaces

(PESs). Specifically, the goal of a GP model in such cases is to predict the continuous PES $E^*(\mathbf{x}^*)$, given the input $\mathbf{x}^*$, discrete training datapoints $\mathbf{x}$ and the corresponding PES $E(\mathbf{x})$.

To build a GP model, we need to construct the joint Gaussian distribution between those two datapoints with an 2-dimensional zero-mean vector, as:

$$\begin{bmatrix} E \\ E^* \end{bmatrix} \sim \mathcal{GP}_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}; \begin{bmatrix} k(\mathbf{x}, \mathbf{x}) + \sigma^2 & k(\mathbf{x}, \mathbf{x}^*) \\ k(\mathbf{x}^*, \mathbf{x}) & k(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix} \right), \quad (2.34)$$

where σ^2 is the observation noise associated with the PES $E(\mathbf{x})$.

This noise arises from electronic structure calculations and can be expressed as

$$\tilde{E}(\mathbf{x}) = E(\mathbf{x}) + \epsilon,$$

where the uncertainty $\epsilon \in \mathbb{R}$ can be presumed to be normal distributed with a mean of zero and a variance of σ^2 . This is defined as:

$$\epsilon \sim \mathcal{N}(0; \sigma^2). \quad (2.35)$$

Therefore, the posterior prediction for E^* is given by

$$E^* | \mathbf{x}^*, \mathbf{x}, E \sim \mathcal{GP}(m^*(\mathbf{x}^*, \mathbf{x}, E); k^*(\mathbf{x}, \mathbf{x}^*)), \quad (2.36)$$

where m^* represents the posterior mean expressed as,

$$m^*(\mathbf{x}, \mathbf{x}^*, E) = k(\mathbf{x}, \mathbf{x}^*) \cdot (k(\mathbf{x}, \mathbf{x}) + \sigma^2)^{-1} \cdot E, \quad (2.37)$$

and k^* is the corresponding posterior covariance derived from

$$k^*(\mathbf{x}, \mathbf{x}^*, E) = k(\mathbf{x}^*, \mathbf{x}^*) + k(\mathbf{x}, \mathbf{x}^*) \cdot (k(\mathbf{x}, \mathbf{x}) + \sigma^2)^{-1} \cdot k(\mathbf{x}^*, \mathbf{x}). \quad (2.38)$$

In physics, Gaussian processes (GPs) have been used extensively. As previously mentioned, GPs can be used for modelling the electronic potential energy surfaces (PESs) of N -atomic systems with the help of previous infor-

mation. Typically, density functional theory (DFT) computations provide potential energies and atomic forces at specific input structures, which form the training data for the GP.

Each structure and its associated potential energy can be represented for GP modelling as a scalar target E_i and a $3N$ -dimensional Cartesian descriptor $\mathbf{x}_i$, respectively. A total of n structures can then be arranged into a $3N \times n$ descriptor matrix $\mathbf{X}$ and an n -dimensional target vector $\mathbf{E}$, yielding a *training* or *test* dataset compatible with GP models.

For N datapoints, exact GPs require inversion of an $N \times N$ covariance matrix when training, which scales as $\mathcal{O}(N^3)$ in time and $\mathcal{O}(N^2)$ in memory. Predictions at N^* new structures scale as $\mathcal{O}(NN^*)$ – $\mathcal{O}(N^2 N^*)$ depending on whether predictive variances are computed. This cubic cost limits the practical application of exact GPs to only a few tens of thousands of training points. Sparse or approximate GP methods alleviate this bottleneck by introducing a much smaller set of $M \ll N$ inducing points; training then scales as $\mathcal{O}(nM^2)$ and prediction as $\mathcal{O}(MN^*)$, making GPs tractable for much larger datasets at the expense of an approximation whose quality depends on how well the inducing points capture the underlying data structure.

Beyond computational cost, another key challenge for GP models of materials is ensuring size transferability. Total energy is an extensive quantity, so to apply a GP model across system sizes different from those used in training it must be reformulated as an additive or local model in which an intensive latent function of each atom’s environment is learned and the total energy is expressed as the sum of these contributions. In our work we use the global Gaussian process approach, which represents each configuration by a single whole-system descriptor and applies a global kernel. Within this framework, the energy prediction cannot be made intensive and size-transferable without restructuring the model. This global choice has the advantage of capturing collective, long-range correlations and being conceptually simple

for fixed-size systems, but at the cost of losing principled size transfer and data efficiency when locality holds.

2.3.2 Reviews on Gaussian Processes for Molecules and Materials

Bartók's and Szlachta's PhD thesis [37; 38] were the first to propose this study of electronic PESs using a GP technique in 2010 and 2013. Simple systems with a few degrees of freedom, such as polarised water [20], a crystalline silicon [23; 39], N_4 [21], methyl chloride [40], or a molecule like H_3O^+ [19], were effectively treated using a GP approach. However, because the GP kernel computation does not scale well, examining a more complicated structures with higher DoFs results in a costly GP calculation.

Several optimisation strategies can be used to lower the computing cost of GP. For instance [1], a GP kernel matrix was reduced in size using a sparse approach, which involves converting it into a smaller kernel matrix with a low rank approximation. For the protonated imidazole dimer ($C_6N_9N_4^+$), Sugisawa et al. successfully implemented this sparse GP to the 51-dimensional PES [22].

A composite GP kernel, which is created by adding or multiplying two or more distinct kernels, is an additional method to increase the complexity of the GP model. By raising the GP model complexity, the number of data points needed for a given accuracy is decreased while the GP model becomes more complex. Global PESs of H_3O^+ were interpolated and extrapolated using this method by J. Dai and R. V. Krems [19].

Furthermore, by differentiating a GP kernel, the relationship between the derivatives of pre and posterior data can be determined [5; 6; 41]. Specifically, the accuracy and dependability of posterior PES predictions can be improved by using forces (energy derivatives), which are often calculated 'for free' from electronic structure calculations by the Hellmann-Feynman theorem, with previous PES data [3].

The mathematical foundation for adding force training to a traditional GP model, which allows for the simultaneous learning of potential energy surfaces and their derivatives, will be covered in the following subsection.

2.3.3 Derivative Model

Atomic forces can be used as prior derivative information to improve the accuracy of Gaussian process (GP) models in forecasting potential energy surfaces (PESs). Atomic forces are often calculated as part of an electronic structure computations. This method improve the model significantly as the force information gives the GP model specific knowledge about the local PES gradient, which enhances its interpolation and extrapolation capabilities.

Since forces are the partial derivatives of energy with respect to atomic displacements, it is necessary to differentiate the Gaussian process kernel function in order to include force information into a GP model in practice [3; 7; 8; 42]. This is accomplished by computing derivatives of the kernel function up to the second order with respect to the selected descriptors, guaranteeing that the model correctly represents the links between force and energy information.

We take the first derivative of the kernel function $k(\mathbf{x}, \mathbf{x}^*)$, as described in Equation (2.33), with respect to the descriptor $\mathbf{x}$ and $\mathbf{x}^*$, where $\mathbf{x}, \mathbf{x}^* \in \mathbb{R}^{3N}$ denote the Cartesian coordinates of N atoms. This differentiation yields first-rank tensors, the **Jacobian matrices**, which measure how sensitive the kernel function is to atomic displacements. The result is a $3N \times 1$ kernel matrix, which are frequently called the **gradient kernel**:

$$[\nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}^*)]^T = \left[\frac{\partial k}{\partial x_1} \quad \frac{\partial k}{\partial x_2} \quad \dots \quad \frac{\partial k}{\partial x_{3N}} \right]^T, \quad (2.39)$$

and

$$[\nabla_{\mathbf{x}^*} k(\mathbf{x}, \mathbf{x}^*)]^T = \left[\frac{\partial k}{\partial x_1^*} \quad \frac{\partial k}{\partial x_2^*} \quad \dots \quad \frac{\partial k}{\partial x_{3N}^*} \right]^T. \quad (2.40)$$

In the derivative kernel, the connection between the potential energy at $\mathbf{x}^*$ and the force at $\mathbf{x}$ is established by the kernel matrix in Equation (2.39), whilst the inverse relationship is defined by Equation (2.40). Finally, to construct the joint distribution similar to Equation (2.34) requires the second derivative, the **Hessian**, of the kernel with respect to those two descriptors. This **Hessian** matrix includes the correlation of the pre-training forces at $\mathbf{x}$, expressed as the $3N \times 3N$ second-rank tensor,

$$\nabla_{\mathbf{x}} \nabla_{\mathbf{x}} k = \begin{bmatrix} \frac{\partial^2 k}{\partial x_1 \partial x_1} & \frac{\partial^2 k}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 k}{\partial x_1 \partial x_{3N}} \\ \frac{\partial^2 k}{\partial x_2 \partial x_1} & \frac{\partial^2 k}{\partial x_2 \partial x_2} & & \vdots \\ \vdots & & \ddots & \\ \frac{\partial^2 k}{\partial x_{3N} \partial x_1} & \cdots & & \frac{\partial^2 k}{\partial x_{3N} \partial x_{3N}} \end{bmatrix}. \quad (2.41)$$

Now, we can build a new joint distribution for the derivative GP model by incorporating these **Jacobian** and **Hessian** kernel matrices into Equation (2.34), yielding,

$$\begin{bmatrix} E \\ \nabla_{\mathbf{x}} E \\ E^* \end{bmatrix} \sim \mathcal{GP}_{(1+3N)+1} \left(\begin{bmatrix} 0 \\ \mathbf{0} \\ 0 \end{bmatrix} ; \begin{bmatrix} k|_{\mathbf{x}, \mathbf{x}} + \sigma^2 & -\nabla_{\mathbf{x}} k & k|_{\mathbf{x}, \mathbf{x}^*} \\ [\nabla_{\mathbf{x}} k]^T & \nabla_{\mathbf{x}} \nabla_{\mathbf{x}} k + \Sigma_f & [\nabla_{\mathbf{x}} k]^T \\ k|_{\mathbf{x}^*, \mathbf{x}} & -\nabla_{\mathbf{x}^*} k & k|_{\mathbf{x}^*, \mathbf{x}^*} \end{bmatrix} \right), \quad (2.42)$$

where $\nabla_{\mathbf{x}} E = -[F_{x,1} \ F_{y,1} \ F_{z,1} \cdots F_{x,N} \ F_{y,N} \ F_{z,N}]^T$ are the pre-training forces at $\mathbf{x}$ and $k|_{\mathbf{x}, \mathbf{x}^*} \equiv k(\mathbf{x}, \mathbf{x}^*)$. We also define $3N$ -dimensional zero-mean vector $\mathbf{0}$ associated with the force components and assign negative signs to ensure that the first order kernel matrices are antisymmetric. Σ_f is the $3N \times 3N$ variance matrix associated with the observation noise of the pre-training forces, defined as:

$$\Sigma_f = \sigma_f^2 \times \mathbb{I}_{3N \times 3N}, \quad (2.43)$$

where $\mathbb{I}_{3N \times 3N}$ denotes a $3N \times 3N$ identity matrix.

The posterior prediction (2.36) can be rewritten as:

$$E^* | \mathbf{x}^*, \mathbf{x}, E, \nabla_{\mathbf{x}} E \sim \mathcal{GP}(m^*|_{\mathbf{x}^*, \mathbf{x}, E, \nabla_{\mathbf{x}} E}; k^*|_{\mathbf{x}, \mathbf{x}^*}), \quad (2.44)$$

where m^* represents the posterior mean redefined as,

$$m^*|_{\mathbf{x}^*, \mathbf{x}, E, \nabla_{\mathbf{x}} E} = \begin{bmatrix} k|_{\mathbf{x}^*, \mathbf{x}} & -\nabla_{\mathbf{x}^*} k \end{bmatrix} \begin{bmatrix} k|_{\mathbf{x}, \mathbf{x}} + \sigma^2 & -\nabla_{\mathbf{x}} k \\ [\nabla_{\mathbf{x}} k]^T & \nabla_{\mathbf{x}} \nabla_{\mathbf{x}} k + \Sigma_f \end{bmatrix}^{-1} \begin{bmatrix} E \\ \nabla_{\mathbf{x}} E \end{bmatrix}, \quad (2.45)$$

and k^* is the corresponding posterior covariance re-expressed as,

$$k^*|_{\mathbf{x}, \mathbf{x}^*} = k|_{\mathbf{x}^*, \mathbf{x}^*} + \begin{bmatrix} k|_{\mathbf{x}^*, \mathbf{x}} & -\nabla_{\mathbf{x}^*} k \end{bmatrix} \begin{bmatrix} k|_{\mathbf{x}, \mathbf{x}} + \sigma^2 & -\nabla_{\mathbf{x}} k \\ [\nabla_{\mathbf{x}} k]^T & \nabla_{\mathbf{x}} \nabla_{\mathbf{x}} k + \Sigma_f \end{bmatrix}^{-1} \begin{bmatrix} k|_{\mathbf{x}, \mathbf{x}^*} \\ [\nabla_{\mathbf{x}^*} k]^T \end{bmatrix}. \quad (2.46)$$

The use of Gaussian process models with differentiation has been investigated in a number of papers in the literature, utilising gradient information to improve prediction efficiency and accuracy. E. G. del Río et al. [7] employed this method as a local optimisation approach to investigate the local minimum of PESs for a gold cluster. In order to find the global minimum of the PES for $[\text{Ta}_2\text{O}_5]_x$ clusters with $x = 1, 2, 3$ and the surface structure of ZrN, S. Kaappa et al.[8] recently used the derivative technique with global optimisation (using all accessible local minima as a guide). Asnaashari and Krems [42] incorporated composite derivative kernels to enhance the accuracy of force and energy predictions for large molecular systems with a global 57-dimensional PES.

Adding **inductive bias** to machine learning models by incorporating physical principles is another approach to increase efficiency of learning. By incorporating previous system knowledge into the model prior to training, this method improves efficiency by removing the need for the model to learn the underlying physics alone from data. Consequently, the amount of training data needed is greatly decreased while preserving the same degree of accuracy.

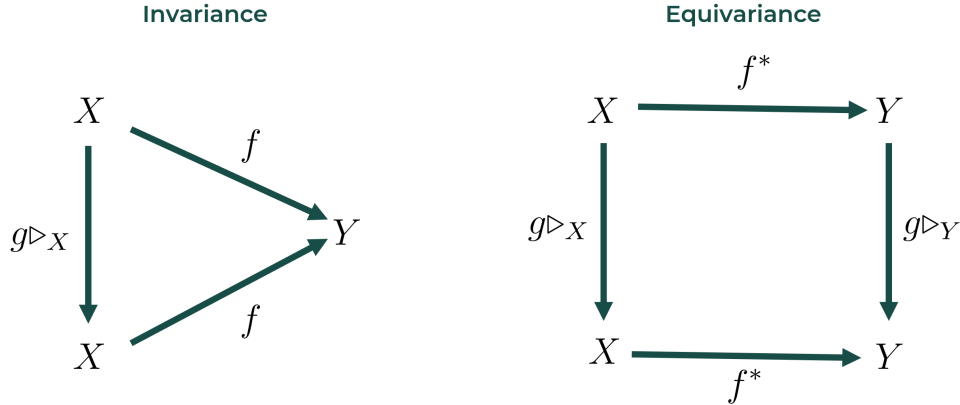


Figure 2.4: Diagrams show invariant and equivariant symmetry properties of functions f and f^* respectively.

2.4 Inductive Bias: Equivariant Descriptors

The set of presumptions incorporated into a machine learning model that directs its learning process and enables it to more successfully generalise from sparse data is known as **inductive bias**. Physical symmetries as inductive biases improve model accuracy and efficiency in the context of machine learning for atomic and molecular systems [43; 44; 45].

Using **Equivariant descriptors**, which imposes symmetry requirements like rotational, translational, and permutational invariance directly inside the model architecture, is an effective strategy. Equivariant models increase generalisation to unseen configurations and minimise the requirement for huge datasets by guaranteeing that the learnt representations follow these basic physical symmetries. This requires the understanding of the equivariance idea which is the symmetry property but different from invariance.

Let consider a map

$$f : X \rightarrow Y$$

which function f maps any input $x \in X$ into $y \in Y$. We need to consider a symmetry group G , e.g. permutation, reflection, or rotation group. In

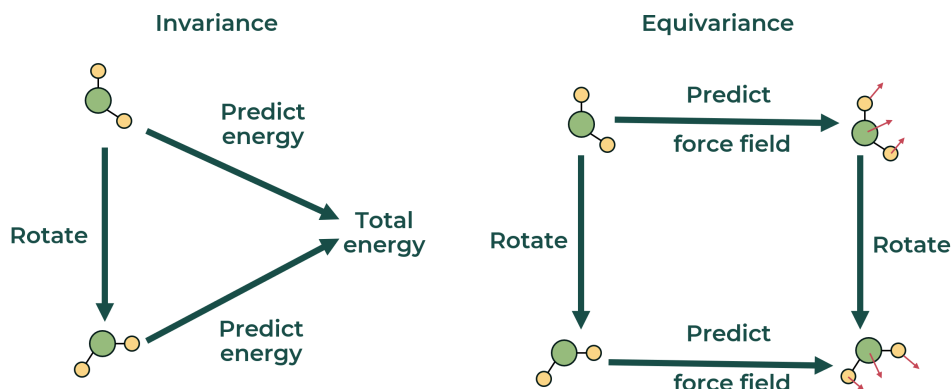


Figure 2.5: Examples of the invariance and equivariance for the prediction of a molecule's properties.

the case of invariance we need a group action on space X . Given a group element $g \in G$, the group action on any $x \in X$ can be written as $g \triangleright_X x$. We can then state the function f being **G-invariant** with respect to this action when the function satisfies

$$f(g \triangleright_X x) = f(x) \quad \forall g \in G, x \in X.$$

This means any transformed input is mapped to the same result and is depicted as the left diagram in Fig. 2.4.

In the case of equivariance, we need another a group action on space Y . It can be written similarly as $g \triangleright_Y y$. The function f^* is then said to be **G-equivariant** with respect to the actions (on space X and Y) when the function satisfies

$$f^*(g \triangleright_X x) = g \triangleright_Y f^*(x) \quad \forall g \in G, x \in X.$$

This is illustrated as the right diagram in Fig. 2.4.

In atomistic modelling, this distinction can be understood intuitively (Fig. 2.5). Predicting a molecule's total energy is an invariant task (left), since the scalar energy does not change under rotations or translations, whereas

predicting a force field is an equivariant task, since the vector forces must rotate in step with the molecule (right).

We discuss how these ideas can be used to create machine learning potentials that faithfully represent interatomic interactions while preserving physical consistency in the following subsections.

2.4.1 Reviews on Atomistic Descriptors

A recent assessment of several descriptors used in machine learning potential models was conducted by Langer et al. [43]. These descriptors fall into two general categories in the context of atomistic systems: local descriptors, which describe specific atomic surroundings [46], and global descriptors, which depict the overall structure of the system.

For modelling local descriptors that rely solely on a finite-size atomic environment, such as forces, nuclear magnetic resonance (NMR) shifts, and core-level excitations [47; 48], local representations are especially well suited. With additive approximations, which get system-wide attributes by adding up atomic contributions [49], extensive global properties, like total energy, can also be modeled using local descriptors. Since local descriptors only rely on a finite atomic environment rather than the whole system, they can be applied immediately to both finite and periodic systems without requiring any adaptation. Examples of local atomistic descriptors are **Atom-Centred Symmetric Functions: ACSF** [47; 50]; **Smooth Overlap Atomic Positions: SOAP** [3; 17; 46]; **(Multi) Atomic Cluster Expansion: (M)ACE** [2; 4; 15; 16]; and the **Message Passing (MP)** version of **ACE**, **MACE** [51; 52].

Note that **MACE** is, however, not a purely descriptor-based model in the traditional sense. While it builds on the ACE formalism for its functional basis, it embeds this within a neural message-passing architecture in which the effective descriptors and their weights are learned jointly from data rather than fixed a priori.

Global descriptors, which can represent either finite systems (molecules and clusters) or periodic systems (2D and 3D crystals), are designed to capture system-wide features such as the band gap, total energy, and polarisability tensor. Examples for finite systems are **Coulomb Matrix: CM** [53], and **Bag of Bonds: BoB** [54], while **Many-body tensor representation: MBTR** [55] can be employed for both finite and periodic systems. Since periodic systems are considered as an infinite system, global descriptors need to be specially created or modified to take into consideration the boundary conditions and repeating nature of the system.

These descriptors meet the requirements of Quantum Mechanics / Machine Learning (QM/ML), except for **CM** [4; 43]. The primary requirement is (1) *Permutational, Translational, and Rotational Invariance*, as the descriptors of the atomistic system should remain unchanged under atom reordering, spatial translation, and rigid 3D rotation.

For interpolation, learning stability, and facilitating gradient-based learning and optimization, descriptors should be smooth functions of atomic positions and differentiable with respect to them. The second prerequisite—(2) *Smoothness, Continuity, and Differentiability*—follows from this. The comparison of these descriptors relied on those transformations and smoothness is illustrated in Table 2.1.

The ability of QM/ML descriptors to discriminate between various atomistic structures in a unique way, hence avoiding degeneracies, is guaranteed by the third condition, (3) *Uniqueness & Completeness*. In addition, they must have enough knowledge to precisely recreate material properties while preserving crucial physical aspects. Completeness is necessary, but so is the fourth condition, (4) *Compactness & Efficiency*. Descriptors should encode all pertinent information required for precise predictions while remaining low-dimensional and computationally efficient. The last prerequisite is (5) *Scalability*, which means that descriptors must effectively scale with

system size to be feasible for high-throughput simulations and large systems. Requirements (3)-(5) of those descriptors are compared in Table 2.2.

ACE and **MACE** descriptors rank among the top options for atomistic simulations, based on our comparison. Yet, since message passing introduces non-linear scaling, **MACE** might be less efficient than **ACE**. As these descriptors meet all necessary conditions—invariance, smoothness, uniqueness, efficiency, and scalability—and provide linear scaling with system size, they are extremely useful for large-scale atomistic simulations. Since **SOAP** descriptors are the foundation for the development of **ACE** models, it is crucial to introduce them in the context of **Gaussian Approximation Potentials (GAP)** before diving into the mathematical underpinnings of **ACE** models. The following subsection will examine this.

2.4.2 GAP & SOAP

Understanding material properties and chemical interactions requires accurate energy calculations in these systems, which include bulk crystals, surfaces, atomic clusters, and molecules. Potential energy datasets derived from DFT calculations are examples of reference electronic structure computations that may be used to predict potential energy surfaces by utilising machine learning. These predictions can be made without the need for detailed physical interpretations [56] if the dataset is large enough.

Many effective machine learning potential approaches have been developed in the last ten years. A. P. Bartók et al. (2010) introduced the **Gaussian Approximation Potential (GAP)** [17; 56], which is a particularly noteworthy approach as the foundation of the current successful approaches, such as a **(M)ACE** model [2; 15; 16; 52].

Desc.	Invariance			Smoothness, Continuity, & Differentiability
	Permutation	Translation	Rotation	
CM	Yes (Require sorting)	No	No	No
BoB	Not Fully (Require sorting)	Not Fully (Require explicit translational symmetry)	No	No
MBTR	Not Inherently (Approximated via kernel methods)	Not Inherently (Approximated using relative distances and angles)	Not Fully (Depending on chosen coordinate system.)	Yes (Using smooth kernel functions.)
ACSF	Yes (Summing over local ACSFs)	Locally Yes (Only considering locally relative atomic distances)	Not Fully (Radial SFs are invariant, but angular SFs are not fully rotation-independent.)	Yes (Employing analytical radial and angular functions.)
SOAP	Fully	Yes (Based on local atomic density)	Yes (Using spherical harmonics)	Fully
ACE	Fully			
MACE	Fully			

Table 2.1: Comparison of atomic descriptors based on invariance properties and smoothness.

Desc.	Uniqueness	Completeness	Compactness & Efficiency	Scalability
CM	No	No	Compact & Efficient	$\mathcal{O}(N^2)$
BoB	No	No	More Compact & Efficient than CM	$\mathcal{O}(N^2)$
MBTR	Yes (Using multi-body correlations)	Yes (Rely on the kernel choices)	Less Compact & Expensive (Descriptors grow quickly with complexity)	$\mathcal{O}(N^3)$
ACSF	Not guaranteed to be Unique. (Different local environments can have similar SFs, leading to possible degeneracies.)	Not guaranteed to be Complete (Certain atomic environments can map to same ACSF.)	Compact & Efficient (Local descriptors controlled by SF parameters & independently evaluated)	$\mathcal{O}(N)$
SOAP	Fully	Complete (Tractable in lower body-order expansions)	Less Compact & Moderately Efficient (Depend on radial/angular resolution)	$\mathcal{O}(N^2)$
ACE	Fully	Fully (only to the chosen body-order)	Compact & Highly Efficient	$\mathcal{O}(N)$
MACE	Fully	Fully (only to the chosen body-order)	Compact but Less Efficient than ACE (Using MP)	$\mathcal{O}(N)$

Table 2.2: Comparison of atomic descriptors based on uniqueness, completeness, compactness, and scalability.

By representing an N -atomic system's overall energy as the sum of its local atomic contributions, this method simulates the interatomic potential of that system. The total energy is expressed by

$$E_{\text{tot}} = \sum_{\alpha \in p, t} \sum_{j \in \alpha} \epsilon_j^\alpha + \mathcal{V}_{\text{long range}}. \quad (2.47)$$

This atomic system's total energy is represented by the sum of its short- and long-range terms [3; 17; 46]. Electrostatic Van der Waals and polarisability contributions are included in the long-range term. The local energy functional contributions inside a certain cutoff radius, which makes up the short-range term. The local contributions within this cutoff zone are classified into various categories, α , based on suitable atomic descriptors. Pair distances for two-body interactions (p) and bond angles for three-body interactions (t) are examples of these descriptors. The index j , which is the sum over the bonds, the bond angles, or the dihedral angles in the system, is used to index each contribution.

Because the GAP framework is additive and local, the approach typically neglects long-range contributions and only includes low-order cluster expansion terms. Consequently, the total energy function's general form can be reduced to

$$E_{\text{tot}}^{\text{GAP}} = \sum_{p \in \forall \text{ pairs}} \epsilon_p^{(2)} + \sum_{t \in \forall \text{ triplets}} \epsilon_t^{(3)}, \quad (2.48)$$

where $\epsilon_p^{(2)}$ and $\epsilon_t^{(3)}$ correspond to two- and three-body contributions.

We can reformulate the system's total energy using an atom-centered method. This method specifies the local atomic environment of each i^{th} atom by positioning it at the centre of a cutoff zone. The nearby atoms in this area are then used to build the atomic descriptors.

The Cartesian coordinates of nearby atoms are transformed into a modified descriptor using a transformation function to represent the local energy contribution of each i^{th} atom. This change guarantees that the local energy

can only be determined by the relative positions of the atoms in the cutoff radius.

The system's total energy can be rewritten as follows using this atom-centered framework:

$$E_{\text{tot}}^{\text{GAP}} = \sum_i^{\text{atoms}} \epsilon^i(\mathbf{d}_i), \quad (2.49)$$

where $\mathbf{d}_i$ is the i^{th} atomic descriptor. In the **GAP** framework, the local energy ϵ^i is expressed as a sum of arbitrary basis functions ϕ_m of the descriptor $\mathbf{d}_i$. This descriptor, more explicitly, represents the coordinates of the atom i surrounded by its neighbour atoms. This local energy can be written as

$$\epsilon^i(\mathbf{d}_i) = \mathbf{c}^T \Phi = \sum_{m=1}^d c_m \phi_m(\mathbf{d}_i), \quad (2.50)$$

where c_m is the weight parameters of a d -dimensional weight vector $\mathbf{c}$ corresponding to basis functions ϕ_m of the basis vector Φ . If each of the weights follows a zero-mean Gaussian distribution, i.e.

$$\mathbf{c} \sim \mathcal{N}_d(\mathbf{c}; \mathbf{0}, \sigma_c \mathbb{I}),$$

then the equation above naturally extends to a Gaussian process (GP). The primary objective of **GAP** is to employ a GP model for predicting atomic energies based on arbitrary descriptors. This prediction is performed using a pre-training dataset obtained from electronic structure calculations, enabling accurate and data-efficient modelling of interatomic interactions.

Employing a GP requires defining similarity measures between two atomic descriptors through a kernel function, $k(\mathbf{d}_i, \mathbf{d}_j)$. This kernel function measures the correlation between two atomic environments, in order to guarantee that structurally similar configurations produce similar predicted energies. The kernel function can be constructed using the covariance be-

tween two atomic energies, $\langle \epsilon^i \epsilon^j \rangle$. This relationship, as deduced in [17], results in:

$$k(\mathbf{d}_i, \mathbf{d}_j) = \sum_{m=1}^d \phi_m(\mathbf{d}_i) \phi_m(\mathbf{d}_j). \quad (2.51)$$

This kernel function must be reconstructed to a covariance matrix for predicting the *local* atomic energy ϵ^* of a descriptor $\mathbf{d}^*$ from a pre-training dataset $\mathbf{E}$. If a pre-training dataset $\mathbf{E} = [\epsilon^1 \epsilon^2 \dots \epsilon^n]^T$ contains n *local* atomic energies corresponding to individual n atomic descriptors in the set of descriptors $\mathbf{D} = [\mathbf{d}_1 \mathbf{d}_2 \dots \mathbf{d}_n]^T$, the $n \times n$ covariance matrix over the pre-training dataset is defined as $K(\mathbf{D}) \propto \langle \mathbf{E} \mathbf{E}^T \rangle$. The predicting posterior mean of the atomic energy, analogous to Eq. (2.37), can be calculated from

$$\bar{\epsilon}^* = k(\mathbf{d}^*, \mathbf{D}) \cdot K^{-1}(\mathbf{D}) \cdot \mathbf{E}, \quad (2.52)$$

where $k \propto \langle \epsilon^* \mathbf{E} \rangle$ is the covariance of between the posterior atomic energy and prior atomic energies.

Using a GP model to predict the *local* atomic energy proves to be a highly effective approach. The prediction relies solely on a set of prior atomic energies obtained from electronic structure calculations and a kernel function that measures the similarity between different descriptors. Unlike traditional methods such as empirical interatomic potentials or cluster expansion methods, there is no need to explicitly define basis functions or determine the corresponding weights. Furthermore, this implies that physical interpretation and understanding are not strictly required to achieve accurate energy predictions.

Although a GP model does not require explicit physical understanding, achieving accurate predictions can be computationally expensive. Since the model is not inherently constrained by underlying physical symmetries and conservation laws, it must learn these properties directly from data through a data-driven approach. This requires training on a large dataset of total energies and forces to capture the correct physical behavior. Consequently,

the number of training datapoints can reach tens of thousands to achieve acceptable predictive accuracy.

To minimize the number of required electronic structure calculations, physical symmetries of the system can be incorporated into the GP model before training. Specifically, the choice of atomic descriptors can be designed to inherently satisfy these symmetries or, mathematically, to serve as representations of the corresponding symmetry groups. By embedding these symmetry-aware descriptors as an inductive bias, the model’s predictive accuracy is enhanced while reducing the dependence on large training datasets.

Bartók and Csányi et al., using the Smooth Overlap of Atomic Positions (SOAP) framework, presented a methodical way to choose a suitable description for a GAP model [17; 46]. In Cartesian coordinates, they first define the atomic descriptor for atom i as a density function that represents its immediate atomic environment, expressed as

$$\rho_i(\mathbf{r}) = \sum_j^{\text{neigh.}} \exp\left(-\frac{\|\mathbf{r} - \mathbf{r}_{ij}\|^2}{2\sigma^2}\right), \quad (2.53)$$

This neighbour density includes Gaussian peaks specifying the spatial distribution of neighboring atoms j . The centre of the neighboring peaks is displaced from atom i by the relative position $\mathbf{r}_{ij}$, capturing the atomic environment in a smooth and differentiable manner. We can employ this neighbour density to rewrite Equation (2.50) in terms of atomic energy ϵ^i as a functional of the atomic density ρ_i ,

$$\epsilon^i[\rho_i] = \int c(\mathbf{r}) \rho_i(\mathbf{r}) d\mathbf{r}, \quad (2.54)$$

where the weight c is also Gaussian distributed. Evaluating the covariance $\langle \epsilon^i \epsilon^j \rangle$ then defines the kernel function as

$$k(\rho_i, \rho_j) = \int \rho_i(\mathbf{r}) \rho_j(\mathbf{r}) d\mathbf{r}. \quad (2.55)$$

By integrating all the associated density functions, this formulation guarantees that the kernel function captures the similarity between atomic environments. However, this kernel function is not invariant under 3D rotations due to the chosen neighbour density ρ_i .

To ensure rotational invariance, the density function in Equation (2.53) can be re-expanded in terms of radial and spherical basis functions to form a **single particle basis function** as follows,

$$\rho^{(i)}(\mathbf{r}) = \sum_{nlm} c_{nlm}^{(i)} R_{nl}(r) Y_l^m(\hat{\mathbf{r}}), \quad n, l \in \mathbb{N}, m \in \{-l, \dots, l\}, \quad (2.56)$$

where $\mathbf{r} = r\hat{\mathbf{r}}$ and $c_{nlm}^{(i)}$ are the coefficients of the expansion for atom i . $R_{nl}(r)$ is a radial basis function depending on the magnitude of vector $\mathbf{r}$ and Y_l^m is a spherical harmonic basis function depending on the direction of vector $\mathbf{r}$.

To use the spherical density function, not only can it be substituted in the above kernel function, it can be used in the form of a **power spectrum** defined as

$$P_{nl}^{(ij_1)} = \sum_{m=-l}^l c_{nlm}^{(i)} \left(c_{nlm}^{(j_1)} \right)^*, \quad (2.57)$$

which captures two-body correlations and serving as the standard SOAP descriptor in **GAP**. The crucial property of Equation (2.57) is that the sum over the magnetic quantum numbers m corresponds to integrating over all possible rotations (i.e. averaging over the orthogonal group $O(3)$ with respect to the Haar measure). As a result, the power spectrum coefficients are invariant under any 3D rotation of the atomic environment, even though the underlying neighbour density $\rho_i(\mathbf{r})$ is not.

To capture three-body correlations, the **Bispectrum**, defined as

$$B_{n_1 l_1 n_2 l_2 l}^{(ij_1 j_2)} = \sum_{m_1 m_2} C_{l_1 m_1 l_2 m_2}^{lm} c_{n_1 l_1 m_1}^{(i)} \left(c_{n_1 l_1 m_1}^{(j_1)} \right)^* c_{n_2 l_2 m_2}^{(j_2)}, \quad (2.58)$$

incorporates Clebsch-Gordan coefficients $C_{l_1 m_1 l_2 m_2}^{lm}$ between two spherical harmonic components. This coupling requires the tensor product of two

density functions in Equation (2.56) leading to more expensive computation than **power spectrum**. While the **power spectrum** efficiently captures structural features and is widely used in GAP, the **bispectrum** provides richer information at the cost of increased complexity. The computational complexity increases factorially with the body order due to the coupling of multiple density components. However, these decompositions make the descriptor more appropriate for machine learning models that adhere to basic physical symmetries by guaranteeing that it stays invariant under 3D rotations. The derivations of these decompositions are stated in [3; 46].

Regarding the power spectrum and bispectrum, we can express the energy of atom i with its environment as follows

$$\epsilon^i = \sum_{j_1} \sum_{k_1} \tilde{c}_{k_1} P_{k_1}^{(i j_1)} + \sum_{j_1 \neq j_2} \sum_{k_2} \tilde{c}_{k_2} B_{k_2}^{(i j_1 j_2)}, \quad (2.59)$$

where $\tilde{c}_{k_1}$ and $\tilde{c}_{k_2}$ are the fitting coefficients indexed by quantum numbers $k_1 = (n, l)$ and $k_2 = (n_1, l_1, n_2, l_2, l)$ respectively. To obtain the total energy, we substitute the atomic energy back into Equations (2.49) and the result is equivalent to Equation (2.48).

To estimate the coefficients $\tilde{\mathbf{c}} = \{\tilde{c}_{k_1}, \tilde{c}_{k_2}\}$ we consider a regularised least-squares loss function with a joint objective over observed total energies $\mathcal{E}_{\mathbf{x}}^{\text{obs}}$, forces $\mathcal{F}_{\mathbf{x}}^{\text{obs}}$, and virials (or stresses) $\mathcal{V}_{\mathbf{x}}^{\text{obs}}$ for each structure $\mathbf{x}$ in the training set $\mathcal{R}$. The objective can be expressed as

$$\begin{aligned} \mathcal{L}(\tilde{\mathbf{c}}) = \sum_{\mathbf{x} \in \mathcal{R}} & \left(w_{E,\mathbf{x}}^2 \|E(\tilde{\mathbf{c}}, \mathbf{x}) - \mathcal{E}_{\mathbf{x}}^{\text{obs}}\|_2^2 + w_{F,\mathbf{x}}^2 \|F(\tilde{\mathbf{c}}, \mathbf{x}) - \mathcal{F}_{\mathbf{x}}^{\text{obs}}\|_2^2 \right. \\ & \left. + w_{V,\mathbf{x}}^2 \|V(\tilde{\mathbf{c}}, \mathbf{x}) - \mathcal{V}_{\mathbf{x}}^{\text{obs}}\|_2^2 \right) + R, \end{aligned} \quad (2.60)$$

where the weights $w_{E,\mathbf{x}}$, $w_{F,\mathbf{x}}$, $w_{V,\mathbf{x}}$ allow differential emphasis on various observables, typically scaled with the number of atoms in each structure R to balance the relative contributions of energies, forces, and virials. The final term represents regularisation for prevent overfitting, occurring when a model learns noise or irrelevant details from the training data instead

of capturing the underlying physical relationships. By adding a penalty term to the loss function, regularisation constrains the model parameters. The detail calculations are provided in section 2 of the Gaussian process review [3].

In the next section, we will provide the extended framework to **GAP** for capturing up to a higher-order body term while maintaining a linear scaling with respect to body order. This computationally tractable framework is known as **Atomic Cluster Expansion (ACE)**.

2.5 Atomic Cluster Expansion

In the previous section, we reviewed atomic descriptors used in machine learning potential models. We also discussed local atomistic descriptors, particularly **SOAP**, and its corresponding Gaussian process model (**GAP**). This model employs a cluster expansion approach, expanding interactions up to the three-body order. It is restricted to low-body orders because the descriptor's complexity scales factorially with increasing body order. To overcome this limitation, the **Atomic Cluster Expansion (ACE)** framework is developed by Ralf Drautz in 2019 [2], providing a more efficient descriptor which achieves linear scaling with the body order.

2.5.1 Canonical Cluster Expansion

In this section, we present a brief mathematical discussion constructing the **ACE** model. The first step is to reformulate the total energy in Equation (2.49) as the sum of the site (atomic) energies [2; 15; 16] as follows

$$E(\{\mathbf{r}_i\}_{i=1}^J) = \sum_{i=1}^J \epsilon(\{\mathbf{r}_{ij}\}_{j=1}^{J-1}), \quad (2.61)$$

where $\mathbf{r}_i$ is the position of the centre atom i and its atomic displacement, from its neighbouring atom j , is defined as $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ for all $j \neq i$. Each site

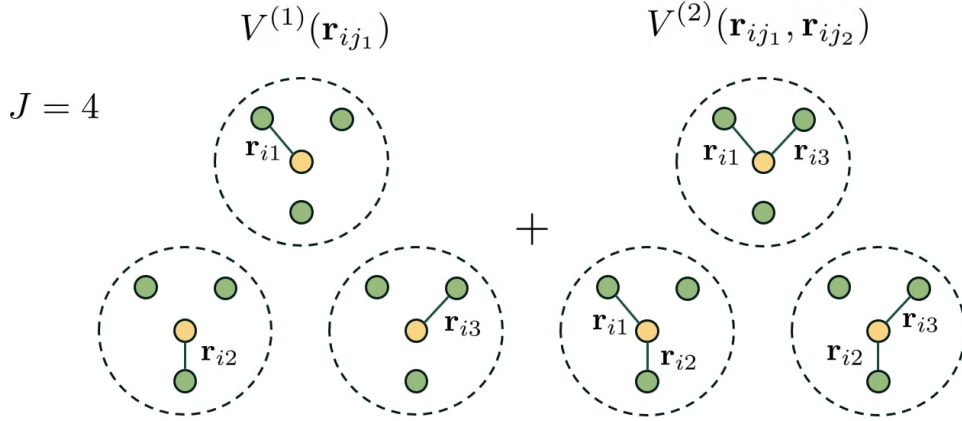


Figure 2.6: This figure shows the 2- and 3-body order clusters for a 4-body system ($J = 4$).

energy ϵ is expanded into a many-body order (*Canonical*) cluster around a centre atom i and expressed as

$$\begin{aligned} \epsilon(\{\mathbf{r}_{ij}\}_{j=1}^{J-1}) = & V^{(0)} + \sum_{j_1} V^{(1)}(\mathbf{r}_{ij_1}) + \sum_{j_1 < j_2} V^{(2)}(\mathbf{r}_{ij_1}, \mathbf{r}_{ij_2}) \\ & + \cdots + \sum_{j_1 < j_2 < \cdots < j_N} V^{(N)}(\mathbf{r}_{ij_1}, \cdots, \mathbf{r}_{ij_N}), \end{aligned} \quad (2.62)$$

where $V^{(N)}$ represents the potential energy function of an $(N+1)$ -body order or an N -interaction order involving a centre atom with its N neighbours [16; 57]. For example, we can redefine the three-body order cluster expansion from Equation (2.59) as

$$\epsilon(\mathbf{r}_{ij_1}, \mathbf{r}_{ij_2}) = \sum_{j_1} V^{(1)}(\mathbf{r}_{ij_1}) + \sum_{j_1 < j_2} V^{(2)}(\mathbf{r}_{ij_1}, \mathbf{r}_{ij_2}). \quad (2.63)$$

This is illustrated in Figure 2.6 which shows the associated clusters of the 2- and 3-body order potential in a 4-body system. In this case, $j_1, j_2 \in \{1, 2, 3\}$ as $J = 4$.

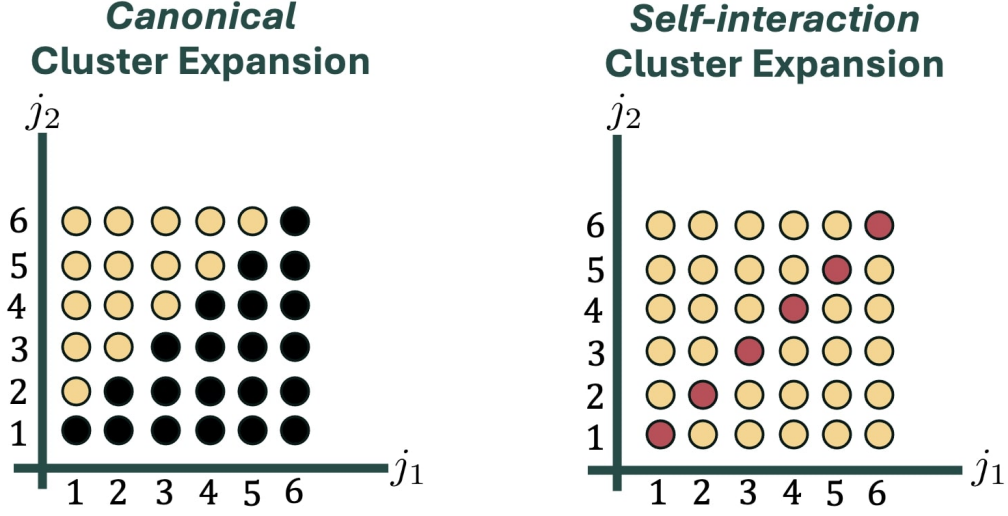


Figure 2.7: These charts, presented by Ortner in GAP & (M)ACE 2024 workshop, show the allowed j_1, j_2 (yellow circles) for the restricted sums in a *Canonical* 3-body's cluster (left chart) and the unrestricted sums in an *Self-interaction* 3-body's cluster (right chart).

2.5.2 Self-interaction Cluster Expansion

To achieve the efficient cluster expansion model, Drautz [2] originally suggested modifying the restricted summations in the *canonical* cluster expansion, defined in Equation (2.62), to the unrestricted summations in the following cluster expansion expressed as

$$\begin{aligned} \epsilon(\{\mathbf{r}_{ij}\}_{j=1}^{J-1}) = & U^{(0)} + \sum_{j_1} U^{(1)}(\mathbf{r}_{ij_1}) + \sum_{j_1 j_2} U^{(2)}(\mathbf{r}_{ij_1}, \mathbf{r}_{ij_2}) \\ & + \cdots + \sum_{j_1 j_2 \cdots j_N} U^{(N)}(\mathbf{r}_{ij_1}, \cdots, \mathbf{r}_{ij_N}). \end{aligned} \quad (2.64)$$

This cluster expansion with these unrestricted summations is called the *Self-interaction* cluster expansion because it includes all (non-physical) self-interactions. For example, in the 3-body order term, there exists the case that $j_1 = j_2$.

To understand these non-physical self-interactions, we can parameterise each interaction $U^{(N)}$ term by the tensor product of N **single particle basis functions** $\phi_k(\mathbf{r}_{ij})$, defined in equation (2.56), as follows

$$\sum_{j_1 j_2 \dots j_N} U^{(N)}(\mathbf{r}_{ij_1}, \dots, \mathbf{r}_{ij_N}) = \sum_{j_1 j_2 \dots j_N} \sum_{k_1 \dots k_N} c_{k_1 \dots k_N}^{(N)} \prod_{t=1}^N \phi_{k_t}(\mathbf{r}_{ij_t}), \quad (2.65)$$

where $c_{k_1 \dots k_N}^{(N)}$ is the $(N+1)$ -body's expansion coefficient and each k_t denotes multi-indices of quantum numbers (n_t, l_t, m_t) .

The 3-body order is, therefore, simply defined as

$$\sum_{j_1 j_2} U^{(2)}(\mathbf{r}_{ij_1}, \mathbf{r}_{ij_2}) = \sum_{j_1 j_2} \sum_{k_1 k_2} c_{k_1 k_2}^{(2)} \phi_{k_1}(\mathbf{r}_{ij_1}) \phi_{k_2}(\mathbf{r}_{ij_2}), \quad (2.66)$$

where $j_1 = j_2 = j$, shown as the red circles in Figure 2.7, introduces the term proportional to the tensor product of the same two single particle basis functions, $\phi_{k_1}(\mathbf{r}_{ij}) \phi_{k_2}(\mathbf{r}_{ij})$. This product is interpreted as a *self-interaction* term and can be re-expanded into a lower order term [2; 57], i.e.

$$\phi_{k_1}(\mathbf{r}_{ij}) \phi_{k_2}(\mathbf{r}_{ij}) = \sum_k a_k^{(1)} \phi_k(\mathbf{r}_{ij}). \quad (2.67)$$

2.5.3 Many-body Product Basis

To avoid factorial scaling from the tensor product of the basis functions requires a few mathematical tricks. We can use *Fubini's theorem* [58] in the discrete case. If each summation over j and each summation over k in Equation (2.65) is convergent, then the order of summation can be interchanged. The equation is rewritten as

$$\begin{aligned} \sum_{j_1 j_2 \dots j_N} U^{(N)}(\{\mathbf{r}_{ij_t}\}_{t=1}^N) &= \sum_{j_1 j_2 \dots j_N} \sum_{k_1 \dots k_N} f_{k_1 \dots k_N}^{(N)}(\{\mathbf{r}_{ij_t}\}_{t=1}^N) \\ &= \sum_{k_1 \dots k_N} \sum_{j_1 j_2 \dots j_N}^{J-1} f_{k_1 \dots k_N}^{(N)}(\{\mathbf{r}_{ij_t}\}_{t=1}^N), \end{aligned} \quad (2.68)$$

where

$$f_{k_1 \dots k_N}^{(N)}(\{\mathbf{r}_{ij_t}\}_{t=1}^N) = c_{k_1 \dots k_N}^{(N)} \prod_{t=1}^N \phi_{k_t}(\mathbf{r}_{ij_t})$$

Moreover, we can use the fact that each of the basis functions in the product is j_t independent. This can refer to the *linearity* and the *factorisation* property of independent sums. These sums over multiple indices, hence, can be swapped and distributed over multiplication. We can re-express Equation (2.68) as follows

$$\begin{aligned} \sum_{j_1 j_2 \dots j_N} U^{(N)}(\{\mathbf{r}_{ij_t}\}_{t=1}^N) &= \sum_{k_1 \dots k_N} \sum_{j_1 j_2 \dots j_N}^{J-1} c_{k_1 \dots k_N}^{(N)} \prod_{t=1}^N \phi_{k_t}(\mathbf{r}_{ij_t}) \\ &= \sum_{k_1 \dots k_N} c_{k_1 \dots k_N}^{(N)} \left(\sum_{j_1} \phi_{k_1}(\mathbf{r}_{ij_1}) \right) \left(\sum_{j_2} \phi_{k_2}(\mathbf{r}_{ij_2}) \right) \\ &\quad \dots \left(\sum_{j_N} \phi_{k_N}(\mathbf{r}_{ij_N}) \right) \\ &= \sum_{k_1 \dots k_N} c_{k_1 \dots k_N}^{(N)} \prod_{t=1}^N \sum_{j_t} \phi_{k_t}(\mathbf{r}_{ij_t}). \end{aligned} \tag{2.69}$$

This makes the evaluation more efficient. Instead of calculating the tensor product of the basis functions before sums over j , we perform a scalar multiplication of $\sum_j^{J-1} \phi_k(\mathbf{r}_{ij})$ leading to an N linear scale for evaluating an N -interaction order term, compared to an $N!$ scaling in the canonical cluster expansion. Each of the sums can be defined as **Atomic neighbour density basis**

$$A_k(\mathbf{r}_{ij}) = \sum_j \phi_k(\mathbf{r}_{ij}). \tag{2.70}$$

The product over N in Equation (2.69) is applied to the basis function, resulting in a **Many body product basis**

$$\mathbf{A}_{\mathbf{k}}^{(N)}(\mathbf{R}) = \prod_{t=1}^N A_{k_t}(\mathbf{r}_{ij_t}), \tag{2.71}$$

where the bold $\mathbf{k}$ index and $\mathbf{R}$ are identified by the collection of $\{k_1 \cdots k_N\}$ and $\{\mathbf{r}_{i_{j_1}}, \dots, \mathbf{r}_{i_{j_N}}\}$. This product basis ensures permutation invariance, yet it is not invariant under rotations in a 3-dimensional space. It can be symmetrised by *Group* averaging over all possible rotations $\hat{Q} \in SO(3)$, the special orthogonal group in three dimensions,

$$\mathbf{B}_{\mathbf{k}}^{(N)} = \int_{SO(3)} \mathbf{A}_{\mathbf{k}}^{(N)}(\hat{Q}\mathbf{R}) d\hat{Q}, \quad (2.72)$$

where $\mathbf{B}_{\mathbf{k}}$ is a rotation-invariant product basis. However, group averaging results in losing all angular information, so only the radial basis functions remain, leading to the model becoming insufficient to represent an atomic system.

2.5.4 Rotation Invariance

To overcome this problem, we can use a foundational concept from representation theory. Under an arbitrary 3D-rotation $\hat{Q}$, the action on a spherical harmonic function Y_l^m , characterised by the angular momentum index l , yields a linear combination of unrotated spherical harmonics Y_l^μ with the same angular momentum [16; 46]. This is expressed as

$$Y_l^m(\hat{Q}\hat{\mathbf{r}}) = \sum_{\mu=-l}^l D_{m\mu}^l(\hat{Q}) Y_l^\mu(\hat{\mathbf{r}}); \quad \forall \hat{Q} \in SO(3), \quad (2.73)$$

where $\hat{\mathbf{r}} \in \mathbb{S}^2$ is a unit vector in the direction of vector $\mathbf{r}$. The $D_{m\mu}^l(\hat{Q})$ defines the *Wigner D-matrices*, which describe how spherical harmonics transform under 3D-rotation. This transformation shows that the spherical harmonics remain consistent under rotational operations, thereby preserving their angular momentum l and capturing rotational effects through the magnetic quantum number m .

We can incorporate the single particle basis function, Equation (2.56) into the above transformation ⁴. It can be re-expressed as

$$\phi_{nlm}(\hat{Q}\mathbf{r}) = \sum_{\mu=-l}^l D_{m\mu}^l(\hat{Q}) \cdot \phi_{nl\mu}(\mathbf{r}) \quad \forall \hat{Q} \in SO(3). \quad (2.74)$$

Substituting this transformation into the **atomic neighbour density** in Equation (2.70) yields the **many body product basis** transforming as follows

$$\begin{aligned} \mathbf{A}_{\mathbf{n}\mathbf{l}\mathbf{m}}^{(N)}(\hat{Q}\mathbf{R}) &= \prod_{t=1}^N \sum_{j_t} \left(\sum_{\mu_t=-l_t}^{l_t} D_{m_t\mu_t}^{l_t}(\hat{Q}) \cdot \phi_{nl\mu}(\mathbf{r}_{ij_t}) \right) \\ &= \prod_{t=1}^N \sum_{\mu_t=-l_t}^{l_t} D_{m_t\mu_t}^{l_t}(\hat{Q}) \cdot A_{n_t l_t \mu_t}(\mathbf{r}_{ij_t}) \\ &= \sum_{\mu \in \mathcal{M}_1} \mathbf{D}_{\mathbf{m}\mu}^{\mathbf{l}}(\hat{Q}) \cdot \mathbf{A}_{\mathbf{n}\mathbf{l}\mu}^{(N)}(\mathbf{R}); \quad \forall \hat{Q} \in SO(3), \end{aligned} \quad (2.75)$$

where the product of *Wigner D-matrices* is defined as

$$\mathbf{D}_{\mathbf{m}\mu}^{\mathbf{l}}(\hat{Q}) = \prod_{t=1}^N D_{m_t\mu_t}^{l_t}(\hat{Q}). \quad (2.76)$$

The set of *principle* and *angular momentum* quantum numbers are given by $\mathbf{n} = \{n_1, \dots, n_N\} \in \mathbb{N}^N$ and $\mathbf{l} = \{l_1, \dots, l_N\} \in \mathbb{N}^N$, respectively, while $\mathbf{m}$ and μ represent the set of *magnetic* quantum numbers within the coupled angular momentum space $\mathcal{M}_1$.

⁴ By coupling two rotational-transformed **single particle basis functions**, we can recover the power spectrum in Eq. (2.57).

Now we can perform 3D-rotational *Group* averaging on Equation (2.75) to achieve the rotation-invariant product basis

$$\begin{aligned} \mathbf{B}_{\mathbf{n}\mathbf{l}\mathbf{m}}^{(N)}(\mathbf{R}) &= \int_{SO(3)} \mathbf{A}_{\mathbf{n}\mathbf{l}\mathbf{m}}^{(N)}(\hat{Q}\mathbf{R}) d\hat{Q} \\ &= \sum_{\mu \in \mathcal{M}_1} \left(\int_{SO(3)} \mathbf{D}_{\mathbf{m}\mu}^{\mathbf{l}}(\hat{Q}) \cdot d\hat{Q} \right) \mathbf{A}_{\mathbf{n}\mathbf{l}\mu}^{(N)}(\mathbf{R}) \\ &= \sum_{\mu \in \mathcal{M}_1} \bar{\mathbf{D}}_{\mathbf{m}\mu}^{\mathbf{l}} \cdot \mathbf{A}_{\mathbf{n}\mathbf{l}\mu}^{(N)}(\mathbf{R}), \end{aligned} \quad (2.77)$$

where $\bar{\mathbf{D}}_{\mathbf{m}\mu}^{\mathbf{l}}$ are the integrated coefficients. These coefficients can be efficiently calculated using a recursion formula involving Clebsch-Gordan coefficients, as derived by Dusson et al. in Appendix A of their work [16]. Furthermore, in section 3.5 of their work, Dusson et al. present a systematic way to incorporate permutation and parity invariance into Equation (2.77), resulting in the symmetrised form,

$$\mathbf{B}_{\mathbf{n}\mathbf{l}\mathbf{m}}^{(N)}(\mathbf{R}) = \sum_{\mu \in \mathcal{M}_1} \mathcal{U}_{\mathbf{m}\mu}^{\mathbf{n}\mathbf{l}} \cdot \mathbf{A}_{\mathbf{n}\mathbf{l}\mu}^{(N)}(\mathbf{R}), \quad (2.78)$$

where $\mathcal{U}_{\mathbf{m}\mu}^{\mathbf{n}\mathbf{l}}$ are the symmetry-adapted (generalised) Clebsch-Gordan coefficients. We can now substitute the symmetrised expression from Equation (2.78) back to Equation (2.69) to obtain the N -interaction order potential

$$\sum_{j_1 j_2 \dots j_N} U^{(N)}(\mathbf{R}) = \sum_{\mathbf{n}\mathbf{l}\mathbf{m}} c_{\mathbf{n}\mathbf{l}\mathbf{m}}^{(N)} \mathbf{B}_{\mathbf{n}\mathbf{l}\mathbf{m}}^{(N)}(\mathbf{R}), \quad (2.79)$$

where $c_{\mathbf{n}\mathbf{l}\mathbf{m}}^{(N)}$ are the N -interaction order expansion coefficients, and $\mathbf{B}_{\mathbf{n}\mathbf{l}\mathbf{m}}^{(N)}(\mathbf{R})$ denotes the symmetrized $(N+1)$ -body basis functions. Eventually, the site energy truncated at N -interaction order can be defined as

$$\epsilon(\{\mathbf{r}_{ij}\}_{j=1}^{J-1}) = \sum_{v=0}^N \sum_{\mathbf{n}\mathbf{l}\mathbf{m}} c_{\mathbf{n}\mathbf{l}\mathbf{m}}^{(v)} \mathbf{B}_{\mathbf{n}\mathbf{l}\mathbf{m}}^{(v)}(\mathbf{r}_{ij_1}, \dots, \mathbf{r}_{ij_v}). \quad (2.80)$$

The set of coefficients $\mathbf{c} = \{c_{\mathbf{n}_0\mathbf{l}_0\mathbf{m}_0}^{(0)}, c_{\mathbf{n}_1\mathbf{l}_1\mathbf{m}_1}^{(1)}, \dots, c_{\mathbf{n}_N\mathbf{l}_N\mathbf{m}_N}^{(N)}\}$ is determined by minimising a least-squares loss function that jointly accounts for the observed total energies $\mathcal{E}_R^{\text{obs}}$, forces $\mathcal{F}_R^{\text{obs}}$, and virials $\mathcal{V}_R^{\text{obs}}$ for each structure R in the training set $\mathcal{R}$. The loss function can be written as

$$\begin{aligned} \mathcal{L}(\mathbf{c}) = \sum_{R \in \mathcal{R}} & (w_{E,R}^2 \|E(\mathbf{c}, R) - \mathcal{E}_R^{\text{obs}}\|_2^2 + w_{F,R}^2 \|F(\mathbf{c}, R) - \mathcal{F}_R^{\text{obs}}\|_2^2 \\ & + w_{V,R}^2 \|V(\mathbf{c}, R) - \mathcal{V}_R^{\text{obs}}\|_2^2), \end{aligned} \quad (2.81)$$

where the weights $w_{E,R}$, $w_{F,R}$, and $w_{V,R}$ control the relative emphasis on each observable. Since all observations are linear, the optimisation is carried out by minimising the above loss function,

$$\arg \min_{\mathbf{c}} \|\mathbf{W}(\mathbf{y} - \mathbf{A}\mathbf{c})\|_2^2 + \lambda \|\Gamma \mathbf{c}\|_2^2, \quad (2.82)$$

where $\mathbf{W}$ is a diagonal matrix consisting of the weights $w_{E,R}$, $w_{F,R}$, $w_{V,R}$, a vector $\mathbf{y}$ concatenates energy, force, virial observations, and $\mathbf{A}$ denotes the design matrix including the ACE basis values associated with the observations. The last term $\lambda \|\Gamma \mathbf{c}\|_2^2$ is a generalised Tikhonov regularised term which prevents overfitting when solving the linear least square system. A scaling parameter λ specifies the relative weight of the regularisation, while Γ determines the regulariser form. Note that Γ is set to zero or the identity as a default and can be calculated via Automatic Relevance determination (ARD) which is distinctive among the ridge regression solvers offered in ACEPOTENTIALS.JL.

The total energy of the system is then obtained by summing over all atomic site energies, as described in Equation (2.61)

$$E(\{\mathbf{r}_i\}_{i=1}^J) = \sum_{i=1}^J \epsilon(\{\mathbf{r}_{ij}\}_{j=1}^{J-1}).$$

2.5.5 Multi-element System & Learnable Radial Basis

In case of a multi-element system, the atomic coordinates $\mathbf{r}_{ij}$ must be extended to account for the chemical identity of the atoms. Each pairwise interaction between the centre atom i and its neighbouring atom j is described by the tuple $(\mathbf{r}_{ij}, \zeta, z_j)$, where ζ and z_j denotes the atomic number of the centre atom i and of the the neighbour j . The single particle basis function can be modified into

$$\begin{aligned}\phi_{\zeta nlm}(\mathbf{r}_{ij}, z_j) &= R_{\zeta nl}(r_{ij}, z_j) Y_l^m(\hat{\mathbf{r}}_{ij}) \\ &= \delta(z_j - \zeta) R_{nl}(r_{ij}) Y_l^m(\hat{\mathbf{r}}_{ij}).\end{aligned}\tag{2.83}$$

Therefore, Equation (2.70), (2.71) and (2.78) are redefined as follows

$$\begin{aligned}A_{\zeta nlm}(\mathbf{r}_{ij}, z_j) &:= \sum_j \delta(z_j - \zeta) \phi_{nlm}(\mathbf{r}_{ij}) \\ \mathbf{A}_{\zeta \mathbf{n} \mathbf{l} \mathbf{m}}^{(N)}(\{\mathbf{r}_{ij_t}, z_{j_t}\}_{t=1}^N) &:= \prod_{t=1}^N A_{\zeta n_t l_t m_t}(\mathbf{r}_{ij_t}, z_{j_t}) \\ \mathbf{B}_{\zeta \mathbf{n} \mathbf{l} \mathbf{m}}^{(N)}(\{\mathbf{r}_{ij_t}, z_{j_t}\}_{t=1}^N) &:= \sum_{\mu \in \mathcal{M}_1} \mathcal{U}_{\mathbf{m} \mu}^{\mathbf{n} \mathbf{l}} \cdot \mathbf{A}_{\zeta \mathbf{n} \mathbf{l} \mu}^{(N)}(\{\mathbf{r}_{ij_t}, z_{j_t}\}_{t=1}^N).\end{aligned}\tag{2.84}$$

Note that the coupling coefficients $\mathcal{U}_{\mathbf{m} \mu}^{\mathbf{n} \mathbf{l}}$ are identical to those employed in the single-species case. In other words, they depend only on the angular momentum coupling structure and are completely independent of the atomic species used in the definition of the basis functions.

Multi-element systems include additional atomic number z -dependence, resulting in the configuration space greater in dimension and can dramatically decrease the density projection's efficiency [59]. The practical implications of this efficiency reduction and solutions of minimising it, such as through compositional structures or sparsification, have not yet been thoroughly investigated.

Witt et al. [15] suggested a choice for the radial basis functions depending on index n only, which means $R_{nl} = R_n$ for all l . They can be defined

$$R_{\zeta n}(r_{ij}, z_j) = f_{\text{env}}(r_{ij}, z_j, \zeta) \cdot P_n(y(r_{ij}, z_j, \zeta)), \quad (2.85)$$

where f_{env} denotes the envelope functions and $P_n(y)$ is an orthogonal basis in y -coordinates.

This functional form in Equation (2.85) is chosen to make sure that the radial basis stays smooth and numerically stable within the finite cutoff domain while also efficiently capturing local structural changes. The envelope function $f_{\text{env}}(r_{ij}, z_j, \zeta)$ imposes physical boundary conditions that ensure the basis smoothly approaches zero as r_{ij} nears the cutoff radius, which is necessary for differentiability needed for force evaluation. The orthogonal polynomial $P_n(y)$, defined in the transformed coordinate $y(r_{ij}, z_j, \zeta) \in [-1, 1]$, provides an orthogonal set of radial functions that minimises correlation between basis channels and improves the conditioning of the least-squares fitting problem. Moreover, using a radial basis that depends only on the index n (i.e., $R_{nl} = R_n$) further reduces the number of fitting parameters while maintaining sufficient flexibility to represent the smooth radial dependence of the atomic density.

In the implementation of this work within ACEPOTENTIALS.JL [15], P_n are chosen to be the *Legendre* polynomials and $y(r_{ij}, z_j, \zeta)$ represents the element-dependent distance transform which is defined as

$$y(r_{ij}, z_j, \zeta) = \left(1 + a \frac{\left(\frac{r_{ij}}{r_o(z_j, \zeta)} \right)^q}{1 + \left(\frac{r_{ij}}{r_o(z_j, \zeta)} \right)^{q-p}} \right)^{-1}, \quad (2.86)$$

where $r_o(z_j, \zeta)$ denotes the equilibrium bond-length between the centre atom (with atomic number ζ) and its neighbour (with atomic number z_j). In order to maximise resolution for the nearest-neighbour interaction, the factor a is selected to maximise the gradient of y at $r_{ij} = r_o(z_j, \zeta)$. The

choice of parameters p and q are introduced to control the resolution of the distance transform at small and large values r_{ij} , respectively. In the case of large r_{ij} , the transform behaves asymptotically as r_{ij}^{-q} , while for small r_{ij} , it can be approximated using binomial expansion as $1 - (r_{ij}^p/a)$.

The envelope function f_{env} , which defines the smooth cutoff area within the cutoff radius r_{cut} , is employed to capture the pairwise interaction $U^{(1)}$. By default, it is defined as

$$f_{\text{env}}(r_{ij}, z_j, \zeta) = \left(\frac{r_{ij}}{r_o}\right)^{-1} - \left(\frac{r_{\text{cut}}}{r_o}\right)^{-1} + \left(\frac{r_{\text{cut}}}{r_o}\right)^{-2} \left(\frac{r_{ij}}{r_o} - \frac{r_{\text{cut}}}{r_o}\right), \quad (2.87)$$

where $r_o \equiv r_o(z_j, \zeta)$. This envelope function obviously resembles a Coulomb potential form which exhibits the short-range repulsion proportional to r_{ij}^{-1} as $r_{ij} \rightarrow 0$. For higher-order interactions, the envelope function f_{env} , by default, is expressed as

$$f_{\text{env}}(r_{ij}, z_j, \zeta) = y_{ij}^2 (y_{ij} - y_{\text{cut}})^2, \quad (2.88)$$

where $y_{ij} = y(r_{ij}, z_j, \zeta)$ and $y_{\text{cut}} = y(r_{\text{cut}}, z_j, \zeta)$.

We will give more discussions on using ACEPOTENTIALS.JL and parameterising p , q and r_{cut} for the specific problems of the incorporation with the lattice force constant prediction in Chapter 3 and of the prediction of charge transfer integrals in Chapter 4.

2.6 Charge Transfer Integrals

The concept of the **transfer integrals**, commonly denoted as J , plays a central role in quantum chemistry and condensed matter physics. It quantifies the electronic coupling between molecular orbitals on adjacent sites, and thus governs key phenomena such as charge and exciton transport in molecular systems. In materials like organic semiconductors, where weak Van der Waals interactions dominate, the magnitude and orientation dependence of J critically influence charge carrier mobility. The accurate evaluation of transfer integrals is essential for the rational design of high-performance electronic and optoelectronic materials. In the first few sections, we will discuss how the transfer integrals are defined mathematically.

2.6.1 Tight-binding Hamiltonian

The microscopic framework for explaining charge transport in organic semiconductors is presented in this section. In these semiconductors, charge motion is well modelled by a tight-binding Hamiltonian that allows hopping between localised molecular orbitals. In second quantisation, the Hamiltonian is defined by [60; 61; 62]

$$\hat{H}_{\text{tb}} = \sum_a \varepsilon_a \hat{a}_a^\dagger \hat{a}_a + \sum_{ab} J_{ab} \hat{a}_a^\dagger \hat{a}_b, \quad (2.89)$$

where $\hat{a}_a^\dagger$ and $\hat{a}_a$ are the creation and annihilation operators for an electron on site a . The **site energies** ε_a represent the energy cost to place a charge on molecule a , and are influenced by disorder, polarisation effects, and the local chemical environment. The **transfer integrals** or **electronic couplings** J_{ab} , which characterise the quantum-mechanical interactions between adjacent molecular orbitals and establish the directionality and rate of charge hopping. The **site energies** are expressed as

$$\varepsilon_a = \langle \psi_a | \hat{H}_{\text{tb}} | \psi_a \rangle, \quad (2.90)$$

and the **transfer integrals** are defined as

$$J_{ab} = \langle \psi_a | \hat{H}_{\text{tb}} | \psi_b \rangle, \quad (2.91)$$

where ψ_a and ψ_b represent the single localised molecular orbital of molecule a and b , respectively. These orbitals correspond to the **Lowest Unoccupied Molecular orbital (LUMO)** for electron transport and to the **Highest Occupied Molecular orbital (HOMO)** for the transport of holes. Equations (2.90) and (2.91) assume that orbitals ψ_a and ψ_b are orthogonal; however, this is typically not the case for **HOMOs** or **LUMOs** positioned on distinct molecules. We will revisit the nonorthogonality problem and discuss how ε_a and J_{ab} change while moving from nonorthogonal to orthogonal representation later in this section.

Another important class of parameters relates to the electron–phonon couplings, which describe how molecular vibrations dynamically modulate both the site energies ε_a and the electronic couplings J_{ab} . These couplings introduce dynamic disorder, a key factor influencing charge transport in organic semiconductors. Unlike in inorganic systems, where electron–phonon interactions are typically weak, in organic semiconductors they are often comparable to—or even stronger than—the electronic interactions. As a result, electron–phonon couplings cannot be treated as a mere perturbation; instead, it leads to the formation of quasiparticles known as "small" polarons, in which the electronic charge is accompanied by a cloud of lattice vibrations [61; 62]. However, electron–phonon couplings will not be discussed in detail in this thesis. A brief overview of their role in charge transport can be found in the review article "*Charge Transport in Organic Semiconductors*" by Coropceanu et al. [60].

2.6.2 Transfer Integrals

Charge transport properties are highly sensitive to the magnitude of the transfer integrals. As introduced in the previous section, transfer integrals are defined analogously to Equation (2.91) as:

$$J_{ab} = \langle \psi_a | \hat{H} | \psi_b \rangle, \quad (2.92)$$

where $\hat{H}$ is the electronic Hamiltonian of the system. The wavefunctions ψ_a and ψ_b correspond to molecular orbital (MO) states in which an electron (or hole) is localized on site a and b , respectively. These MO states are assumed to be non-interacting, meaning there is no electronic coupling between them; hence, they are referred to as **diabatic** states [62].

In practice, diabatic states ($|\psi_a\rangle$, $|\psi_b\rangle$) cannot be directly obtained from standard density functional theory (DFT) or quantum chemistry calculations on a molecular dimer. Instead, such calculations yield adiabatic states ($|\chi_{\pm}\rangle$), in which the charge is delocalized over both molecules. These adiabatic states are the true eigenstates of the electronic Hamiltonian—that is, unlike diabatic states, they are diagonal with respect to the electronic Hamiltonian.

One approach to approximate transfer integrals from the adiabatic states is the "energy-splitting-in-dimer (ESD)" method [60; 63; 64]. This method is particularly effective for symmetric inversion point dimers, where the transfer integrals are not significantly perturbed by molecular orbital overlap. This method assumes the electronic Hamiltonian of the orthogonal diabatic states, i.e. $\langle \psi_a | \psi_b \rangle = \delta_{ab}$, is defined via the tight-binding model as

$$\hat{H} = \begin{bmatrix} \varepsilon_a & J_{ab} \\ J_{ab} & \varepsilon_b \end{bmatrix} = \begin{bmatrix} \varepsilon & J \\ J & \varepsilon \end{bmatrix}, \quad (2.93)$$

which refers to the bonding and antibonding combinations, similar to in molecular orbital theory in which the Schödinger's equation is defined as

$$\hat{H}|\Psi\rangle = E|\Psi\rangle. \quad (2.94)$$

where $|\Psi\rangle = \sum_i c_i |\psi_i\rangle$ for all $i \in a, b$. Solving the eigenvalue problem results in the energy splitting

$$E^- = \varepsilon - J \quad (\text{for bonding}) \text{ and} \quad (2.95a)$$

$$E^+ = \varepsilon + J \quad (\text{for antibonding}), \quad (2.95b)$$

which corresponds to the adiabatic eigenstates

$$|\chi_-\rangle = \frac{1}{\sqrt{2}}(|\psi_a\rangle - |\psi_b\rangle) \quad (\text{for bonding}) \text{ and} \quad (2.96a)$$

$$|\chi_+\rangle = \frac{1}{\sqrt{2}}(|\psi_a\rangle + |\psi_b\rangle) \quad (\text{for antibonding}). \quad (2.96b)$$

Regarding Equation (2.95a) and (2.95b), the transfer integrals can be determined by

$$J = \frac{\Delta E}{2} = \frac{E_2 - E_1}{2}, \quad (2.97)$$

where ΔE is the energy splitting between the two relevant molecular orbitals arising from intermolecular interaction. In the context of estimating the transfer integral for electron (or hole) transfer from molecule a to b , this splitting can be approximated from the orbital energy differences obtained from a dimer calculation. Accordingly, Equation (2.97) can be approximated as

$$J_e = \frac{\varepsilon_{\text{LUMO}+1} - \varepsilon_{\text{LUMO}}}{2} \quad (2.98a)$$

$$J_h = \frac{\varepsilon_{\text{HOMO}} - \varepsilon_{\text{HOMO}-1}}{2}, \quad (2.98b)$$

where J_e and J_h represent the transfer integrals for electron and hole transport, respectively. The energy splitting between the adiabatic dimer's frontier molecular orbitals (LUMO and HOMO) is used to estimate these.

Although this method performs efficiently for the symmetric dimers where the Hamiltonian assumes orthogonal molecular orbital states, it remains limited when applied to the systems in which the molecular orbitals overlap—that is, when the overlap matrix component

$$\langle \tilde{\psi}_a | \tilde{\psi}_b \rangle = S_{ab} \neq 0.$$

The non-orthogonal molecular orbitals that arise from this non-zero overlap component S_{ab} have an additional impact on the energy splitting between the adiabatic states as well as the transfer integrals. The ESD approach becomes inaccurate and could result in incorrect predictions of charge transport properties if this overlap is not appropriately considered for account.

To address this overlap molecular orbital, we examine the following effective time-independent Schrödinger's (TISE) equation

$$\hat{\hat{H}}|\tilde{\Psi}\rangle = E\hat{\hat{S}}|\tilde{\Psi}\rangle, \quad (2.99)$$

where $|\tilde{\Psi}\rangle = \sum_i \tilde{c}_i |\tilde{\psi}_i\rangle$ for all $i \in a, b$, and the Hamiltonian and the overlap matrix are defined as

$$\hat{\hat{H}} = \begin{bmatrix} \tilde{\varepsilon}_a & \tilde{J}_{ab} \\ \tilde{J}_{ab} & \tilde{\varepsilon}_b \end{bmatrix} \quad \text{and} \quad \hat{\hat{S}} = \begin{bmatrix} 1 & S_{ab} \\ S_{ab} & 1 \end{bmatrix}, \quad (2.100)$$

respectively. Those matrix elements are defined as follows

$$\tilde{\varepsilon}_a = \langle \tilde{\psi}_a | \hat{\hat{H}} | \tilde{\psi}_a \rangle, \quad (2.101a)$$

$$\tilde{J}_{ab} = \langle \tilde{\psi}_a | \hat{\hat{H}} | \tilde{\psi}_b \rangle, \quad (2.101b)$$

$$S_{ab} = \langle \tilde{\psi}_a | \tilde{\psi}_b \rangle. \quad (2.101c)$$

After solving this Schrödinger equation, one can use the resulting eigenvalues to calculate the dimer energy splitting given by

$$\Delta E_{ab} = \frac{\sqrt{(\tilde{\epsilon}_b - \tilde{\epsilon}_a)^2 + 4 \left(\tilde{J}_{ab}^2 - \tilde{J}_{ab} S_{ab} (\tilde{\epsilon}_b + \tilde{\epsilon}_a) + \tilde{\epsilon}_a \tilde{\epsilon}_b S_{ab}^2 \right)}}{1 - S_{ab}^2}. \quad (2.102)$$

In the case of electron transfer, this energy splitting and the transfer integrals $\tilde{J}_{ab}$ are equivalent to $\epsilon_{\text{LUMO}+1} - \epsilon_{\text{LUMO}}$ and the electron transfer integrals J_e , respectively, as defined in Equation (2.98a). However, extracting the electron transfer integrals directly from ΔE_{12} is challenging due to the complexity introduced by the non-zero overlap S_{ab} unlike the simpler expression available in the orthogonal case.

We can employ the Löwdin's symmetric orthogonalisation [65] to transform the non-orthogonal states $|\tilde{\psi}_a\rangle$ into the orthogonal states $|\psi_a\rangle$, i.e.

$$|\psi_i\rangle = \sum_j (\hat{S}^{-1/2})_{ij} |\tilde{\psi}_j\rangle; \quad \forall i, j \in \{a, b\}, \quad (2.103)$$

where $\hat{S}^{-1/2}$ depicts the inverse square root of the overlap matrix. In many organic semiconductors, the overlap matrix element $\hat{S}_{ab} \ll 1$, so we can expand the inverse square root of the overlap matrix as

$$\hat{S}^{-1/2} \approx \frac{1}{\sqrt{1 - S_{ab}^2}} \begin{bmatrix} 1 & -S_{ab}/2 \\ -S_{ab}/2 & 1 \end{bmatrix}. \quad (2.104)$$

Now, we can apply this transformation to calculate the Hamiltonian matrix elements in the orthogonal basis as follows

$$\begin{aligned} \hat{H}_{ij} &= \langle \psi_i | \hat{H} | \psi_j \rangle \\ &= \sum_{mn} (\hat{S}^{-1/2})_{im}^\dagger \langle \tilde{\psi}_m | \hat{H} | \tilde{\psi}_n \rangle (\hat{S}^{-1/2})_{nj}, \end{aligned} \quad (2.105)$$

where $i, j \in \{a, b\}$. The diagonal site energies are defined as

$$\varepsilon_{a(b)} \approx \frac{\tilde{\varepsilon}_{a(b)} - \tilde{J}_{ab} S_{ab}}{1 - S_{ab}^2}, \quad (2.106)$$

while the off-diagonal transfer integrals are

$$J_{ab} \approx \frac{\tilde{J}_{ab} - \frac{1}{2}(\tilde{\varepsilon}_a + \tilde{\varepsilon}_b) S_{ab}}{1 - S_{ab}^2}. \quad (2.107)$$

These are equivalent to Equation (28) and (29) in the review by Coropceanu et al. [60], where S_{ab}^2 are small. The dimer energy splitting can be determined by solving the Schrödinger's equation of the Hamiltonian of the orthogonal states and is

$$\Delta E_{ab} = \frac{\sqrt{(\varepsilon_b - \varepsilon_a)^2 + 4J_{ab}^2}}{1 - S_{ab}^2}. \quad (2.108)$$

Substituting Equation (2.106) and (2.107) into this dimer energy splitting recovers Equation (2.102)

$$\Delta E_{ab} = \frac{\sqrt{(\tilde{\varepsilon}_b - \tilde{\varepsilon}_a)^2 + 4\left(\tilde{J}_{ab}^2 - \tilde{J}_{ab} S_{ab}(\tilde{\varepsilon}_b + \tilde{\varepsilon}_a) + \tilde{\varepsilon}_a \tilde{\varepsilon}_b S_{ab}^2\right)}}{1 - S_{ab}^2}.$$

2.6.3 Computational Approaches

Transfer integrals can be evaluated using a variety of calculation methods, each with differing degrees of precision and complexity:

Dimer Projection Method

The dimer projection (DIPRO) method is the foundational method to study the transfer integrals or electronic couplings of molecular dimers. DIPRO projects dimer molecular orbitals $|\tilde{\Psi}\rangle$ onto isolated monomer molecular

orbitals $\{|\psi_a\rangle, |\psi_b\rangle\}$ in order to address non-orthogonality. The DIPRO procedure can be divided into four following steps:

- Employing quantum chemistry software (e.g., GAUSSIAN, ORCA, or Q-CHEM) to perform a separately standard DFT calculation to obtain isolated monomer MOs $\{|\psi_a^{\text{iso}}\rangle, |\psi_b^{\text{iso}}\rangle\}$ of molecule a and b .
- Performing a single-point DFT calculation on the corresponding molecular dimer to achieve the dimer Hamiltonian $\tilde{H}^{\text{dim}}$ in Equation (2.100).
- Projecting the dimer Hamiltonian onto the isolated monomer MOs the calculate Hamiltonian matrix elements $\tilde{H}_{ij}^{\text{dim}}$ and determining the overlap matrix elements S_{ij} , for all $i, j \in \{a, b\}$;

$$\tilde{\epsilon}_i = \langle \psi_i^{\text{iso}} | \tilde{H}^{\text{dim}} | \psi_i^{\text{iso}} \rangle, \quad (2.109a)$$

$$\tilde{J}_{ij} = \langle \psi_i^{\text{iso}} | \tilde{H}^{\text{dim}} | \psi_j^{\text{iso}} \rangle, \quad (2.109b)$$

$$S_{ij} = \langle \psi_i^{\text{iso}} | \psi_j^{\text{iso}} \rangle, \quad (2.109c)$$

Note that, isolated monomer molecular orbitals (MOs), obtained from separate DFT calculations, are initially not defined with respect to orbital overlap. Only after projecting the dimer Hamiltonian onto these isolated MOs does the orbital overlap become explicit. This projection procedure enables us to computationally evaluate Equations (2.101a)-(2.101c), leading directly to Equations (2.109a)-(2.109c).

- Using the computed quantities $\tilde{\epsilon}_i$, $\tilde{J}_{ij}$ and S_{ij} as inputs, one can calculate the site energies and transfer integrals via Equations (2.106) and (2.107), respectively. These results can then be employed to determine the energy splitting according to Equation (2.108).

The DIPRO method is highly effective for accurately calculating electronic couplings in molecular dimers, particularly benefiting from its ability to handle asymmetry and orbital non-orthogonality explicitly. While it provides

physically clear and robust transfer integrals, DIPRO does require careful orbital localization, precise monomer calculations, and additional computational effort compared to simpler approaches like the ESD method. Nevertheless, its accuracy and flexibility, rooted in projecting the dimer Hamiltonian onto isolated monomer orbitals as detailed in Equations (2.109a)-(2.109c), make it the foundational and preferred method for studying transfer integrals in complex molecular systems.

Recently, Kohn et al. (2023) significantly accelerates the standard DIPRO method by incorporating a novel semi-empirical potential called the density matrix tight-binding potential [66]. This incorporation significantly decrease computational cost, since PTB achieves a roughly 300-fold computational speed-up compared to standard DFT [67]. Unlike full DFT, which explicitly solves the complex Kohn–Sham equations iteratively (requiring multiple SCF cycles), PTB employs a parametrised, semiempirical Hamiltonian. This PTB-based DIPRO enables efficient calculations for much larger molecular systems, such as metal-organic frameworks and large molecular aggregates.

ZINDO with Molecular Orbital Overlap

Another approach to evaluate the transfer integrals was presented by Kirkpatrick in 2007 [68]. He introduced the method, based on the Zerner’s Independent Neglect of Differential Overlap (ZINDO) Hamiltonian [69], which closely relates to the DIPRO (Dimer Projection) method. Both methods involve projecting molecular orbitals (MOs) of a dimer onto MOs defined on isolated monomers to calculate the transfer integrals.

The DIPRO method explicitly calculates the Hamiltonian (or Fock) matrix elements and orbital overlaps by performing a full DFT-based self-consistent field (SCF) calculation on the dimer. In contrast, the ZINDO-based approximation, developed by Kirkpatrick, avoids these computationally intensive dimer-level SCF calculations. It estimates transfer integrals using a simpli-

fied analytical expression for atomic orbital (AO) overlaps, derived from a semiempirical framework. The AO Fock matrix elements are defined as

$$F_{\mu\nu}^{\text{AO}} = \bar{S}_{\mu\nu}^{\text{AO}} \frac{\beta_i + \beta_j}{2} + P_{\mu\nu} \frac{\gamma_{ij}}{2}, \quad (2.110)$$

where $\bar{S}_{\mu\nu}^{\text{AO}}$ represents the weighted atomic orbital overlap matrix elements between atomic orbital μ and ν , which are centered on atom i and j , respectively. The parameters β_i and β_j represent atomic bonding parameters specific to atoms i and j , and γ_{ij} is Mataga–Nashimoto potential describing electron-electron repulsion between these atoms. $P_{\mu\nu}$ is the density matrix element, accounting for electron density coupling between AOs.

A key simplification of Zindo is the assumption that the density matrix P is block diagonal, meaning $P_{\mu\nu} \approx 0$ for inter-molecular pairs. This effectively neglects electron density contributions between atoms on different monomers, reducing the Fock matrix elements to

$$F_{\mu\nu}^{\text{AO}} \approx \bar{S}_{\mu\nu}^{\text{AO}} \frac{\beta_i + \beta_j}{2}. \quad (2.111)$$

Each monomer's molecular orbitals are expressed in the linear combination of atomic orbitals (LCAO) form

$$|\psi_a^{\text{iso}}\rangle = \sum_{\mu \in a} C_{\mu}^{(a)} |\phi_{\mu}\rangle, \quad (2.112)$$

where $C_{\mu}^{(a)}$ is the coefficient describing the contribution of atomic orbital μ to the monomer MO of molecule a . The transfer integral between orbitals $|\psi_a^{\text{iso}}\rangle$ and $|\psi_b^{\text{iso}}\rangle$ on monomers a and b is then calculated as

$$\begin{aligned} J_{ab} &= \langle \psi_a^{\text{iso}} | \hat{F} | \psi_b^{\text{iso}} \rangle \\ &= \sum_{\mu \in a} \sum_{\nu \in b} C_{\mu}^{(a)} C_{\nu}^{(b)} F_{\mu\nu}^{\text{AO}}. \end{aligned} \quad (2.113)$$

The work in [68] shows the ZINDO-based method produces transfer integrals in good agreement with those obtained from the full DIPRO method.

This validation makes the ZINDO-based approximation a valuable tool for large-scale simulations, where the computational cost of full SCF calculations would be prohibitive. This approach is particularly useful for high-throughput screening of materials.

2.6.4 Machine Learning-Based Approaches

In recent years, several machine-learning-based approaches have been developed to predict transfer integrals. The remarkable work, done by Lederer et al. in 2018 [24], is the prediction of the pentacene dimer's transfer integrals using a Gaussian process (GP) model which we discuss in section 2.3. They employed a Gaussian kernel, as defined in Equation (2.33), and a simple geometric descriptor containing relative orientations and positions of the molecules in the system. In 2020, they released the following work that improves the model accuracy by changing the model descriptor into Coulomb matrix (CM)-based descriptors and performed them with a Gaussian kernel and a Laplacian kernel [70].

During the same period, Wang et al. [25] also predicted the transfer integrals of ethylene dimers via GP regression characterised by a Gaussian kernel and a Laplacian kernel. Additionally, they investigated the effect of different CM descriptors—such as intramolecular, intermolecular, and individual-atom contributions—on the prediction. Their results clearly demonstrated that the intermolecular contributions alone are sufficient to accurately describe the system, as transfer integrals inherently represent quantum-mechanical intermolecular interactions. In their following work [71], they replaced the GP regression model with artificial neural networks (ANNs). The results show ANNs are capable of capturing complex nonlinear relationships of transfer integrals more effectively. This allows them to predict the transfer integrals of more complex molecular systems, i.e. the naphthalene dimers in amorphous environments.

In contrast to solely geometric or Coulomb-matrix-based methods, Hafizi et al. [72] created a neural-network-based model that estimates transfer integrals by directly adding up atomic contributions. The complexities of local atomic interactions are adequately expressed by their model’s use of atomic local-environment descriptors built from symmetry functions (SF). To enable them to select training data as efficiently as possible, they also introduced farthest-point sampling (FPS) and active learning techniques (such as query-by-committee, or QbC). Compared to earlier approaches, this strategy significantly lowers the amount of the training set that is needed. Additionally, their work integrates physics-based models through a Δ -machine learning (Δ -ML) framework, where the neural network learns to correct predictions from an existing analytic baseline model. This combination of data efficiency and physical grounding enhances both the accuracy and generalizability of the model.

Building on the idea of incorporating physical structure into machine learning, the Atomic Cluster Expansion (ACE) framework [2; 15; 16], discussed in section 2.5, may offer a powerful and systematically improvable approach for predicting transfer integrals. ACE can provide a symmetry-preserving descriptor of atomic environments, crucial for capturing subtle electronic couplings that influence charge transport. Its capability to include higher-order many-body terms with linear computational scaling, along with its flexibility to encode both geometric and chemical features, can make ACE a highly efficient and accurate tool for modelling quantum properties such as transfer integrals.

Learning transfer integrals directly, instead of entire Hamiltonians or density matrices, has both conceptual and computational benefits. Transfer integrals are scalar quantities that have physical meaning and measure how well electronic coupling works between molecular sites. This means that they only store the part of the Hamiltonian information associated to charge transport.

This dimensional reduction simplifies the learning problem, makes data more useful, and makes it easier to understand by directly targeting the quantities that go into rate theories like Marcus or hopping transport models. This simplification, however, sacrifices generality; models trained exclusively on transfer integrals are not readily extendable to retrieve comprehensive electronic structure information or to characterise properties necessitating the complete Hamiltonian or density matrix (e.g., band dispersion or polarisation).

On the other hand, learning the Hamiltonian or density matrix gives a more complete and transferable picture of the electronic system. However, this needs much bigger training datasets, stricter phase-consistency rules, and a higher chance of overparameterisation. So, learning transfer integrals is a tractable way to balance physical relevance and computational tractability. This makes it an efficient and focused strategy when charge transport is the main property of interest.

3

Anharmonic Phonons with Gaussian Processes

In this chapter, we introduce an approach for predicting harmonic (FC2) and cubic anharmonic (FC3) force constants using differentiable Gaussian process (GP) models. This approach uses the smoothness and uncertainty quantification offered by GPs to efficiently estimate force constants from a small number of atomic configurations. We describe how this methodology is implemented in the GPFC.JL package and show how it may be used for calculations involving phonons and lattice dynamics.

We begin by establishing the connection between force constants and differentiable Gaussian process (GP) models. Building on this foundation, we derive the formulation of a derivative GP model using Cartesian descriptors for predicting harmonic (FC2) and cubic anharmonic (FC3) force constants. The results and discussions are presented in Section 3.1. Section 3.2 introduces the use of phonon descriptors in GPFC.JL. These descriptors naturally impose translational symmetry and reduce the model size, leading to an improvement in the accuracy and efficiency of the model. We compare the learning behaviour of models trained with phonon descriptors versus Cartesian descriptors and outline the procedure for recovering FC2s and FC3s from the phonon basis to the Cartesian basis. Section 3.3 discusses possible extensions, including the integration of atomic cluster expansion (ACE) descriptors and the use of active learning strategies for enlarging

training sets, facilitated by ACEPOTENTIALS.JL. Finally, Section 3.4 outlines directions for future work.

3.1 GPFC.JL with Cartesian descriptors

In this section, we describe the construction of the differentiable Gaussian process (GP) model used in GPFC.JL for predicting force constants from atomic forces, using Cartesian descriptors. Cartesian descriptors represent atomic environments directly in terms of relative atomic positions, preserving full spatial resolution without imposing any symmetry constraints. This choice allows us to directly construct the GP model derivatives with respect to atomic displacements, enabling the prediction of harmonic and cubic anharmonic force constants (FC2s and FC3s) from prior energies and forces of materials. We begin by outlining the theoretical formulation of the GP model and its derivatives, followed by implementation details and validation results.

3.1.1 Higher Derivative GP Model

To extract harmonic and anharmonic force constants from a derivative Gaussian process (GP) model of the force field, it is necessary to compute higher-order derivatives of the predicted forces with respect to atomic displacements. Specifically, the second derivatives correspond to the force constant matrix (FC2), and the third derivatives yield the cubic anharmonic force constants (FC3), as defined in the Taylor expansion of the potential energy in Equation (2.2).

We begin by revisiting the derivative GP model with a kernel function $k(\mathbf{x}, \mathbf{x}^*)$ which defines the similarity between the descriptor $\mathbf{x}$ and $\mathbf{x}^*$ of N atoms in Cartesian representations, i.e. $\mathbf{x}, \mathbf{x}^* \in \mathbb{R}^{3N}$. In our implementation, we

mainly use the square exponential (radial basis function: RBF) kernel, equivalent to (2.33), generically defined as

$$k(\mathbf{x}, \mathbf{x}^*) = \sigma_o \exp \left[-\frac{1}{2\ell^2} d(\mathbf{x}, \mathbf{x}^*)^2 \right], \quad (3.1)$$

where $d(\mathbf{x}, \mathbf{x}^*) = \|\mathbf{x} - \mathbf{x}^*\|_2^2$ is the Euclidean metric. The kernel scale σ_o and the length scale ℓ can be obtained from hyperparameter optimisation.

The SE kernel was chosen because it assumes a smooth, infinitely differentiable latent function. This is physically well-motivated for electronic-structure-related quantities like potential energies that change smoothly with small changes in geometry. Its isotropic form also provides a straightforward way to encode geometric similarity without introducing additional directional parameters, making it computationally efficient and numerically stable during training.

A systematic grid-based scan was used to find the kernel hyperparameters σ_o and ℓ instead of doing full evidence maximisation. This manual optimisation method was used to make sure that the smoothness and amplitude scales were under control. This way, the chosen values would match the physically meaningful changes in the data without causing numerical problems that can happen when automatic optimisation is used on small or noise-sensitive datasets.

Conditioning this GP model with the energy E and forces ∇E at $\mathbf{x}$ yields the posterior potential energy at $\mathbf{x}^*$, analogous to (2.45) and (2.52), expressed as

$$E(\mathbf{x}^*) = \begin{bmatrix} k & (\nabla_{\mathbf{x}} k)^T \end{bmatrix}_{\mathbf{x}, \mathbf{x}^*} \cdot \mathcal{K}_{\mathbf{x}, \mathbf{x}}^{-1} \cdot \begin{bmatrix} E \\ \nabla E \end{bmatrix}_{\mathbf{x}}, \quad (3.2)$$

where the subscripts indicate the functions evaluated at the descriptors. The first derivative of the kernel function $(\nabla_{\mathbf{x}} k)$ with respect to the descriptor $\mathbf{x}$ correlates the potential energy at $\mathbf{x}^*$ to the force fields at $\mathbf{x}$. This derivative,

as mentioned in section 2.3.3, can be represented as a $3N \times 1$ covariance matrix (Jacobian matrix as a first rank tensor)

$$[\nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}^*)]^T = \left[\frac{\partial k}{\partial x_1} \quad \frac{\partial k}{\partial x_2} \cdots \frac{\partial k}{\partial x_{3N}} \right]^T. \quad (3.3)$$

Marginal likelihood covariance matrix $\mathcal{K}$ which defines the probability distribution $(1 + 3N) \times (1 + 3N)$ matrix of the priors, is

$$\mathcal{K}(\mathbf{x}, \mathbf{x}) = \begin{bmatrix} k & (\nabla_{\mathbf{x}} k)^T \\ \nabla_{\mathbf{x}^*} k & \nabla_{\mathbf{x}^*} \nabla_{\mathbf{x}} k \end{bmatrix}_{\mathbf{x}^*, \mathbf{x}} + \begin{bmatrix} \sigma_e^2 & \mathbf{0} \\ \mathbf{0} & \sigma_f^2 \cdot \mathbb{1}_{3N \times 3N} \end{bmatrix}. \quad (3.4)$$

The first matrix contains the **kernel**, the **Gradient kernel**, and the **Hessian kernel** for constructing the correlation between the prior energy and the prior forces. The expressions of those derivatives of the kernel functions are illustrated in (3.3) and (2.41). The second matrix is the observation noise matrix, which includes σ_e and σ_f indicates observation noises for prior energy and force, respectively. One of the choices for these noises is the energy and force convergence factor in the underlying electronic structure calculations used for training.

The recent works [3; 7; 8; 42] on the derivative GPs suggested that the prediction of potential energy can be potentially extended to predict forces at $\mathbf{x}^*$. This force prediction requires evaluating the first and second order derivatives of the kernel function with respect to $\mathbf{x}$ and $\mathbf{x}^*$ as follows:

$$\nabla_{\mathbf{x}^*} k \quad \text{and} \quad \nabla_{\mathbf{x}^*} \nabla_{\mathbf{x}} k.$$

The first derivative is already expressed in Equation (2.40). The second derivative is defined as the $(3N \times 3N)$ rank-2 tensor, linking the force fields at $\mathbf{x}$ to the force fields at $\mathbf{x}^*$. This allows us to reconstruct (3.2) with changes in those covariance matrices:

$$\nabla E(\mathbf{x}^*) = \begin{bmatrix} \nabla_{\mathbf{x}^*} k & \nabla_{\mathbf{x}^*} \nabla_{\mathbf{x}} k \end{bmatrix}_{\mathbf{x}^*, \mathbf{x}} \cdot \mathcal{K}_{\mathbf{x}, \mathbf{x}}^{-1} \cdot \begin{bmatrix} E \\ \nabla E \end{bmatrix}_{\mathbf{x}}. \quad (3.5)$$

Motivated by this extension, we perform the higher-order differentiation to predict the harmonic (FC2) and cubic anharmonic (FC3) force constants and present the model construction in our work [9]. To predict FC2 requires extra differentiation on those covariance matrices with respect to $\mathbf{x}^*$ which are equivalent to the second and third order kernel derivative with respect to $\mathbf{x}$ and $\mathbf{x}^*$, given by

$$(\nabla_{\mathbf{x}^*})^2 k \quad \text{and} \quad (\nabla_{\mathbf{x}^*})^2 \nabla_{\mathbf{x}} k. \quad (3.6)$$

These $(3N \times 3N)$ rank-2 and $(3N \times 3N \times 3N)$ rank-3 tensors are interpreted as the covariance correlating the harmonic force constant at $\mathbf{x}^*$ to the energy and to the forces, respectively at $\mathbf{x}$. Equation (3.5) can, then, be modified into

GPFC2 prediction

$$\Phi^{(2)}(\mathbf{x}^*) = \left[(\nabla_{\mathbf{x}^*})^2 k \quad (\nabla_{\mathbf{x}^*})^2 \nabla_{\mathbf{x}} k \right]_{\mathbf{x}, \mathbf{x}^*} \cdot \mathcal{K}_{\mathbf{x}, \mathbf{x}}^{-1} \cdot \begin{bmatrix} E \\ \nabla E \end{bmatrix}_{\mathbf{x}}, \quad (3.7)$$

where $\Phi^{(2)}$ is the Gaussian process $(3N \times 3N)$ harmonic force constant tensor (GPFC2) at descriptor $\mathbf{x}^*$, constructed by correlating all degrees of freedom of $\mathbf{x}^*$ via the derivative GP model.

It is important to emphasise that this formulation follows the same fundamental learning framework as the original Gaussian Approximation Potential (GAP) approach [56], which already incorporates total energy and derivative (force) observations as training targets, combined with sparsification for computational efficiency. As reviewed comprehensively in *Chemical Reviews* (2021) [3], this joint energy–force learning strategy underpins nearly all GAP-based applications to molecular and condensed-phase systems. The present Gaussian Process Force Constant (GPFC) model adopts this established foundation but extends it through an additional analytical differentiation of the trained Gaussian process posterior, allowing direct com-

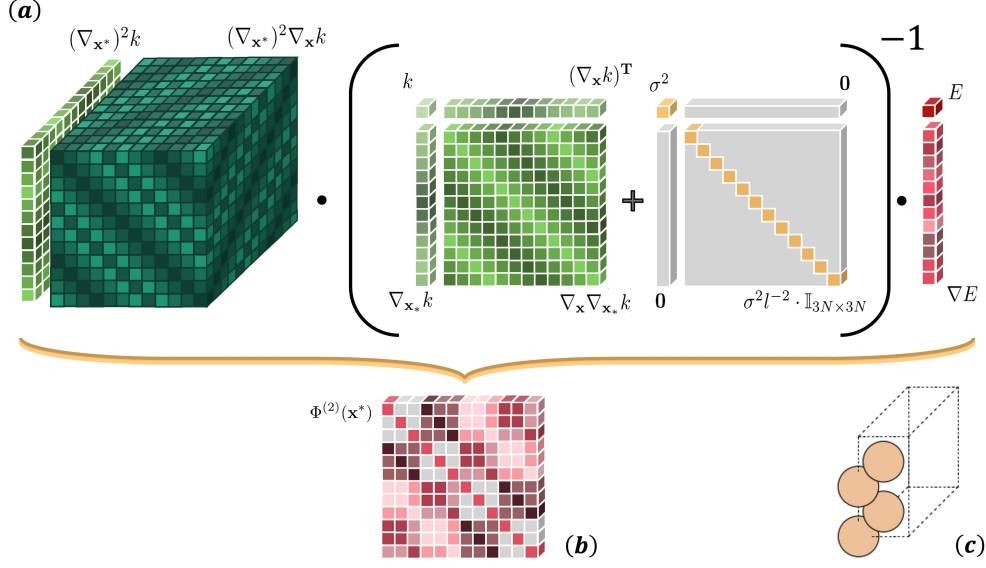


Figure 3.1: Figure illustrates the tensor contraction in Equation (3.7), where the derivative GP model is trained on a single data point at descriptor $\mathbf{x}$.

putation of higher-order force constants (harmonic and cubic anharmonic) from the same learned potential energy surface.

This post-processing differentiation step—absent from the original GAP formalism—constitutes the principal methodological novelty of the GPFC framework.

Equation (3.7) can be constructed as a tensor contraction (instead of the matrix multiplication) as illustrated in Figure 3.1a. The resulting harmonic force constant tensor $\Phi^{(2)}$ is predicted at descriptor $\mathbf{x}^*$, as shown in Figure 3.1b, for a $2 \times 1 \times 1$ diamond-structure silicon supercell depicted in Figure 3.1c [9].

The prediction of the cubic anharmonic force constant (FC3), similarly, requires evaluating the rank-3 and rank-4 covariance tensors correlating the FC3 to the potential energy and to the force fields, respectively. These covariance tensors are defined as

$$(\nabla_{\mathbf{x}^*})^3 k \quad \text{and} \quad (\nabla_{\mathbf{x}^*})^3 \nabla_{\mathbf{x}} k. \quad (3.8)$$

```

1 using KernelFunctions           # Kernel methods for ML
2 using ForwardDiff               # Forward-mode AD
3 using Zygote                    # Reverse-mode AD
4
5 kernel = SqExponentialKernel() # Define RBF kernel
6
7 function diff3_kernel(kernel, x, x_star)
8     ForwardDiff.jacobian(
9         a -> Zygote.hessian(
10             b -> kernel(b, x_star),
11             a
12         ),
13         x
14     )
15 end

```

Figure 3.2: Julia implementation of a third-order derivative of a kernel function, computed by applying forward-mode automatic differentiation (AD) via `ForwardDiff.jacobian` on the Hessian obtained from reverse-mode AD via `Zygote.hessian`.

Evaluating these allows us to recast Equation (3.7) as

GPFC3 prediction

$$\Phi^{(3)}(\mathbf{x}^*) = \left[(\nabla_{\mathbf{x}^*})^3 k \quad (\nabla_{\mathbf{x}^*})^3 \nabla_{\mathbf{x}} k \right]_{\mathbf{x}, \mathbf{x}^*} \cdot \mathcal{K}_{\mathbf{x}, \mathbf{x}}^{-1} \cdot \begin{bmatrix} E \\ \nabla E \end{bmatrix}_{\mathbf{x}}, \quad (3.9)$$

where $\Phi^{(3)}$ the predicted cubic anharmonic force constant tensor (GPFC3) is the size of $(3N \times 3N \times 3N)$.

With these methods for directly evaluating force constants from the fitted Gaussian process model, we now have all the essential components for performing lattice dynamics calculations, including anharmonic effects.

To efficiently evaluate those derivatives of the kernel function, our implementation employs `FORWARDDIFF.JL` and `ZYGOTE.JL`, *Automatic Differentiation (AD)* packages in `JULIA`, for doing differentiation on the kernel function

defined from `KERNELFUNCTIONS.JL`. The example of the evaluation of the third-order derivative of the square exponential kernel with respect to $\mathbf{x}$, $(\nabla_{\mathbf{x}})^3 k$, is given by the code in Figure 3.2.

Combining these two AD packages results in the efficient computation of higher-order derivatives compared to implementing `FORWARDDIFF.JL` purely. This is due to the fact that `ZYGOTE.JL`’s reverse-mode AD effectively manages scalar-output functions with substantially reduced memory allocations, avoiding the computational cost of repeating nested forward-mode dual number evaluations, which are necessary for higher-order derivatives.

Moreover, the full derivations for differentiating the kernel function to arbitrary order with respect to the input descriptors are provided in Appendix A.1. These derivations reveal the symmetry and antisymmetry properties of derivative kernels in (3.6) and (3.8), at a given order, which can be exploited to significantly reduce the number of required kernel derivative evaluations.

3.1.2 Dataset Preparation & Training

To evaluate the performance of our method and compare it with standard approaches, we consider a set of representative solid-state materials. We benchmark our approach—referred to as GPFC (Gaussian Process Force Constants)—against two established methods: the traditional finite displacement method, as implemented in `PHONOPY` [13; 14], and third-order many-body perturbation theory, as implemented in `PHONO3PY` [12; 13]. For consistency and comparability, we process force constants predicted by our model using both `PHONOPY` and `PHONO3PY`.

Dataset Preparation

Details of the training datasets used for each method are provided below:

- **PHONO(3)PY-dataset:** We use the PHONO(3)PY package to generate displaced structures of $2 \times 2 \times 2$ supercells using the finite difference method (FDM). Energies and forces for these structures are computed using density functional theory (DFT) via VASP, with the following parameters: plane-wave cutoff energy (ENMAX = 800 eV), self-consistent field energy convergence (EDIFF = 10^{-8} eV), and a Monkhorst–Pack k -point mesh of [6,6,6]. The resulting forces are processed with PHONOPY to generate FC2s and FC3s using space group symmetry operations.
- **GPFC-dataset:** Rattled $2 \times 2 \times 2$ supercell structures are generated using the ASE package, with random displacements sampled from a normal distribution with a standard deviation of 0.01 Å. Using the same computational parameters earlier mentioned, DFT is used in VASP to calculate the corresponding energies and forces. Equations (3.5) and (3.6) describe how the derivative GP model is trained using these data without imposing any symmetry constraints.

Training

In the training process, we aim to optimise the hyperparameters of the GPFC model—namely, the kernel output scale σ_o , the length scale l , and the observation noise parameters for the potential energy σ_e and forces σ_f —by maximising the logarithm of the marginal likelihood of the prior dataset.

Let the prior dataset consist of m total-energy values and corresponding force vectors, each associated with a descriptor $\mathbf{x}_i$ for a "global" system of N atoms. We denote

$$\mathbf{E} := [E_1 \ E_2 \cdots E_m]^T, \quad (3.10a)$$

$$\mathbf{f} := \{\nabla_{\mathbf{x}_1} E_1 \ \nabla_{\mathbf{x}_2} E_2 \cdots \nabla_{\mathbf{x}_m} E_m\}, \quad (3.10b)$$

$$\mathbf{X} := \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}, \quad (3.10c)$$

where $\mathbf{E}$ and $\mathbf{f}$ represent the training targets for energy and force, respectively. Each energy E_i is subject to observation noise σ_e arising from electronic structure calculation errors or convergence limitations and can be expressed as:

$$E_i = E(\mathbf{x}_i) + \epsilon_e, \quad \epsilon_e \sim \mathcal{N}(0, \sigma_e^2). \quad (3.11)$$

Similarly, each force vector $\nabla_{\mathbf{x}_i} E_i$ carries its own observation noise σ_f :

$$\nabla_{\mathbf{x}_i} E_i = \nabla_{\mathbf{x}_i} E(\mathbf{x}_i) + \epsilon_f, \quad \epsilon_f \sim \mathcal{N}_{3N}(\mathbf{0}, \Sigma_f), \quad (3.12)$$

where $\Sigma_f = \sigma_f^2 \cdot \mathbb{I}_{3N \times 3N}$.

Assuming (3.11) and (3.12) allows us to construct an independent $(1 + 3N)$ -multivariate distribution for a pair $\mathbf{y}_i$ of the energy E_i and the corresponding force fields $\nabla_{\mathbf{x}_i} E_i$, i.e.,

$$\mathbf{y}_i = \begin{bmatrix} E_i \\ \nabla_{\mathbf{x}_i} E_i \end{bmatrix}_{(1+3N) \times 1}.$$

The covariance of any two of the prior data pairs can be expressed as

$$\langle \mathbf{y}_i \mathbf{y}_j^T \rangle = \begin{bmatrix} k & (\nabla_{\mathbf{x}_j} k)^T \\ \nabla_{\mathbf{x}_i} k & \nabla_{\mathbf{x}_i} \nabla_{\mathbf{x}_j} k \end{bmatrix}_{\mathbf{x}_i, \mathbf{x}_j} + \delta_{ij} \begin{bmatrix} \sigma_e^2 & \mathbf{0} \\ \mathbf{0} & \sigma_f^2 \cdot \mathbb{I}_{3N \times 3N} \end{bmatrix}. \quad (3.13)$$

Note that, in the case $i = j$, we can recover the marginal likelihood covariance matrix $\mathcal{K}_{\mathbf{x}_i, \mathbf{x}_i}$, defined in (3.4). This $(1 + 3N) \times (1 + 3N)$ covariance matrix allows the multivariate probability distribution (marginal likelihood) of each prior $\mathbf{y}_i$ to be stated with zero mean as

$$P(\mathbf{y}_i) = \frac{1}{\sqrt{(2\pi)^{1+3N} \det[\mathcal{K}_{\mathbf{x}_i, \mathbf{x}_i}]}} \exp \left[-\frac{1}{2} \mathbf{y}_i^T \cdot \mathcal{K}_{\mathbf{x}_i, \mathbf{x}_i}^{-1} \cdot \mathbf{y}_i \right]. \quad (3.14)$$

Hence, the marginal likelihood across all priors,

$$\mathbf{Y} = \begin{bmatrix} \mathbf{E} \\ \mathbf{f} \end{bmatrix}_{m \cdot (1+3N) \times 1},$$

can be expressed as a *multivariate* Gaussian distribution with zero mean and the $m \cdot (1 + 3N) \times m \cdot (1 + 3N)$ covariance matrix, $\langle \mathbf{Y}\mathbf{Y}^T \rangle = \mathcal{K}_{\mathbf{X}, \mathbf{X}}$:

$$P(\mathbf{Y}) = \frac{1}{\sqrt{(2\pi)^{m \cdot (1+3N)} \det[\mathcal{K}_{\mathbf{X}, \mathbf{X}}]}} \exp \left[-\frac{1}{2} \mathbf{Y}^T \cdot \mathcal{K}_{\mathbf{X}, \mathbf{X}}^{-1} \cdot \mathbf{Y} \right]. \quad (3.15)$$

Now, we can take the logarithm of this marginal likelihood as

Marginal Log-likelihood

$$\begin{aligned} \log P(\mathbf{Y}) = & -\frac{1}{2} \mathbf{Y}^T \cdot [\mathcal{K}_{\mathbf{X}, \mathbf{X}}] \cdot \mathbf{Y} \\ & -\frac{1}{2} \log(\det[\mathcal{K}_{\mathbf{X}, \mathbf{X}}]) - \frac{m \cdot (1 + 3N)}{2} \log 2\pi, \end{aligned} \quad (3.16)$$

and then maximise it to compute those hyperparameters. However, in this work we set the hyperparameters to fixed, physically reasonable values. The kernel scale σ_o and length scale l are set to 1.0 and 0.3-0.5, respectively, which are appropriate for atomic displacements sampled from a normal distribution with a standard deviation in the range of 0.01–0.05 Å, as suggested in Ref. [7].

The observation noise parameters σ_e and σ_f are both set to 10^{-8} , matching the self-consistent field (SCF) energy convergence threshold (EDIFF) used in our VASP calculations. This choice reflects the assumption that the reference energies and forces obtained from DFT are effectively noise-free within the numerical precision of the electronic structure method. In practice, such a tight noise level ensures that the Gaussian process regression treats the DFT data as highly reliable, enforcing a near-exact interpolation of the training points. Larger noise values were tested but resulted in marginally reduced accuracy and less stable hyperparameter optimisation, confirming that the 10^{-8} setting provides the best balance between numerical stability and physical consistency with the underlying quantum-mechanical data.

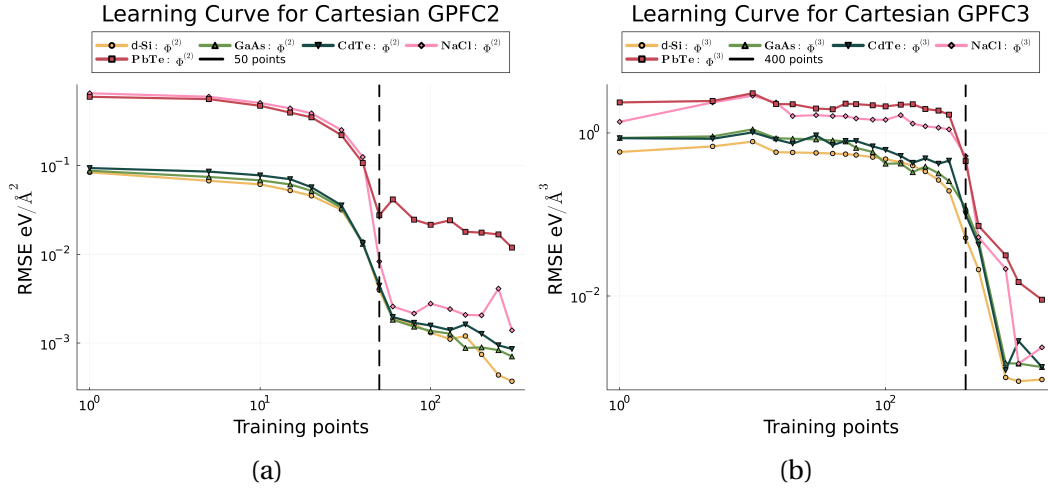


Figure 3.3: Learning curves in (a) and (b) showing the RMSE of predicted GPFC2 and GPFC3 in Cartesian descriptors compared to FC2 and FC3 values, respectively, obtained via the finite displacement method (FDM) using PHONOPY.

3.1.3 Anharmonic Phonons from GPFCs

As shown in our previous work [9], we evaluate the accuracy of our predicted harmonic (second-order) force constants by comparing the GPFC2 results obtained from our model against those from the standard finite displacement method (FDM) as implemented in PHONOPY. The comparison is performed for five representative materials: diamond-structure silicon (d-Si, FCC), **GaAs**, **CdTe**, **NaCl**, and **PbTe**, all expressed in Cartesian descriptors.

The model accuracy is quantified using root mean square errors (RMSE), shown as learning curves in Figure 3.3 (a). For all five materials, the RMSE curves in the Cartesian basis converge to near-zero values with approximately 50 training data points. This value aligns with the number of degrees of freedom in the descriptor: for a $2 \times 2 \times 2$ supercell of a two-atom primitive cell, there are 48 Cartesian components. Remarkably, these materials' learning curves show comparable convergence behaviour with respect to the number of training samples, though their chemistry and bonding differ.

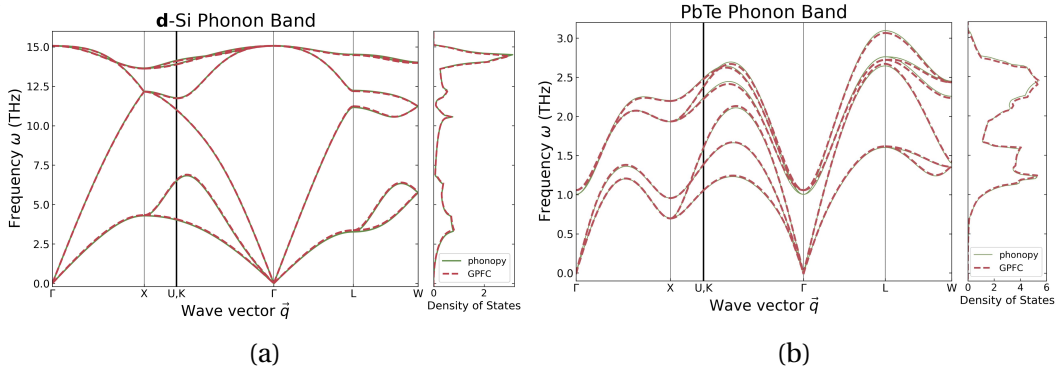


Figure 3.4: Panels (a) and (b) show the phonon band structures and density of states for **d-Si** and **PbTe**, respectively.

Moreover, the convergence of these RMSE learning curves can be confirmed by analysing the convergence of the acoustic sumrule across those five materials, shown in Fig. A.7a in Appendix A.2.3. This demonstrates that our Cartesian-based GP model learns translational symmetry from the training data, even though it employs Cartesian descriptors that do not explicitly encode the system’s symmetries into the model.

To further validate the physical reliability of our predicted GPFC2 tensors, we use them to compute the phonon band structures and phonon density of states (DOS) for **d-Si** and **PbTe**. These results are shown in Figure 3.4. The predicted phonon properties from GPFC2 closely match those obtained from FDM-based force constants. Quantitatively, the root mean square errors of the phonon band structures are approximately 1.13% for **d-Si** and 3.89% for **PbTe**.

We include supplementary comparisons of phonon band structures and phonon density of states, calculated using our GPFC2, with those computed from the FDM-based FC2s for the remaining materials (**GaAs**, **CdTe**, and **NaCl**) in Appendix A.2.1. Additionally, in Appendix A.2.4 we compare our **d-Si** phonon results with the phonon band structures and phonon density of states evaluated using the FC2 based on the cluster expansion method, HiPHIVE [73].

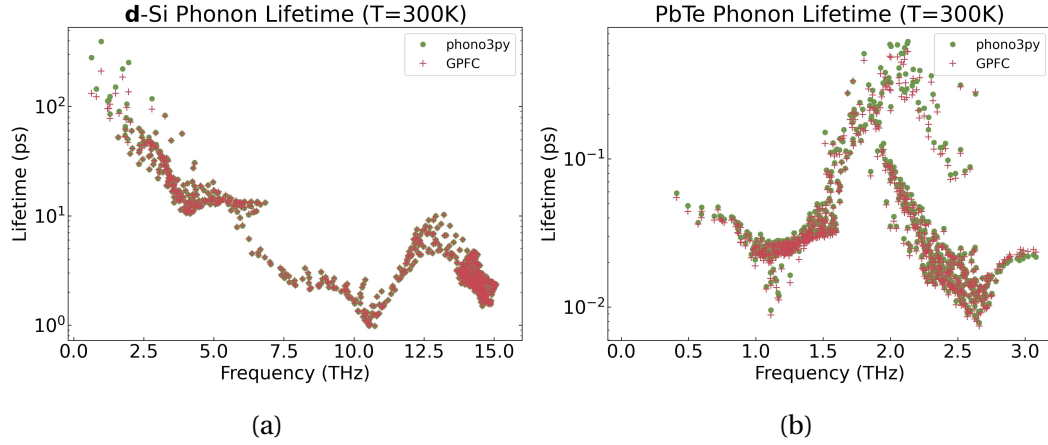


Figure 3.5: These figures show the predicted phonon lifetimes at 300 K of **d-Si** and **PbTe**.

To evaluate the accuracy of cubic anharmonic (third-order) force constants predicted by our model, we present RMSE learning curves analogous to those used for GPFC2 in Figure 3.3 (b). Based on the convergence behaviour, accurate prediction of third-order force constants in a Cartesian basis requires approximately 400 energy and force evaluations.

We further assess the physical reliability of these predicted cubic force constants by computing finite-temperature phonon lifetimes and lattice thermal conductivities for five materials: **d-Si**, **NaCl**, **PbTe**, **GaAs**, and **CdTe**, over a temperature range of 300–1200 K. These quantities are calculated using third-order many-body perturbation theory as implemented in PHONO3PY, with input FC3s generated either via finite displacement methods (FDM) or our GP-based approach (GPFC). The cubic anharmonic GP force constants are predicted using Equation (3.9), with the same training data used for the harmonic GPFC2 model.

Figure 3.5 presents the predicted phonon lifetimes for **d-Si** and **PbTe** at 300 K in panels (a) and (b), respectively. The corresponding lattice thermal conductivities are shown in Figure 3.6. To quantify differences in phonon lifetimes across temperatures, we compute the Earth Mover Distance (EMD)

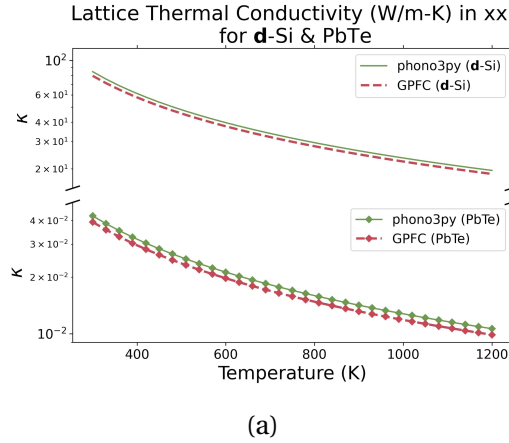


Figure 3.6: This figure illustrates the lattice thermal conductivities of **d-Si** and **PbTe** over the temperature range from 300 K to 1200 K.

between the lifetime density of states at 30 K intervals from 300 K to 1200 K. The mean EMD values are 3.9×10^{-1} ps for **d-Si** and 8.6×10^{-3} ps for **PbTe**, indicating high fidelity of the GPFC3 predictions, particularly for the latter.

The mean absolute errors in lattice thermal conductivity predictions using GPFC3 are $2.03 \text{ W}\cdot\text{m}^{-1}\text{K}^{-1}$ ($\text{MAE} \approx 2.87\%$ error) for **d-Si**, and $1.4 \times 10^{-3} \text{ W}\cdot\text{m}^{-1}\text{K}^{-1}$ ($\text{MAE} \approx 7.48\%$ error) for **PbTe**. Additional results for **NaCl**, **GaAs**, and **CdTe**, also computed with GPFC3, exhibit relative errors in the range of 2–7%.

Among all studied systems, the lowest accuracy of GPFC3 is observed in **PbTe**, which is known to exhibit strong anharmonicity and pronounced phonon–phonon interactions that shorten phonon lifetimes and challenge model generalisation.

Supplementary comparisons of phonon lifetimes and lattice thermal conductivities for **GaAs**, **CdTe**, and **NaCl** are also provided in Appendix A.2.2.

It is important to note that the GPFC framework does not offer a computational speed advantage over conventional finite-displacement approaches implemented in PHONOPY or PHONO3PY. In all applications presented here, the GPFC calculations are in fact slower, primarily due to the analytical

evaluation of higher-order kernel derivatives and the absence of symmetry exploitation, which substantially increases the number of degrees of freedom. However, the purpose of GPFC is methodological generality rather than speed. Because the present implementation does not impose crystal symmetry, the model can be directly applied to systems with low or broken symmetry—such as amorphous materials, molecular crystals, or interfacial structures—where traditional finite-displacement or symmetry-based lattice-dynamics approaches become inefficient or ambiguous. Therefore, although slower for highly symmetric crystals, the GPFC method provides a unified, symmetry-agnostic framework that can extract harmonic and anharmonic force constants consistently from general atomic configurations.

While the Cartesian descriptor provides a straightforward and general representation of atomic displacements, it does not exploit the underlying symmetries of the crystal lattice—such as translational invariance and dynamical matrix sparsity—which can lead to redundancy in the model and increased data requirements. To address this, we will introduce a phonon-mode-based descriptor that inherently incorporates lattice symmetries and reduces the dimensionality of the learning problem. In the following section, we present how this descriptor is implemented in GPFC.JL, analyse its impact on model efficiency and accuracy, and compare its learning behaviour with that of the Cartesian approach.

3.2 GPFC.JL with Phonon descriptors

In another section of our previous work [9], we investigate a different strategy based on phonon coordinates, or normal modes, as input descriptors for the Gaussian process model rather than straight Cartesian representations. These normal modes form a natural basis for atomic motion in a crystalline material or molecule and inherently encode the symmetries of the system. Moreover, they serve as the foundation for higher-order theoretical frameworks in lattice dynamics, such as many-body perturbation theory, which couples normal modes through anharmonic interactions.

3.2.1 Phonon descriptors

Our motivation for using phonon-mode descriptors is to improve the efficiency of the machine learning model—achieving higher accuracy with fewer training data—without explicitly encoding symmetry or equivariance in the model architecture. The normal modes are obtained by diagonalising the mass-weighted second-order force constant matrix, also known as the dynamical matrix.

Before constructing phonon descriptors, we now revisit the expression of the “*Dynamical Matrix*” [12; 13; 14] at wave vector $\mathbf{q}$:

$$D_{\alpha\beta, \kappa\kappa'}(\mathbf{q}) = \frac{1}{\sqrt{m_{(\kappa)}m_{(\kappa')}}} \sum_{l'} \Phi_{\alpha l\kappa, \beta l'\kappa'}^{(2)} \mathbf{e}^{i\mathbf{q} \cdot (\mathbf{R}_{\kappa'l'}^0 - \mathbf{R}_{\kappa l}^0)}, \quad (3.17)$$

where the masses $m_{(\kappa)}$ and $m_{(\kappa')}$ correspond to atoms κ and κ' in unit cells l and l' , respectively. The term $\Phi_{\alpha l\kappa, \beta l'\kappa'}^{(2)}$ is the harmonic force constant between components α and β of atoms κ and κ' in cells l and l' . The vectors $\mathbf{R}_{\kappa l}^0$ and $\mathbf{R}_{\kappa'l'}^0$ denote their respective equilibrium positions. The summation runs over all primitive cells l' , with $l = 0$ referring to the reference cell.

Equation (3.17) can be interpreted as the discrete Fourier transform of the harmonic force constants expressed in the Cartesian basis. By analogy, we

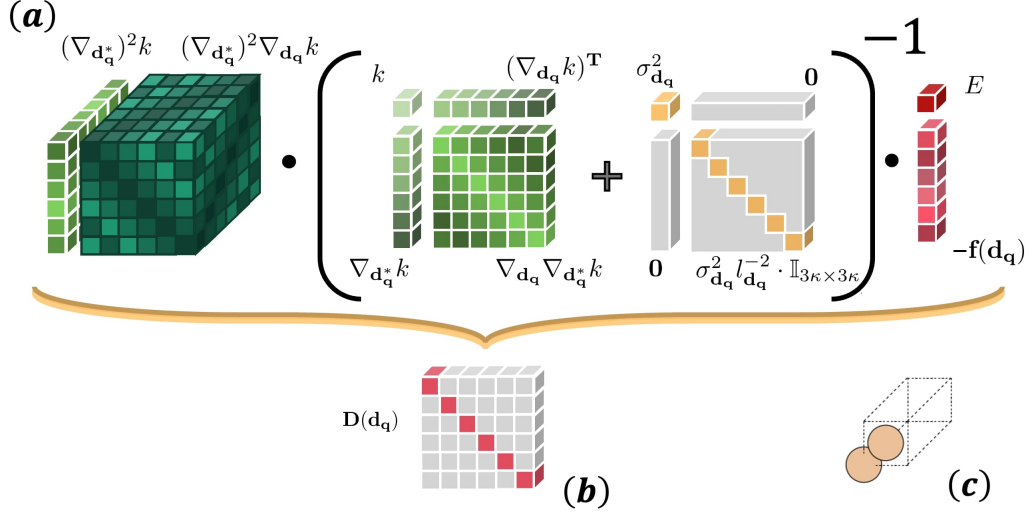


Figure 3.7: Figure illustrates the tensor contraction in Equation (3.7) using phonon-mode descriptors.

can apply a similar transformation to the atomic displacement descriptors themselves. Specifically, we recast the Cartesian descriptors in reciprocal space and obtain the phonon-mode descriptor as

Phonon descriptor

$$d_{\beta\kappa'}(\mathbf{q}) = \sqrt{m_{\kappa'}} \sum_{l'} \mathbf{x}_{\beta l' \kappa'} e^{i\mathbf{q} \cdot (\mathbf{R}_{\kappa' l'}^0 - \mathbf{R}_{\kappa l}^0)}, \quad (3.18)$$

where $\mathbf{x}_{\beta l' \kappa'}$ are the atomic displacement descriptors in Cartesian coordinates, $m_{\kappa'}$ is the mass of atom κ' , and $\mathbf{R}_{\kappa' l'}^0$ denotes the equilibrium position of the atom κ' in the unit cell l' . The resulting quantity $d_{\beta\kappa'}(\mathbf{q})$ represents the projection of atomic displacements onto the vibrational normal modes at wavevector $\mathbf{q}$, and is analogous to the phonon eigenvector components obtained from diagonalising the dynamical matrix.

Integrating the phonon descriptor directly imposes the system's symmetry in our GP model, which removes the requirement for the model to infer

these symmetries from data. Moreover, the descriptor dimensionality is significantly decreased: the model performs on a primitive cell that includes just κ atoms, unlike a complete $l \times l \times l$ supercell. This means that while there are $N = l^3 \kappa$ atoms in the atomic environment, the phonon-based descriptors reduce this information to the size of the primitive cell. As seen in Figure 3.7, the equivalent tensor contraction, which was first represented in Equation (3.7), can be redefined appropriately as

Phonon-basis GPFC2 prediction

$$\mathbf{D}(\mathbf{d}_q^*) = \left[(\nabla_{\mathbf{d}_q^*})^2 k \quad (\nabla_{\mathbf{d}_q^*})^2 \nabla_{\mathbf{d}_q} k \right]_{\mathbf{d}_q, \mathbf{d}_q^*} \cdot \mathcal{K}_{\mathbf{d}_q, \mathbf{d}_q}^{-1} \cdot \begin{bmatrix} E \\ -\mathbf{f} \end{bmatrix}_{\mathbf{d}_q}, \quad (3.19)$$

where $\mathbf{D}(\mathbf{d}_q^*)$ denotes the dynamical matrix evaluated at wavevector $\mathbf{q}$ and $\mathbf{f}$ is the phonon-basis force obeyed the discrete Fourier transform in (3.18). The prediction of GPFC3 can be recast in a form analogous to the GPFC2 expression derived above.

For a training set of size m , the computational cost of learning—dominated by the inversion of a $(m \cdot (1 + 3N)) \times (m \cdot (1 + 3N))$ covariance matrix—is $\mathcal{O}(m^3 \cdot (1 + 3N)^3)$. Similarly, the cost of prediction—arising from the dot product of two vectors of dimension $m \cdot (1 + 3N)$ —is $\mathcal{O}(m \cdot (1 + 3N))$ [1; 74]. By employing phonon-mode descriptors, where the input dimensionality is reduced from the full atomic environment (N atoms) to that of the primitive cell (κ atoms), these costs are reduced to $\mathcal{O}(m^3 \cdot (1 + 3\kappa)^3)$ for training and $\mathcal{O}(m \cdot (1 + 3\kappa))$ for prediction, respectively.

3.2.2 Phonon-basis GPFC

We also compare our phonon-based GPFC2 and GPFC3 predictions with the dynamical matrix and third-order force constants (FC3) obtained from the finite displacement method (FDM), as implemented in PHONO(3)PY. The

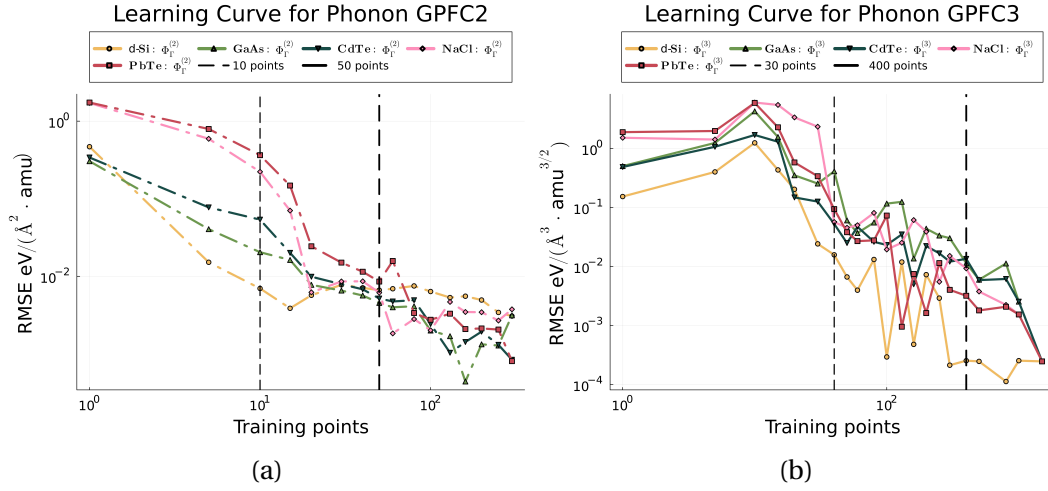


Figure 3.8: Learning curves in (a) and (b) illustrate the direct comparison of GPFC2 and GPFC3 in phonon descriptors to FC2 and FC3 from FDM, respectively, using PHONO(3)PY.

same dataset used for the Cartesian-basis models in Section 3.2.2—comprising five representative materials (**d-Si** (FCC), **GaAs**, **CdTe**, **NaCl**, and **PbTe**)—is employed to train the derivative GP model using phonon-mode descriptors, enabling the prediction of Gaussian process force constants in the phonon basis.

All GPFC predictions are evaluated at the Γ -point, as the current implementation of automatic differentiation limits predictions at arbitrary $\mathbf{q}$ -points. We will discuss the extended procedure to address this limitation in the next subsection. The accuracy of the GPFC2 and GPFC3 models is shown in Figures 3.8(a) and (b), respectively, in terms of root mean square error (RMSE) learning curves. The RMSE of GPFC2 predictions converges rapidly to near-zero values using only 6–10 training points, while the GPFC3 model requires approximately 30–50 data points to achieve comparable accuracy. Compared to the Cartesian-basis results in Figure 3.3, these learning curves again reflect the dimensionality of the descriptor: the number of required training samples closely matches the number of degrees of freedom in the

phonon descriptor, which corresponds to the number of atoms in the primitive cell.

Moreover, the phonon-based GPFC2 and GPFC3 models also exhibit learning of translational symmetry, as illustrated in Fig. A.7b. The convergence of the acoustic sum rule becomes evident in the learning curves after approximately 10 and 30–50 data points, respectively.

We also observe that predicting the cubic anharmonic force constants (GPFC3) generally requires roughly eight times more training data than is needed to achieve similar accuracy for the harmonic case (GPFC2).

3.2.3 Recovering Real Space FCs

Recovering phonon-based force constants into the Cartesian basis is essential for calculating full phonon dispersion relations and lattice thermal conductivity. However, a limitation arises due to current automatic differentiation frameworks, which do not support taking derivatives with respect to complex-valued phonon descriptors beyond second order. As a result, both the harmonic and cubic anharmonic dynamical matrices can only be evaluated at the Γ -point, where the phonon descriptors are real-valued.

To address this limitation, we propose the construction and implementation of a differentiable Gaussian process model capable of handling complex-valued descriptors. A discussion of this extension, along with the derivation of the relevant complex derivatives, is provided in Appendix A.1.2.

Once the harmonic (GPFC2) and cubic anharmonic (GPFC3) Gaussian process force constants have been computed in the phonon basis, we follow a well-defined procedure to transform them back into the Cartesian basis [12; 13; 14]. For the harmonic case, the inverse discrete Fourier transform of the set of dynamical matrices over all $N_{\mathbf{q}}$ wavevectors can be used to recover the real-space force constants as:

$$\Phi_{\alpha l \kappa, \beta l' \kappa'}^{(2)} = \frac{\sqrt{m_\kappa m_{\kappa'}}}{N_{\mathbf{q}}} \sum_{\mathbf{q}} D_{\alpha \beta, \kappa \kappa'}(\mathbf{q}) e^{-i \mathbf{q} \cdot (\mathbf{R}_{\kappa' l'}^0 - \mathbf{R}_{\kappa l}^0)}. \quad (3.20)$$

Here, $N_{\mathbf{q}} = N_x \times N_y \times N_z$ corresponds to the number of $\mathbf{q}$ -points sampled from a Monkhorst–Pack grid covering the Brillouin zone. For a $2 \times 2 \times 2$ supercell, we require only 8 $\mathbf{q}$ -points.

Similarly, GPFC3 values in the phonon basis can be transformed back to Cartesian coordinates using an expression analogous to the inverse Fourier transform of Equation (2.19):

$$\begin{aligned} \Phi_{\alpha l_1 \kappa_1, \beta l_2 \kappa_2, \gamma l_3 \kappa_3}^{(3)} &= \frac{\sqrt{m_{\kappa_1} m_{\kappa_2} m_{\kappa_3}}}{N_{\mathbf{q}}^2} \sum_{\mathbf{q}_2, \mathbf{q}_3} \Phi_{\alpha \kappa_1, \beta \kappa_2, \gamma \kappa_3}^{(3)}(\mathbf{q}_2, \mathbf{q}_3) \\ &\quad \times e^{-i \mathbf{q}_2 \cdot (\mathbf{R}_{\kappa_2 l_2}^0 - \mathbf{R}_{\kappa_1 l_1}^0)} \\ &\quad \times e^{-i \mathbf{q}_3 \cdot (\mathbf{R}_{\kappa_3 l_3}^0 - \mathbf{R}_{\kappa_1 l_1}^0)} \\ &\quad \times \delta^*(\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3), \end{aligned} \quad (3.21)$$

where $\delta^*(\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3)$ enforces momentum conservation and is defined as

$$\delta^*(\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3) = \begin{cases} 1, & \text{if } \mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3 \text{ is a reciprocal lattice vector,} \\ 0, & \text{otherwise.} \end{cases} \quad (3.22)$$

Note that predicting $\Phi^{(3)}(\mathbf{q}_2, \mathbf{q}_3)$ using our current GP model is limited to specific $\mathbf{q}$ -points. Nevertheless, the GPFC3 predicted at wavevectors $(\mathbf{q}_2, \mathbf{q}_3)$ corresponds to the cubic anharmonic dynamical matrix in reciprocal space. This quantity is the Fourier transform of the real-space third-order force constants, and represents the linear response of the force on atom κ_1 to displacements with wavevectors $\mathbf{q}_2$ and $\mathbf{q}_3$.

The relationship between this response and the force in reciprocal space is given by:

$$F_{\alpha\kappa_1}(\mathbf{q}_1) = -\frac{1}{\sqrt{l^3}} \sum_{\beta\kappa_2} \sum_{\gamma\kappa_3} \Phi_{\alpha\kappa_1, \beta\kappa_2, \gamma\kappa_3}^{(3)}(\mathbf{q}_2, \mathbf{q}_3) d_{\beta\kappa_2}(\mathbf{q}_2) d_{\gamma\kappa_3}(\mathbf{q}_3) \delta^*(\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3), \quad (3.23)$$

allowing us to obtain the phonon descriptors $d_{\alpha\kappa_1}(\mathbf{q}_1)$ from Equation (3.18) and use the corresponding force $F_{\alpha\kappa_1}(\mathbf{q}_1)$ as training data for GPFC3.

For a $2 \times 2 \times 2$ supercell lattice, recovering the full GPFC3 tensor in the Cartesian basis requires evaluating $N_{\mathbf{q}}^2 = 64$ cubic anharmonic dynamical matrices, assuming both $\mathbf{q}_2$ and $\mathbf{q}_3$ are sampled from the same mesh.

3.3 Atomic Cluster Expansion Model

In this section, we discuss possible approaches to further improve the efficiency of our differentiable Gaussian process (GP) model. Incorporating physical symmetries of atomic structures has proven to be a sensible and effective strategy in several machine learning potentials [43].

As reviewed in Chapters 2.4 and 2.5, Atomic Cluster Expansion (ACE) descriptors offer one of the most efficient local basis sets for representing individual atoms within atomic structures prior to training the GP model. This is because they are systematically constructed to capture the atomic environment while inherently satisfying permutation, translation, reflection, and rotation symmetries, as shown in Equations (2.70), (2.71), and (2.78).

To integrate ACE descriptors with our model, we employ a JULIA-language implementation of the ACE framework, the ACEPOTENTIALS.JL package [15]. ACEPOTENTIALS.JL builds upon the foundational concepts of ACE [2; 16], incorporating several extensions such as smoothness priors, robust parameter estimation, and an extended design space for capturing atomic species through radial basis functions. These features make ACEPOTENTIALS.JL

a simple, interpretable, and robust framework, well-suited for use in our differentiable GP model.

3.3.1 ACE descriptors

This section aims to investigate the construction of ACE descriptors using the ACEPOTENTIALS.JL package and to assess their compatibility and efficiency when incorporated into our differentiable GP model.

In an N -atom structure, ACE descriptors which represent the centred atom i as a **symmetrised many-body product basis** set $\mathbf{B}_i$. This basis set can be written as the set consisting of the many-body product basis from 1-body order to $(v + 1)$ -body order expressed as

$$\mathbf{B}_i = \left[\mathbf{B}_i^{(0)} \quad \mathbf{B}_i^{(1)} \quad \mathbf{B}_i^{(2)} \quad \dots \quad \mathbf{B}_i^{(v)} \right]^T, \quad (3.24)$$

where $\mathbf{B}_i^{(v)}$ are defined in (2.78) for a single-element system and (2.84) for a multi-element system. The set of the many-body product bases can be evaluated using the code shown in Figure 3.9. The `acemodel` constructor specifies a set of hyperparameters that control the structure and behaviour of the basis set. Within `acemodel`, the `element` parameter specifies the chemical species, the `order` parameter sets the interaction order N (corresponding to $N + 1$ -body correlations), and the `totaldegree` parameter constrains the maximum total polynomial degree of the basis functions. The parameters `r0` and `rcut` define the equilibrium bond length, specific to the chemical species, and the cutoff radius beyond which basis functions vanish smoothly, respectively.

The `pair_transform` parameter corresponds to the transformation defined in Equation (2.86), typically using the Agnesi function (`:agnesi`), which enhances resolution near the nearest-neighbour distance and improves the accuracy of the interatomic potential. The `pair_envelope` parameter, defined in Equations (2.87) and (2.88), modifies the radial behaviour for

```
1 model = acemodel(elements = [:Si,],  
2                 order = 2,                #3-body order  
3                 totaldegree = 5,  
4                 r0 = rnn(:Si),  
5                 rcut = 3.0,  
6                 pair_transform = (:agnesi, p = 1, q = 4),  
7                 pair_envelope = (:r, 2, 2)  
8             )  
9             @show length(model.basis)  
10 > length(model.basis)> = 12
```

Figure 3.9: Construction of a 3-body order ACE descriptor for a silicon structure using a total degree constraint of 5, an Agnesi radial transform, and a smooth polynomial envelope. The code prints the resulting size of basis functions.

ensuring that functions vanish at the cutoff and behave smoothly at short range.

Note that only the number of chemical species, the order parameter, and the totaldegree parameter determine the size of the many-body product basis, which is printed as the length of the model basis in the code shown above.

To determine `r0`, `rcut`, `pair_transform`, and `pair_envelope` parameters for a given dataset, it is necessary to analyse the radial distribution function (RDF) of the centred atom in each configuration and then modify the radial basis functions (RBFs), defined in Equation (2.85), as these RBFs serve as building blocks for constructing the many-body product basis.

Let us consider the top panel of Figure 3.10, which shows the RDF histogram of the input d-Si dataset, with dashed black lines indicating the set of equilibrium bond lengths of silicon ($r_1, r_2, \dots, r_5$). The bottom panel illustrates the possible choices of radial basis sets (centred at the average of these distances within the cutoff radius 3 Å (red), 5 Å (green), 6 Å (yellow)), extracted from the code shown in Figure 3.9. We construct three sets of radial basis functions, shown in red, green, and yellow, designed to cover the first peak, the first three peaks, and all peaks in the RDF, respectively. These three

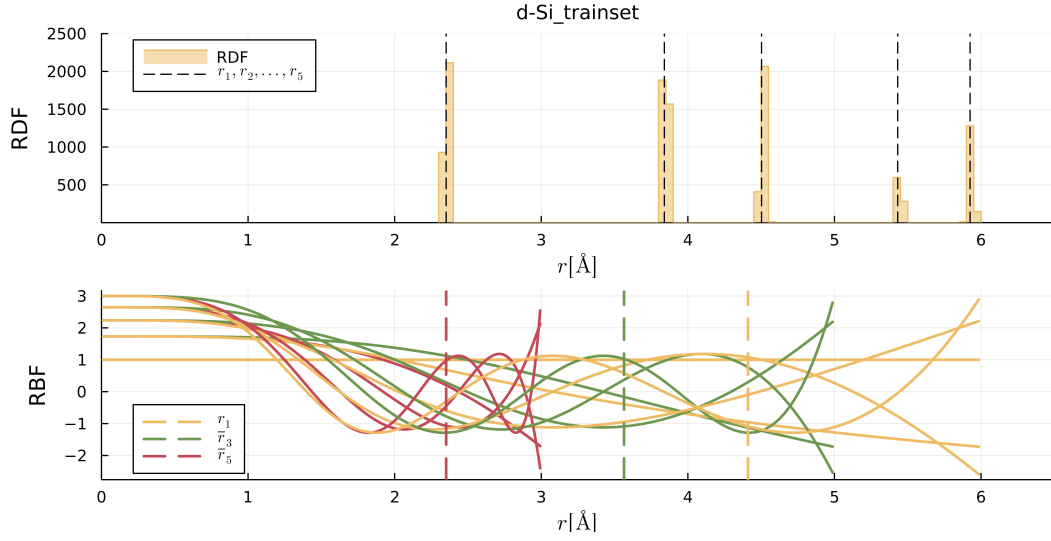


Figure 3.10: Radial distribution function (RDF) and radial basis functions (RBF) for the d-Si training set.

sets are characterised by different cutoff radii and corresponding average equilibrium bond lengths, i.e., $\{3 \text{ \AA}, r_1\}$, $\{5 \text{ \AA}, \bar{r}_3\}$, and $\{6 \text{ \AA}, \bar{r}_5\}$, where $\bar{r}_3$ and $\bar{r}_5$ denote the averages of $\{r_1, r_2, r_3\}$ and $\{r_1, r_2, \dots, r_5\}$, respectively.

After obtaining the appropriate radial basis set for the individual centred atom, we can then form the descriptor of the N -atom structure by concatenating the many-body product basis for all atoms in the structure. This total structure descriptor is expressed as

$$\mathbf{B}_{\text{tot}} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{B}_2 & \cdots & \mathbf{B}_N \end{bmatrix}^T, \quad (3.25)$$

where each $\mathbf{B}_i$ is defined in Equation (3.24). Therefore, the size of the full descriptor $\mathbf{B}_{\text{tot}}$ is given by the size of the ACE model basis, $\dim(\mathbf{B}_i)$, multiplied by the number of atoms N in the structure. For example, using the ACE basis constructed by the code in Figure 3.9, the descriptor size for a $2 \times 2 \times 2$ d-Si supercell is $12 \times 16 = 192$ degrees of freedom.

We also investigate the construction of ACE descriptors for a **PbTe** structure as an example of a multi-element system. The many-body product basis

```

1 model = acemodel(elements = [:Pb, :Te,],
2                   order = 2,                      #3-body order
3                   totaldegree = 5,
4                   r0 = (rnn(:Pb)+rnn(:Te))/2,
5                   rcut = 6.5,
6                   pair_transform = (:agnesi, p = 1, q = 4),
7                   pair_envelope = (:r, 2, 2)
8               )
9         @show length(model.basis)
10 > length(model.basis) = 65

```

Figure 3.11: Construction of a 3-body order ACE descriptor for a **PbTe**-structure using a total degree constraint of 5, an Agnesi radial transform, and a smooth polynomial envelope.

set can be generated using the code shown in Figure 3.11, which follows a similar procedure to the d-**Si** case, with the change for accounting for the two input elements and the average equilibrium bond length of the **Pb–Te** pair.

The hyperparameters r_0 , $rcut$, $pair_transform$, and $pair_envelope$ are determined by analysing the RDFs and RBFs of the input **PbTe** structure, as illustrated in Figure 3.12. Since **PbTe** is a two-element system, it is necessary to consider the radial distributions between **Pb–Pb**, **Pb–Te**, and **Te–Te** atom pairs, which are shown as yellow, green, and red lines, respectively, in the RDF histogram in the top panel.

The r_0 parameter is selected as the average position of the first three RDF peaks, as the fourth peak, associated with the **Pb–Te** bond length, is relatively small compared to the others. The $rcut$ parameter is chosen to cover the fourth peak, while $pair_transform$ and $pair_envelope$ are employed identically to the d-**Si** case. The resulting radial basis functions (RBFs) extracted from the ACE model with these parameters are shown in the bottom panel of Figure 3.12.

To construct the full descriptor $\mathbf{B}_{tot}$ for the **PbTe** structure, we can now perform Equation (3.25), resulting in a $65 \times 16 = 1040$ -dimensional feature

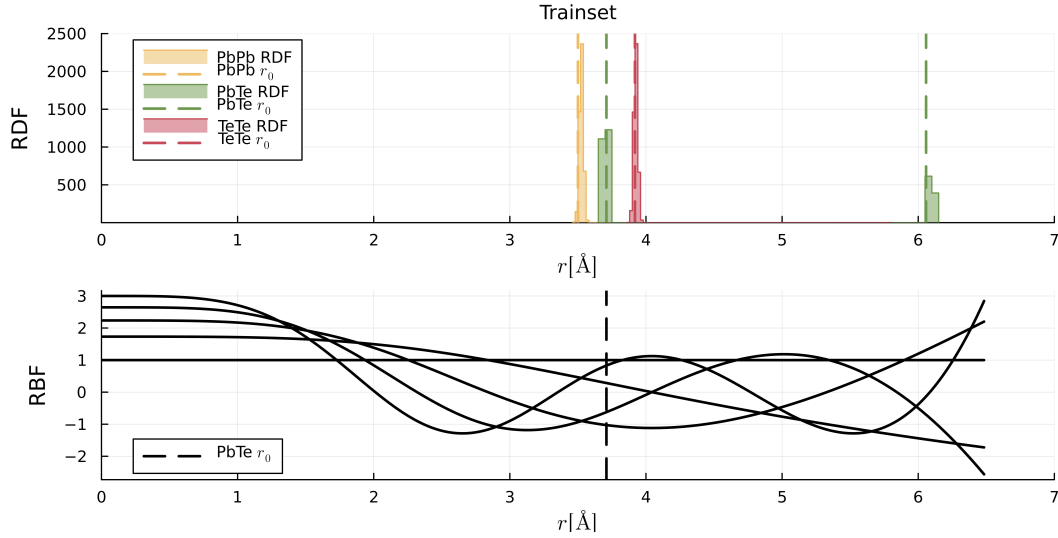


Figure 3.12: Radial distribution function (RDF) and radial basis functions (RBF) for the **PbTe** training set.

vector. This descriptor size of 1040 degrees of freedom also applies to other two-element systems, such as **NaCl**, **GaAs**, and **CdTe**, when using the same hyperparameter settings.

However, we observe that incorporating ACE descriptors directly into our differentiable GP model may not be sufficiently efficient, as the size of the ACE descriptors is significantly larger than that of the Cartesian and phonon descriptors. This increased descriptor size impacts both the computational cost and memory requirements for calculating higher-order derivatives with the current version of automatic differentiation. Therefore, instead of using ACE descriptors directly in the GP model, we propose to incorporate the ACE framework in an alternative way, namely for data generation and active learning, which will be discussed in the next subsection.

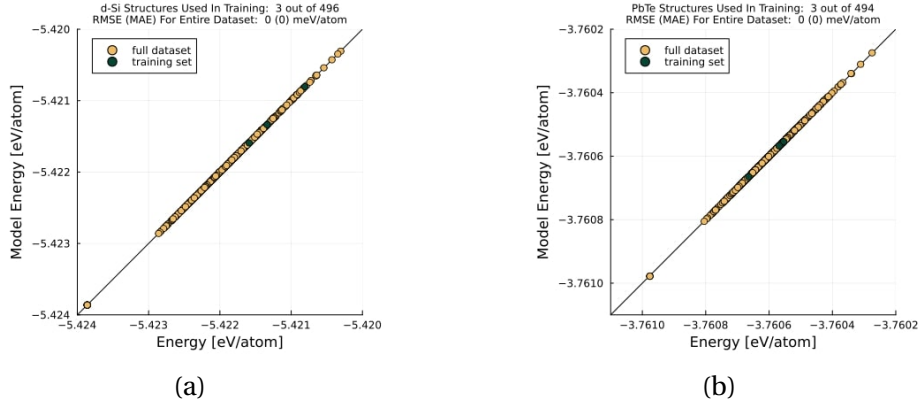


Figure 3.13: Panels (a) and (b) show the ACE-fitted potentials for an input **d-Si** and **PbTe** training data point (blue) alongside the corresponding target energies (yellow).

3.3.2 ACE fitting

In this section, we discuss an alternative approach for incorporating the ACE framework into our differentiable GP model. To generate datasets efficiently, we use the ACEPOTENTIALS instead of ACE descriptors as inputs.

The ACEPOTENTIALS enables highly accurate energy fitting for both **d-Si** and **PbTe** structures. Using the ACE model settings shown in Figures 3.9 and 3.11, we train the ACE model with only three data points obtained from DFT calculations¹.

With the minimal training set, the ACE model achieves excellent predictive accuracy: it recovers all 500 energy values with RMSEs (MAEs) of 2.37(1.84) $\mu\text{eV}/\text{atom}$ for the **d-Si** structure and 136.18(6.68) $\mu\text{eV}/\text{atom}$ for the **PbTe** structure, as shown in Figure 3.13. The RMSEs (MAEs) of the predicted forces are similarly low, at 0.89(0.68) $\text{meV}/\text{\AA}$ for **d-Si** and 0.11(0.09) $\text{meV}/\text{\AA}$ for **PbTe**.

The strong predictive performance of the ACE model enables efficient generation of energies and forces for new data points, which can be used to further

¹ The DFT calculations are performed using VASP, with convergence settings detailed in Subsection 3.1.2.

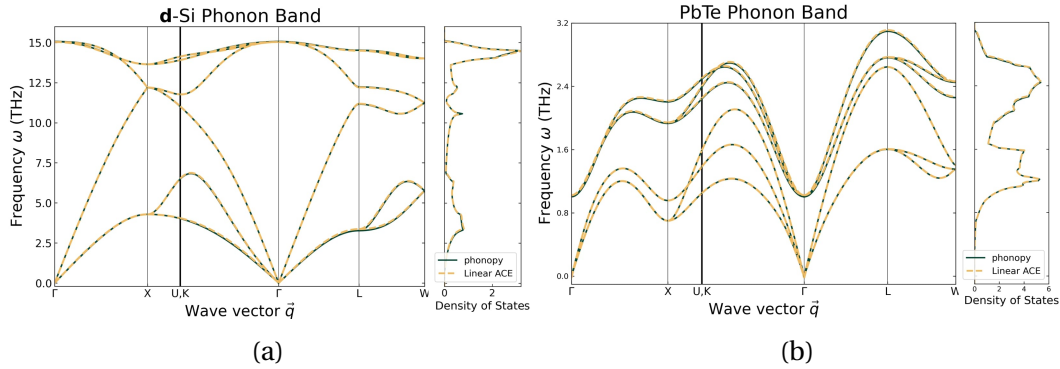


Figure 3.14: Panels (a) and (b) show the phonon band structures and density of states for **d-Si** and **PbTe**, respectively.

train our GP model. These new configurations, denoted as $\{\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_n\}$, can initially be sampled from a normal distribution around the equilibrium structure.

We offer two approaches to benefit from using these Linear ACE data. One is directly doing finite different methods to calculate force constants for phonon calculations. This will be discussed in the next subsection. Another is doing Active learning on top of GPFC models which the guidance is provide in subsection 3.3.4.

3.3.3 Phonons from Linear ACE data

The highly accurate forces obtained from the linear ACE model enable accurate phonon calculations using standard approaches such as Finite Difference Methods (FDM).

To evaluate the fitting results against our model, it is necessary to disregard the crystal symmetry (i.e., deactivate the symmetry in PHONOPY). We compute the harmonic force constants using the data point number of $6N$, where N indicates the number of atoms in the system as mentioned in section 2.1.1. For a $2 \times 2 \times 2$ diamond-structured silicon and a $2 \times 2 \times 2$ lead

telluride, the symmetry-breaking force distribution matrices necessitate 96 data points (16 atoms per system).

The results are illustrated in figure 3.14. Their RMSEs are roughly 1.01

This shows that the Linear ACE model is potentially powerful for using generating dataset for the study of phonons. For the future work, we will propose this to capture the anharmonic phonons.

3.3.4 Active learning with Linear ACE data

Moreover, we can apply an active learning strategy [75; 76; 77; 78] to select the next data points by maximising the predictive variance of the Gaussian process force constants. The predictive variance of the GPFC2 model is defined as

$$\begin{aligned} \Sigma^{(2)}(\mathbf{x}^*) &= [(\nabla_{\mathbf{x}^*})^2 (\nabla_{\mathbf{x}})^2 k]_{\mathbf{x}^*, \mathbf{x}^*} \\ &\quad - \begin{bmatrix} (\nabla_{\mathbf{x}^*})^2 k & (\nabla_{\mathbf{x}^*})^2 \nabla_{\mathbf{x}} k \end{bmatrix}_{\mathbf{x}^*, \mathbf{x}} \cdot \mathcal{K}_{\mathbf{x}, \mathbf{x}}^{-1} \cdot \begin{bmatrix} (\nabla_{\mathbf{x}^*})^2 k \\ (\nabla_{\mathbf{x}^*})^2 \nabla_{\mathbf{x}} k \end{bmatrix}_{\mathbf{x}, \mathbf{x}^*}, \end{aligned} \quad (3.26)$$

which corresponds to the predictive mean of GPFC2s given in Equation (3.7) and is analogous to the predictive variance of the energy in a differentiable GP model, as defined in Equation (2.46). The next data point can be selected by maximising the predictive variance according to the following expression:

Maximization of Predictive GPFC2 Variance

$$\mathbf{x}'_{\text{next}} = \arg \max_{\mathbf{x}^*} \Sigma^{(2)}(\mathbf{x}^*), \quad (3.27)$$

where $\Sigma^{(2)}(\mathbf{x}^*)$ is defined in Equation (3.26). Finding $\mathbf{x}'_{\text{next}}$ requires differentiating $\Sigma^{(2)}(\mathbf{x}^*)$ with respect to $\mathbf{x}^*$, which in turn demands evaluation of up to the fifth-order derivatives of the kernel function.

Meanwhile, active learning for the GPFC3 model would require constructing an analogous expression for the predictive variance of GPFC3, followed by maximisation to select the next data point. In this case, kernel derivatives up to the seventh order would be required.

At present, we have implemented only the random sampling method in our model. Full implementation of the active learning strategy remains challenging, as performing high-order automatic differentiation leads to substantial memory demands. However, analytic expressions for arbitrary-order derivatives of the radial basis function (RBF) kernel are provided in Appendix A.1, which may facilitate future improvements.

3.4 Conclusion and Future work

We developed a differentiable Gaussian process (GP) model for predicting harmonic (second-order) and cubic anharmonic (third-order) force constants. Our approach, which is incorporated into the GPFC.JL package, uses higher-order kernel derivatives to predict force constants directly from density functional theory (DFT)-calculated energies and forces.

Initially, we employed Cartesian descriptors, validating the model by comparing the predicted GPFC2 and GPFC3 tensors against finite displacement method (FDM) results for several representative materials. Accurate phonon properties, including band structures, density of states, phonon lifetimes, and lattice thermal conductivities, were obtained from the GPFC predictions with excellent agreement relative to traditional FDM approaches.

To improve model efficiency, we introduced phonon-mode descriptors, which inherently impose system symmetries and reduce the dimensionality of the learning problem. Using phonon descriptors significantly decreased the number of training data required, particularly for GPFC2 predictions, while maintaining or improving model accuracy.

We also outlined procedures for transforming phonon-based force constants back into the Cartesian basis, necessary for full Brillouin zone sampling and thermal property calculations. Although our current implementation is restricted to real-valued descriptors (thus, the Γ -point), we proposed strategies for handling complex descriptors in future extensions.

Furthermore, we investigated how the Atomic Cluster Expansion (ACE) framework might be integrated. While direct incorporation of ACE descriptors into our GP model proved inefficient due to large descriptor sizes, we proposed and tested an alternative: using ACE models for rapid and accurate data generation. With minimal DFT training, ACE potentials were able to reproduce reference datasets with very low errors, making them ideal for augmenting GPFC training through either random sampling or active learning.

Finally, we outlined an active learning strategy based on maximising the predictive variance of the GPFCs to help guide the selection of new training data. The full implementation of the active learning procedure is still challenging due to the memory demands of high-order derivatives, but the analytic expressions provided in the Appendices pave the way for future work.

In summary, this study introduces a flexible and powerful framework for learning force constants with differentiable Gaussian processes. By improving both accuracy and data efficiency, it creates new possibilities for modelling lattice thermal properties and anharmonic phonons in complex materials.

4

ACE-Transfer Integrals

In this chapter, we present an alternative approach for predicting transfer integrals (electronic couplings) of organic molecular dimers using ACEPOTENTIALS.JL. Instead of employing standard atomic neighbour representations, we introduce two alternative schemes, namely, *Bead* (*Coarse-Grained*) and *Heavy-atom* representations.

This methodology is motivated by the physical origin of transfer integrals, which are determined by atomic and molecular orbital couplings. It is therefore reasonable to group the transfer integrals associated with individual atomic orbitals into effective molecular orbitals of a coarse-grained system or to focus specifically on molecular orbitals localised on carbon atoms (e.g., C–H coupling orbitals).

We outline how to build ACE descriptors that are customised to handle various input datasets and analyse which representations are well suited for different classes of molecular dimers.

Section 4.1 highlights the significance of transfer integrals in organic semiconductors. We then discuss the symmetries of the transfer integrals and how we incorporate them into ACE models in section 4.2. Section 4.3 provides an analysis of the naphthalene dimer’s transfer integral energy surfaces. The transfer integrals of its uniform configurations are calculated from the quantum chemistry package PYMOO, ZINDO-based method [68]. It also covers the construction of ACE descriptors within the *Bead* and *Heavy-atom*

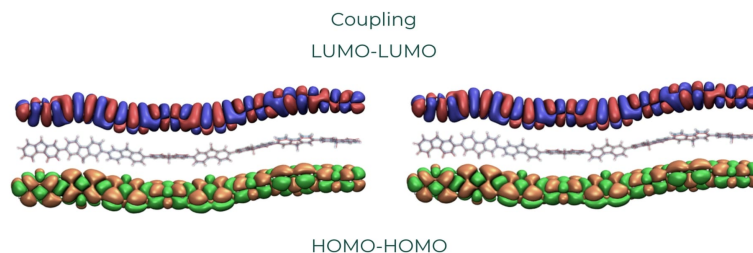


Figure 4.1: Coupling between frontier molecular orbitals (FMOs) of two conjugated organic molecules. The top and bottom FMOs represent the coupling between the lowest unoccupied molecular orbitals (LUMO–LUMO) and the coupling between the highest occupied molecular orbitals (HOMO–HOMO), respectively.

frameworks, along with the fitting results and discussions that follow. In Section 4.4, we apply the approach to other simpler molecular dimers (ethylene and thiophene). Their random configuration’s transfer integrals, generated from XTB-DIPRO package [66], are used to train the ACE model. We then evaluate the predictive capability of the *Heavy-atom* representations within the ACE framework. Finally, we conclude with a summary of the main findings and outline potential directions for future work in Section 4.5.

4.1 Transfer Integrals in Organic Semiconductors

Organic semiconductors are now essential components for lightweight, flexible electronics. Charge carriers in organic materials are frequently localised, unlike in inorganic semiconductors, where charge transport occurs through band-like conduction.

The active layer of the organic semiconductors is typically composed of π -conjugated molecular systems, in which overlapping π -orbitals allow for the delocalisation of π -electrons along the molecular backbone. This π -electron delocalisation facilitates the formation of frontier molecular orbitals, shown in Figure 4.1, that are critical for transport processes.

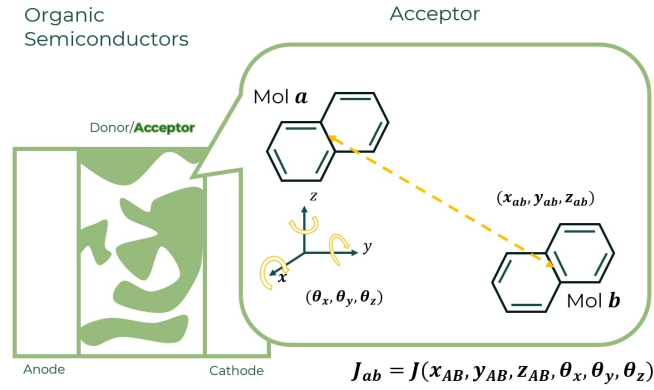


Figure 4.2: Schematic illustration of charge transport in organic semiconductors. The active layer consists of an intermixed donor/acceptor morphology between anode and cathode electrodes. A zoomed-in view shows the electronic coupling between two π -conjugated molecules (Mol *a* and Mol *b*), characterised by their relative centre-of-mass displacement (x_{ab}, y_{ab}, z_{ab}) and relative rotational orientations $(\theta_x, \theta_y, \theta_z)$. The transfer integral J_{ab} , governing the charge transfer rate between molecules, is a function of these relative coordinates.

Regarding the discussion in section 2.6, transport processes can be determined by the electronic coupling between two molecular sites and influence the rate of charge hopping transports, defined as the transfer integrals J

$$J_{ab} = \langle \psi_a | \hat{H} | \psi_b \rangle,$$

where ψ_a and ψ_b denote the frontier molecular orbitals on the respective molecular sites, and $\hat{H}$ is the electronic Hamiltonian. Practically, we can parameterise J_{ab} using the relative centre-of-mass displacement (x_{ab}, y_{ab}, z_{ab}) and the relative rotational orientations $(\theta_x, \theta_y, \theta_z)$ between two molecules, illustrated in Figure 4.2.

In Marcus theory, the transfer integrals can be employed to calculate the charge transfer rate k between two molecular sites, given by

$$k = \frac{2\pi}{\hbar} |J|^2 \frac{1}{\sqrt{4\pi\lambda k_B T}} \exp\left(-\frac{(\Delta G + \lambda)^2}{4\lambda k_B T}\right), \quad (4.1)$$

where λ is the reorganisation energy, ΔG is the Gibbs free energy change, k_B is Boltzmann's constant, and T is the temperature. The rate depends quadratically on the magnitude of the transfer integral $|J|^2$, implying that precise predictions of J are critical for reliable modelling of charge transport phenomena in organic materials.

However, predicting transfer integrals presents several challenges.

- Direct quantum chemical calculations are computationally expensive, particularly for large datasets of molecular dimers [79]. Accurately computing J typically requires electronic structure methods such as density functional theory (DFT) or fragment orbital approaches, which become prohibitively costly when thousands of molecular configurations must be evaluated, or the molecular assemblies contain thousands of atoms.
- Transfer integrals are highly sensitive to molecular geometry [80]; even small structural perturbations can lead to significant variations in J . Thermal fluctuations, molecular vibrations, and slight changes in intermolecular orientation can all induce large deviations in the electronic coupling, leading to the difficulty of generalising results across different configurations.

Applying a machine learning potential approach, particularly in frameworks based on the atomic cluster expansion (ACE), offers a promising alternative. By learning the mapping between molecular configurations and transfer integrals, ACE-based models can achieve high predictive accuracy across diverse geometries while significantly reducing computational costs compared to traditional *ab initio* methods.

4.2 ACE Model Incorporation

The purpose of this study is to estimate the transfer integrals J_{ab} between molecules a and b utilising the Atomic Cluster Expansion (ACE) model.

We first recast the site enegy in Eq. (2.80) for the multi-element system

$$\epsilon(\{\mathbf{x}_{ij}\}_{j\sim i}) = \sum_{\zeta \mathbf{n} \mathbf{l} \mathbf{m}} c_{\zeta \mathbf{n} \mathbf{l} \mathbf{m}}^{(Z_i)} \mathbf{B}_{\zeta \mathbf{n} \mathbf{l} \mathbf{m}}^{(i)}, \quad (4.2)$$

Similar to Equation (4.2), the *site* transfer integral may be articulated by determining the parameter $\tilde{c}$ as follows:

$$\mathcal{J}_{ab}(\{\mathbf{x}_{ij}\}_{j\sim i}) = \sum_{\zeta \mathbf{n} \mathbf{l} \mathbf{m}} \tilde{c}_{\zeta \mathbf{n} \mathbf{l} \mathbf{m}, ab}^{(Z_i)} \mathbf{B}_{\zeta \mathbf{n} \mathbf{l} \mathbf{m}}^{(i)}, \quad (4.3)$$

where the chemical species Z_i and the index ζ need not correspond to actual atomic species but may be designated as a *pseudo-atomic species* (any categorical variable representing an abstract particle), as elaborated in next section.

Upon ascertaining the parameters $\tilde{c}$ for the model in Equation (4.3), the overall transfer integral of the molecular dimer system is subsequently formulated as a summation of the *site* contributions.

$$J_{ab}(\{\mathbf{x}_i\}_{i=1}^J) = \sum_{i=1}^J \mathcal{J}_{ab}(\{\mathbf{x}_{ij}\}_{j\sim i}). \quad (4.4)$$

Text in bold Symmetry concerns. The transfer integrals J_{ab} adhere to three essential symmetry constraints dictated by the electronic Hamiltonian: (i) Hermitian symmetry ($J_{ab} = J_{ba}^*$), (ii) spatial invariance under rotations and translations, and (iii) localisation of electronic interactions. In developing the ACE model in Equations (4.3) and (4.4), we expressly guarantee the preservation of all these qualities.

The transfer integral between molecular orbitals ψ_a and ψ_b is defined as

$$J_{ab} = \langle \psi_a | \hat{H} | \psi_b \rangle,$$

which, due to the Hermiticity of the Hamiltonian $\hat{H} = \hat{H}^\dagger$, adheres to the following condition:

$$J_{ab} = J_{ba}^*. \quad (4.5)$$

In systems with real-valued molecular orbitals, this simplifies to $J_{ab} = J_{ba}$, signifying that the transfer integral is a *symmetric scalar quantity* upon the interchange of monomers a and b . This property is inherently maintained in the ACE representation via symmetric parameterisation of the coefficients $\tilde{c}_{\zeta \mathbf{n} \mathbf{l} \mathbf{m}, ab}^{(Z_i)}$ or through training on symmetrised data, hence ensuring that

$$\mathcal{J}_{ab}(\{\mathbf{x}_{ij}\}_{j \sim i}) = \mathcal{J}_{ba}(\{\mathbf{x}_{ij}\}_{j \sim i}). \quad (4.6)$$

Given that the transfer integrals are contingent solely upon the relative geometry of the molecular dimer, rather than its absolute position or orientation, the ACE basis functions $\mathbf{B}_{\zeta \mathbf{n} \mathbf{l} \mathbf{m}}^{(i)}$ are formulated to exhibit both rotational and translational invariance, thereby ensuring that

$$J_{ab}(\{\mathbf{x}_i\}_{i=1}^J) = J_{ab}(\{\mathbf{x}'_i\}_{i=1}^J), \quad \text{for } \mathbf{x}'_i = \hat{\mathbf{Q}}\mathbf{x}_i + \mathbf{t}, \quad (4.7)$$

where $\hat{\mathbf{Q}}$ denotes an arbitrary rotation matrix and $\mathbf{t}$ represents a translation vector. The ACE framework explicitly encodes these symmetry requirements by constructing descriptors and basis functions that remain invariant under the Euclidean group $E(3)$, which includes all rotations, translations, and reflections. Consequently, the resultant model parameters $\tilde{\mathbf{c}}$ inherently ensure rotational and translational invariance without requiring explicit data augmentation.

The spatial decomposition of the **single particle basis function** in Equation (2.83) illustrates the location of electronic interactions. In quantum

mechanics, the total coupling can be expressed as a summation of local orbital-orbital interactions,

$$J_{ab} = \sum_{i \in a} \sum_{j \in b} \langle \phi_i | \hat{H} | \phi_j \rangle, \quad (4.8)$$

Because each term diminishes swiftly with increasing interatomic distance due to the localised characteristics of atomic orbitals. Analogously, each site contribution $\mathcal{J}_{ab}(\{\mathbf{x}_{ij}\})$ in the ACE model relies solely on adjacent atoms within a defined cut-off radius, guaranteeing that J_{ab} is represented as a summation of local, symmetry-preserving contributions:

$$\mathcal{J}_{ab}(\{\mathbf{x}_{ij}\}_{j \sim i}) = \mathcal{J}_{ab}(\{\mathbf{x}_{ij}\}_{|\mathbf{x}_{ij}| < r_{\text{cut}}}). \quad (4.9)$$

4.3 Naphthalene Dimer

This section presents the transfer integral energy surface analysis of the naphthalene dimer. This naphthalene dimer are rigid and fully parameterised with 6 DoF. Moreover, the dimer is relatively small, leading to rapid transfer integral calculations. Therefore, it is sensible to begin modelling their transfer integrals as a starting point.

Two alternative frameworks, the *Coarse-Grained* and *Heavy-atom* representations, are employed to construct ACE descriptors. Each representation's predictive performance is assessed by fitting the transfer integrals derived from calculations involving quantum chemistry, and the outcomes are then discussed.

We intend to evaluate the ACE methodology's ability to capture the dependence of electronic couplings on molecular configurations by looking at the naphthalene systems.

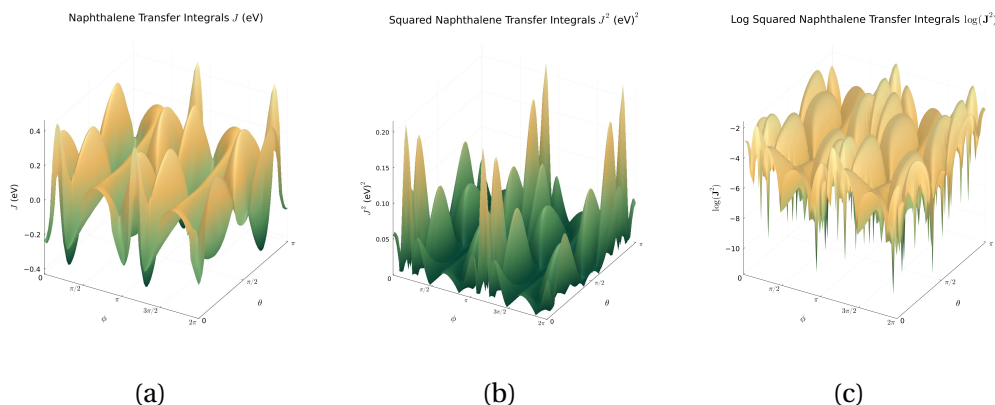


Figure 4.3: (a), (b), and (c) show the surface plots of the transfer integral $J(\theta, \phi)$ of a naphthalene dimer, the squared transfer integral $J^2(\theta, \phi)$, and the logarithm of the squared transfer integral.

Analysis of Transfer Integral Surface

We begin by investigating the surfaces of the transfer integrals. As discussed in Section 2.6.4, many studies predict transfer integrals based on either the square or the absolute value of the transfer integrals, since these quantities correlate linearly with the charge transfer rate. However, the surface of the squared transfer integrals J^2 is typically discontinuous and highly skewed, containing many small values and a few large peaks, resulting in non-smooth behaviour that poses challenges for machine learning models based on smooth basis sets.

Since ACE models are constructed from a set of smooth basis functions, fitting highly non-smooth quantities such as J^2 may be particularly sensitive to the choice of basis parameters. In such cases, training models directly on the transfer integrals J or on the logarithm of the squared transfer integrals, $\log(J^2)$, can improve predictive capability. Both J and $\log(J^2)$ produce distributions that are smoother and more Gaussian-like, which are well suited for regression tasks.

Figures 4.3 and 4.4 illustrate the three $2d$ -surfaces and heatmaps corresponding to the transfer integral of the naphthalene dimer: the transfer

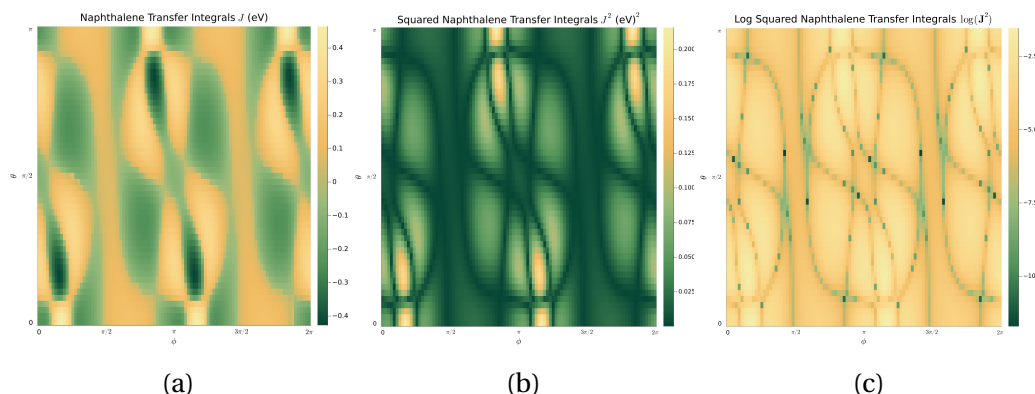


Figure 4.4: (a), (b), and (c) show the heatmaps of the naphthalene dimer’s transfer integral, the squared transfer integral, and the logarithm of the squared transfer integral.

integral itself, its squared value, and the logarithm of its squared value. These surfaces are parameterised by the orientational angles $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi]$ of molecule b with respect to molecule a , as depicted in Figure 4.2. In other words, molecule b is rotated over the surface of the sphere around molecule a , at a fixed centre-of-mass distance of 12 Å. It is evident that the surfaces of the squared transfer integrals (Figure 4.3b and 4.4b) and the logarithm of the squared transfer integrals (Figures 4.3c and 4.4c) exhibit sharp singularities, whereas the transfer integrals J (Figure 4.3a and 4.4a) are smoother by comparison. This observation suggests that choosing appropriate target quantities, such as J , is critical for achieving efficient and accurate ACE model fitting.

Representing Naphthalene Dimers

Before constructing the ACE descriptors, we introduce an efficient representation of the naphthalene dimer aimed at reducing the dimensionality of the descriptor space while preserving key geometric and electronic features relevant to transfer integral prediction.

One such approach is the *Bead* or *Coarse-Grained Representation*, illustrated on the left pathway of Figure 4.5. This method is inspired by the recent

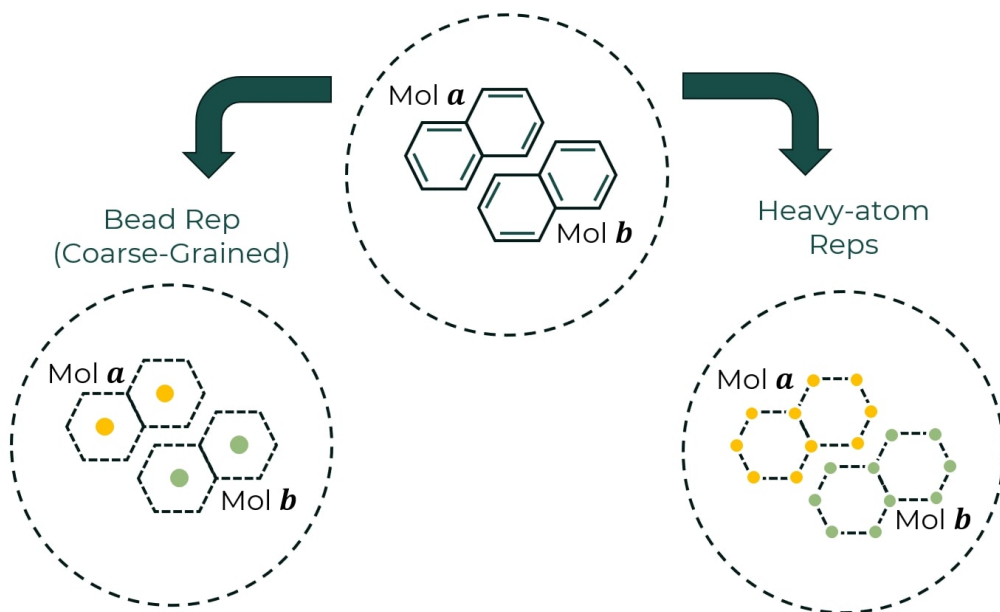


Figure 4.5: Illustration of two types of molecular representations used for constructing ACE descriptors.

work of Wang et al. [81], in which ACE models were successfully applied to coarse-grained molecular dynamics. We adapt this idea to model electronic coupling (transfer integrals) by constructing Bead ACE descriptors tailored to dimer interactions.

Since the transfer integral is an intermolecular property, it is important to distinguish between atoms (or Beads) belonging to different molecules. As shown in Figure 4.5, we assign distinct element labels to each molecule: the yellow Beads in Mol *a* are treated as pseudo-C elements, while the green Beads in Mol *b* are labelled as pseudo-Si elements. This labelling enforces a separation of molecular environments and ensures that intra- and intermolecular interactions are treated distinctly within the ACE framework.

The next step is to analyse the intra- and intermolecular radial distribution functions (RDFs), plotted as a function of Bead-pairwise distance in the top panel of Figure 4.6. The yellow histogram represents the intra-molecular RDF, while the red histogram shows the distribution of intermolecular pair

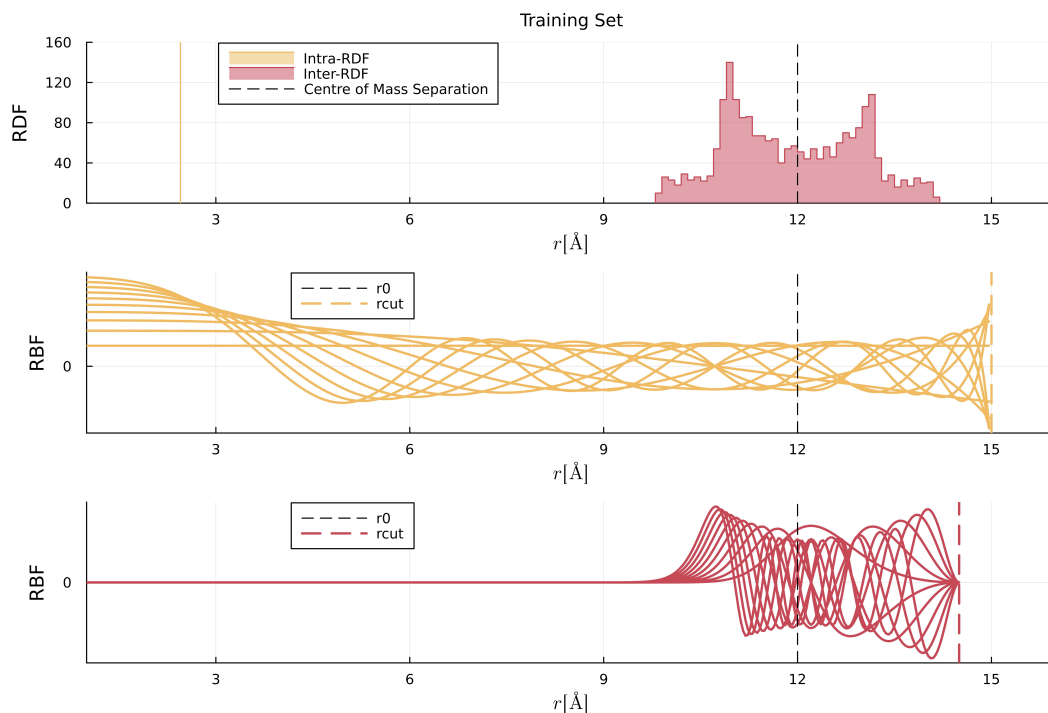


Figure 4.6: Analysis of radial distribution functions (RDFs) and radial basis functions (RBFs) used in the ACE descriptor construction for the coarse-grained (Bead) representation.

distances. The dashed vertical line indicates the fixed centre-of-mass separation (12 Å) used in our dimer configurations.

The analysis of the RDF guides the construction of radial basis functions (RBFs) using the `acemodel()` function, as illustrated in Figure 4.7. A discussion of the relevant hyperparameters is provided in Section 3.3.1.

Here, we emphasise the role of the envelope function used in constructing the RBFs. The default envelope `pair_envelope = (:r, 2, 2)` leads to basis functions extending across both intra- and intermolecular regions (middle panel of Figure 4.6). As shown in the bottom panel, modifying the envelope to `(:x, 2, 2)` localises the RBFs to the intermolecular range only. In other words, RBFs are adapted to the relevant intermolecular RDF support ranges, enabling resolution where features are concentrated. This

```
1 model = acemodel(elements = [:C, :Si, ],  
2                 order = 2,                      #3-body order  
3                 totaldegree = 10,  
4                 r0 = 12.0,  
5                 rcut = 13.6,  
6                 pair_transform = (:agnesi, p = 19, q = 43),  
7                 pair_envelope = (:x, 2, 2),  
8                 Eref = [:C => 0.0, :Si => 0.0]  
9             )
```

Figure 4.7: Construction of a 3-body order ACE descriptor for a Bead structure using a total degree constraint of 10, an Agnesi radial transform, and a smooth polynomial envelope.

ensures that the descriptor focuses on features relevant to transfer integrals, which are by nature intermolecular.

An additional hyperparameter, E_{ref} , defines the one-body (atomic reference) energy. By setting $E_{\text{ref}} = 0$ for each Bead type, we enforce the physical constraint that the transfer integral vanishes for isolated molecules or Beads.

For the *Heavy-atom Representation* shown on the right pathway in Figure 4.5, a similar analysis can be performed to construct the radial basis set for the corresponding ACE descriptors. In this representation, only heavy atoms (typically carbon) are included, thereby retaining chemically relevant geometric features while reducing the descriptor size.

The radial distribution functions (RDFs) for this representation are shown in the top panel of Figure 4.8, plotted as a function of pairwise distances between carbon atoms. As in the coarse-grained case, the yellow histogram corresponds to intramolecular RDFs, while the red histogram captures intermolecular distances. The fixed centre-of-mass separation is again marked by a dashed vertical line.

The choice of envelope functions for the radial basis can follow the same design as used in the Bead representation. Specifically, by applying an

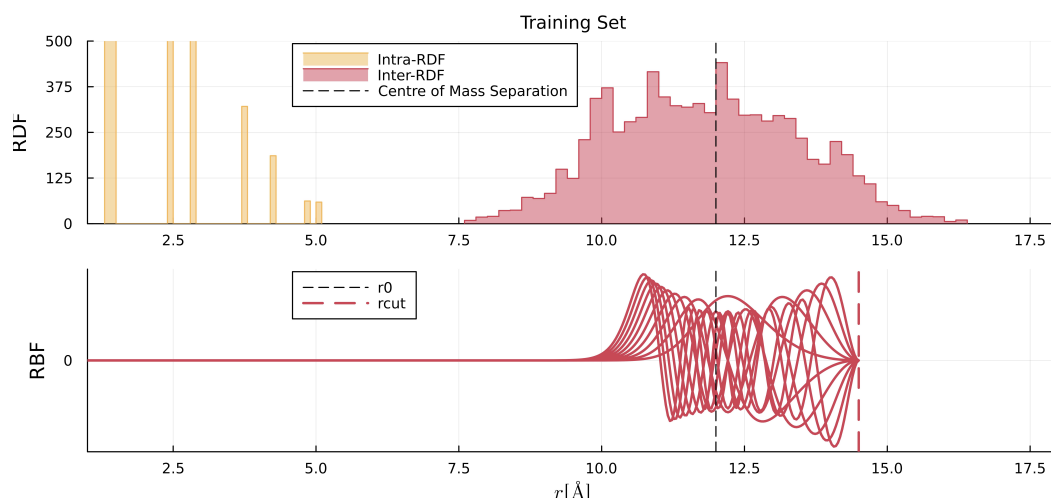


Figure 4.8: Analysis of naphthalene dimer RDFs and RBFs used in the ACE descriptor construction for the Heavy-atom representation.

intermolecular envelope (e.g., $(:x, 2, 2)$), the radial basis functions can be restricted to the intermolecular distance range, thereby focusing the model on the features most relevant for predicting transfer integrals.

ACEfit!

With all descriptors and hyperparameters defined, we now proceed to fit the ACE model to the transfer integral data of the naphthalene dimer. The reference transfer integrals were computed using the PyMOO, ZINDO-based method [68], package as implemented by J. M. Frost.

The ACE models constructed using the descriptors in Figure 4.7 were trained on 1,350 dimer configurations out of a total of 4,753. The resulting fits using the *Bead* and *Heavy-atom* representations are shown in Figures 4.9a and 4.9b, respectively. Both models achieve high accuracy in predicting transfer integrals. The Bead-based model yields a root-mean-square error (RMSE) of 23.85 meV/molecule and a mean absolute error (MAE) of 16.86 meV/molecule. In comparison, the Heavy-atom model achieves even lower errors: RMSE = 13.63 meV/molecule and MAE = 9.84 meV/molecule.

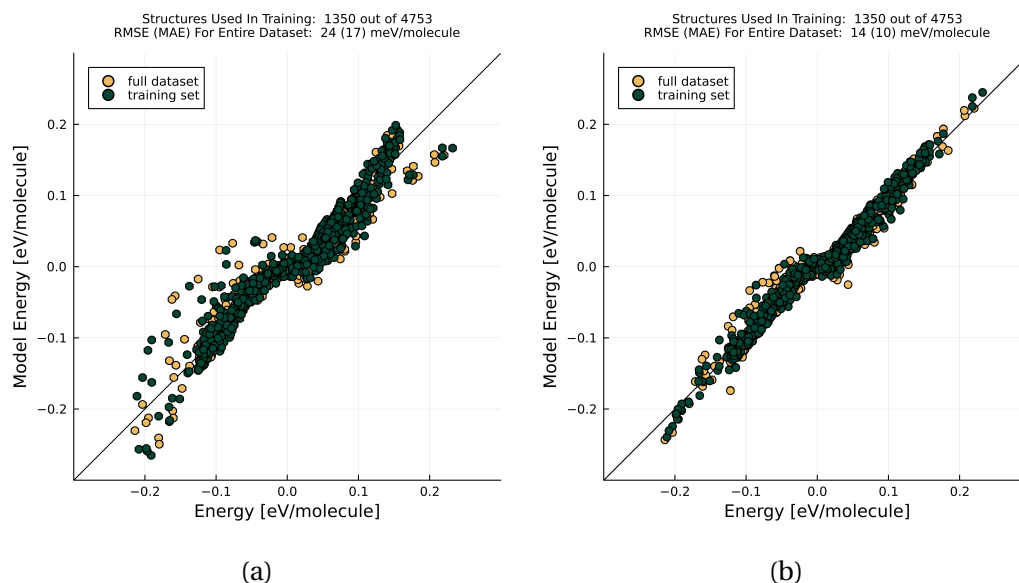


Figure 4.9: The plots comparing predicted and reference energies for ACE models trained on 1350 structures out of a total of 4,753 dimer configurations. (a) Model trained using the *Bead* representation. (b) Model trained using the *Heavy-atom* representation.

These results are notable in that the predictive errors are of similar magnitude across the two representations, yet the computational cost of the Bead-based model is significantly lower—by approximately a factor of five. The Bead representation uses only four coarse-grained Beads per naphthalene dimer, while the Heavy-atom representation employs twenty carbon atoms per configuration. This efficiency gain results from the reduced number of sites per configuration.

Therefore, the Bead (or coarse-grained) representation may offer an advantage in high-throughput screening scenarios where computational efficiency is critical. On the other hand, when maximal predictive accuracy is required—for example, in detailed mechanistic studies or high-fidelity simulations—the Heavy-atom model provides superior performance.

We also provide an analysis of hyperparameter fine-tuning with respect to the order (interaction order) and totaldegree parameters. Figure 4.10

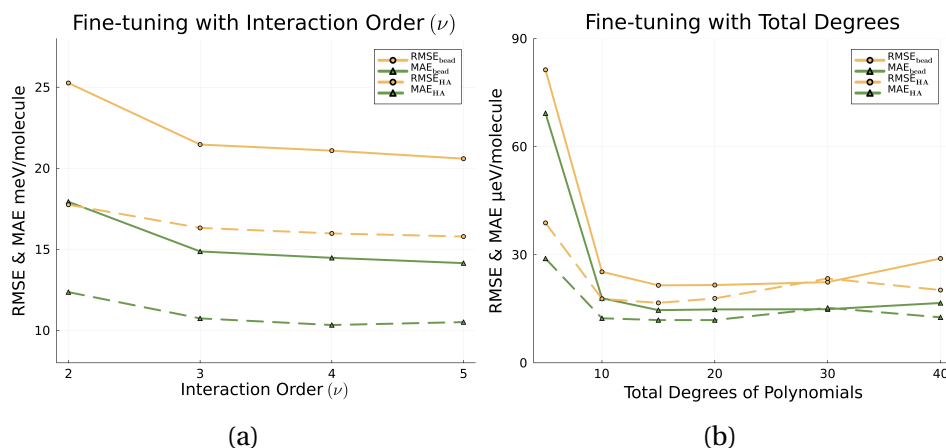


Figure 4.10: Hyperparameter sensitivity analysis of ACE models fitted to transfer integrals using Bead (solid) and Heavy-atom representations (dashed).

presents the performance of ACE models fitted to transfer integrals using Bead (solid lines) and Heavy-atom (dashed lines) representations. Additional fine-tuning on these two hyperparameters is provided in Appendix A.3.1 and A.3.2, respectively.

Panel (a) shows the root-mean-square error (RMSE) and mean absolute error (MAE) as a function of the interaction order ν , which determines the maximum $(\nu + 1)$ -body correlation captured in the ACE descriptor. As expected, increasing the interaction order improves accuracy for both representations, although with diminishing returns beyond $\nu = 4$.

Panel (b) illustrates the effect of increasing the total polynomial degree. Both representations benefit from increasing `totaldegree` up to a point of saturation (`totaldegree = 20`), after which performance plateaus or slightly worsens due to overfitting or numerical instability.

We can also see that this fine-tuning analysis supports the earlier discussion for Figure 4.9, confirming that the Heavy-atom representation consistently achieves lower errors than the Bead-based model across these two tested hyperparameter settings.

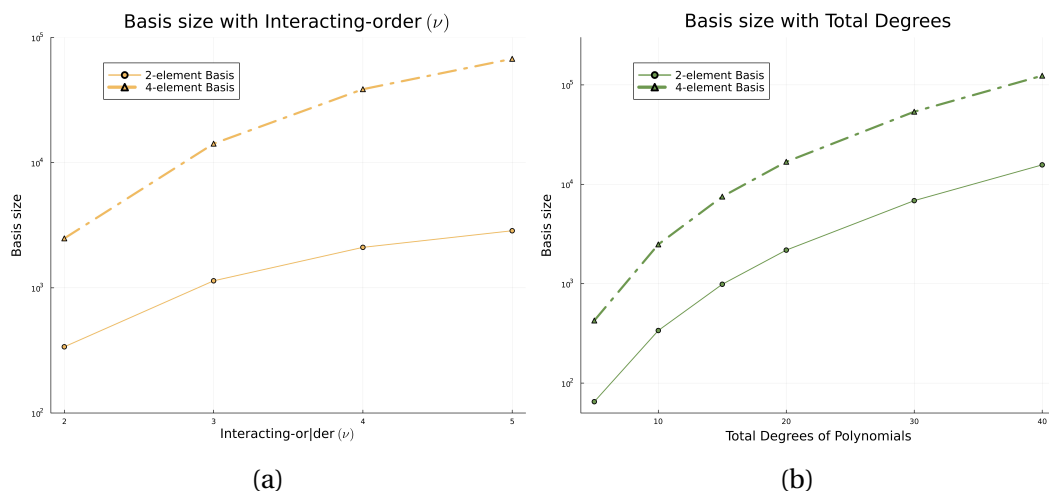


Figure 4.11: Growth of ACE basis size as a function of key hyperparameters for 2- (solid) and 4- (dashdot) chemical species.

However, these improvements in accuracy come with an increased computational cost. As shown in Figure 4.11, the size of the ACE basis set for 2-chemical species grows rapidly with both order and totaldegree values. Higher values lead to a substantial increase in both memory usage and model training time. These trends emphasise the compensation between model expressivity and computational efficiency in ACE model construction.

We also further analysed and fine-tuned additional hyperparameters, including cutoff radius r_{cut} , parameters p , and q for controlling the envelop function of RBFs. The corresponding ACE fitting results and discussions, provided in Appendix A.3, will justify the suitable range of these hyperparameters for both Bead and Heavy-atom representation.

In the following sections, we apply the same analysis workflow to predict the transfer integrals of two additional molecular dimers: ethylene and thiophene. For these systems, the reference transfer integrals were generated using the XTB-DIPRO package [66], rather than PYMOO. As discussed in Section 2.6.3, XTB-DIPRO leverages a semi-empirical tight-binding framework, offering significantly lower computational costs compared to conventional

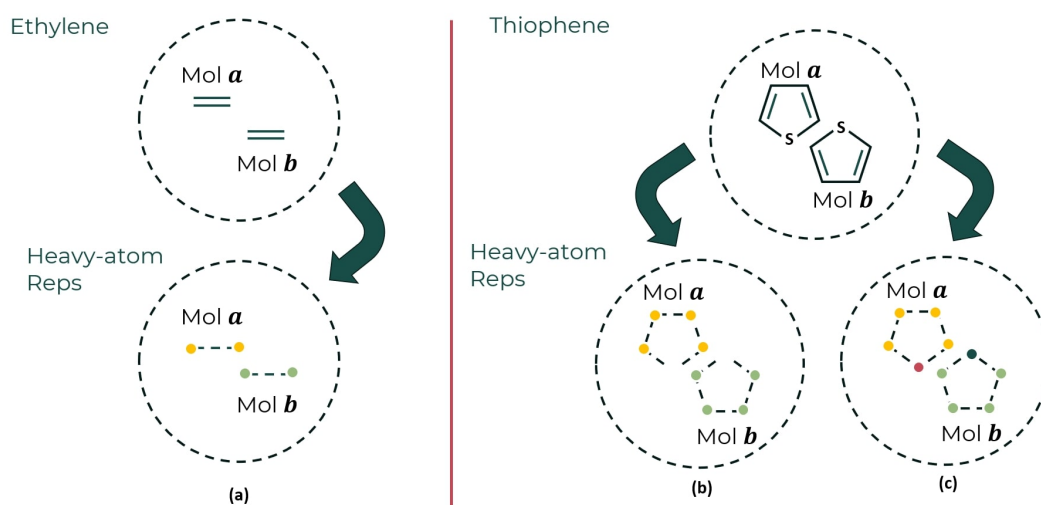


Figure 4.12: Illustration of three types of molecular representations used for constructing ACE descriptors for the ethylene and thiophene dimer (Mol *a* and Mol *b*).

density functional theory (DFT) methods. This makes it particularly well-suited for generating large datasets required for ACE model training.

4.4 Ethylene and Thiophene Dimers

In this section, we examine suitable molecular representations for the ethylene and thiophene dimers and evaluate the corresponding ACE model predictions of their transfer integrals.

We, here, exclude the *Bead* (coarse-grained) representation from consideration. For small molecules, this representation results in insufficient geometric resolution as a molecule turning into a point (pseudo-atom). This makes it difficult to build effective radial basis sets and may lead to poor predictive performance.

We focus on the *Heavy-atom* representation, as illustrated in Figure 4.12. On the left of the figure (a), the ethylene dimer is depicted using the standard Heavy-atom representation, incorporating only carbon atoms. On the right,

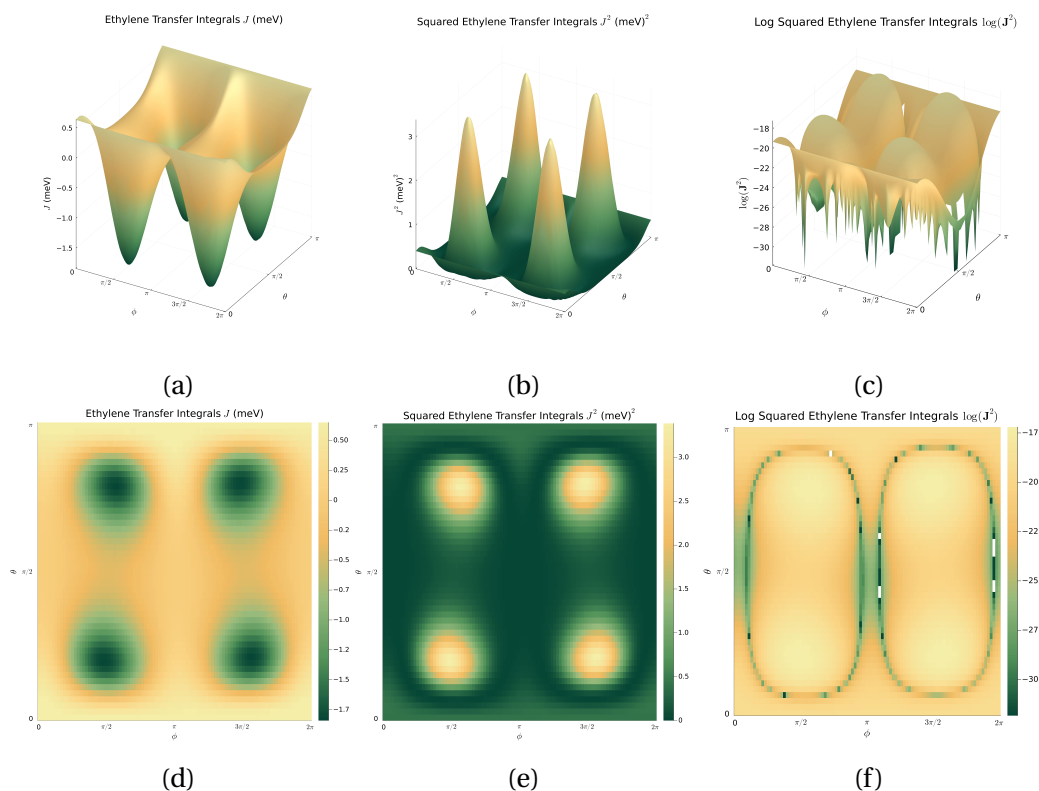


Figure 4.13: (a), (b), and (c) show the heatmaps of the ethylene dimer's transfer integral, the squared transfer integral, and the logarithm of the squared transfer integral. (d), (e), and (f) illustrate the corresponding heatmap.

we consider two variants for representing the thiophene dimer: pathway (b) includes only carbon atoms, while pathway (c) incorporates both carbon and sulphur atoms as distinct species.

Analysis of Transfer Integral Surface and Training Sets

Figure 4.13 and 4.14 presents the transfer integral surfaces and heatmaps for the ethylene and thiophene dimers. Panels (a)–(f) correspond to the ethylene and thiophene dimer show the transfer integrals J , their squared values J^2 , and the logarithm of the squared values $\log(J^2)$, respectively. These transfer integrals are generated by systematically rotating molecule b over the upper hemisphere around molecule a , while maintaining a fixed

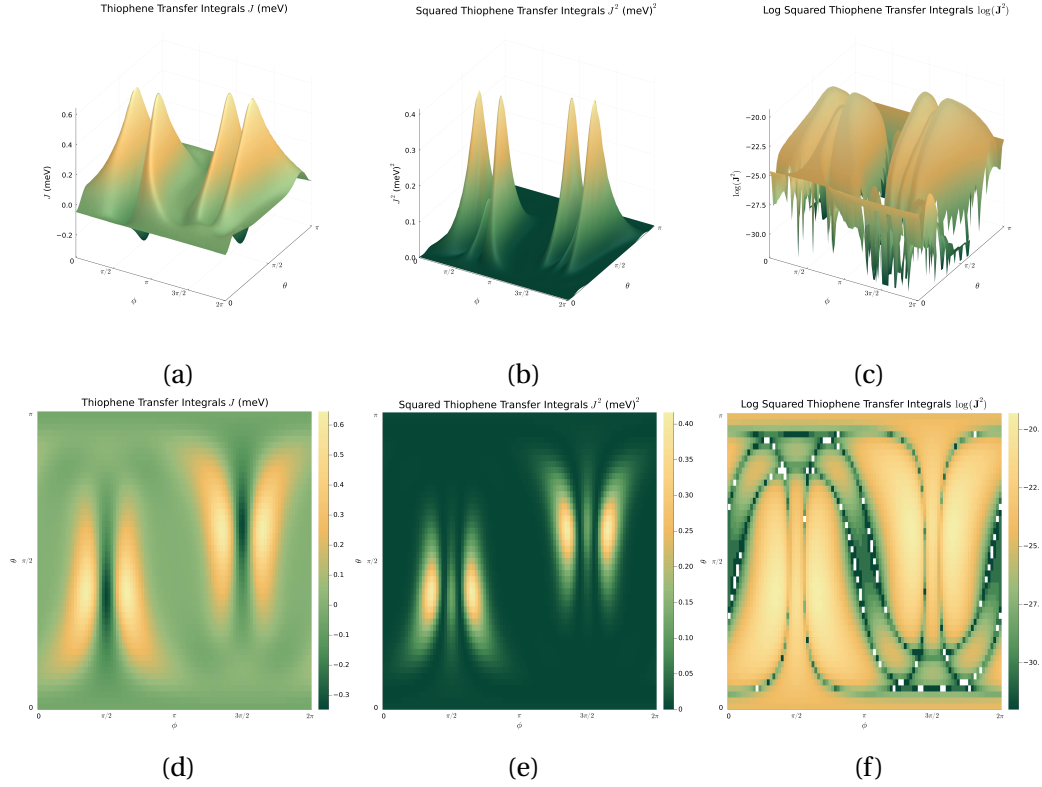


Figure 4.14: (a), (b), and (c) show the heatmaps of the thiophene dimer’s transfer integral, the squared transfer integral, and the logarithm of the squared transfer integral. (d), (e), and (f) illustrate the corresponding heatmap.

centre-of-mass distance of 6.25 Å for the ethylene dimer and of 7.50 Å for the thiophene dimer. The orientations are parameterised by the angular coordinates $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi]$.

This visualisation helps to determine the most suitable target quantity for ACE model training. Despite some local skewness and discontinuities, the transfer integral surface J appears noticeably smoother than those of J^2 and $\log(J^2)$, supporting its selection as a more appropriate target for regression with smooth basis function models such as ACE.

For generating the ethylene’s training set, we sampled a three-dimensional surface,

$$J(x_{ab}, y_{ab}, z_{ab}) := J(r_{ab}, \theta_{ab}, \phi_{ab}),$$

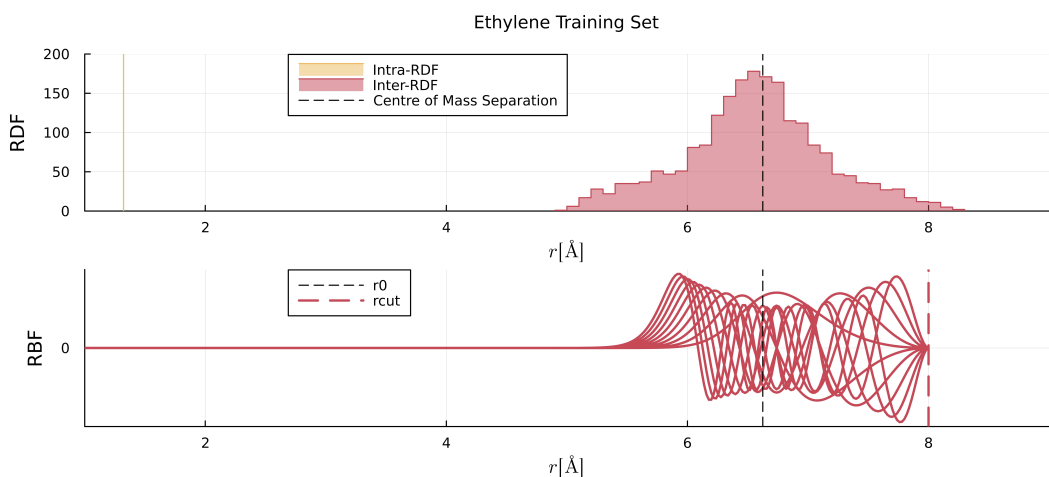


Figure 4.15: Ethylene dimer’s RDFs and RBFs used in the ACE descriptor construction for the Heavy-atom representation.

where the intermolecular separation $r_{ab} \in [6.25, 7.00] \text{ \AA}$. This separation range guarantees the transfer integral results being smooth and continuous. The polar angle $\theta_{ab} \in [0, \pi]$, while the azimuthal angle $\phi_{ab} \in [0, 2\pi]$. The relative orientation of molecule B is fixed at $\theta_z = \pi/6$ with respect to molecule A . This θ_z will provide normal distributed radial distribution functions (RDFs) which are shown in next section.

For the thiophene’s training set, we use the same setting as the ethylene’s one but change the intermolecular separation $r_{ab} \in [7.25, 7.75] \text{ \AA}$.

Representing Dimers

We now proceed to construct the ACE descriptors for the ethylene and thiophene dimer, following the molecular representations illustrated in Figure 4.12.

The left pathway (panel a) shows the Heavy-atom representation for the ethylene dimer. Similar to the naphthalene’s case, to distinguish the two molecules, the atoms in Mol a is labelled as a pseudo-**C** element (yellow), while the atoms in Mol b is labelled as a pseudo-**Si** element (green). This

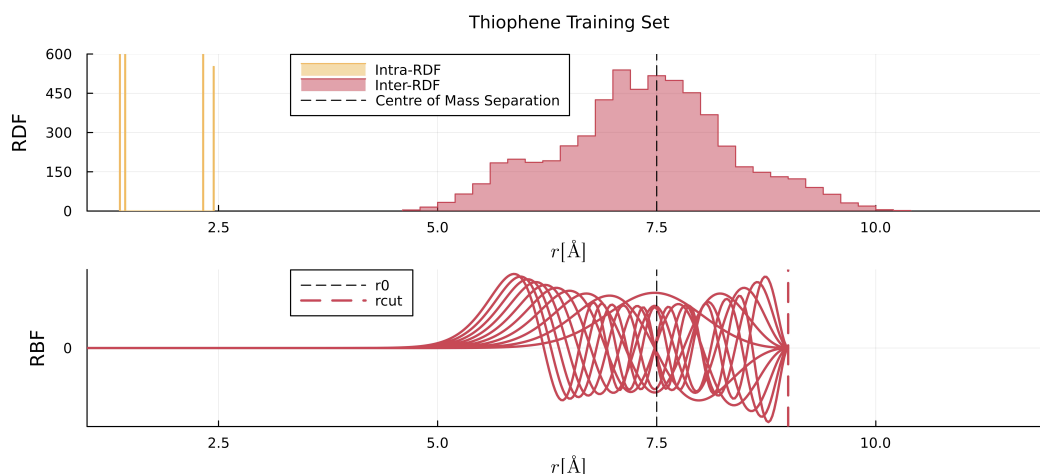


Figure 4.16: Thiophene dimer RDFs and RBFs used in the ACE descriptor construction for the Heavy-atom representation.

labelling enforces molecule-specific environments, allowing the ACE model to distinguish between intra- and intermolecular interactions.

The right pathway (panels b and c) illustrates two variations of the *Heavy-atom* representation. Panel (b) includes only carbon atoms, while panel (c) includes both carbon and sulphur atoms. The sulphur atoms are assigned distinct element labels (e.g., red and dark green) to ensure that atomic species and molecular origin are treated separately.

Again these representations enable a reduction in descriptor dimensionality while preserving essential geometric and electronic features relevant to transfer integral prediction. Moreover, the labelling strategy ensures clear separation of intra- and intermolecular contributions in the ACE basis.

As a next step, we analyse the radial distribution functions (RDFs) for intra- and intermolecular interactions. These are plotted as a function of pairwise distance between the atoms in the ethylene and thiophene dimer in the top panel of Figure 4.15 and of Figure 4.16, respectively. The yellow histogram corresponds to intra-molecular RDFs, while the red histogram represents intermolecular RDFs. The dashed vertical line indicates the fixed centre-of-mass separation (6.5 $\text{\AA}$ and 7.5 $\text{\AA}$) used in the ethylene and thiophene's

```
1 model = acemodel(elements = [:C, :Si, :S, :O, ],  
2     ..., # c.f. Figure 4.7  
3     Eref = [:C => 0.0, :Si => 0.0, :S => 0.0, :O => 0.0, ]  
4     )
```

Figure 4.17: The extension of the construction of the ACE basis to support 4-chemical element structures.

dataset. The both bottom panels illustrate the RBFs which can be extracted from the ACE model in the following discussion.

For the construction of the ACE model, we adopt the framework illustrated in Figure 4.7, applying it to the Heavy-atom representation containing only carbon atoms (Figure 4.13a and 4.13b). In each case, the parameters r_0 and r_{cut} are appropriately adjusted to reflect the relevant intermolecular distance ranges observed in the thiophene dimer dataset.

Extending the model to the Heavy-atom representation that includes both carbon and sulphur atoms (Figure 4.13c) requires additional species which must be introduced to the ACE descriptor. In particular, sulphur atoms in molecules *a* and *b* are treated as distinct atomic types. For example, the species may be defined as in the code in Fig. 4.17.

In this configuration, a pseudo-oxygen label ($:O$) is assigned to the sulphur atom in Mol *b* to ensure that the ACE model distinguishes between chemically identical atoms residing in different molecular environments. This strategy is consistent with the intermolecular nature of transfer integrals and enables explicit encoding of cross-molecular interactions in the basis construction.

ACEfit!

The fitting performance of the ACE models for the ethylene dimer is shown in Figure 4.18a. The model was trained on 2,000 structures out of a total of 8,000 and exhibit strong predictive performance, with the RMSEs (MAEs)

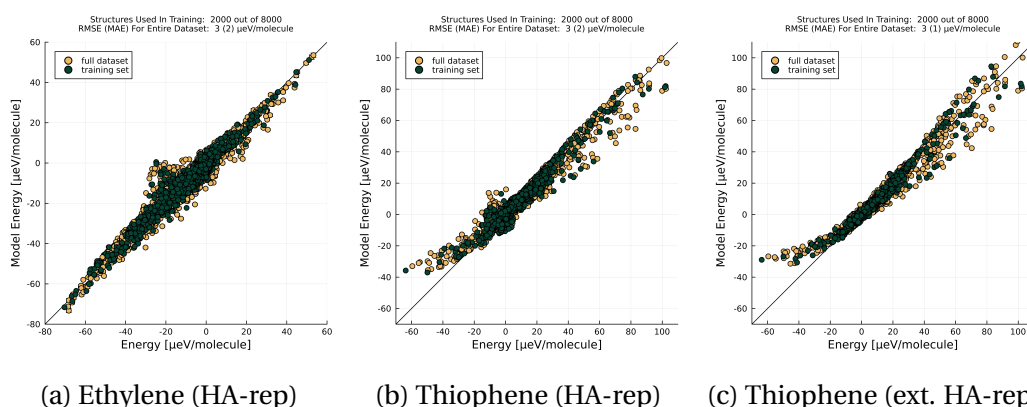


Figure 4.18: The plots comparing predicted and reference transfer integrals for ACE models trained on 2,000 structures out of a total of 8,000 dimer configurations.

of 3.20 (2.30) $\mu\text{eV}/\text{molecule}$ for the purely carbon (ethylene’s Heavy-atom) representations.

The thiophene’s fitting results are illustrated in Figures 4.18b and 4.18c. Both of the results are fitted using the models were trained on 2,000 structures out of a total of 8,000 as the ethylene’s case. The fitting performance is also high with the RMSEs (MAEs) of 3.10 (2.03) $\mu\text{eV}/\text{molecule}$ for the purely carbon (thiophene’s Heavy-atom) and of 2.65 (1.40) $\mu\text{eV}/\text{molecule}$ for the carbon-sulphur (thiophene’s extended Heavy-atom) representations.

In comparing the two models, the incorporation of sulphur reduces the model error by around 15% for the RMSE and approximately 40% for the MAE, based on 2,000 training structures. This change, however, increases the computational cost associated with assessing the model descriptors (the basis set) and extends the duration required for fitting the ACE model: the training wall-clock time increases from 15 to 20 minutes (2,000 training structures).

Moreover, we perform a systematic hyperparameter study focusing on the order (interaction order) and totaldegree parameters. Figure 4.19 illustrates the performance of ACE models trained to predict transfer integrals for the purely carbon (ethylene heavy-atom) representation.

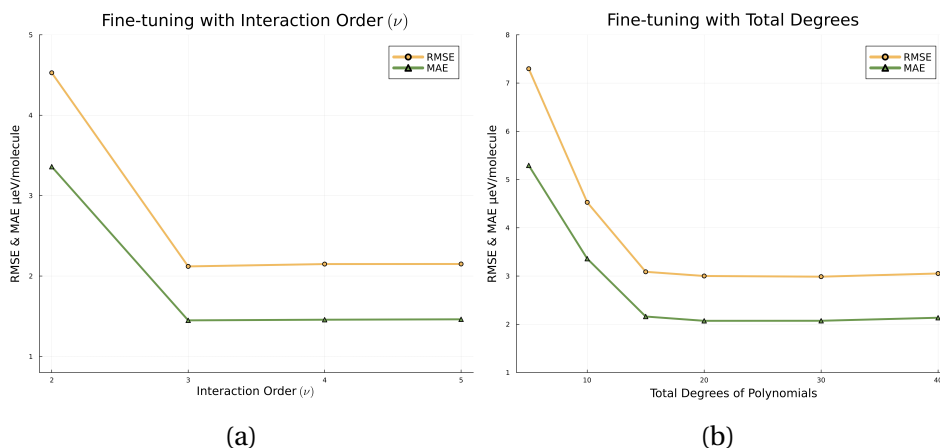


Figure 4.19: Hyperparameter sensitivity analysis of ACE models fitted to transfer integrals using Heavy-atom representations for fitting the ethylene’s transfer integrals.

Panel (a) presents the RMSE and MAE as functions of the interaction order ν , which determines the maximum $(\nu + 1)$ -body correlations represented in the ACE descriptor. Increasing the interaction order improves model accuracy, with the optimal performance observed at $\nu = 3$. Panel (b) shows the effect of the total polynomial degree, `totaldegree`, on fitting performance. Increasing `totaldegree` enhances accuracy up to a saturation point at `totaldegree = 20`, beyond which the performance plateaus or slightly degrades due to overfitting and numerical instabilities.

The results for thiophene are presented in Figure 4.20. For the purely carbon (thiophene heavy-atom) representation, the trends are consistent with those observed for ethylene, while increasing the interaction order has little effect for the richer carbon–sulphur (extended heavy-atom) representation. In both cases, increasing the total polynomial degree improves accuracy, with optimal performance also achieved at `totaldegree = 20`.

We evaluated our model’s efficiency across a variety of semiconducting substances. The naphthalene model has an acceptable inaccuracy for charge-transfer simulations, with inaccuracies less than 2% (9.84 meV/molecule). The error for ethylene and thiophene is considerably elevated (5–10%) when

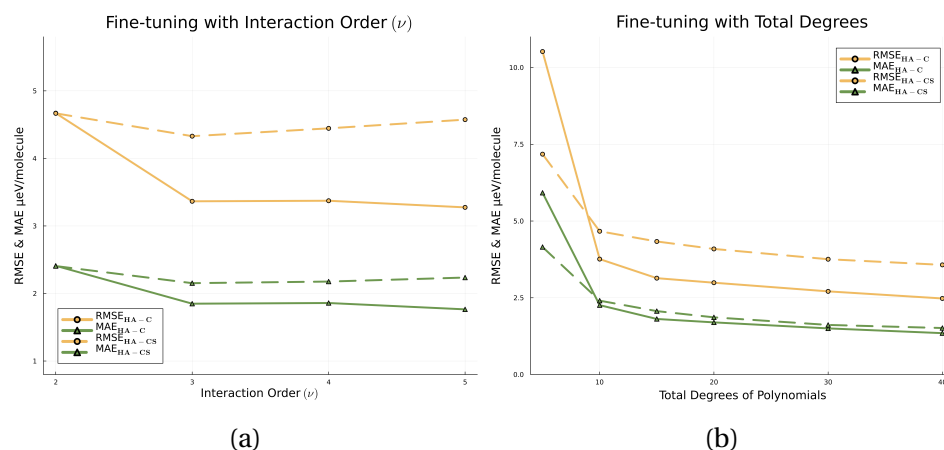


Figure 4.20: Hyperparameter sensitivity analysis of ACE models fitted to transfer integrals using Heavy-atom representations (green) for fitting the ethylene’s transfer integrals.

trained on 2,000 samples, suggesting a necessity for enhanced inductive biases for these systems. We propose that this decreased performance results from the model’s capacity to deduce the orientation in the case of ethylene (heavy atoms) and the distinct chemistry of the C_2 and C_3 carbons in thiophene, which we drive the model to combine. There is substantial valuable work to be carried out experimenting with the fitting of complicated organic electronic molecules of contemporary technical significance, as well as with other structural representations.

To contextualise our findings, we compare them to existing literature. This comparison is constrained by the use of different datasets. The establishment of standard standards for organic electronic transfer integrals will greatly assist the community in the development and validation of these methodologies, rendering them efficient tools for modelling real materials.

We emphasise the data efficiency of our approach. Recent studies[71; 82] reveal mean absolute errors (MAEs) of 1.99 meV for naphthalene and 0.27 meV for ethylene using multi-stage kernel ridge regression (KRR) models with 15,000 to 40,000 training samples [71; 82], and approximately 6.5 meV for naphthalene utilising artificial neural networks (ANNs) trained on 120,000

samples [71]. Our naphthalene model achieves comparable accuracy with 1,350 samples.

4.5 Conclusions and Future Work

This chapter has designed and extensively assessed an Atomic Cluster Expansion (ACE) framework for predicting electronic transfer integrals in organic molecular dimers. By expressing J_{ab} as a summation of localised, symmetry-conserving site contributions, the proposed methodology guarantees Hermitian symmetry, rotational and translational invariance, and spatial locality—attributes that are crucial for accurately depicting intermolecular electronic couplings in a physically relevant context.

Two molecular representations were presented: a coarse-grained *Bead* representation and a chemically informed *Heavy-atom* representation. The Bead method delivers notable computational performance, especially for extensive aromatic systems, whilst the Heavy-atom representation ensures enhanced accuracy by preserving critical geometric and chemical information. ACE models demonstrated high predictive accuracy across naphthalene, ethylene, and thiophene dimers with relatively limited training datasets, showcasing superior data efficiency compared to traditional kernel- and neural-network-based methods.

An hyperparameter investigation indicated that model performance enhances with rising interaction order and polynomial degree until a saturation threshold is reached, after which further complexity results in decreasing returns or causes numerical instability. In chemically simple systems like ethylene, increased interaction orders significantly improve accuracy; yet, in more heterogeneous systems such as thiophene with an extensive carbon-sulphur representation, the advantage becomes negligible due to the existing chemical variety enriching the descriptor. The ideal polynomial degree was consistently identified at $\text{totaldegree} = 20$.

The comparative investigation of various molecular representations and hyperparameter configurations demonstrates that ACE is highly effective at capturing the geometric dependency of transfer integrals, even in chemically diverse systems. The study simultaneously identifies specific limitations: in systems where electronic coupling fluctuates substantially with geometry or displays significant anisotropy, the continuous polynomial framework of the ACE basis might require larger basis sets or more focused sampling of configuration space to attain equivalent accuracy. These observations point toward promising directions for future development, including hybrid descriptor schemes, improved target transformations (such as modelling $\log J^2$ to stabilise learning), and tighter integration with graph-based neural descriptors or message-passing schemes [51; 52].

This chapter’s results indicate that ACE offers a robust, symmetry-preserving, and computationally scalable framework for simulating intermolecular electrical interactions. The technique provides reliable predictions while maintaining sufficient flexibility to address the complicated nature of real molecular materials. This offers a robust basis for future extensive charge-transport simulations in organic semiconductors and provides an opportunity to expand ACE-based models to more sophisticated chemical and structural contexts.

5

Conclusion and Future directions

This thesis demonstrates that physically informed machine learning can be effectively used to model complex phenomena in functional materials, substantially improving the predictive capability and efficiency of conventional first-principles methods. By integrating solid physical theories with machine learning methods, we have created sophisticated surrogate models that maintain accuracy while significantly decreasing computational demands.

Our development of higher-derivative Gaussian process models allowed us to accurately reconstruct harmonic and cubic anharmonic force constants directly from DFT-calculated energies and forces. These Gaussian process force constants (GPFCs) can then be used to perform phonon calculations, such as dispersion, lifetimes, and thermal conductivity, considerably surpassing standard finite-displacement techniques in computational efficiency. Nevertheless, the current implementation encountered practical limitations, notably high memory requirements for higher-order kernel derivatives, restricting the full realisation of advanced active learning strategies. Future research should explicitly address these computational constraints by exploring improved computational techniques and alternative implementations. Moreover, extending our methodology to complex-valued phonon descriptors beyond the Γ -point, and incorporating the Atomic Cluster Expansion (ACE) framework for accelerated data generation represent promising paths forward.

In parallel, our adaptation of the ACE framework for modelling charge transfer integrals in organic molecular dimers has shown strong potential. The ACE-based method, utilising systematically constructed symmetry-adapted descriptors, showed its ability to precisely model the complex relationship between electronic couplings, molecular orientation, and separation. Challenges still remain—especially regarding model performance when confronted with sharply peaked or discontinuous transfer integral surfaces characteristic of particular molecular systems. Addressing these challenges requires additional studies aimed at enhancing radial sampling techniques for obtaining improved descriptor resolution and distribution. Altering prediction targets, such as using log-scaled values, may result in more uniform distributions and enhanced model accuracy. Moreover, incorporating ACE descriptors with sophisticated neural architectures, such as graph neural networks or message-passing frameworks [51; 52], offers a successful technique for capturing complicated anisotropic and long-range interactions that characterise real-world systems.

Ultimately, applying these enhanced models to large-scale and realistic simulations of device-level phenomena is a crucial next step, validating their practical utility and opening new avenues for the discovery and optimisation of novel functional materials.



Detailed Derivations and Results

A.1 Differentiable GP Kernels

A key property of many kernel functions used in Gaussian process regression is smoothness, which is directly related to differentiability. In this section, we examine the differentiability properties of the widely-used squared exponential (SE) or a radial basis function (RBF) kernel, which is infinitely differentiable and thus yields a Gaussian process prior over smooth functions [1; 5].

Before examining a squared exponential kernel and its derivative, we initially define the squared exponential kernel as

$$k(\mathbf{x}, \mathbf{x}^*) = \sigma_o \exp\left(-\frac{1}{2\ell^2} \|\mathbf{x} - \mathbf{x}^*\|^2\right), \quad (\text{A.1})$$

where σ_o is the , ℓ is the length scale, and $\|\cdot\|$ denotes the Euclidean norm. The kernel depends on the squared distance between input vectors, making it isotropic and stationary.

A.1.1 Arbitrary Order Derivatives

Since the SE kernel is infinitely differentiable with respect to both arguments $\mathbf{x}$ and $\mathbf{x}^*$, this property allows us to compute analytic gradients and higher-order derivatives, which are crucial for modelling derivatives of functions

(e.g., force fields, gradient, and higher-order derivative observations) or constructing physically informed models.

I begin with deriving from the first- up to sixth-order derivatives which are essential for the predictive means and variances of Gaussian process forces constant, i.e., defined in Equation (3.7), (3.9), and (3.26).

The first order derivative of the kernel with respect to $\mathbf{x}$ can be derived To compute the first-order partial derivative of $k(\mathbf{x}, \mathbf{x}^*)$ with respect to component x_i of $\mathbf{x}$, we first define $\mathbf{r} = \mathbf{x} - \mathbf{x}^*$ and its Euclidean distance

$$\mathbf{r}^2(\mathbf{x}, \mathbf{x}^*) = \|\mathbf{x} - \mathbf{x}^*\|^2 = \sum_{j=1}^d (x_j - x_j^*)^2,$$

where $\mathbf{x}, \mathbf{x}^* \in \mathbb{R}^d$. The kernel then becomes:

$$k(\mathbf{x}, \mathbf{x}^*) = \sigma_o \exp\left(-\frac{1}{2\ell^2} \mathbf{r}^2\right),$$

which is analougous to (3.1). Now the gradient of the kernel can be evaluated using the chain rule:

$$\begin{aligned} \frac{\partial k}{\partial x_i} &= \frac{dk}{d\mathbf{r}^2} \cdot \frac{\partial \mathbf{r}^2}{\partial x_i} \\ &= \left(-\frac{1}{2\ell^2} k(\mathbf{x}, \mathbf{x}^*)\right) \cdot 2(x_i - x_i^*) \\ &= -\frac{1}{\ell^2} k(\mathbf{x}, \mathbf{x}^*) (x_i - x_i^*). \end{aligned} \tag{A.2}$$

Thus, we can represent it in the form of the d -dimensional vector given by

$$\boxed{\nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}^*) = -\frac{1}{\ell^2} k(\mathbf{x}, \mathbf{x}^*) \cdot \mathbf{r}.} \tag{A.3}$$

Likewise, the gradient with respect to $\mathbf{x}^*$ is expressed as

$$\boxed{\nabla_{\mathbf{x}^*} k(\mathbf{x}, \mathbf{x}^*) = \frac{1}{\ell^2} k(\mathbf{x}, \mathbf{x}^*) \cdot \mathbf{r},} \tag{A.4}$$

which is the antisymmetric form of Eq. (A.3).

We can now compute the second-order derivative of the squared exponential kernel with respect to the j -th component of $\mathbf{x}$:

$$\frac{\partial^2 k}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_j} \left(\frac{\partial k}{\partial x_i} \right).$$

This calculation is divided into two cases: $i = j$ and $i \neq j$. Using the first derivative solution in (A.3), we compute the second-order partial derivative of the kernel with respect to $x_j = x_i$ for the first case, resulting in

$$\begin{aligned} \frac{\partial^2 k(\mathbf{x}, \mathbf{x}^*)}{\partial x_i^2} &= \frac{\partial}{\partial x_i} \left[-\frac{1}{\ell^2} (x_i - x_i^*) k(\mathbf{x}, \mathbf{x}^*) \right] \\ &= -\frac{1}{\ell^2} \left[\frac{\partial}{\partial x_i} (x_i - x_i^*) \cdot k(\mathbf{x}, \mathbf{x}^*) + (x_i - x_i^*) \cdot \frac{\partial k(\mathbf{x}, \mathbf{x}^*)}{\partial x_i} \right] \\ &= -\frac{1}{\ell^2} \left[k(\mathbf{x}, \mathbf{x}^*) - \frac{1}{\ell^2} (x_i - x_i^*)^2 k(\mathbf{x}, \mathbf{x}^*) \right] \\ &= \left(\frac{1}{\ell^4} (x_i - x_i^*)^2 - \frac{1}{\ell^2} \right) k(\mathbf{x}, \mathbf{x}^*). \end{aligned} \tag{A.5}$$

For the second case, cross-differentiating with respect to $x_j \neq x_i$ yields

$$\begin{aligned} \frac{\partial^2 k(\mathbf{x}, \mathbf{x}^*)}{\partial x_i \partial x_j} &= \frac{\partial}{\partial x_j} \left(-\frac{1}{\ell^2} (x_i - x_i^*) k(\mathbf{x}, \mathbf{x}^*) \right) \\ &= -\frac{1}{\ell^2} (x_i - x_i^*) \cdot \frac{\partial k(\mathbf{x}, \mathbf{x}^*)}{\partial x_j} \\ &= -\frac{1}{\ell^2} (x_i - x_i^*) \cdot \left(-\frac{1}{\ell^2} (x_j - x_j^*) k(\mathbf{x}, \mathbf{x}^*) \right) \\ &= \frac{1}{\ell^4} (x_i - x_i^*) (x_j - x_j^*) k(\mathbf{x}, \mathbf{x}^*). \end{aligned} \tag{A.6}$$

Combining the solutions in (A.5) and (A.6), the second derivative is expressed as a rank-2 tensor or a $d \times d$ matrix

$$\boxed{(\nabla_{\mathbf{x}})^2 k(\mathbf{x}, \mathbf{x}^*) = \left(\frac{1}{\ell^4} \mathbf{r} \otimes \mathbf{r} - \frac{1}{\ell^2} \mathbb{I} \right) k(\mathbf{x}, \mathbf{x}^*),} \tag{A.7}$$

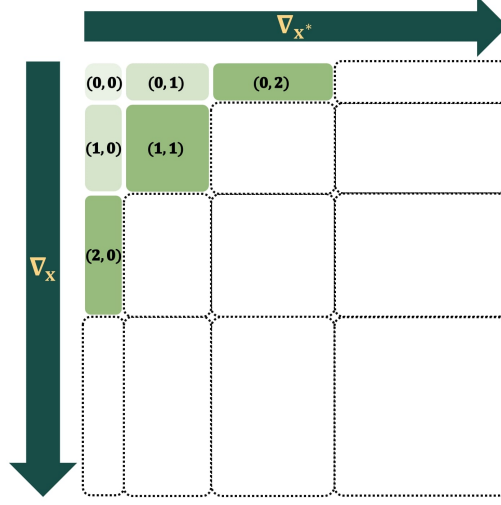


Figure A.1: Construction of the derivative kernel (covariance) matrix used in our GP model.

where $\mathbb{1}$ is a $(d \times d)$ identity matrix. Similarly, the mixed partial derivative of the squared exponential kernel with respect to $\mathbf{x}$ and $\mathbf{x}^*$ is expressed as

$$\nabla_{\mathbf{x}} \nabla_{\mathbf{x}^*} k(\mathbf{x}, \mathbf{x}^*) = \left(-\frac{1}{\ell^4} \mathbf{r} \otimes \mathbf{r} + \frac{1}{\ell^2} \mathbb{1} \right) k(\mathbf{x}, \mathbf{x}^*), \quad (\text{A.8})$$

and the second derivative with respect to $\mathbf{x}^*$ and $\mathbf{x}^*$ is given by

$$(\nabla_{\mathbf{x}^*})^2 k(\mathbf{x}, \mathbf{x}^*) = \left(\frac{1}{\ell^4} \mathbf{r} \otimes \mathbf{r} - \frac{1}{\ell^2} \mathbb{1} \right) k(\mathbf{x}, \mathbf{x}^*). \quad (\text{A.9})$$

By examining these three second-order derivative solutions, we can form the symmetric relation among them as follows

$$(\nabla_{\mathbf{x}})^2 k(\mathbf{x}, \mathbf{x}^*) = -\nabla_{\mathbf{x}} \nabla_{\mathbf{x}^*} k(\mathbf{x}, \mathbf{x}^*) = (\nabla_{\mathbf{x}^*})^2 k(\mathbf{x}, \mathbf{x}^*). \quad (\text{A.10})$$

The SE kernel, along with the solutions of its first- and second-order derivatives, can be used to construct the covariance matrix depicted in Figure A.1. This matrix construction provides the systematic classification of the $(m+n)$ th-order kernel derivative $(\nabla_{\mathbf{x}})^m (\nabla_{\mathbf{x}^*})^n k(\mathbf{x}, \mathbf{x}^*)$, where the order m

and n increase along the diagonal direction. Each $(\nabla_{\mathbf{x}})^m (\nabla_{\mathbf{x}^*})^n k(\mathbf{x}, \mathbf{x}^*)$ is denoted by the block (m, n) . For example, the SE kernel is the block $(0, 0)$ whereas the block $(2, 0)$ represents the second-order derivative given in (A.7).

Moreover, this matrix assists us in identifying the (anti)symmetry of the kernel derivative in the same order. Specifically, block matrices of identical colour that align along the same antidiagonal exhibit correlation due to their (anti)symmetric property. The blocks $(2, 0)$, $(1, 1)$, and $(0, 2)$ exemplify the (anti)symmetric property defined in (A.10). This matrix construction will become more evident after completing all the blocks.

Let us consider the third-order derivative of the kernel. Correlating the symmetric property of the relevant blocks—specifically, $(3, 0)$, $(2, 1)$, $(1, 2)$, and $(0, 3)$ —allow us to formulate

$$\boxed{(\nabla_{\mathbf{x}})^3 k = -(\nabla_{\mathbf{x}})^2 \nabla_{\mathbf{x}^*} k = \nabla_{\mathbf{x}} (\nabla_{\mathbf{x}^*})^2 k = -(\nabla_{\mathbf{x}^*})^3 k.} \quad (\text{A.11})$$

Accordingly, the construction of all third-order derivative blocks requires the calculation of only a single representative term, from which the remaining components follow by symmetry. We then solve the solution of the third-order derivative with respect to components x_i , x_j , and x_k of $\mathbf{x}$:

$$\frac{\partial^3 k}{\partial x_i \partial x_j \partial x_k}.$$

We begin with applying the product rule to differentiate the Hessian component of Equation (A.7), yielding

$$\frac{\partial^3 k}{\partial x_i \partial x_j \partial x_k} = \frac{\partial}{\partial x_k} \left(\frac{\partial^2 k}{\partial x_i \partial x_j} \right) = \frac{\partial}{\partial x_k} \left[\left(\frac{1}{\ell^4} r_i r_j - \frac{1}{\ell^2} \delta_{ij} \right) k \right].$$

The product rule is, now, employed to obtain

$$\frac{\partial^3 k}{\partial x_i \partial x_j \partial x_k} = \left(\frac{1}{\ell^4} \delta_{ik} r_j + \frac{1}{\ell^4} r_i \delta_{jk} \right) k + \left(\frac{1}{\ell^4} r_i r_j - \frac{1}{\ell^2} \delta_{ij} \right) \frac{\partial k}{\partial x_k}. \quad (\text{A.12})$$

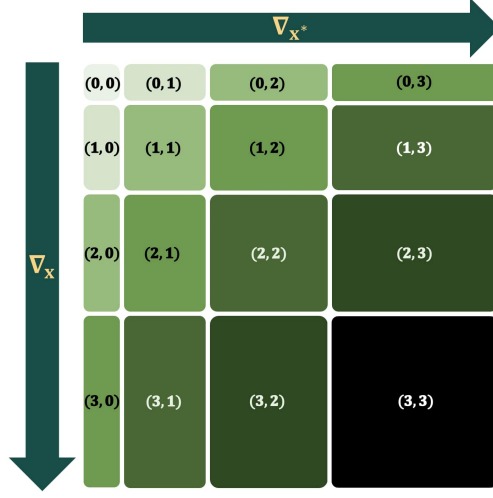


Figure A.2: Complete construction of the derivative kernel matrix used in our GP model.

We substitute the gradient solution given in (A.2), resulting in

$$\begin{aligned} \frac{\partial^3 k}{\partial x_i \partial x_j \partial x_k} &= \left(\frac{1}{\ell^4} \delta_{ik} r_j + \frac{1}{\ell^4} r_i \delta_{jk} \right) k - \left(\frac{1}{\ell^4} r_i r_j - \frac{1}{\ell^2} \delta_{ij} \right) \frac{1}{\ell^2} r_k k \\ &= \left[\frac{1}{\ell^4} (\delta_{ik} r_j + \delta_{jk} r_i) - \frac{1}{\ell^6} r_i r_j r_k + \frac{1}{\ell^4} \delta_{ij} r_k \right] k. \end{aligned} \quad (\text{A.13})$$

Therefore, the final expression is given by

$$\boxed{\frac{\partial^3 k}{\partial x_i \partial x_j \partial x_k} = \left[-\frac{1}{\ell^6} r_i r_j r_k + \frac{1}{\ell^4} (\delta_{ij} r_k + \delta_{ik} r_j + \delta_{jk} r_i) \right] k(\mathbf{x}, \mathbf{x}^*)}. \quad (\text{A.14})$$

We now proceed to populate the third-order derivative blocks (3,0), (2,1), (1,2), and (0,3), as illustrated in Figure A.2. To evaluate the remaining higher-order derivative blocks—namely the fourth-order (3,1), (2,2), (1,3); the fifth-order (3,2), (2,3); and the sixth-order (3,3)—we similarly employ the (anti)symmetry relations between the mixed Wirtinger derivatives, as follows:

- **Fourth-order** (involving $(\nabla_{\mathbf{x}})^4 k$):

$$(\nabla_{\mathbf{x}})^3 \nabla_{\mathbf{x}^*} k = -(\nabla_{\mathbf{x}})^2 (\nabla_{\mathbf{x}^*})^2 k = \nabla_{\mathbf{x}} (\nabla_{\mathbf{x}^*})^3 k := -(\nabla_{\mathbf{x}})^4 k.$$

- **Fifth-order** (involving $(\nabla_{\mathbf{x}})^5 k$):

$$(\nabla_{\mathbf{x}})^3 (\nabla_{\mathbf{x}^*})^2 k = -(\nabla_{\mathbf{x}})^2 (\nabla_{\mathbf{x}^*})^3 k = (\nabla_{\mathbf{x}})^5 k.$$

- **Sixth-order** (involving $(\nabla_{\mathbf{x}})^6 k$):

$$(\nabla_{\mathbf{x}})^3 (\nabla_{\mathbf{x}^*})^3 k := -(\nabla_{\mathbf{x}})^6 k.$$

To evaluate the required derivatives— $(\nabla_{\mathbf{x}})^4 k$, $(\nabla_{\mathbf{x}})^5 k$, and $(\nabla_{\mathbf{x}})^6 k$ —one must additionally apply both the product rule and the chain rule, in a manner analogous to the third-order case discussed in Equations (A.12)–(A.14).

Rather than presenting full derivations, I will provide the solutions associated to those three cases. Beginning with the fourth-order, its components are given by

$$\begin{aligned} \frac{\partial^4 k}{\partial x_i \partial x_j \partial x_k \partial x_a} = & \left[\frac{1}{\ell^8} r_i r_j r_k r_a \right. \\ & - \frac{1}{\ell^6} (\delta_{ij} r_k r_a + \delta_{ik} r_j r_a + \delta_{ia} r_j r_k \\ & + \delta_{jk} r_i r_a + \delta_{ja} r_i r_k + \delta_{ka} r_i r_j) \\ & \left. + \frac{1}{\ell^4} (\delta_{ij} \delta_{ka} + \delta_{ik} \delta_{ja} + \delta_{ia} \delta_{jk}) \right] k(\mathbf{x}, \mathbf{x}^*). \end{aligned} \quad (\text{A.15})$$

We can rewrite the expression in terms of symmetrised permutation notation, yielding

$$\frac{\partial^4 k}{\partial x_i \partial x_j \partial x_k \partial x_a} = \left[\frac{1}{\ell^8} r_i r_j r_k r_a - \frac{6}{\ell^6} \delta_{(ij} r_k r_a) + \frac{3}{\ell^4} \delta_{(ij} \delta_{ka)} \right] k(\mathbf{x}, \mathbf{x}^*) \quad (\text{A.16})$$

where the parentheses $(\cdot)$ denote symmetrisation over the enclosed indices.

Finally, the fifth- and the sixth-order derivatives can be expressed in the symmetrised permutation notation as

$$\frac{\partial^5 k}{\partial x_i \partial x_j \partial x_k \partial x_a \partial x_b} = \left[-\frac{1}{\ell^{10}} r_i r_j r_k r_a r_b + \frac{10}{\ell^8} \delta_{(ij} r_k r_a r_b) - \frac{15}{\ell^6} \delta_{(ij} \delta_{ka} r_b) \right] k(\mathbf{x}, \mathbf{x}^*), \quad (\text{A.17})$$

and

$$\frac{\partial^6 k}{\partial x_i \partial x_j \partial x_k \partial x_a \partial x_b \partial x_c} = \left[\frac{1}{\ell^{12}} r_i r_j r_k r_a r_b r_c - \frac{15}{\ell^{10}} \delta_{(ij} r_k r_a r_b r_c) + \frac{45}{\ell^8} \delta_{(ij} \delta_{ka} r_b r_c) - \frac{15}{\ell^6} \delta_{(ij} \delta_{ka} \delta_{bc}) \right] k(\mathbf{x}, \mathbf{x}^*), \quad (\text{A.18})$$

respectively.

In the following section, I will present a sophisticated framework for developing the differentiable Gaussian Process model to accommodate complex representations, such as phonon descriptors.

A.1.2 Complex Gaussian Process

In this section, we investigate the kernel function, which supports complex vectors, and its derivative with respect to these vectors. Deriving these derivatives allows us to build the differentiable GP model supporting phonon descriptors evaluated at arbitrary wavevectors $\mathbf{q}$. These phonon descriptors are currently not compatible with our GP model, while the phonon descriptor evaluated at Γ -point ($\mathbf{q} = 0$) is only the case that can be applied to our model, as discussed in section 3.2.

We begin with consider a squared exponential kernel defined on complex vectors $\mathbf{z}, \mathbf{z}^* \in \mathbb{C}^d$, where

$$\mathbf{z} = \mathbf{a} + i\mathbf{b}, \quad \mathbf{z}^* = \mathbf{a}^* + i\mathbf{b}^*, \quad \text{with } \mathbf{a}, \mathbf{b} \in \mathbb{R}^d.$$

The squared exponential kernel can be rewritten as

$$k(\mathbf{z}, \mathbf{z}^*) = \sigma_o \exp\left(-\frac{1}{2\ell^2} \|\mathbf{z} - \mathbf{z}^*\|^2\right), \quad (\text{A.19})$$

where

$$\|\mathbf{z} - \mathbf{z}^*\|^2 := (\mathbf{z} - \mathbf{z}^*)^\dagger (\mathbf{z} - \mathbf{z}^*) = \sum_{j=1}^d |z_j - z_j^*|^2,$$

ensuring that k is real-valued and positive definite.

To differentiate the kernel requires to define the *Wirtinger* derivatives with respect to the complex coordinates:

$$\frac{\partial}{\partial z_i} := \frac{1}{2} \left(\frac{\partial}{\partial a_i} - i \frac{\partial}{\partial b_i} \right), \quad \frac{\partial}{\partial \bar{z}_i} := \frac{1}{2} \left(\frac{\partial}{\partial a_i} + i \frac{\partial}{\partial b_i} \right).$$

The first-order derivatives of the kernel are given by:

$$\boxed{\frac{\partial k(\mathbf{z}, \mathbf{z}^*)}{\partial z_i} = -\frac{1}{\ell^2} (z_i - z_i^*) k(\mathbf{z}, \mathbf{z}^*)}, \quad (\text{A.20})$$

$$\boxed{\frac{\partial k(\mathbf{z}, \mathbf{z}^*)}{\partial \bar{z}_i} = -\frac{1}{\ell^2}(\bar{z}_i - \bar{z}_i^*) k(\mathbf{z}, \mathbf{z}^*)}. \quad (\text{A.21})$$

These expressions respect *Holomorphic*¹ structure when treating $\mathbf{z}$ and $\mathbf{z}^*$ as independent formal variables. Higher-order Wirtinger derivatives follow similarly and preserve symmetry under complex conjugation.

We can also take the *Wirtinger* derivative of the kernel with respect to $\mathbf{z}^*$ and $\bar{\mathbf{z}}^*$, resulting in the similar results as Equation (A.4).

Computing the second-order derivatives can be divided into two cases. One is the pure (holomorphic) second derivative with respect to $\mathbf{z}$ and $\mathbf{z}$:

$$\begin{aligned} \frac{\partial^2 k}{\partial z_i \partial z_j} &= \frac{\partial}{\partial z_j} \left[-\frac{1}{\ell^2} (z_i - z_i^*) k \right] \\ &= -\frac{1}{\ell^2} \delta_{ij} k + \frac{1}{\ell^4} (z_i - z_i^*) (z_j - z_j^*) k, \end{aligned} \quad (\text{A.22})$$

which can be expressed as

$$\boxed{\frac{\partial^2 k}{\partial z_i \partial z_j} = \left(\frac{1}{\ell^4} (z_i - z_i^*) (z_j - z_j^*) - \frac{1}{\ell^2} \delta_{ij} \right) k}. \quad (\text{A.23})$$

This result is defined as the component i, j of the rank-2 tensor $\nabla_{\mathbf{z}\mathbf{z}}^2 k$. Meanwhile, the mixed derivative with respect to $\mathbf{z}$ and $\bar{\mathbf{z}}$ can be solved as follows

$$\begin{aligned} \frac{\partial^2 k}{\partial z_i \partial \bar{z}_j} &= \frac{\partial}{\partial \bar{z}_j} \left[-\frac{1}{\ell^2} (z_i - z_i^*) k \right] \\ &= -\frac{1}{\ell^2} (z_i - z_i^*) \cdot \frac{\partial k}{\partial \bar{z}_j} \\ &= -\frac{1}{\ell^2} (z_i - z_i^*) \cdot \left(-\frac{1}{\ell^2} (\bar{z}_j - \bar{z}_j^*) k \right) \\ &= \frac{1}{\ell^4} (z_i - z_i^*) (\bar{z}_j - \bar{z}_j^*) k, \end{aligned} \quad (\text{A.24})$$

¹ A function is considered holomorphic if it is complex differentiable in an open neighbourhood, a characteristic that entails infinite differentiability and analyticity as a result of fulfilling the Cauchy-Riemann equations. This concept is fundamental to complex analysis and is essential for deriving derivatives through Wirtinger calculus [83].

yielding

$$\boxed{\frac{\partial^2 k}{\partial z_i \partial \bar{z}_j} = \frac{1}{\ell^4} (z_i - z_i^*) (\bar{z}_j - \bar{z}_j^*) k} \quad (\text{A.25})$$

as the component i, j of the rank-2 tensor $\nabla_{\bar{\mathbf{z}}}^2 k$.

To obtain the remaining second order derivatives, i.e. $\nabla_{\bar{\mathbf{z}}}^2 k$ and $\nabla_{\mathbf{z}}^2 k$, requires to use the conjugation of derivative identity defined as

$$\overline{\left(\frac{\partial f}{\partial z_i} \right)} = \frac{\partial f}{\partial \bar{z}_i}, \quad (\text{A.26})$$

for $f : \mathbb{C}^d \rightarrow \mathbb{R}$ be a real-valued, smooth function of complex variables. They can, therefore, evaluated through the relations

$$\nabla_{\bar{\mathbf{z}}}^2 k = \overline{\nabla_{\mathbf{z}}^2 k}, \quad (\text{A.27})$$

$$\nabla_{\mathbf{z}}^2 k = \overline{\nabla_{\bar{\mathbf{z}}}^2 k}. \quad (\text{A.28})$$

Combining Equation (A.23), (A.25), (A.27), and (A.28) allow us to form the full description of the second order derivative with respect to $\mathbf{z}$ and $\bar{\mathbf{z}}$, defined as a $(2d \times 2d)$ -matrix:

$$\tilde{D}_{\mathbf{z}\bar{\mathbf{z}}}^2 k = \begin{bmatrix} \nabla_{\mathbf{z}\mathbf{z}}^2 k & \nabla_{\mathbf{z}\bar{\mathbf{z}}}^2 k \\ \nabla_{\bar{\mathbf{z}}\mathbf{z}}^2 k & \nabla_{\bar{\mathbf{z}}\bar{\mathbf{z}}}^2 k \end{bmatrix}. \quad (\text{A.29})$$

The diagonal block matrices $\nabla_{\mathbf{z}\mathbf{z}}^2 k$ and $\nabla_{\bar{\mathbf{z}}\bar{\mathbf{z}}}^2 k$ denote the standard second-order holomorphic derivative and its anti-holomorphic counterpart, respectively. The existence of both real and imaginary values in these two blocks causes themselves generally unphysical. Meanwhile, the off-diagonal blocks $\nabla_{\mathbf{z}\bar{\mathbf{z}}}^2 k$ and $\nabla_{\bar{\mathbf{z}}\mathbf{z}}^2 k$ exhibit physical properties, as they are Hermitian conjugates of each other.

To construct the differential GP model for predicting harmonic force constants, analogous to (3.7), also requires to evaluate the third-order derivatives with respect to those complex vectors.

Regarding the symmetric conjugation property, only three cases are adequate to populate all solutions for the third order derivative.

We begin with defining $r_i := z_i - z_i^*$, and $\bar{r}_i := \bar{z}_i - \bar{z}_i^*$. The third-order Wirtinger derivatives of the complex squared exponential kernel are given by

(i) Holomorphic third derivative:

$$\frac{\partial^3 k}{\partial z_i \partial z_j \partial z_k} = \left(-\frac{1}{\ell^6} r_i r_j r_k + \frac{3}{\ell^4} \delta_{(ij} r_k \right) k, \quad (\text{A.30})$$

(ii) Mixed two holomorphic, one antiholomorphic:

$$\frac{\partial^3 k}{\partial z_i \partial z_j \partial \bar{z}_k} = \left(-\frac{1}{\ell^6} r_i r_j \bar{r}_k + \frac{1}{\ell^4} \delta_{ij} \bar{r}_k \right) k, \quad (\text{A.31})$$

(iii) Mixed one holomorphic, two antiholomorphic:

$$\frac{\partial^3 k}{\partial z_i \partial \bar{z}_j \partial \bar{z}_k} = \left(-\frac{1}{\ell^6} r_i \bar{r}_j \bar{r}_k + \frac{1}{\ell^4} \delta_{jk} r_i \right) k. \quad (\text{A.32})$$

These are the components i, j, k of $\nabla_{\underline{zzz}}^3 k$, $\nabla_{\underline{zz}\bar{z}}^3 k$ and $\nabla_{\underline{z}\bar{z}\bar{z}}^3 k$, respectively. We can employ the conjugation of derivative identity and the permutation symmetry to achieve the following relation

$$\overline{\nabla_{\underline{zzz}}^3 k} = \nabla_{\underline{zz}\bar{z}}^3 k, \quad (\text{A.33})$$

$$\nabla_{\underline{zz}\bar{z}}^3 k = \nabla_{\underline{z}\bar{z}\bar{z}}^3 k = \nabla_{\bar{z}\bar{z}\underline{z}}^3 k, \quad (\text{A.34})$$

$$\nabla_{\bar{z}\bar{z}\underline{z}}^3 k = \nabla_{\bar{z}\underline{z}\bar{z}}^3 k = \nabla_{\underline{z}\bar{z}\bar{z}}^3 k. \quad (\text{A.35})$$

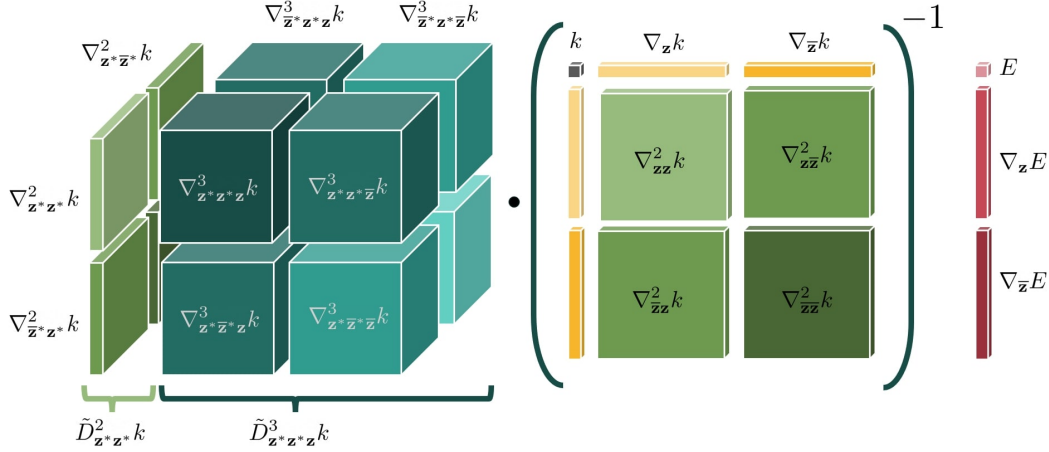


Figure A.3: Figure illustrates the tensor contraction in Equation (A.37), where the derivative GP model is trained on a single data point at descriptor $\mathbf{z}$. This serve as an extension to the tensor contraction shown in Fig. 3.1 and 3.7.

We can now construct the full third order derivative with respect to $\mathbf{z}$ and $\bar{\mathbf{z}}$, defined as a rank-3 tensor with size $2d \times 2d \times 2d$:

$$\tilde{D}_{\mathbf{z}\mathbf{z}\mathbf{z}}^3 k = \left\{ \begin{bmatrix} \nabla_{\mathbf{z}\mathbf{z}\mathbf{z}}^3 k & \nabla_{\mathbf{z}\mathbf{z}\mathbf{z}}^3 k \\ \nabla_{\mathbf{z}\mathbf{z}\mathbf{z}}^3 k & \nabla_{\mathbf{z}\mathbf{z}\mathbf{z}}^3 k \end{bmatrix}, \begin{bmatrix} \nabla_{\mathbf{z}\mathbf{z}\mathbf{z}}^3 k & \nabla_{\mathbf{z}\mathbf{z}\mathbf{z}}^3 k \\ \nabla_{\mathbf{z}\mathbf{z}\mathbf{z}}^3 k & \nabla_{\mathbf{z}\mathbf{z}\mathbf{z}}^3 k \end{bmatrix} \right\}. \quad (\text{A.36})$$

Therefore, our differential GP model, for predicting the second order force constants in both the Cartesian (3.7) and the phonon basis (3.19), can be rewritten to support the complex vectors as

Wirtinger GPFC2

$$\Phi^{(2)}(\mathbf{z}^*) = \begin{bmatrix} \tilde{D}_{\mathbf{z}^* \mathbf{z}^*}^2 k \\ \tilde{D}_{\mathbf{z}^* \mathbf{z}^* \mathbf{z}}^3 k \end{bmatrix}_{\mathbf{z}, \mathbf{z}^*}^T \cdot \begin{bmatrix} k & (\tilde{D}_{\mathbf{z}} k)^T \\ \tilde{D}_{\mathbf{z}} k & \tilde{D}_{\mathbf{z}^* \mathbf{z}^*}^2 k \end{bmatrix}_{\mathbf{z}, \mathbf{z}}^{-1} \cdot \begin{bmatrix} E \\ \nabla_{\mathbf{z}} E \\ \nabla_{\bar{\mathbf{z}}} E \end{bmatrix}_{\mathbf{z}}, \quad (\text{A.37})$$

where

$$\tilde{D}_{\mathbf{z}} k = \begin{bmatrix} \nabla_{\mathbf{z}} k & \nabla_{\bar{\mathbf{z}}} k \end{bmatrix}.$$

Here, $\nabla_{\mathbf{z}}E$ and $\nabla_{\bar{\mathbf{z}}}E$ correspond to the forces defined on complex vector $\mathbf{z}$. They can, particularly, be obtained using the transformation to transform Cartesian forces into the phonon coordinate at arbitrary $\mathbf{q}$ -wavevector.

Subsequent to the prediction, we will acquire the $2d \times 2d$ complex GPFC2

$$\Phi^{(2)}(\mathbf{z}^*) := \Phi^{(2)}(\mathbf{z}^*, \bar{\mathbf{z}}^*).$$

This corresponds to the 12×12 dynamical matrix for a $2 \times 2 \times 2$ supercell of a diatomic crystal structure, such as **d-Si**, **PbTe**, and **NaCl**, which is inconsistent. However, we can use the fact that the 6×6 off-diagonal block matrices are non-holomorphic (or physical). This is due to the fact that these off-diagonal block matrices depend on both $\mathbf{z}^*$ and $\bar{\mathbf{z}}^*$, while the diagonal block matrices depend only on $\mathbf{z}^*$ (or $\bar{\mathbf{z}}^*$). This $\mathbf{z}^*$ - (or $\bar{\mathbf{z}}^*$ -) dependence leads to (anti)holomorphism, preventing the definition of the Hermitian property. Therefore, the real (physical) dynamical matrix should be equivalent one of the 6×6 off-diagonal block matrices, with the other block being its Hermitian matrix.

In the case of the cubic anharmonic force constants, this Wirtinger derivative GP model requires the fourth-order derivative for the extension to support phonon-based GPFC3 prediction.

A.2 GPFC Results

This section provides additional results that support Chapter 3. Initially, we present a comparison of the phonon dispersion relations and phonon density of states for the three remaining materials: **NaCl**, **GaAs**, and **CdTe**. The subsequent data includes the associated phonon lifetimes and lattice thermal conductivities, assessed at 300 K.

To further evaluate the convergence of our Gaussian Process (GP) model, we present learning curves derived from the violation of the acoustic sum rule in the predicted force constants. This measure shows the model’s capacity to learn the translational symmetry of crystal structures during the training process. We, finally, compare our model performance with HIPHIVE [73], the force constant prediction model based on the cluster expansion method. The phonon properties are used as the standard of comparison.

A.2.1 Phonon Dispersion and Density of States

Here, I first present the phonon dispersion relations and phonon density of states for **NaCl**, **GaAs**, and **CdTe**, as shown in Fig. A.4. The dashed red lines indicate the results obtained using Gaussian Process-predicted harmonic force constants (GPFC2), while the solid green lines represent the reference results computed using second-order force constants (FC2) derived from the finite displacement method (FDM) as implemented in PHONOPY.

The GPFC2 results show excellent agreement with those from the standard FDM approach. The root mean square errors (RMSE) of the phonon band structures of **NaCl**, **GaAs**, and **CdTe** are 1.87%, 1.42%, and 2.01%, respectively. Wasserstein’s distance is here employed to measure the dissimilarity of two phonon density of states between different methods over a frequency ω region: we calculate these as 0.071, 0.037, and 0.021 $THz \cdot eV^{-1} \text{\AA}^{-3}$ respectively.

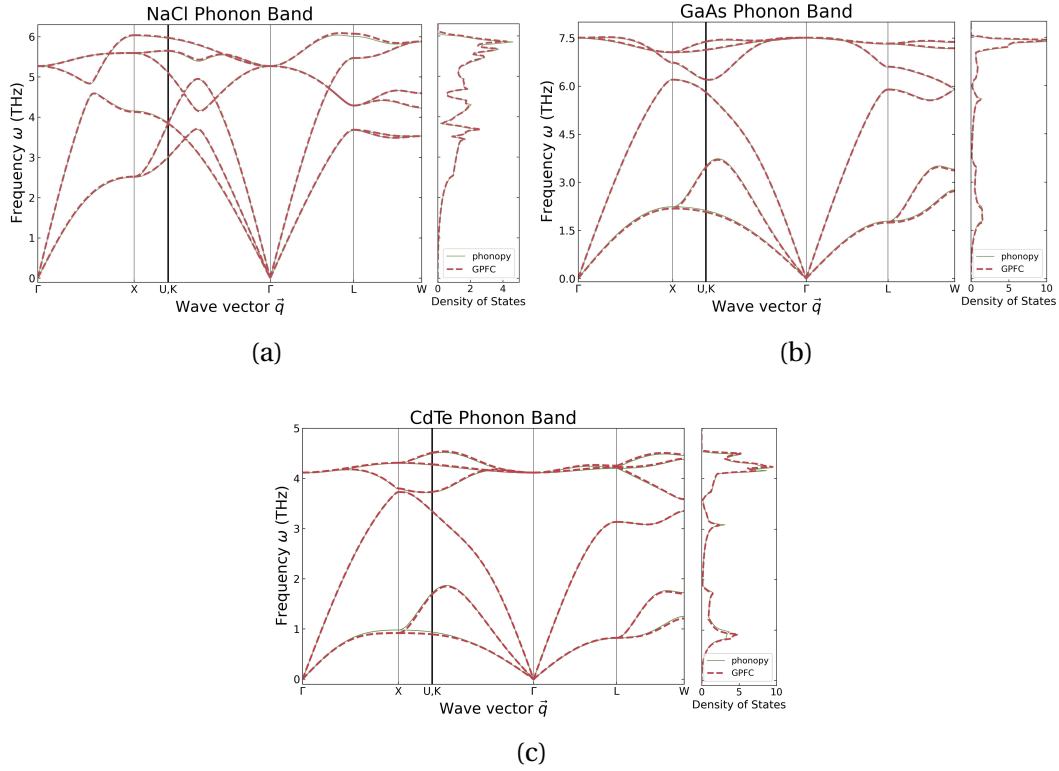


Figure A.4: Panels (a), (b) and (c) show the phonon band structures and density of states for **NaCl**, **GaAs** and **CdTe**, respectively.

A.2.2 Phonon Lifetime and Lattice Thermal Conductivity

We examine the phonon lifetime and lattice thermal conductivity of **NaCl**, **GaAs**, and **CdTe** at finite temperatures ranging from 300 to 1200 K. They are computed utilising a third-order many-body perturbation theory methodology, incorporating distinct cubic anharmonic force constants—one derived from FDM in PHONO3PY and the other from kernel regression in our GPFC.

The predicted phonon lifetimes of **NaCl**, **GaAs**, and **CdTe** at 300 K are presented in panels A.5 (a), (b), and (c) respectively, while their associated lattice thermal conductivity is shown in Figure A.6 (a), (b), and (c). EMD is again employed to measure the dissimilarity of the distribution of the predicted lifetime for each 30 K increment from 300 K to 1200 K. The average EMD are $3.12 \times 10^{-2} ps$ (NaCl), $2.99 \times 10^{-2} ps$ (GaAs), $8.07 \times 10^{-2} ps$ (CdTe).

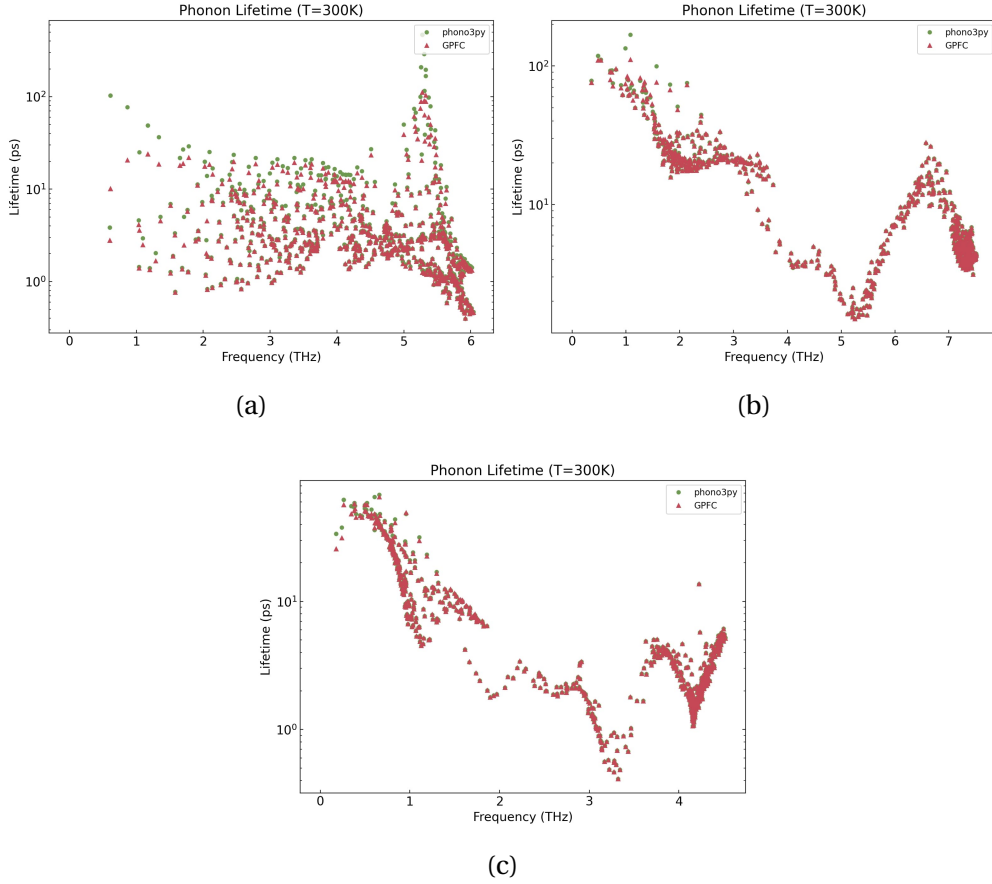


Figure A.5: Figures (a), (b), and (c) present the predicted phonon lifetimes at 300 K for **NaCl**, **GaAs**, and **CdTe**, respectively.

The mean absolute errors of the lattice thermal conductivity, computed using the GPFC3, are 4.33%, 3.56% and 3.77%, respectively.

A.2.3 Acoustic Sumrule

The acoustic sum rule [73; 84; 85] is a physical constraint that indicates whether the harmonic (second-order, FC2) and cubic anharmonic (third-order, FC3) force constants satisfy translational symmetry. The sumrule for FC2 is defined as

$$\sum_{\kappa'} \Phi^{(2)} \alpha \kappa, \beta \kappa' = 0, \quad (\text{A.38})$$

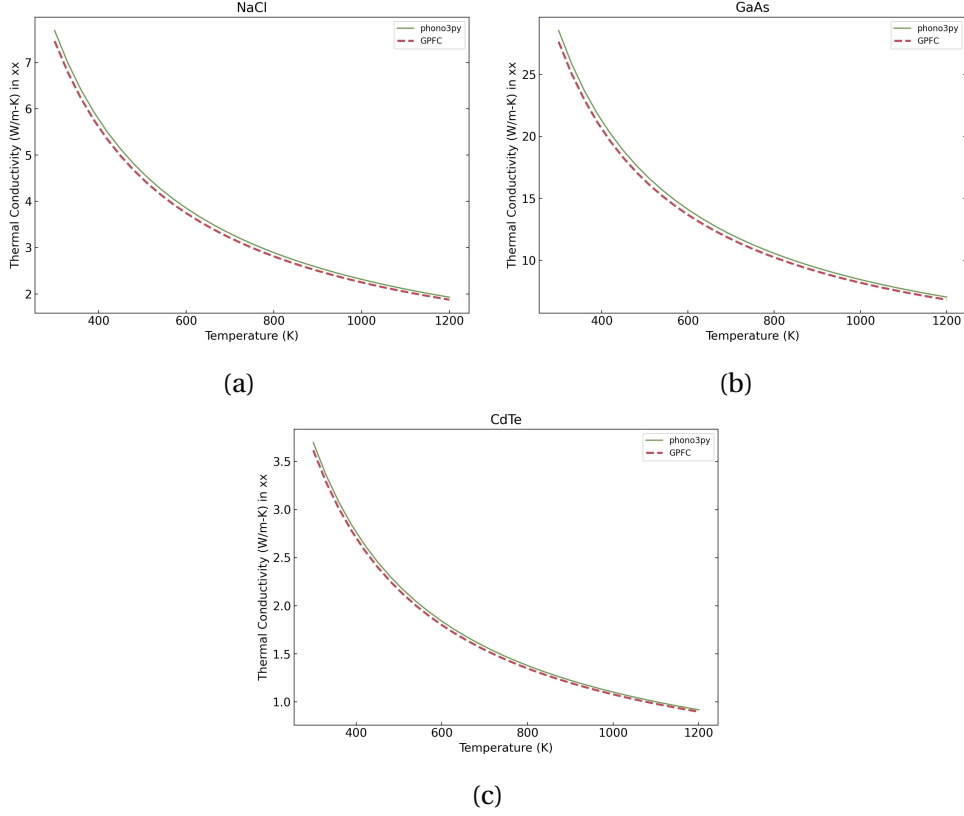


Figure A.6: Figures (a), (b), and (c) illustrate the lattice thermal conductivities of these materials across the temperature range of 300 K to 1200 K.

while the corresponding sum rule for FC3 is given by

$$\sum_{\kappa''} \Phi_{\alpha\kappa, \beta\kappa', \gamma\kappa''}^{(3)} = 0, \quad (\text{A.39})$$

where α , β , and γ denote Cartesian components of atoms κ , κ' , and κ'' , respectively.

Since translational symmetry is not explicitly imposed in our GP model, analysing the learning curves associated with violations of the acoustic sum rule offers a meaningful way to assess the model's capability to learn this physical constraint. These learning curves provide insight into how well the

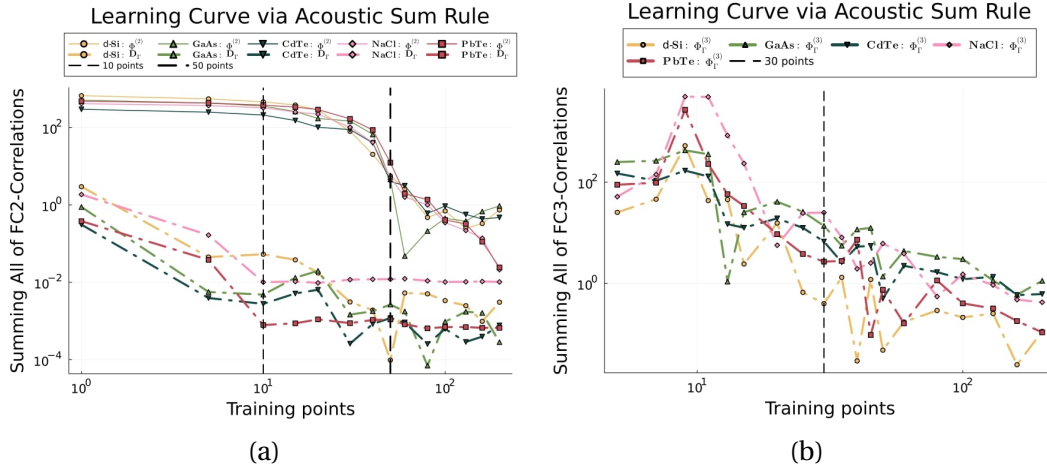


Figure A.7: Learning curves showing violations of the acoustic sum rule for GP-predicted force constants (GPFCs) across various materials. (a) Translational symmetry violations in GPFC2, quantified by the total magnitude of the residuals in Eq. (A.38), as a function of the number of training points. (b) Translational symmetry violations in GPFC3, computed from the residuals in Eq. (A.39), using phonon-based descriptors.

GP model captures the underlying symmetry as a function of the training data size or model complexity.

The learning curves are illustrated in Figure A.7. In panel (a), solid lines represent predictions based on Cartesian descriptors, whereas dashed lines correspond to those based on phonon descriptors. The learning curves converge within approximately 50 and 10 training points, respectively, which is consistent with the RMSEs of the GPFC2s shown in Figures 3.3a and 3.8a for Cartesian and phonon descriptors. Phonon-based GPFC3s also satisfy the acoustic sum rule around 30–50 training points, in agreement with the RMSE convergence observed in Figure 3.8b.

In contrast, the Cartesian-based GPFC3s do not appear to preserve translational symmetry. This may be attributed to the inherent noise in the Gaussian Process model, which can lead to small but non-negligible violations of symmetry. During training, this noise can perturb trivial (symmetry-

forbidden) off-diagonal components of the GPFC3 tensor, causing them to deviate from zero.

A.2.4 HiPHIVE

In this final section, we compare the phonon properties calculated using our GP-predicted second-order force constants (GPFC2) with those obtained using second-order force constants predicted by the HiPHIVE package.

HiPHIVE [73] is a recent machine learning model designed to extract lattice force constants from electronic structure data. The model is based on the cluster expansion formalism combined with compressed sensing techniques. As such, it provides a suitable benchmark for evaluating the performance of our GP-based model against an established machine learning method.

To evaluate the force constants using HiPHIVE, we first construct a training dataset. For consistency, we use electronic structure calculations with settings identical to those described in Section 3.1.2. These calculations are carried out on $2 \times 2 \times 2$, $3 \times 3 \times 3$, and $4 \times 4 \times 4$ supercells of **d-Si**, **NaCl**, **GaAs**, **CdTe**, and **PbTe**, using the same random displacement sampling employed in generating the GPFC training data.

The choice of supercell size directly determines the maximum allowable cutoff distance for clusters included in the expansion. For example, $2 \times 2 \times 2$, $3 \times 3 \times 3$, and $4 \times 4 \times 4$ supercells of the **d-Si** structure permit maximum cutoff distances of approximately 3Å, 5Å, and 7 Å, respectively, for two-body clusters, which are required for constructing the FC2 model. Consequently, these configurations yield 6, 30, and 72 two-body representative clusters per structure, which are then used to train the HiPHIVE model.

The comparison results are illustrated through the phonon dispersion relations and phonon density of states (DOS) for **d-Si**, as shown in Figure A.8. Solid green and dashed red lines represent results computed using second-

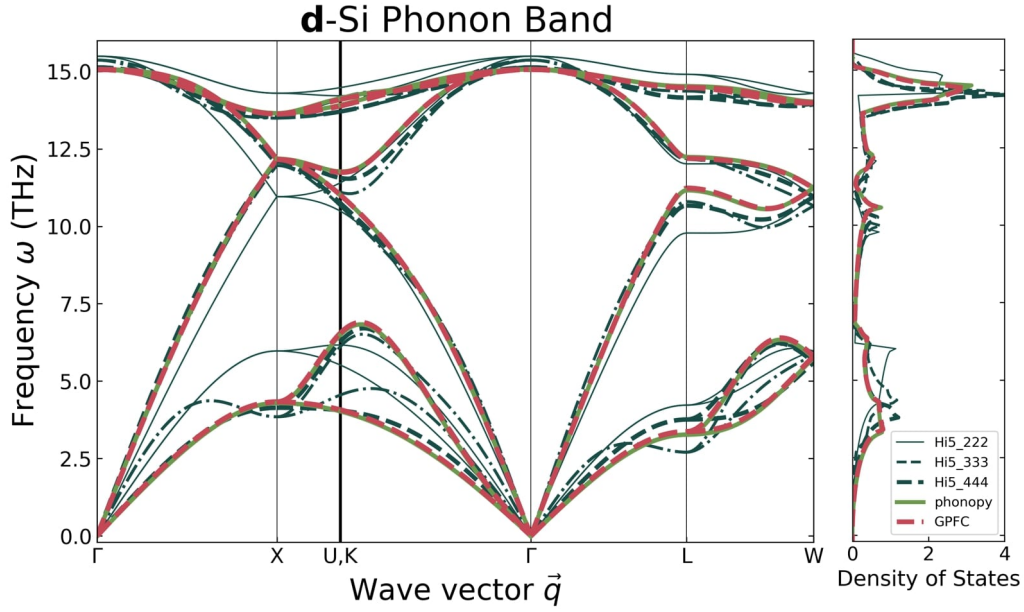


Figure A.8: Figure shows the phonon band structures and density of states for d-Si.

order force constants (FC2) obtained from PHONOPY (via the finite displacement method, FDM) and our GPFC2 predictions, respectively—similar to those shown in Figure 3.4a. The additional solid, dash-dotted, and dashed blue lines correspond to phonon properties calculated using FC2 predicted by HiPHIVE, trained on $2 \times 2 \times 2$, $3 \times 3 \times 3$, and $4 \times 4 \times 4$ supercell structures, respectively.

For the HiPHIVE results, we observe that increasing the supercell size improves model accuracy. The RMSEs of the phonon band structures, compared to the FDM reference, are 6.85%, 3.97%, and 2.26% for the $2 \times 2 \times 2$, $3 \times 3 \times 3$, and $4 \times 4 \times 4$ supercells, respectively.

We employ the Wasserstein distance (also known as the Earth Mover’s Distance, EMD) to quantify the dissimilarity between the phonon DOS calculated using FDM-based FC2 and those obtained from HiPHIVE. The resulting EMD values are 0.469, 0.376, and 0.200 $\text{THz} \cdot \text{eV}^{-1} \text{\AA}^{-3}$ for the increasing

supercell sizes, respectively, further confirming the improved agreement with the larger number of the two-body representative clusters.

Moreover, Zhu et al. [86] propose that employing an 8-10 Å cutoff radius, acceptable within a $6 \times 6 \times 6$ supercell of the d-Si structure, is appropriate for predicting the FC2 and the FC3. Therefore, to further substantiate this observation, it would be necessary to train the HIPHIVE model using larger supercell structures, extend the comparison to include the remaining materials, or even explore higher-order force constants—for example, by employing the predicted FC3 to compute phonon lifetimes and lattice thermal conductivity.

A.3 Fine-tuning ACE Transfer Integrals

This section outlines the hyperparameters, including interaction-order, `totaldegree`, `rcut`, `p`, and `q`, involved in the fine-tuning of the Atomic Cluster Expansion (ACE) model for predicting transfer integrals (electronic couplings) in a naphthalene dimer.

All evaluations are based on ACE models trained on 900 dimer configurations, represented by green markers in the subsequent plots. Model performance is tested on an independent set of 3,853 configurations, denoted by yellow markers. This setup enables a detailed investigation of not only model accuracy but also the risk of overfitting.

A.3.1 Interaction-order

We begin by examining the ACE fitting performance with respect to variations in two hyperparameters: the interaction-order and the `totaldegree`, for both bead and Heavy-atom representations, as depicted in Fig. 4.5. Varying these hyperparameters affects the size of the model basis, as illustrated in Figure 4.11. It is therefore important to assess not only the model's pre-

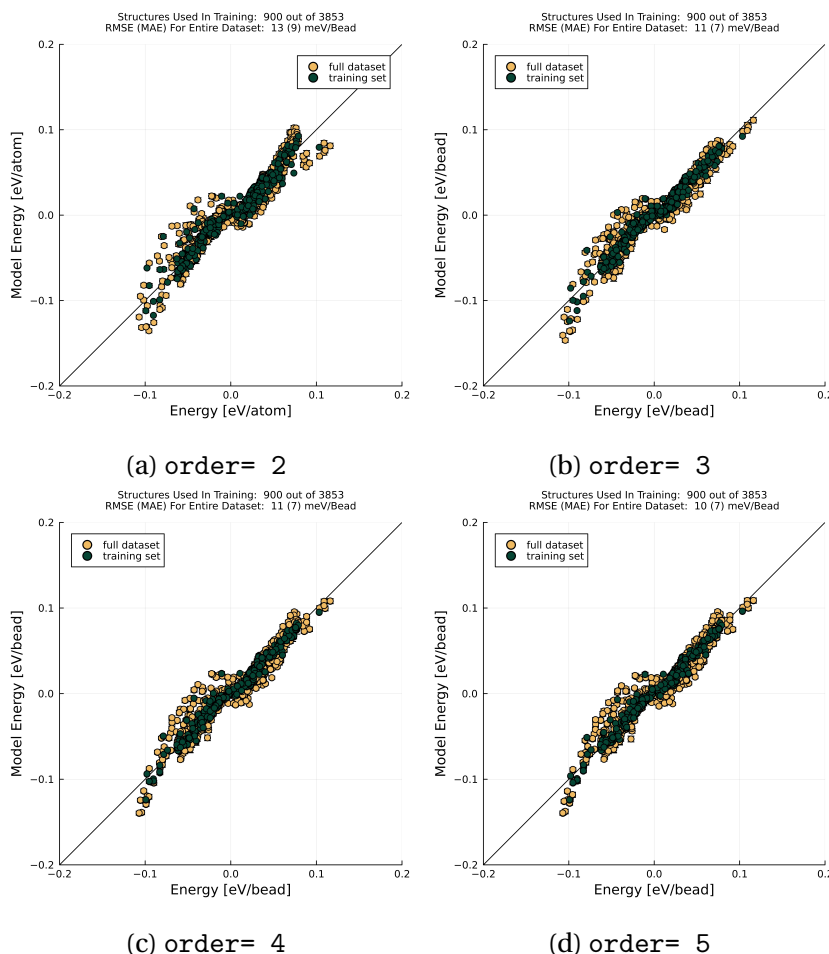


Figure A.9: Panel (a), (b), (c) and (d) show the model trained using the *Bead* representation parameterising with order = 2, 3, 4 and 5, respectively.

dictive accuracy, but also its fitting behaviour, particularly with regard to potential overfitting.

The trade-off is clearly illustrated in Figure A.9, which shows that an increase in the interaction-order enhances the fitting performance of the bead-based model. The Heavy-atom-based model demonstrates a minimal improvement in predictive accuracy at zero-valued transfer integrals, yet it also reveals obvious signs of overfitting, as illustrated in Figure A.10. The convergence behaviour of RMSE and MAE, as illustrated in Figure 4.10a, indicates that increasing the interaction-order beyond 3 results in diminish-

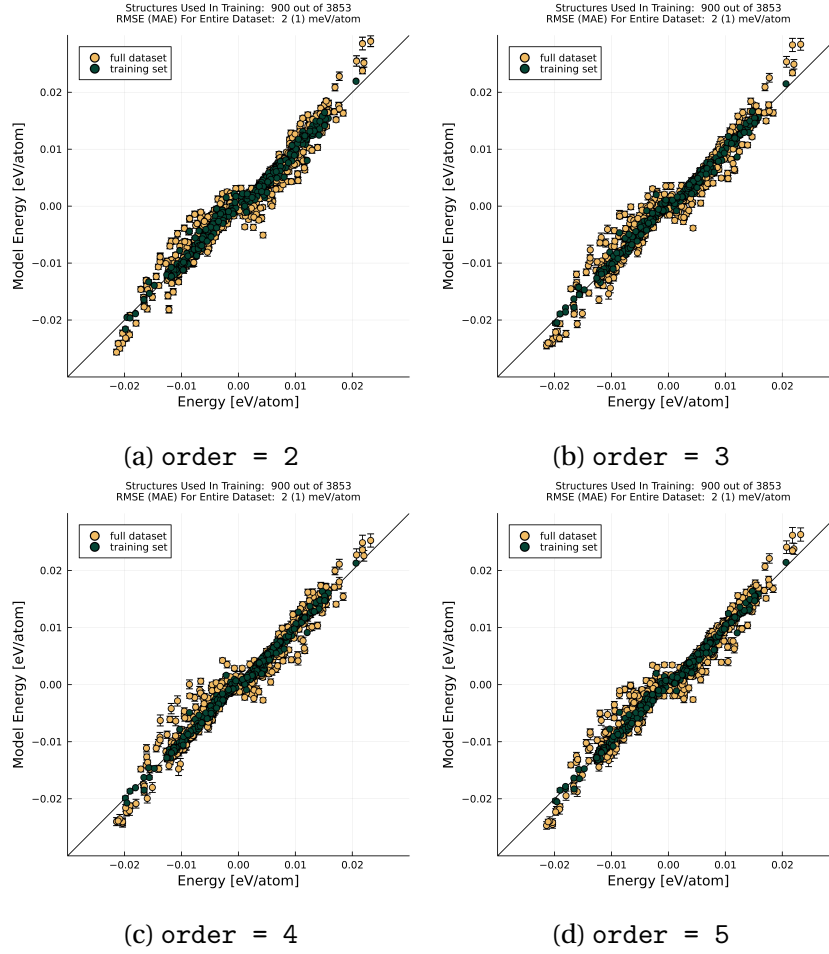


Figure A.10: Panel (a), (b), (c) and (d) show the model trained using the *Heavy-atom* representation parameterising with order = 2, 3, 4 and 5, respectively.

ing returns. The improvements in accuracy are not sufficient to warrant the increased computational cost and memory needed to assess the expanded model basis.

A.3.2 totaldegree

With similar investigation, Figures A.11 and A.12 show ACE fitting results for the bead-based and Heavy-atom-based models, respectively, by varying the totaldegree hyperparameter. These hyperparameter typically controls

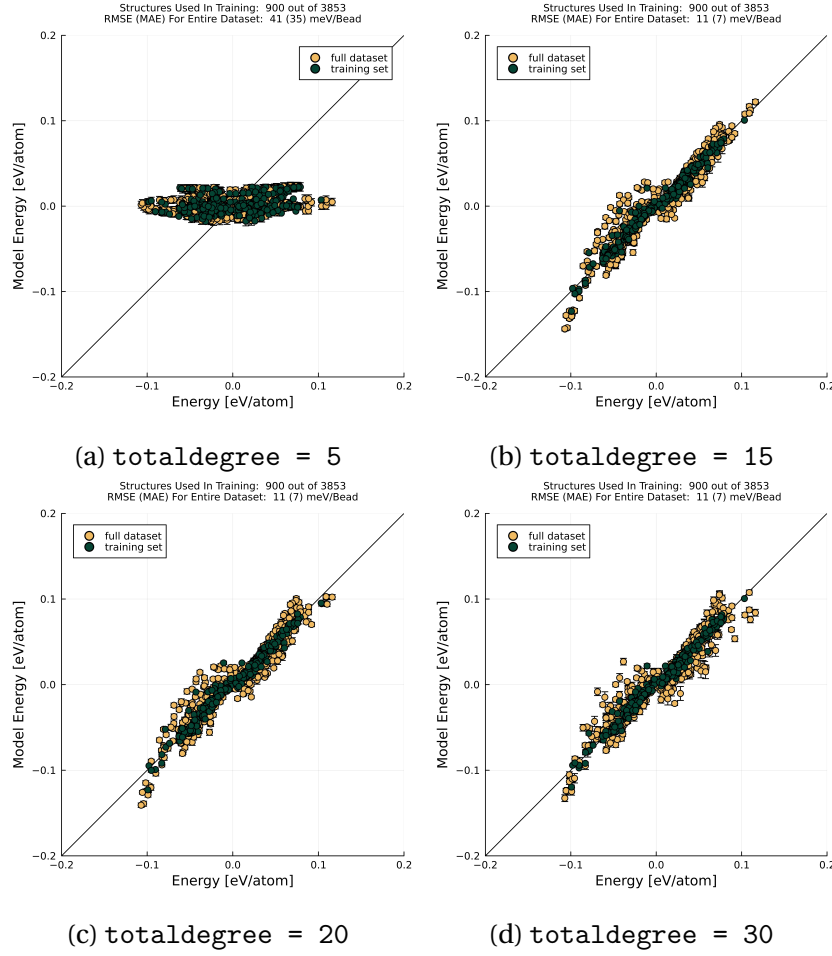


Figure A.11: Panel (a), (b), (c) and (d) show the model trained using the *Bead* representation parameterising with totaldegree = 5, 15, 20 and 30, respectively.

the maximum degree of the polynomial basis. The totaldegree values are set to 5 (a), 15 (b), 20 (c), and 30 (d). As shown in the figures, it is evident that increasing totaldegree beyond 15 yields diminishing returns in terms of fitting accuracy. This observation is consistent with the convergence behavior of the RMSEs and MAEs shown in Figure 4.10b.

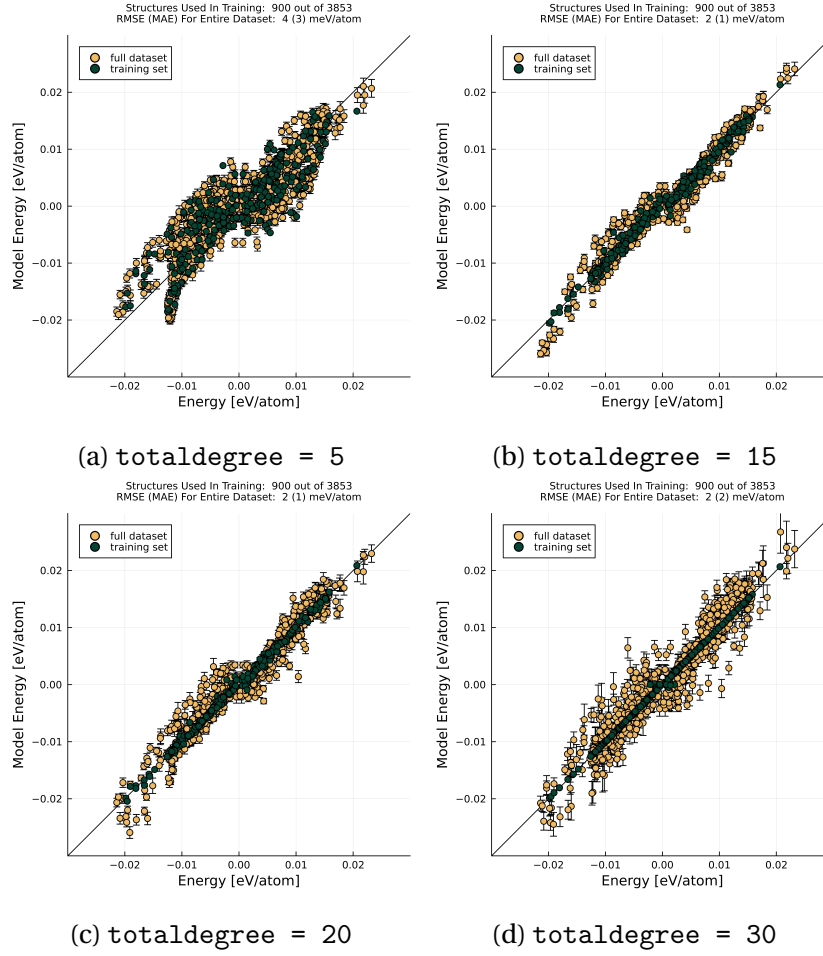


Figure A.12: Panel (a), (b), (c) and (d) show the model trained using the Heavy-atom representation parameterising with `totaldegree` = 5, 15, 20 and 30, respectively.

A.3.3 `rcut`

In this section, we now examine how variations in the cutoff radius (`rcut`) hyperparameter affect the ACE model's performance. The bead-based model, in particular, exhibits strong sensitivity to changes in the cutoff radius, as shown in Figure A.13. The predictive capability of the model is confined to a narrow range of `rcuts`, i.e., between 13.3 and 13.9 Å. The predictions decrease outside of this range, showing a horizontal distribution pattern that

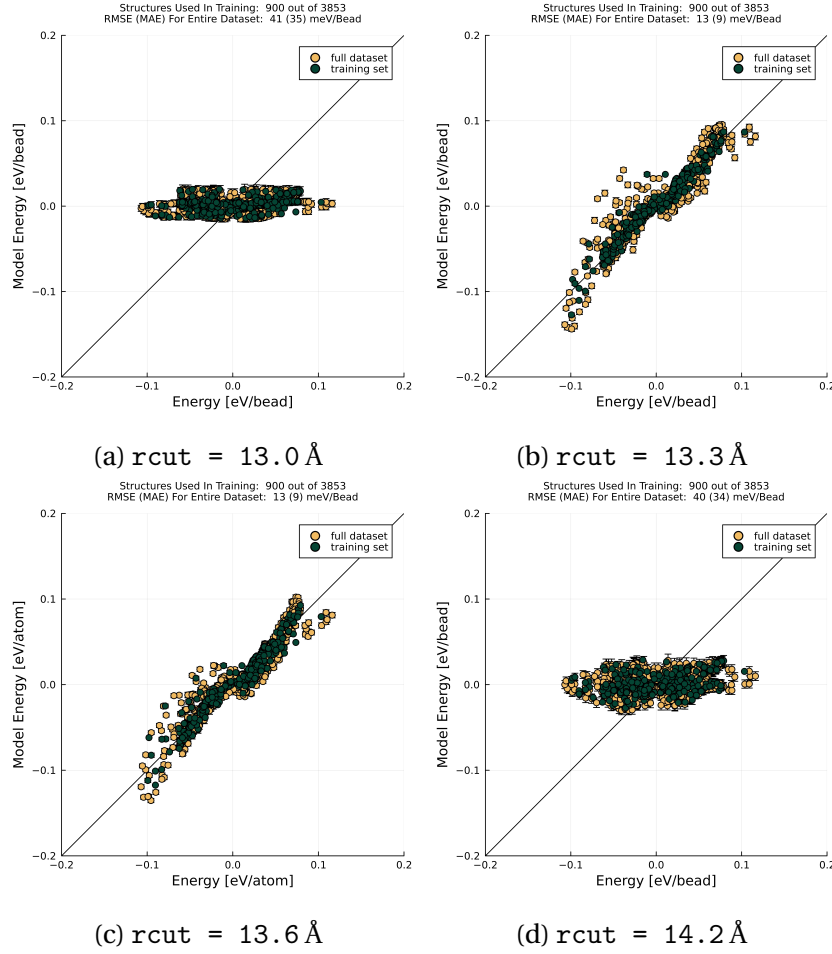


Figure A.13: Panel (a), (b), (c) and (d) show the model trained using the *Bead* representation parameterising with $r_{\text{cut}} = 13.0 \text{ \AA}$, $13.3 \text{ \AA}$, $13.6 \text{ \AA}$, and $14.2 \text{ \AA}$, respectively.

suggests the model is unable to discriminate between various geometric configurations.

This can be confirmed by examining the first panel of Figure A.14. Selecting a cutoff radius of $13.3 \text{ \AA}$ effectively avoids including the second peak of the RDFs of the bead training set, where the region around the peak encompasses a variety of distinct geometric configurations. Meanwhile, increasing the cutoff beyond $13.9 \text{ \AA}$ captures the right tail of the RDF, which includes atomic environments with a wide range of relative positions and orienta-

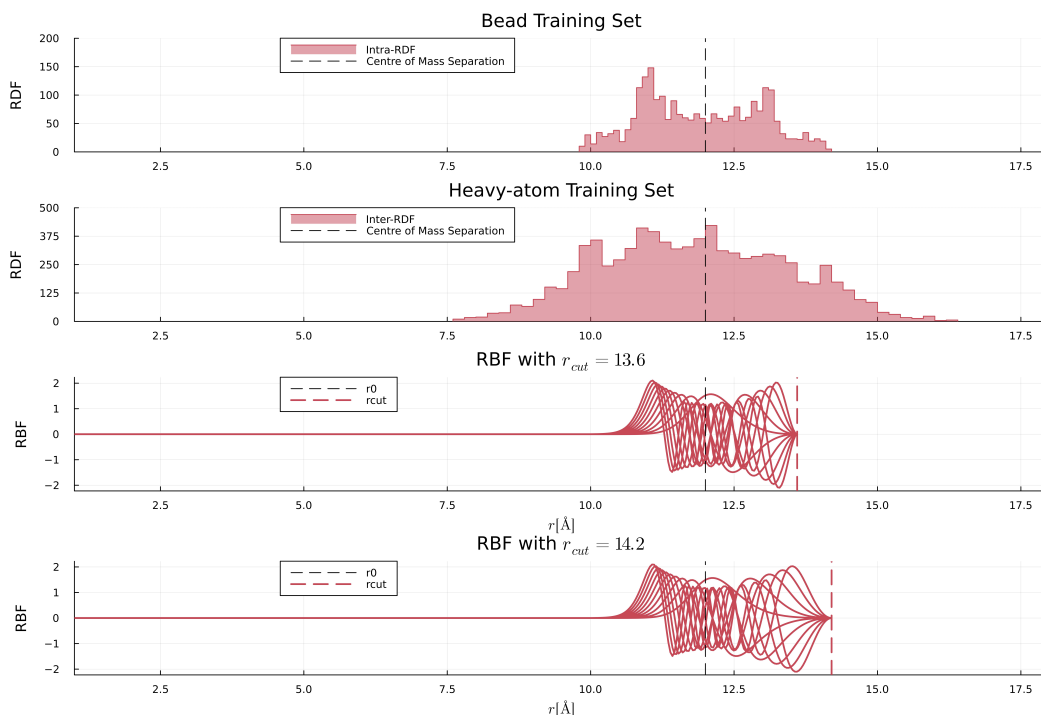


Figure A.14: Radial distribution functions (RDFs) and radial basis functions (RBFs) used in the ACE descriptor construction for the bead and Heavy-atom representation.

tions. These long-range configurations are structurally diverse—meaning they differ significantly in geometry—and are only coarsely represented by the model basis. This constrained resolution probably heightens predictive uncertainty and results in diminished performance. Unlike the sensitivity observed in the bead-based model, modifications to the r_{cut} parameter do not significantly affect the performance of the Heavy-atom-based model. The accuracy of predictions improves, particularly for transfer integrals around zero as the cutoff radius increases. This enhancement is ascribed to the Coulombic characteristics of the transfer integrals, which diminish inversely with distance. Furthermore, as shown in the second panel of Figure A.15, the distribution of interatomic distances in the Heavy-atom training set is more uniform at larger separations (i.e., in the right tail of the distribution), in contrast to the more structured radial distribution function

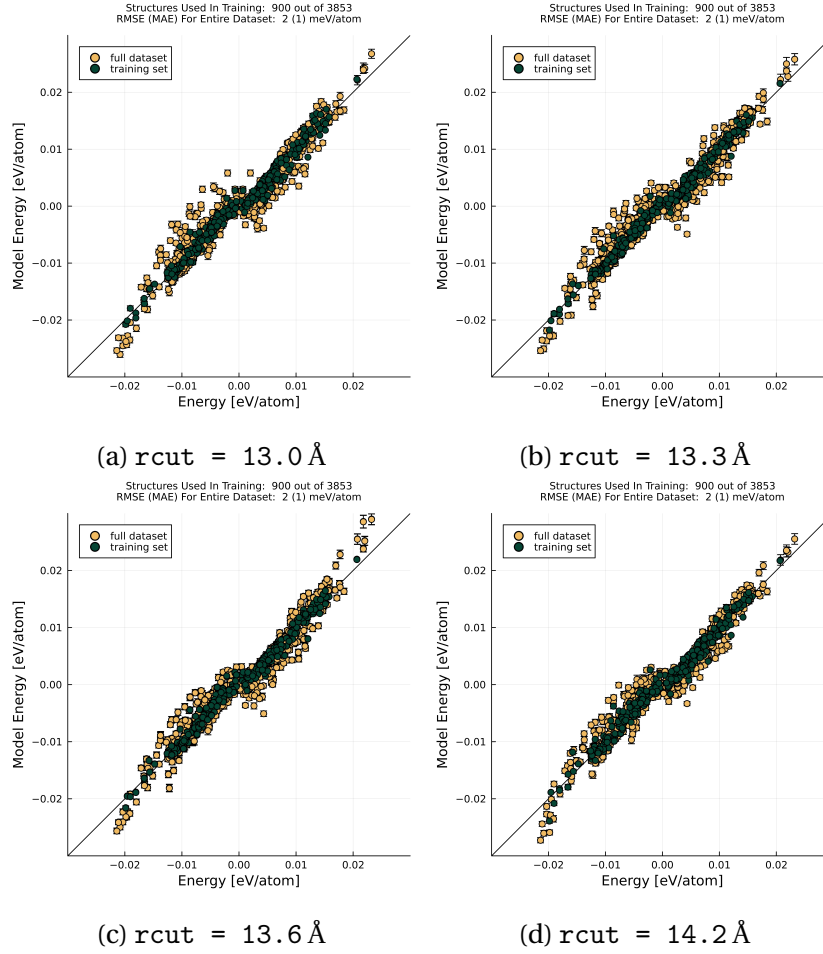


Figure A.15: Panel (a), (b), (c) and (d) show the model trained using the Heavy-atom representation parameterising with $rcut = 13.0 \text{ \AA}$, $13.3 \text{ \AA}$, $13.6 \text{ \AA}$, and $14.2 \text{ \AA}$, respectively.

observed for the bead-based dataset. This uniformity facilitates learning of long-range interactions without introducing ambiguity or overfitting.

A.3.4 q parameter

As mentioned in the theory chapter (Section 2.5.5), the parameter q specifically controls the shape of the left side of the radial basis functions (RBFs). The RBFs are extended further toward shorter interatomic distances with a

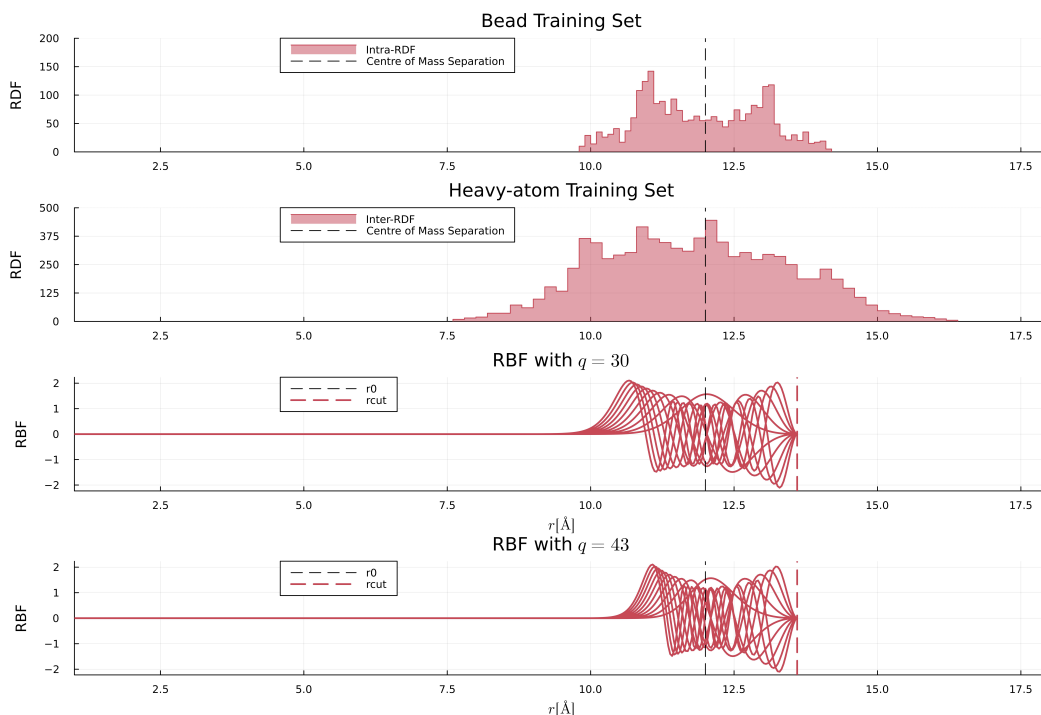


Figure A.16: Radial distribution functions (RDFs) and radial basis functions (RBFs) used in the ACE descriptor construction for the bead and Heavy-atom representation.

smaller value of q , while a larger value contracts the basis functions closer to the equilibrium bond length r_0 . These behaviours are illustrated in the third and fourth panels of Figure A.16.

This adjustment of the q parameter has a significant impact on the ACE model's fitting performance for the bead-based training set. An excessively small q value expands the RBFs too far into the short-range domain, thereby capturing structurally diverse configurations. This is similar to the earlier discussion on the effects of modifying the cutoff radius (r_{cut}). This overrepresentation of complex local environments may lead to unstable behaviour in the model, as reflected in the unphysical horizontal prediction patterns observed when q is set outside the optimal range of 41 to 43 (see Figure A.17). In contrast, in Figure A.18, for the Heavy-atom-based model, adjusting q within the range of 39 to 45 does not significantly affect prediction perfor-

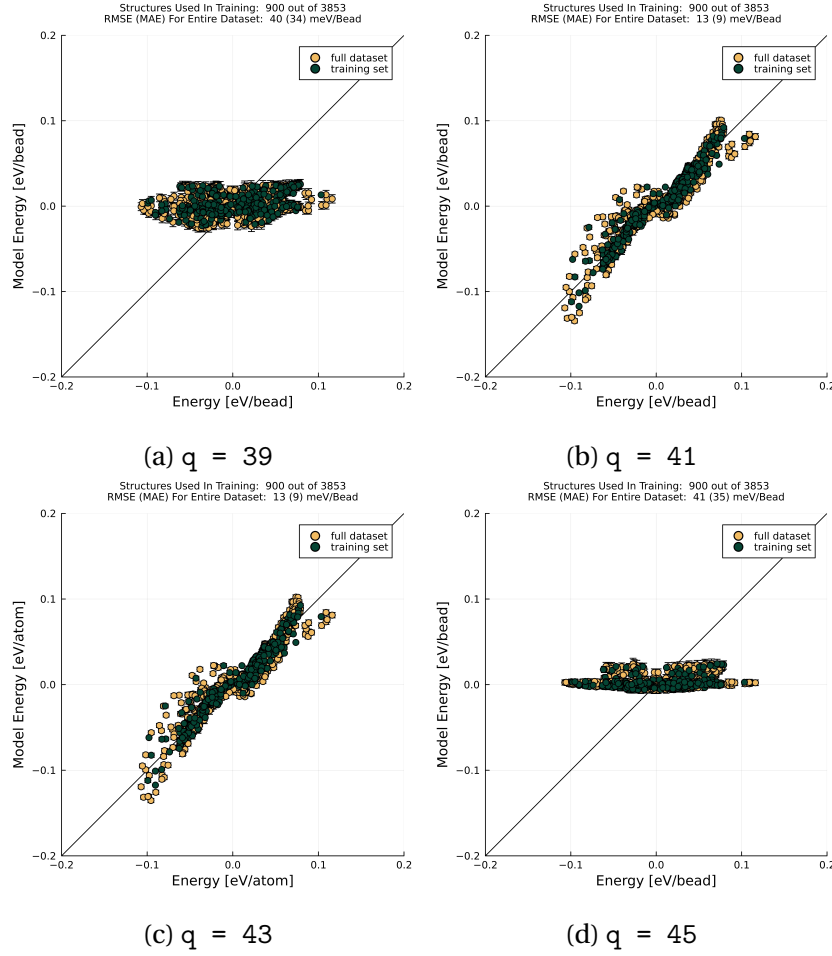


Figure A.17: Panel (a), (b), (c) and (d) show the model trained using the *Bead* representation parameterising with $q = 39, 41, 43$, and 45 , respectively.

mance. However, we hypothesise that performance in this case could be improved by choosing a sufficiently small q value to allow the RBFs to better capture short-range interactions associated with large-magnitude transfer integrals. This possibility is supported by the left-side radial distribution of the Heavy-atom dataset, shown in the second panel of Figure A.16.

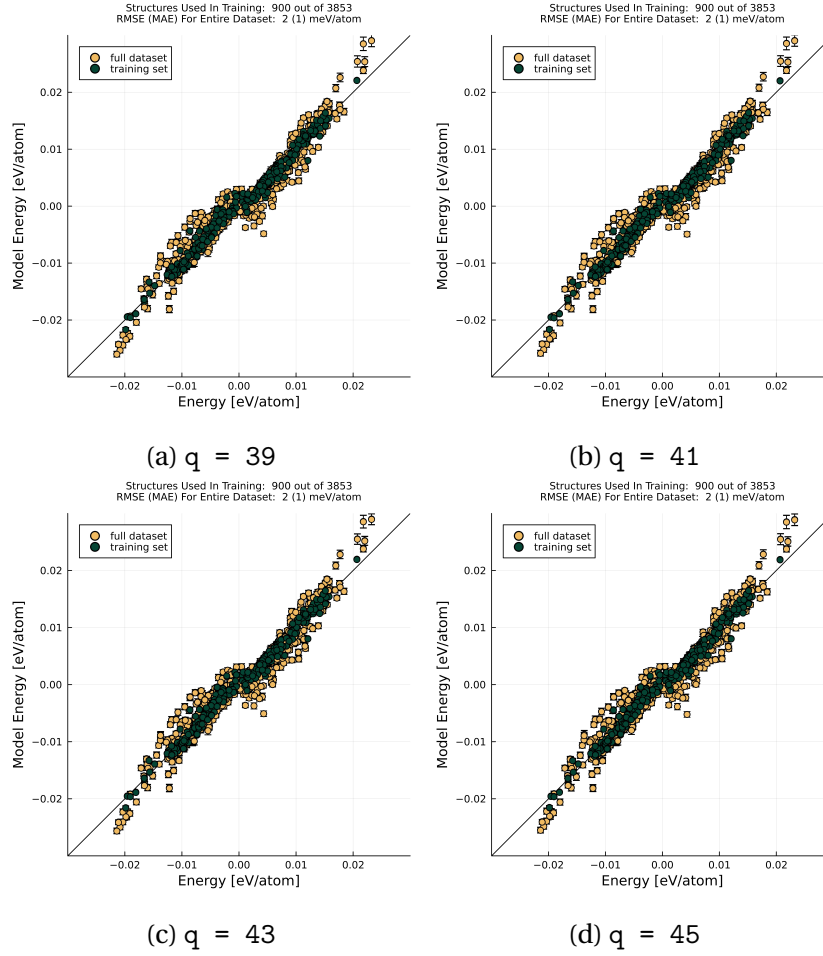


Figure A.18: Panel (a), (b), (c) and (d) show the model trained using the *Heavy-atom* representation parameterising with $q = 39, 41, 43$, and 45 , respectively.

A.3.5 p parameter

This final section evaluates the ACE model's fitting performance with respect to variations in the fine-tunable hyperparameter p . Unlike q , which controls the left side of the radial basis function (RBF) envelope, the parameter p governs the sharpness and localization of the envelope near the equilibrium bond length r_0 , particularly on the right-hand side. A reduced p results in a more gradual ascent of the envelope function near r_{cut} , permitting the RBFs to oscillate further from r_0 . Concurrently, an increased p yields a more

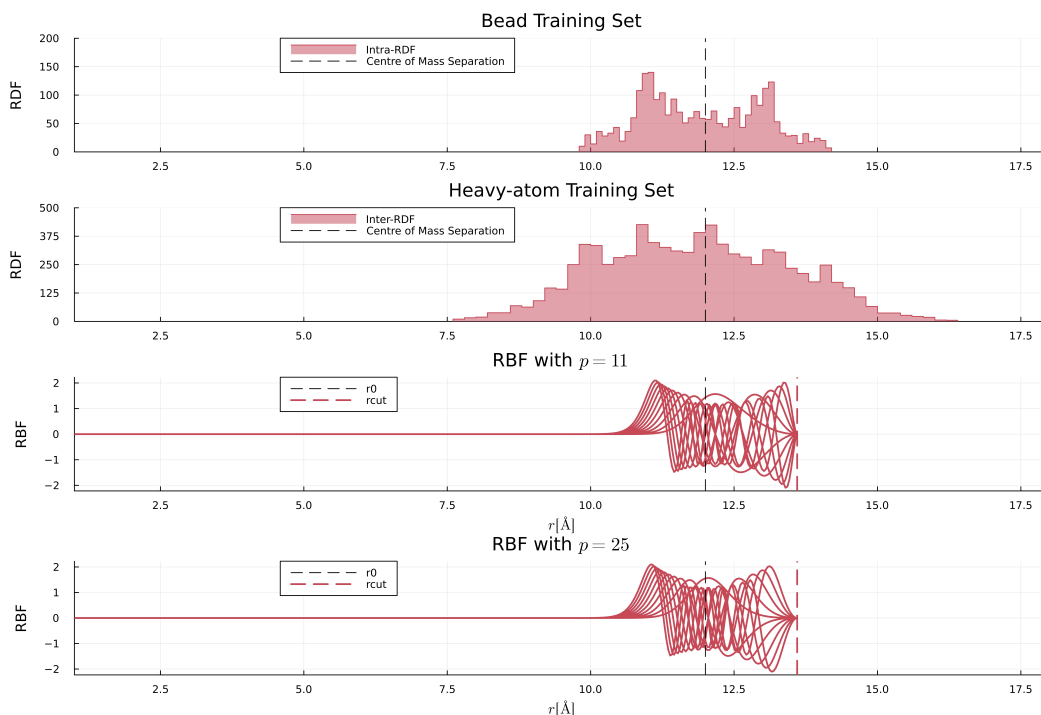


Figure A.19: Radial distribution functions (RDFs) and radial basis functions (RBFs) used in the ACE descriptor construction for the bead and Heavy-atom representation.

acute transition and heightened localisation around r_0 . The alteration in envelope shape is depicted in the third and fourth panels of Figure A.19.

As demonstrated in Figure A.20, modifying p does not seem to influence the model's prediction effectiveness, because of the closeness of r_0 and r_{cut} in the model formulation. Consequently, the modification of p minimally changes the resolution of the RBFs inside the constrained interval between r_0 and r_{cut} , thus preserving the RBFs' ability to obtain structural information.

In the Heavy-atom-based model, the change of p similarly exhibits minimal impact on predicted accuracy, as illustrated in Figure A.21. To achieve a more significant effect, it may be essential to increase the cutoff radius r_{cut} to a value further from the equilibrium bond length. This would expand the

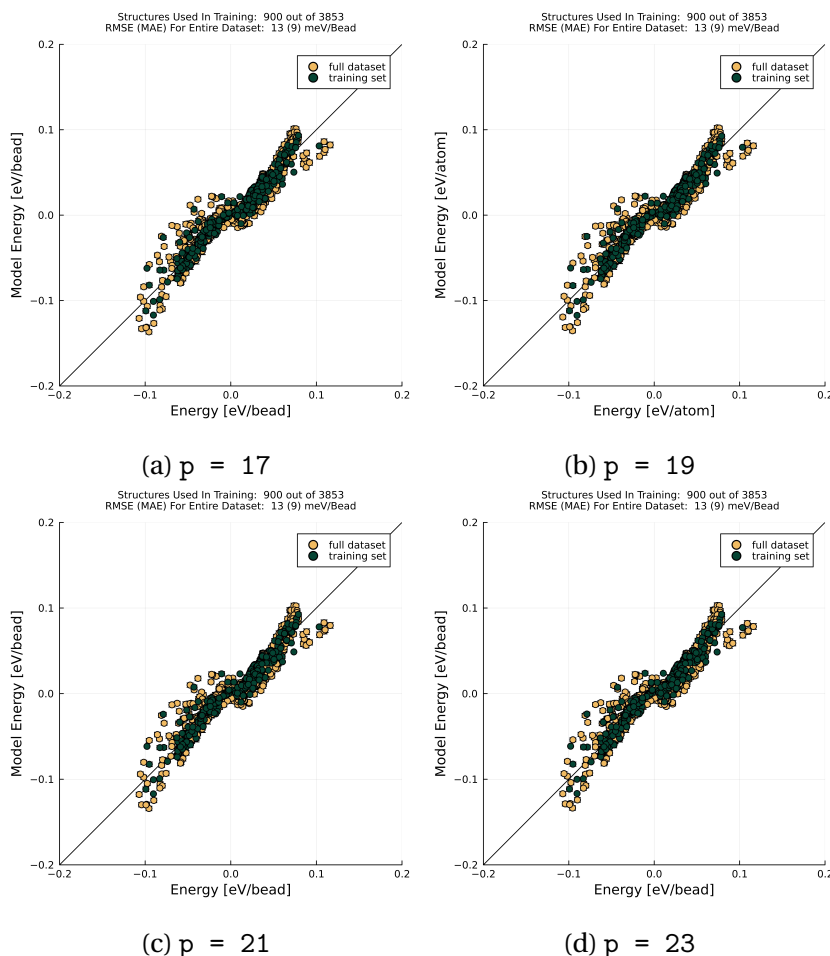


Figure A.20: Panel (a), (b), (c) and (d) show the model trained using the *Bead* representation parameterising with $p = 17, 19, 21$, and 23 , respectively.

radial domain, so enhancing the impact of p on the resolution of the RBFs and perhaps influencing the model's performance.

To sum up, this section examines the optimisation of the Atomic Cluster Expansion (ACE) model for predicting electronic transfer integrals in molecular dimers, using naphthalene as a model system. The investigation employs models trained on 900 configurations and evaluated on an independent set of 3,853 configurations. This configuration enables the assessment of accuracy and generalisation, specifically aimed at minimis-

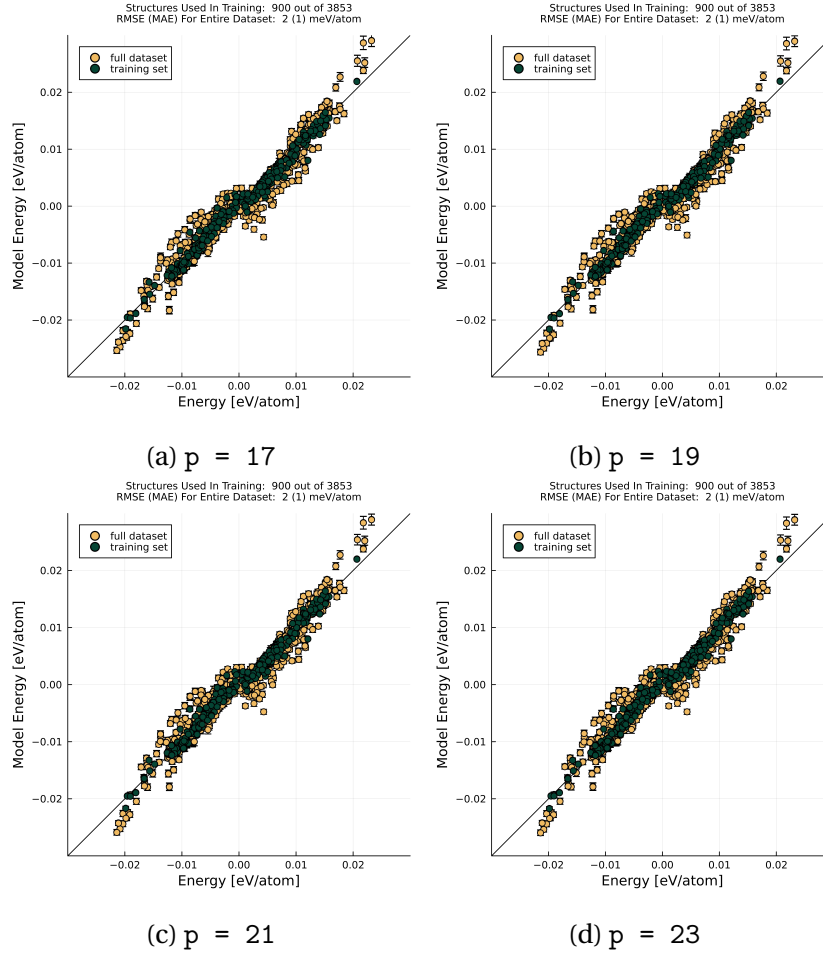


Figure A.21: Panel (a), (b), (c) and (d) show the model trained using the *Heavy-atom* representation parameterising with $p = 17, 19, 21$, and 23 , respectively.

ing overfitting. Five major hyperparameters are analysed comprehensively: interaction-order, totaldegree, cutoff radius (r_{cut}), and the envelope parameters q and p .

The interaction-order defines the complexity of the many-body interactions represented in the ACE basis. Increasing the interaction order in the bead representation enhances fitting performance. However, in the Heavy-atom representation, improvements are constrained to nearly negligible transfer integrals, and higher orders result in overfitting. Analyses of

RMSE and MAE indicate that increasing the interaction order beyond three provides negligible advantages compared to the raised computational cost. The `totaldegree` parameter establishes the maximum value of the polynomial base degree. Raising `totaldegree` to 15 can enhance the performance of both bead-based and Heavy-atom-based models, beyond which improvements are negligible. This can be confirmed by the convergence patterns in RMSE and MAE, indicating that elevated polynomial degrees add unneeded complexity without substantial improvements in accuracy.

The effect of `rcut` is significant in bead-based models. Optimal performance occurs within a limited range of 13.3 to 13.9 Å. Beyond this range, particularly at higher cutoffs, the model includes geometrically varied configurations that the current basis cannot sufficiently resolve, resulting in diminished predictive accuracy. This sensitivity is exemplified by changes in the radial distribution function (RDF). The Heavy-atom representation exhibits more resilience to fluctuations in `rcut`. As the cutoff grows, accuracy enhances—especially for small transfer integrals—due to the model’s improved capture of long-range Coulombic interactions. The Heavy-atom dataset demonstrates a more consistent RDF at extended distances, hence promoting accurate learning of long-range effects.

The `q` parameter governs the left side of the radial basis function envelope, influencing the extent of short-range structure integration. In bead-based models, a smaller `q` extends the radial basis functions (RBFs) over short distances, hence encapsulating highly variable local environments, which frequently results in instability or overfitting. Optimal outcomes are attained within the range `q` = 41–43. The Heavy-atom representation exhibits reduced sensitivity to this parameter; yet, the model may gain from minor `q` values to enhance the resolution of high-magnitude transfer integrals resulting from short-range interactions.

Finally, the `p` parameter affects the accuracy and localisation of the RBF envelope in proximity to the equilibrium bond length `r0`, especially on

the right side. Modifying p does not substantially affect model accuracy in either representation, primarily due to the closeness of r_0 and r_{cut} values. This restricted domain constrains the impact of changing p , while the RBF resolution stays fundamentally unaltered. A larger r_{cut} is necessary to expand the radial domain and allow p to significantly impact the envelope.

B

Bibliography

- [1] C.E. Rasmussen and C.K.I. Williams, *Gaussian processes for machine learning*, Adaptive computation and machine learning, MIT Press (2006).
- [2] R. Drautz, *Atomic cluster expansion for accurate and transferable interatomic potentials*, *Physical Review B* **99** (2019) .
- [3] V.L. Deringer, A.P. Bartók, N. Bernstein, D.M. Wilkins, M. Ceriotti and G. Csányi, *Gaussian process regression for materials and molecules*, *Chemical Reviews* **121** (2021) 10073–10141.
- [4] F. Musil, A. Grisafi, A.P. Bartók, C. Ortner, G. Csányi and M. Ceriotti, *Physics-inspired structural representations for molecules and materials*, *Chemical Reviews* **121** (2021) 9759–9815.
- [5] E. Solak, R. Murray-Smith, W.E. Leithead, D.J. Leith and C.E. Rasmussen, *Derivative observations in gaussian process models of dynamic systems*, in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, (Cambridge, MA, USA), p. 1057–1064, MIT Press, 2002.
- [6] A. McHutchon, “Differentiating gaussian processes.” <http://mlg.eng.cam.ac.uk/mchutchon/DifferentiatingGPs.pdf>, 2013.

- [7] E.G. del Río, J.J. Mortensen and K.W. Jacobsen, *Local bayesian optimizer for atomic structures*, *Physical Review B* **100** (2019) .
- [8] S. Kaappa, E.G. del Río and K.W. Jacobsen, *Global optimization of atomic structures with gradient-enhanced gaussian process regression*, *Physical Review B* **103** (2021) .
- [9] K. Keeratikarn and J.M. Frost, *Anharmonic phonons with gaussian processes*, 2024. 10.48550/ARXIV.2405.05113.
- [10] K. Keeratikarn and J.M. Frost.
<https://github.com/Frost-group/GPFC.jl>, 2021–2024.
- [11] A. Togo and I. Tanaka, *First principles phonon calculations in materials science*, *Scripta Materialia* **108** (2015) 1–5.
- [12] A. Togo, L. Chaput and I. Tanaka, *Distributions of phonon lifetimes in brillouin zones*, *Phys. Rev. B Condens. Matter Mater. Phys.* **91** (2015) .
- [13] A. Togo, L. Chaput, T. Tadano and I. Tanaka, *Implementation strategies in phonopy and phono3py*, *J. Phys. Condens. Matter* **35** (2023) .
- [14] A. Togo, *First-principles phonon calculations with phonopy and phono3py*, *J. Phys. Soc. Jpn.* **92** (2023) 012001.
- [15] W.C. Witt, C. van der Oord, E. Gelžinytė, T. Järvinen, A. Ross, J.P. Darby et al., *Acepotentials.jl: A julia implementation of the atomic cluster expansion*, *The Journal of Chemical Physics* **159** (2023) .
- [16] G. Dusson, M. Bachmayr, G. Csányi, R. Drautz, S. Etter, C. van der Oord et al., *Atomic cluster expansion: Completeness, efficiency and stability*, *Journal of Computational Physics* **454** (2022) 110946.
- [17] A.P. Bartók and G. Csányi, *Gaussian approximation potentials: A brief tutorial introduction*, *International Journal of Quantum Chemistry* **115** (2015) 1051–1057.

-
- [18] A.P. Bartók, S. De, C. Poelking, N. Bernstein, J.R. Kermode, G. Csányi et al., *Machine learning unifies the modeling of materials and molecules*, *Science Advances* **3** (2017) .
- [19] J. Dai and R.V. Krems, *Interpolation and extrapolation of global potential energy surfaces for polyatomic systems by gaussian processes with composite kernels*, *Journal of Chemical Theory and Computation* **16** (2020) 1386.
- [20] C.M. Handley, G.I. Hawe, D.B. Kell and P.L.A. Popelier, *Optimal construction of a fast and accurate polarisable water potential based on multipole moments trained by machine learning*, *Physical Chemistry Chemical Physics* **11** (2009) 6365.
- [21] J. Cui and R.V. Krems, *Efficient non-parametric fitting of potential energy surfaces for polyatomic molecules with gaussian processes*, *Journal of Physics B: Atomic, Molecular and Optical Physics* **49** (2016) 224001.
- [22] H. Sugisawa, T. Ida and R.V. Krems, *Gaussian process model of 51-dimensional potential energy surface for protonated imidazole dimer*, *The Journal of Chemical Physics* **153** (2020) 114101.
- [23] A.P. Bartók, J. Kermode, N. Bernstein and G. Csányi, *Machine learning a general-purpose interatomic potential for silicon*, *Physical Review X* **8** (2018) .
- [24] J. Lederer, W. Kaiser, A. Mattoni and A. Gagliardi, *Machine learning-based charge transport computation for pentacene*, *Advanced Theory and Simulations* **2** (2018) .
- [25] C.-I. Wang, M.K.E. Braza, G.C. Claudio, R.B. Nellas and C.-P. Hsu, *Machine learning for predicting electron transfer coupling*, *The Journal of Physical Chemistry A* **123** (2019) 7792–7802.

- [26] C. Kittel, *Introduction to Solid State Physics*, John Wiley & Sons, Hoboken, NJ, 8 ed. (2005).
- [27] J.M. Ziman, *Electrons and Phonons: The Theory of Transport Phenomena in Solids*, Clarendon Press (2007).
- [28] L. Kantorovich, *Quantum Theory of the Solid State: An Introduction*, KLUWER ACADEMIC (MA) (2004).
- [29] D.C. Wallace, *Thermodynamics of Crystals*, Dover Publications (1998).
- [30] M.T. Dove, *Introduction to Lattice Dynamics*, Cambridge University Press (2005).
- [31] N.W. Ashcroft and N.D. Mermin, *Solid State Physics*, Holt, Rinehart and Winston, New York (1976).
- [32] T. Tadano and S. Tsuneyuki, *First-principles lattice dynamics method for strongly anharmonic crystals*, *Journal of the Physical Society of Japan* **87** (2018) 041015.
- [33] R. Cowley, *Raman scattering from crystals of the diamond structure*, *Journal de Physique* **26** (1965) 659.
- [34] G.P. Srivastava, *The Physics of Phonons.*, Hilger (1990).
- [35] J. Sun, A. Ruzsinszky and J. Perdew, *Strongly constrained and appropriately normed semilocal density functional*, *Physical Review Letters* **115** (2015) .
- [36] J. Tao, J.P. Perdew, V.N. Staroverov and G.E. Scuseria, *Climbing the density functional ladder: Nonempirical meta-generalized gradient approximation designed for molecules and solids*, *Physical Review Letters* **91** (2003) .

-
- [37] A.P. Bartók, *Gaussian Approximation Potential: an interatomic potential derived from first principles Quantum Mechanics*, Ph.D. thesis, University of Cambridge, 2009.
- [38] W.J. Szlachta, *First principles interatomic potential for tungsten based on Gaussian process regression*, Ph.D. thesis, University of Cambridge, 2013.
- [39] O. Strickson, N. Nikiforakis and E. Artacho, *Dynamical continuum simulation of condensed matter from first principles*, *Physical Review Research* **1** (2019) .
- [40] P.O. Dral, A. Owens, S.N. Yurchenko and W. Thiel, *Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels*, *The Journal of Chemical Physics* **146** (2017) .
- [41] A. Wu, M.C. Aoi and J.W. Pillow, *Exploiting gradients and Hessians in Bayesian optimization and Bayesian quadrature*, 2017. 10.48550/ARXIV.1704.00060.
- [42] K. Asnaashari and R.V. Krems, *Gradient domain machine learning with composite kernels: improving the accuracy of PES and force fields for large molecules*, *Machine Learning: Science and Technology* **3** (2021) 015005.
- [43] M.F. Langer, A. Goeßmann and M. Rupp, *Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning*, *npj Computational Materials* **8** (2022) .
- [44] J.J. Meyer, M. Mularski, E. Gil-Fuster, A.A. Mele, F. Arzani, A. Wilms et al., *Exploiting symmetry in variational quantum machine learning*, *PRX Quantum* **4** (2023) .

- [45] M.F. Langer, S.N. Pozdnyakov and M. Ceriotti, *Probing the effects of broken symmetries in machine learning*, *Machine Learning: Science and Technology* **5** (2024) 04LT01.
- [46] A.P. Bartók, R. Kondor and G. Csányi, *On representing chemical environments*, *Physical Review B* **87** (2013) .
- [47] J. Behler and M. Parrinello, *Generalized neural-network representation of high-dimensional potential-energy surfaces*, *Physical Review Letters* **98** (2007) .
- [48] M. Rupp, R. Ramakrishnan and O.A. von Lilienfeld, *Machine learning for quantum mechanical properties of atoms in molecules*, *The Journal of Physical Chemistry Letters* **6** (2015) 3309–3313.
- [49] L. Himanen, M.O. Jäger, E.V. Morooka, F. Federici Canova, Y.S. Ranawat, D.Z. Gao et al., *Dscribe: Library of descriptors for machine learning in materials science*, *Computer Physics Communications* **247** (2020) 106949.
- [50] J. Behler, *Atom-centered symmetry functions for constructing high-dimensional neural network potentials*, *The Journal of Chemical Physics* **134** (2011) .
- [51] I. Batatia, D.P. Kovács, G.N.C. Simm, C. Ortner and G. Csányi, *Mace: Higher order equivariant message passing neural networks for fast and accurate force fields*, 2022. 10.48550/ARXIV.2206.07697.
- [52] I. Batatia, S. Batzner, D.P. Kovács, A. Musaelian, G.N.C. Simm, R. Drautz et al., *The design space of $e(3)$ -equivariant atom-centred interatomic potentials*, *Nature Machine Intelligence* **7** (2025) 56–67.
- [53] M. Rupp, A. Tkatchenko, K.-R. Müller and O.A. von Lilienfeld, *Fast and accurate modeling of molecular atomization energies with machine learning*, *Physical Review Letters* **108** (2012) .

-
- [54] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O.A. von Lilienfeld, K.-R. Müller et al., *Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space*, *The Journal of Physical Chemistry Letters* **6** (2015) 2326–2331.
- [55] H. Huo and M. Rupp, *Unified representation of molecules and crystals for machine learning*, *Machine Learning: Science and Technology* **3** (2022) 045017.
- [56] A.P. Bartók, M.C. Payne, R. Kondor and G. Csányi, *Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons*, *Physical Review Letters* **104** (2010) .
- [57] C.H. Ho, T.S. Gutleb and C. Ortner, *Atomic cluster expansion without self-interaction*, *Journal of Computational Physics* **515** (2024) 113271.
- [58] W. Rudin, *Principles of Mathematical Analysis*, McGraw-Hill, 3rd ed. (1976).
- [59] J.P. Darby, D.P. Kovács, I. Batatia, M.A. Caro, G.L. Hart, C. Ortner et al., *Tensor-reduced atomic density representations*, *Physical Review Letters* **131** (2023) .
- [60] V. Coropceanu, J. Cornil, D.A. da Silva Filho, Y. Olivier, R. Silbey and J.-L. Brédas, *Charge transport in organic semiconductors*, *Chemical Reviews* **107** (2007) 926–952.
- [61] M. Pope and C.E. Swenberg, *Electronic processes in polysilane (ps)*, in *Electronic Processes in Organic Crystals and Polymers*, p. 877–901, Oxford University Press New York, NY (1999), DOI.
- [62] E.A. Silinsh and V. Cápek, *Organic molecular crystals : interaction, localization, and transport phenomena*, American Inst. of Physics, NY (1994), <https://api.semanticscholar.org/CorpusID:92619605>.

- [63] V. Lemaure, D.A. da Silva Filho, V. Coropceanu, M. Lehmann, Y. Geerts, J. Piris et al., *Charge transport properties in discotic liquid crystals: A quantum-chemical insight into structure-property relationships*, *Journal of the American Chemical Society* **126** (2004) 3271–3279.
- [64] G.R. Hutchison, M.A. Ratner and T.J. Marks, *Intermolecular charge transfer between heterocyclic oligomers. effects of heteroatom and molecular packing on hopping transport in organic semiconductors*, *Journal of the American Chemical Society* **127** (2005) 16866–16881.
- [65] P.-O. Löwdin, *On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals*, *The Journal of Chemical Physics* **18** (1950) 365–375.
- [66] J.T. Kohn, N. Gildemeister, S. Grimme, D. Fazzi and A. Hansen, *Efficient calculation of electronic coupling integrals with the dimer projection method via a density matrix tight-binding potential*, *The Journal of Chemical Physics* **159** (2023) .
- [67] S. Grimme, M. Müller and A. Hansen, *A non-self-consistent tight-binding electronic structure potential in a polarized double-basis set for all spd-block elements up to $z = 86$* , *The Journal of Chemical Physics* **158** (2023) .
- [68] J. Kirkpatrick, *An approximate method for calculating transfer integrals based on the zindo hamiltonian*, *International Journal of Quantum Chemistry* **108** (2007) 51–56.
- [69] J. Ridley and M. Zerner, *An intermediate neglect of differential overlap technique for spectroscopy: Pyrrole and the azines*, *Theoretica Chimica Acta* **32** (1973) 111–134.
- [70] M. Rinderle, W. Kaiser, A. Mattoni and A. Gagliardi, *Machine-learned charge transfer integrals for multiscale simulations in organic thin films*, *The Journal of Physical Chemistry C* **124** (2020) 17733–17743.

-
- [71] C.-I. Wang, I. Joanito, C.-F. Lan and C.-P. Hsu, *Artificial neural networks for predicting charge transfer coupling*, *The Journal of Chemical Physics* **153** (2020) .
- [72] R. Hafizi, J. Elsner and J. Blumberger, *Ultrafast electronic coupling estimators: Neural networks versus physics-based approaches*, *Journal of Chemical Theory and Computation* **19** (2023) 4232–4242.
- [73] F. Eriksson, E. Fransson and P. Erhart, *The hiphive package for the extraction of high-order force constants by machine learning*, *Advanced Theory and Simulations* **2** (2019) .
- [74] D. Eriksson, K. Dong, E.H. Lee, D. Bindel and A.G. Wilson, “Scaling gaussian process regression with derivatives.”
<https://dl.acm.org/doi/pdf/10.5555/3327757.3327791>.
- [75] T.N. Hoang, K.H. Low, P. Jaillet and M. Kankanhalli, *Active learning is planning: Nonmyopic -bayes-optimal active learning of gaussian processes*, in *Machine Learning and Knowledge Discovery in Databases*, p. 494–498, Springer Berlin Heidelberg (2014), DOI.
- [76] D. Khatamsaz, B. Vela, P. Singh, D.D. Johnson, D. Allaire and R. Arróyave, *Bayesian optimization with active learning of design constraints using an entropy-based approach*, *npj Computational Materials* **9** (2023) .
- [77] M. Ghorbani, M. Boley, P.N.H. Nakashima and N. Birbilis, *An active machine learning approach for optimal design of magnesium alloys using bayesian optimisation*, *Scientific Reports* **14** (2024) .
- [78] M. Buisson-Fenet, F. Solowjow and S. Trimpe, *Actively learning gaussian process dynamics*, 2019. 10.48550/ARXIV.1911.09946.
- [79] A. Troisi, *Charge transport in high mobility molecular semiconductors: classical models and new theories*, *Chemical Society Reviews* **40** (2011) 2347.

- [80] A. Troisi and G. Orlandi, *Charge-transport regime of crystalline organic semiconductors: Diffusion limited by thermal off-diagonal electronic disorder*, *Physical Review Letters* **96** (2006) .
- [81] Y. Wang, G. Csanyi and C. Ortner, *Many-body coarse-grained molecular dynamics with the atomic cluster expansion*, 2025.
10.48550/ARXIV.2502.04661.
- [82] H.-H. Lin, C.-I. Wang, C.-H. Yang, M.K. Secario and C.-P. Hsu, *Two-step machine learning approach for charge-transfer coupling with structurally diverse data*, *The Journal of Physical Chemistry A* **128** (2023) 271–280.
- [83] E.M. Stein and R. Shakarchi, *Complex Analysis*, Princeton University Press (2003).
- [84] P.N. Keating, *Sum rules for the dielectric screening properties of crystals*, *Physical Review* **187** (1969) 1190–1192.
- [85] K. Shindo and H. Hara, *Local field correction and acoustic sum rule for si and ge*, *Journal of the Physical Society of Japan* **43** (1977) 899–902.
- [86] Z. Zhu, J. Park, H. Sahasrabuddhe, A. Ganose, R. Chang, J. Lawson et al., *A high-throughput framework for lattice dynamics*, Mar., 2024.
10.26434/chemrxiv-2024-c82v8.

Declaration of Academic Integrity

I hereby declare that I have written the present work myself and did not use any sources or tools other than the ones indicated.

Date: 21 11 2025

.....Keerati...Keeratikarn.....