UNIVERSITÉ DU LUXEMBOURG

PhD-No PhD-FSTM-101
The Faculty of Science, Technology and Medicine
Institute of Advanced Studies

DISSERTATION

Presented on 24.10.2025 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN PHYSIQUE

by

Mirela Puleva

Towards quantum-accurate description of molecular interactions in biochemical systems

Dissertation defence committee

Prof. Dr. Alexandre Tkatchenko, Supervisor
University of Luxembourg

Prof. Dr. Alexander Skupin, Internal jury member
University of Luxembourg

Prof. Dr. Reinhard Maurer, External jury member
University of Warwick

Prof. Dr. Jan Řezáč, External jury member
Academy of Sciences of the Czech Republic

Prof. Dr. Etienne Fodor, Chair
University of Luxembourg

Affidavit / Statement of originality

I declare that this thesis:

- is the result of my own work. Any contribution from any other party, and any use of generative artificial intelligence technologies have been duly cited and acknowledged;
- is not substantially the same as any other that I have submitted, and;
- is not being concurrently submitted for a degree, diploma or other qualification at the University of Luxembourg or any other University or similar institution except as specified in the text.

With my approval I furthermore confirm the following:

- I have adhered to the rules set out in the University of Luxembourg's Code of Conduct and the Doctoral Education Agreement (DEA)[1], in particular with regard to Research Integrity.
- I have documented all methods, data, and processes truthfully and fully.
- I have mentioned all the significant contributors to the work.
- I am aware that the work may be screened electronically for originality.

I acknowledge that if any issues are raised regarding good research practices based on the review of the thesis, the examination may be postponed pending the outcome of any investigation of such issues. If a degree was conferred, any such subsequently discovered issues may result in the cancellation of the degree.

Approved on 2025-07-31

[1] If applicable (DEA is compulsory since August 2020)

Abstract

The accurate prediction of molecular properties is essential for understanding physicochemical phenomena and enabling impactful applications such as acceleration of the drug discovery pipeline. To this end, accurate description of quantum mechanical (QM) effects is crucial, especially in non-covalently bound systems such as protein-ligand interactions, due to collective long-range van der Waals (vdW) interactions. While high-level *ab initio* techniques like Coupled Cluster (CC) and diffusion Monte Carlo (DMC) can provide accurate estimations of relevant molecular properties, such as interaction energies, they are computationally feasible only for systems up to hundreds, and in rare cases thousands of atoms. With the increase of system size, approximate methods, such as density functional theory (DFT) with vdW corrections, semi-empirical models, and classical force fields (FF), must be employed depending on the achievable balance between computational cost and accuracy.

In the last decade, a new family of alternative approaches, *i.e.*, Machine Learning (ML) methods, has gained growing attention due to the promise of delivering *ab initio* level estimations at the cost of coarse-grained approaches such as classical FFs. As a matter of fact, ML methods, 'learn' complex relationships between atomic configurations and molecular properties from data, ideally producing scalable and transferable models, which enable accurate predictions at reduced computational cost, even for large-scale systems.

Yet, until now, ML methods have not been sufficiently generalized to tackle in a reliable way protein-ligand interactions. To obtain reliable models for such large conformers, in fact, two elements are required. Firstly, it is necessary to have accurate data on large molecular compounds representatives for amino acids, *i.e.*, the main components of proteins, and for their reciprocal interactions. Secondly, well defined test cases for the models are necessary to verify their stability in performing dynamical simulations at different temperatures, and to test their portability.

This thesis has thus the objective of addressing both these questions, by defining a protocol to construct high-quality reference data – that can be used to training data-driven ML potentials and eventually for validating approximate QM methods – and by proposing a set of benchmark tests that serve the purpose of validating the stability and accuracy of the models. To answer the former question, here a novel dataset is introduced, *i.e.*, the Quan-

tum Interacting Dimer (QUID) dataset, that is specifically created using state-of-the-art PBE0+MBD calculations, on large, complex and chemically diverse pocket-like molecular dimers, in and out of equilibrium, serving as prototype reference data for protein-ligand systems that interact via vdW interactions. The non-covalent interactions within the dataset are investigated in-depth, and compared to gold standard *ab initio* calculations obtained via CC and QMC. These accurate references have afterwards been used to test existing DFT exchange-correlation functionals, semi-empirical and classical Force Field (FF) methods, and for the ablation studies of a ML model. Several dispersion-inclusive density functionals are proven to provide accurate interaction energy predictions in the study, while the investigation of the semiempirical methods and FFs indicates towards a need for further improvements.

On the other hand, to answer the latter question of ML models' stability here, the TEA Challenge 2023 is presented as a reliable verification to the performance and stability of a representative sample of ML approaches, both kernel and neural network (NN), on MD simulations of large organic molecules. As a result, the region of reliable performance for the models was identified, avenues for improvement were proposed, and a set of guidelines to the field were suggested. Based on the outcome of the TEA Challenge 2023, a ML model was chosen for a protein-protein interactions test of the Sars-Cov-2 protein.
Furthermore, as an alternative strategy towards accurate calculations of large (bio)molecular systems, a ML-corrected DFTB method was proposed in the EquiDTB framework, aiming to improve the accuracy of the semi-empirical method by replacing its standard parametrised repulsive potential with a physics-inspired equivariant NN.

In summary, the systematic investigations presented in this thesis lay a rigorous foundation for the development of next-generation models capable of predicting binding energies and performing molecular dynamics simulations of protein–ligand interactions with quantum chemical accuracy. By integrating high-fidelity reference data, physically grounded approximations, and machine learning techniques, this work contributes essential methodological insights and benchmarks that will support future efforts to model complex biomolecular systems with both accuracy and scalability.

Keywords: Molecular dimers, Protein-ligand interactions, Machine Learning, Method benchmarking

Preface

Note on publications

This thesis is partly based on the previously published articles:

[1] **M. Puleva** et al. “Extending quantum-mechanical benchmark accuracy to biological ligand-pocket interactions" *Nature Communications* **16**, 8583, 2025.

[2] I. Poltavsky, **M. Puleva**, et. al. “Crash Testing Machine Learning Force Fields for Molecules, Materials, and Interfaces: Model Analysis in the TEA Challenge 2023" *Chemical Science* **16**, 3720-3737, 2025.

[3] I. Poltavsky, **M. Puleva**, et. al “Crash Testing Machine Learning Force Fields for Molecules, Materials, and Interfaces: Molecular Dynamics in the TEA Challenge 2023" *Chemical Science* **16**, 3738-3754, 2025.

[4] L. Medrano Sandonas, **M. Puleva**, et. al. “Advancing Density Functional Tight-Binding method for Large Organic Molecules through Equivariant Neural Networks", submitted 2025, pre-print on ChemRxiv 10.26434/chemrxiv-2025-z3mhh.

The author’s contribution to each paper is given at the beginning of a chapter or section where the main results are included.

Contents

List of Figures

List of Tables

List of Abbreviations

1. (in)com – (in)complete
2. (un)fold – (un)folded
3. ACE2 - Angiotensin-Converting Enzyme 2
4. CC - Coupled Cluster
5. CCS - Chemical Compound Space
6. CCSD(T) - Coupled Cluster Singles Doubles (Triplets)
7. DFT - Density Functional Theory
8. DFTB - Density Functional Tight Binding
9. DMC - Diffusion Monte Carlo
10. EquiDTB - Equivariant networks for Delta Tight Binding
11. F, SF, L - Folded, Semi-Folded, Linear
12. FCHL19* – Faber–Christensen–Huang–Lilienfeld 19*
13. FF - Force Field
14. GGA - Generalized Gradient Approximation
15. HF - Hartree-Fock
16. HPC - High Performance Computing
17. KRR - Kernel Ridge Regression
18. LDA - Local Density Approximation
19. LNO-CCSD(T) - Local Natural Orbital CCSD(T)
20. MAE – Mean Absolute Error
21. MBD - Many Body Dispersion
22. MD - Molecular Dynamics
23. ML - Machine Learning
24. MM - Molecular Mechanics
25. MP2 - Møller–Plesset perturbation theory of second order
26. NCI - Non-Covalent Interactions
27. NN - Neural Networks
28. PBE - Perdew–Burke-Ernzerhof
29. PES - Potential Energy Surface

30. QM - Quantum Mechanics
31. QMC - Quantum Monte Carlo
32. QUID - QUantum Interacting Dimer
33. RBD - Receptor Binding Domain
34. RMSE – Root Mean Square Error
35. SAPT - Symmetry Adapted Perturbation Theory
36. SARS-Cov-2 - Severe Acute Respiratory Syndrome CoronaVirus 2
37. SE - Semi-empirical
38. sGDML – Symmetric Gradient Domain Machine Learning
39. SO3 – SO3krates
40. SOAP/GAP – Smooth Overlap of Atomic Position / Gaussian Approximation Potential
41. TEA - crash TEsting machine learning force fields for molecules, materials, and interfaces
42. vdW - van der Waals
43. XDM - eXchange (hole) Dipole Moment

List of Symbols and Notation

a_0	Atomic distance unit
E_h	Atomic energy unit
$\in$	Belongs to
ΔG	Binding free energy
$\langle \vert, \vert \rangle$	Bra,ket
q	Charge
R	Classical nuclei position
$\mathbb{C}$	Complex numbers space
det	Determinant of a matrix
μ	Dipole moment
ρ	Electronic density
E	Energy
Eq.	Equation
err	Error function
exp	Exponential function
$\forall$	For all
∇	Gradient
H	Hamiltonian
$\int dx$	Integral over x
E_{int}	Interaction energy
$\mathcal{J}$	Jastrow factor
Δ, ∇^2	Laplacian
m	Mass
α	Molecular polarisability
∂/∂_x	Partial derivative with respect to x
r	Particle position
$\prod$	Product of a sequence
$\mathbb{R}$	Real numbers space
$\sum$	Summatory
$\otimes$	Tensor Product

Chapter 1
Introduction

The modelling of molecular compounds and of their interactions is an area of enquiry involving many disciplines, such as physics, chemistry, biochemistry and biology. In particular, it focuses on the prediction of the stability, the interactions, the dynamics and the reactivity of various systems spanning from small molecules, to proteins of millions of atoms, and materials. Since the equations that describe the systems of atoms and electrons are not analytically solvable, various approximations have to be introduced, with different levels of accuracy, in order to tackle the different energetic and temporal scales of these phenomena. In the last decades, with the further development of High-Performance Computing (HPC), numerous computational methods have been introduced with the aim of revealing these intriguing interaction mechanisms and assist novel discoveries. A particular area of interest for the latter is the use of computational models for accelerating the drug design pipeline and improving its efficacy with *in silico* development [1]. Currently, it takes ∼10-15 years as well as hundreds to thousands of millions of dollars in research for one novel drug out of every 5-10k identified promising candidates to reach the market [2], hence the need for a cheaper and faster method for selecting candidates with desired pharmacological properties. This has become a particularly salient point in light of the recent COVID-19 pandemic and the increased spread of antibiotic resistance among critical classes of antibiotics [3–6].

In order to improve the drug candidate screening process, computational models need to be capable of reliable and highly accurate predictions of physicochemical properties in organic molecular systems, *e.g.*, drug compounds, peptides, and proteins. In particular, the robust prediction of the interactions of drug candidates and their targets, most commonly protein pockets [7], requires the accurate description of both covalent and non-covalent bonds. While covalent bonds involve shared electrons between atoms, the weaker non-covalent ones involve a broad set of electromagnetic interactions broadly classified in chemistry as electrostatic, van der Waals (vdW), π interactions, and Hydrogen bonds [8, 9]. Given the long-standing concerns of toxicity and off-target effects associated with highly reactive covalently binding drugs, the pharmaceutical compounds currently on the market predominantly interact with their targets via non-covalent interactions [7], with some notable exceptions in recent years in the research of inhibitors [10]. Non-covalent interactions also

govern protein folding, protein-protein interactions, and solvation effects, and are thus a key ingredient for any computational method aiming to accurately describe and predict the dynamic behaviour of biochemical systems. In fact, although single non-covalent interactions are rather weak when compared to the covalent ones, their collective effects, in large conformers, and their long-range impact, are a driving force in biochemical processes [8]. In systems such as DNA, the collective contribution of many vdW contacts—alongside hydrogen bonding—significantly stabilises base pair stacking, enhancing the overall structural integrity and binding affinity of the double helix.

Two paradigms have emerged historically in the attempt to describe interactions in molecular systems. One has arisen in the field of computational biochemistry, driven by the challenges posed by the size and complexity of biomolecules and proteins. It relies on heuristics - empirical parametrisations and classical mechanics; *i.e.*, the classical Molecular Mechanics (MM) Force Fields (FF) for rapid calculations of large biomolecules at the cost of inherent limitations in accuracy. The second paradigm, in the field of computational chemical physics, has been fuelled by the drive to understand and predict molecular interactions at a high degree of reliability and accuracy from first principles. This has motivated the development of various approaches for the quantum mechanical description of molecular systems at different levels of approximation. Due to the limitations of current technologies for finding accurate numerical solutions to the complex physical equations governing molecular systems, the approximated electronic structure theory methods have become the go-to approach. Density Functional Theory (DFT) has been a particularly popular approach among them since the 1980s, shortly after its inception [11], due to its ability to tackle large chemical compounds with reasonable accuracy. Since then, advancements in research and computational capabilities, *e.g.*, as parallel computations and HPC centers, have enabled the property prediction and simulation of molecular systems spanning up to thousands of atoms. This has been achieved via two main research lines. Firstly, by improving the approximations and algorithm implementations of the quantum chemistry methods in both DFT and the more accurate wave function *ab initio* methods, like Coupled Cluster theory[12] and Quantum Monte Carlo[13, 14]. In DFT, through developing new functionals[15] and dispersion corrections[16], and in the wave function methods by local approximations and code optimisation implementations. Secondly, another research line has been the improvement of quantum effects incorporation in semi-empirical approaches as the efficient middle ground between the empirical classical force fields and the aforementioned quantum chemistry methods.

However, in the last decade, a third paradigm has been introduced with the aim to bridge the gap between empirical approximations able to simulate large systems and the accurate quantum chemistry methods that can only be used on smaller molecules: Machine Learning (ML) models. ML models trained on molecular data, have allowed for faster predictions of molecular properties [17–19], the *de novo* molecular structure generation [20–22], and even simulations of the dynamics of biomolecular systems [23–26] with quantum chemical accuracy. Further progress has been driven by the development of graphic-processing units (GPU) for scientific calculations, and by the introduction of convolutional neural networks,

the attention mechanism and transformer architecture [27], and by the latest equivariance paradigm applied to Neural-Networks (NN)s [28]. A success story of these developments for the representation and modelling of biomolecular systems came in 2020 with the success of AlphaFold 2 [29] at the Critical Assessment of Structure Prediction, CASP14 challenge [30] for protein structure prediction. However, that significant step in progress provides only a single snapshot of the protein systems and still lacks a key ingredient for molecular modelling for the drug design pipeline, such as the prediction of dynamical processes, like ligand binding that is characterised by physicochemical properties out of equilibrium. The latest development in the quest for property prediction by ML models has been the creation of physically-inspired equivariant NN [31–43] as a way to already encode some physics laws inside the models, without having to learn them from data, as well as to provide more interpretability to the epistemologically opaque NNs. This has also led to improvements towards generation of stable and accurate molecular dynamics simulations by corresponding MLFFs.

In ML development, increased efficiency and transferability in learning are critically important given the high cost of highly accurate reference data for training procured from experiments or quantum chemistry simulations [44–46]. Beyond training datasets, high-fidelity benchmarks are required for rigorous testing of the ML models to establish their reliability in the relevant regions of Chemical Compound Space (CCS). However, there is also the question of the reliability of the benchmark data itself. To this end, (semi-)blind test challenges are devised and conducted regularly to assess the state of the predictive power by computational methods for molecular systems, which serve as a barometer of the advances in the corresponding fields and identifiers of remaining challenges. Famous examples include the SAMPL challenge [47] for prediction of properties in (bio)molecules, the CACHE challenge [48] to identify ligands binding to a given protein target, and the CSP blind tests for crystal structure prediction [49].

Furthermore, while we can see these blind tests as reviews of a specific field, many other benchmark tests [50, 51] are constantly published to assess the performance of state-of-the-art (SOTA) computational methods. For example, in the field of quantum chemistry, the yearly output of papers for DFT benchmarks alone is now in the hundreds [52]. The assessment of these various methods relies on the creation of increasingly comprehensive datasets of highly accurate calculations, including more diverse molecular systemsh and including more complex interactions [53–56].

In summary, a delicate balance of accuracy, efficiency, and scale accessibility is sought by both quantum chemistry and ML methods for the prediction of properties and behaviour of molecular systems. While quantum chemistry methods have dominated the small molecules regime, ML's efficiency and transferability offers a bridge to bring that accuracy to larger systems of real-life relevance like druggable protein targets interacting with drug candidate ligands. In the context of this outstanding challenge, **the goal of this thesis is to shed light on the construction of appropriate study cases and test for physical, ML, and ML-corrected physical models including important quantum effects, for**

the description of dynamical processes in organic systems towards protein-ligand interactions. In particular, two of the main questions that are required for establishing reliable models are addressed: firstly, the **creation of benchmark and training datasets that can be used for ML models for protein-ligand dynamics**; and secondly, the need to **define tests that can estimate the accuracy, stability, and portability of the models, and thus their applicability to general systems.**

1.1 Thesis aims

To answer the former question, a protocol for constructing an accurate dataset of large molecular as proxies for protein pocket interacting with a ligand. To this end, a high-fidelity dataset of Quantum Interacting Dimers (QUID) is presented as one of the key contributions of this thesis [57]. The dataset, optimised at high-level quantum chemistry for large molecules, *i.e.*, DFT with the PBE0 exchange-correlation functional [58] combined with Many-Body dispersion (MBD) [59, 60], comprises of larger organic chemically diverse and flexible dimers. The binding energies of the dimers, obtained for structures in and out of equilibrium, are compared against gold standard *ab initio* calculations from Coupled Cluster (CC) and quantum Monte Carlo (QMC), displaying a great degree of compatibility. QUID is then used as a probe for different DFT, semi-empirical methods, and classical and MLFFs.

To answer the latter question of assessing ML models, the TEA Challenge 2023 [61, 62] was created and conducted, with the aim to test the performance of a representative sample of ML approaches, from well-established kernel methods to novel equivariant neural networks. They are tested in terms of accuracy, stability of molecular dynamics simulations and efficiency. The work presented in this thesis involves Challenges I and II of the TEA Challenge 2023. They comprise of a test of molecular dynamics simulations on large organic molecules in two regimes - where high-level reference data is present and where it is not, charting the way towards gaining insight and trust into ML models' performances for larger (bio)molecular systems.

While the QUID work established the limitations of semi-empirical methods' accuracy for such systems, and the TEA Challenge 2023 project assessed the promise and drawbacks of ML methods, a middle ground strategy is also suggested here. Namely, the proposed EquiTB framework [63] aims to combine the best of both worlds with ML-corrected Density-Functional Tight binding method [64, 65]. The models are probed extensively on a series of tests in different environments, including non-covalent, organic molecular systems.

Finally, the last project of this thesis is linked to the outlook and future work with application of a ML model to the prediction of interaction energy in protein-protein interactions between the SARS-Cov-2 protein [66] and a human receptor.

In conclusion, the works in this thesis offer insights and advances towards the goal of constructing ML models for simulating the dynamical processes in proteins at the quantum

level. In particular, this work identifies the domains and limits of reliability of quantum chemical and ML methods for organic systems of interest as building blocks to modelling interactions between protein pockets and ligands, paving the way to the construction and testing of new and more accurate ML models.

1.2 Thesis outline

This thesis is divided into seven Chapters, starting with the current Chapter 1.2 providing a brief introduction to the topics of research and the motivation for their combination in this interdisciplinary work. The following Chapter 2 contains the theoretical background of the utilised computational methods; further supporting information for methods touched upon in this work but not directly explored is available in Appendix A- E. The results obtained during this PhD project are presented in four thesis chapters: Chapter 3 to Chapter 6, followed by a Summary and Outlook for future work in Chapter 7. Among the results, Chapter 3 provides a detailed study of new complex molecular dimers as proxy examples for the interactions between protein pockets and ligand molecules. While Chapter 3 provides a snapshot of the quantum chemistry and classical FF methods for predicting properties in interacting organic systems, Chapter 4 investigates the ML models' stability and accuracy in predicting the dynamics of organic molecules. Specific venue for utilisation of the SOTA NNs and conclusions on good practices going forward have been identified in Chapter 4. The project in Chapter 5 builds upon the findings of Chapters 3 and 4 by employing the neural networks class found to perform best in the TEA Challenge 2023 for the improvement of a semi-empirical method, which is found insufficiently accurate in Chapter 5 when tested on larger molecular dimers. Finally, an application on realistic protein systems has been demonstrated in Chapter 6. Some of the work presented in this thesis has already been submitted or published in peer-reviewed journals, and thus has a partially cumulative character.

Chapter 2
Theoretical Background

The physical theory describing atomic and molecular systems in the non-relativistic limit is Quantum Mechanics, having its roots in the early 20th century. Although this led to the definition of one of the most successful scientific frameworks of all time, exact solutions to the theory cannot be obtained, besides few simple cases such as a single particle in a central potential: the most notable examples being the quantum harmonic oscillator and the hydrogen atom [67, 68]. In general, no exact analytic solutions exist for the quantum many-body problem due to the pairwise Coulomb interactions present in the general theory, which introduce explicit correlation between the coordinates of all the quantum particles in the systems. Hence, many different numerical methods have been introduced in order to find approximate estimations for the system observables, *e.g.*, total energy, structural geometries, and interaction energies between fragments.

This chapter introduces the quantum chemistry methods used to approximate solutions to the Schrödinger equation, laying the basis for the calculations and analyses presented in this thesis. It provides an overview of the electronic structure methods employed in the individual projects, with an emphasis on their theoretical foundations and thereby relevant domains of applicability and shortcomings in the organic molecular systems of interest. While this is not an exhaustive account of all available methods, it aims to inform the subsequent discussions of the results obtained from a variety of computational approaches used in this thesis.

2.1 The Schrödinger equation for molecular systems

2.1.1 General quantum mechanical framework

The state of a quantum system is, represented as an element $|\Psi\rangle$ of a given Hilbert space $\mathcal{H}$. The time evolution of such a state is described by the *time-dependent Schrödinger equation* (TDSE) [69–73]

$$i\hbar\frac{\partial|\Psi(t)\rangle}{\partial t} = \hat{H}|\Psi(t)\rangle, \tag{2.1}$$

where $\hbar$ is the reduced Planck constant, and $\hat{H}$ is the Hamiltonian operator belonging to the set of Hermitian operators acting on $\mathcal{H}$. When the Hamiltonian is a time-independent operator, the knowledge of the eigenvalues and eigenvectors of the Hamiltonian operator $\hat{H}$ fully determines the solutions of Eq. (2.1), once the initial condition are fixed. Let us now consider a real number $E \in \mathbb{R}$ and an associated quantum state $|\psi_E\rangle$, such that the so-called *time-independent Schrödinger equation* (TISE) is satisfied, *i.e.*,

$$\hat{H}|\psi_E\rangle = E|\psi_E\rangle \ . \tag{2.2}$$

From Eq. (2.1) the eigenvalue of the Hamiltonian operator E can be interpreted physically as the energy of the corresponding time-independent eigenstate[1] $|\psi_E\rangle$. For a fixed initial state $|\Psi(0)\rangle = \sum_{E\in\mathcal{S}_{\hat{H}}} c_E(0)|\psi_E\rangle$, where $c_E \in \mathbb{C}$ and $\mathcal{S}_{\hat{H}}$ is the spectrum of $\hat{H}$, the solution to the Eq. (2.1) is given by

$$|\Psi(t)\rangle = \sum_{E\in\mathcal{S}_{\hat{H}}} \exp\left[-i\frac{Et}{\hbar}\right] c_E(0)|\psi_E\rangle \ . \tag{2.3}$$

The problem of solving the dynamics of a quantum system is reduced to determining the eigenspectrum — that is, the eigenvalues and eigenstates — of the Hamiltonian operator in a Hilbert space. The Hilbert space is the mathematical complete space in which quantum states reside, where each state is represented by a vector, and physical observables are represented by linear operators acting on this space. Analytical solutions are available only for a few simple cases with limited number of degrees of freedom and high symmetry, such as a single particle in a central potential—the classic examples being the quantum harmonic oscillator and the Hydrogen-like atom (the Coulomb potential). For realistic systems with many quantum degrees of freedom, as in the case of molecular systems composed by electrons and nuclei mutually interacting through Coulomb interactions, the Hamiltonian operator cannot be similarly decomposed into a sum of independent integrable operators and thus analytic solutions are not viable. Modelling dynamic systems evolving in time is correspondingly even more complex and further approximations need to be applied as will be discussed later.

2.1.2 The quantum molecular system

Let us consider a system composed of M nuclei and n electrons, and define the set of operators

$$\{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{n+M}\} = \{\hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_n, \hat{\mathbf{R}}_1, \ldots, \hat{\mathbf{R}}_M\}, \tag{2.4}$$

where $\hat{\mathbf{R}}_A$ and $\hat{\mathbf{r}}_a$ are three-dimensional vectors of Hermitian operators describing the position of the A-th atomic nucleus and the a-th electron[2] in $\mathbb{R}^3$, respectively. By introducing the position operators eigenstate $|\mathbf{x}\rangle = |\mathbf{r}, \mathbf{R}\rangle$, the wavefunction $\Psi(\mathbf{x}; t) = \langle\mathbf{x}|\Psi(t)\rangle$

[1] For simplicity, a non-degenerate spectrum of $\hat{H}$ is assumed.

[2] At the level of operators, electrons are treated as distinguishable particles. To account for their indistinguishability and the Pauli exclusion principle, specific constraints must be imposed on the space of quantum states, which will be discussed later in detail.

can be introduced, associated to a given quantum state. The wavefunction is, in general, a square integrable complex-valued function on the configuration space, *i.e.*, $\Psi(\mathbf{x}) \in L^2(\mathbb{C}, \mathbb{R}^{3(M+n)})$.

The Hamiltonian of a molecular system is a Hermitian operator given by

$$\hat{H}_{\text{mol}} = \hat{T}_N + \hat{T}_e + \hat{V}_{Ne} + \hat{V}_{ee} + \hat{V}_{NN}, \tag{2.5}$$

where $\hat{T}_N$ and $\hat{T}_e$ are the total kinetic energy operators of the nuclei and electrons, respectively, while $\hat{V}_{Ne}$, $\hat{V}_{ee}$, and $\hat{V}_{NN}$ denote the potential energy operators corresponding to nucleus–electron, electron–electron, and nucleus–nucleus interactions. Such operators can be written in terms of their action on the molecular wavefunction. The kinetic energy operators can be expressed in terms of the Laplacian operator $\nabla^2_{\mathbf{r}_a} = \partial^2_{r_{a_x}} + \partial^2_{r_{a_y}} + \partial^2_{r_{a_z}}$ (with $\mathbf{R}_A$ instead of $\mathbf{r}_a$ for the nuclear coordinates)

$$\langle \mathbf{x}|\hat{T}_N|\Psi(t)\rangle = \sum_{A=1}^{M} \frac{(-\hbar^2)}{2M_A} \nabla^2_{\mathbf{R}_A} \Psi(\mathbf{x}, t) \quad \langle \mathbf{x}|\hat{T}_e|\Psi(t)\rangle = \sum_{a=1}^{n} \frac{(-\hbar^2)}{2m} \nabla^2_{\mathbf{r}_a} \Psi(\mathbf{x}, t), \tag{2.6}$$

while the potential energy terms reads:

$$\langle \mathbf{x}|\hat{V}_{ee}|\Psi(t)\rangle = \frac{e^2}{4\pi\epsilon_0} \sum_{a=1}^{n} \sum_{b<a} \frac{1}{|\mathbf{r}_a - \mathbf{r}_b|} \Psi(\mathbf{x}, t) \tag{2.7}$$

$$\langle \mathbf{x}|\hat{V}_{Ne}|\Psi(t)\rangle = \frac{e^2}{4\pi\epsilon_0} \sum_{A=1}^{M} \sum_{a=1}^{n} \frac{Z_A}{|\mathbf{R}_A - \mathbf{r}_a|} \Psi(\mathbf{x}, t) \tag{2.8}$$

$$\langle \mathbf{x}|\hat{V}_{NN}|\Psi(t)\rangle = \frac{e^2}{4\pi\epsilon_0} \sum_{A=1}^{M} \sum_{B<A} \frac{Z_A Z_B}{|\mathbf{R}_a - \mathbf{R}_b|} \Psi(\mathbf{x}, t), \tag{2.9}$$

$$\tag{2.10}$$

where m is the electron mass, e is the electron charge, ϵ_0 is the vacuum permittivity constant, Z_A is the atomic number of the A-th atom and M_A its mass.

Following Eq. (2.1), the exact quantum dynamics is given by the solution of a partial differential equation for the wavefunction $\Psi(\mathbf{x}, t)$, *i.e.*,

$$i\hbar \frac{\partial \Psi(\mathbf{x}, t)}{\partial t} = \langle \mathbf{x}|\hat{H}_{\text{mol}}|\Psi(t)\rangle \qquad . \tag{2.11}$$

Since the Hamiltonian does not exhibit explicit time dependence, the quantum dynamics of the molecular system is entirely governed by the TISE associated with the molecular Hamiltonian operator $\hat{H}_{\text{mol}}$. However, even in this case, one must solve a partial differential equation defined over a $3(M + n)$-dimensional space, where M and n denote the number of nuclei and electrons, respectively. To mitigate the computational complexity inherent in

this task, various approximations have been developed within the framework of quantum chemistry. Many of these approximations can be expressed through specific ansätze for the molecular wavefunction. In the following sections, we review some of these approaches, which form the foundation of the quantum chemical computational methods employed throughout this thesis.

2.1.3 Born-Oppenheimer approximation

The first approximation introduced to tackle the solution of the time-independent Schrödinger equation of a molecular system is known as the Born-Oppenheimer approximation (BOA) [74], or adiabatic approximation. It is based on the assumption that, while the Coulomb forces between nuclei and electrons are of the same order of magnitude (for the third principle of dynamics), the kinetic energy of the nuclei is ≈ 1900 times smaller than that of the electrons, due to the differences in their masses $M_A \approx 1900 m_e$. Thus, it is possible to decouple the nuclear and electronic degrees of freedom, assuming the latter to be always in equilibrium with the former.

In this case, the total wavefunction $\Psi(\mathbf{r}, \mathbf{R})$ can thus be expressed as a linear combination of the electronic eigenfunctions of the orthonormal basis set ψ_k

$$\Psi(\mathbf{r}, \mathbf{R}) = \sum_k C_k(\mathbf{R}) \psi_k(\mathbf{r}; \mathbf{R}), \quad \text{where} \quad \psi_k(\mathbf{r}; \mathbf{R}) = \langle \mathbf{r} | \psi_k(\mathbf{R}) \rangle, \tag{2.12}$$

where $C_k(\mathbf{R})$ are the coefficients of the wavefunctions that give the roto-vibrational motion of the nuclei in a molecular system for a system with electrons in state k. The wave functions of the electrons and of the nuclei are obtained through the system of equations for each electron state k that is given by

$$\begin{cases} \hat{H}_e(\mathbf{R}) \left| \Psi_k(\mathbf{R}) \right\rangle = E_k(\mathbf{R}) \left| \Psi_k(\mathbf{R}) \right\rangle \\ \left[-\sum_{a=1}^{M} \dfrac{\hbar^2}{2M_A} \nabla^2_{\mathbf{R}_A} + E_k(\mathbf{R}) + \hat{V}_{NN}(\mathbf{R}) \right] C_k(\mathbf{R}) = E C_k(\mathbf{R}), \end{cases} \tag{2.13}$$

in which the second equation for the nuclei, describes the particles moving in the potential energy determined by the electronic state $E_k(\mathbf{R})$.

Therefore, an electronic energy can be obtained from the above equation per nuclear geometry, and the collection of the electronic energies for all possible molecular configurations *i.e.*, nuclear arrangements, produces the Potential Energy Surface (PES) of the molecule. In this work, only PES for electronic ground states are considered.

Various electronic structure methods arise from the different approximations and techniques used to solve the electronic Schrödinger equation within the BOA framework. These methods range from simpler, computationally less expensive ones like Hartree-Fock and Density Functional Theory (DFT), to more sophisticated, accurate methods like coupled cluster and quantum Monte Carlo.

2.2 Hartree-Fock method

The Hartree-Fock (HF) [75] is a mean-field quantum chemistry approximation method that models a many-electron system, for which electrons are subjected to a mean field potential created by all the other electrons. Thus reducing the many-body problem into an effective one-body problem. HF is a variational method where the approximate ground-state energy is obtained by applying the effective one-electron Fock operator on an approximate wavefunction, the HF wavefunction, which is a single Slater determinant [76] for an n-electron system. HF explicitly considers only individual electron wavefunctions and neglects electron correlation effects beyond the exchange interaction, rendering it less accurate for strongly correlated molecular systems [77].

$$\Psi_{\mathrm{HF}}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \psi_1(\mathbf{x}_1) & \psi_2(\mathbf{x}_1) & \cdots & \psi_n(\mathbf{x}_1) \\ \psi_1(\mathbf{x}_2) & \psi_2(\mathbf{x}_2) & \cdots & \psi_n(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(\mathbf{x}_n) & \psi_2(\mathbf{x}_n) & \cdots & \psi_n(\mathbf{x}_n) \end{vmatrix}, \tag{2.14}$$

where $\mathbf{x}_i = (\mathbf{r}_i, s_i)$ are orthonomal combined spin molecular orbitals with spatial coordinates $\mathbf{r}_i$ and spin s_i of electron i. The determinant enforces antisymmetry with respect to electron exchange to satisfy the Pauli exclusion principle, which states states under the exchange of two identical particles, the total wave function should be antisymmetric for fermions, causing that two or more identical particles with half-integer spins cannot simultaneously occupy the same quantum state.

The solution of the HF wavefunction is obtained by solving the Fock equation

$$\hat{F}(\{\psi_l\})\psi_k(\mathbf{x}) = \epsilon_k \psi_k(\mathbf{x}), \tag{2.15}$$

where ψ_l and ψ_k are two spin molecular orbitals for electron j and $\{\psi_l\}$ is the full set of orbitals ψ_l. The Fock operator $\hat{F}(i)$ given in Eq. 2.16 is composed of three main terms: firstly, the core Hamiltonian $\hat{h}(i)$, which includes the kinetic energy of electron i and its electrostatic attraction to the nuclei, secondly, the Coulomb operator $\hat{J}_j(i)$, representing the electrostatic repulsion between electrons i and j, and lastly the exchange operator $\hat{K}_j(i)$, which arises from the antisymmetry of the total wavefunction and accounts for the exchange between electrons with the same spin. Mathematically, this is expressed as

$$\hat{F}(i) = \hat{h}(i) + \sum_{j \neq i} \hat{J}_j(i) - \hat{K}_j(i). \tag{2.16}$$

The $\hat{J}_j(i)$ operator acts on molecular orbitals and produces $J_{k,l}$, the so-called direct term, as the mean Coulomb interaction between the i^{th} and j^{th} electrons, given by

$$J_{k,l} = \frac{e^2}{4\pi\epsilon_0} \left\langle \psi_k(\mathbf{x}_i)\psi_l(\mathbf{x}_j) \left| \frac{1}{|r_i - r_j|} \right| \psi_k(\mathbf{x}_i)\psi_l(\mathbf{x}_j) \right\rangle. \tag{2.17}$$

The $K_{k,l}$, the exchange term, also called Fock term, is

$$K_{k,l} = \frac{e^2}{4\pi\epsilon_0} \left\langle \psi_k(\mathbf{x}_i)\psi_l(\mathbf{x}_j) \left| \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \right| \psi_l(\mathbf{x}_i)\psi_k(\mathbf{x}_j) \right\rangle . \tag{2.18}$$

The core Hamiltonian $\hat{h}(i)$ is evaluated on single-particle Slater determinants, which produces $\sum_k I_k$ (the one-electron energy integral for orbital ψ_k) as the sum over all the occupied eigenstates k by the n electrons. Taking the following expression

$$\sum_{k=1}^{n} \langle \psi_k(\mathbf{x}_i) | \hat{h}_i | \psi_k(\mathbf{x}_i) \rangle = \langle \Psi_{HF} | \left| -\frac{\hbar^2}{2m_e} \sum_{i=1}^{n} \nabla_{r_i}^2 - \frac{e}{4\pi\epsilon_0} \sum_{A=1}^{M} \sum_{i=1}^{N_e} \frac{Z_A}{|r_i - R_A|} \right| \Psi_{HF} \rangle = \sum_k I_k \tag{2.19}$$

as well as Eq. 2.17 and Eq. 2.18 and plugging them into the Fock operator in Eq. 2.16, and using the Slater determinant orbitals (Eq. 2.14), while avoiding double counting of terms due to $J_{k,l} = J_{l,k}$ and $K_{k,l} = K_{l,k}$, the energy is obtained as

$$E[\Psi_{HF}] = \sum_k I_k + \frac{1}{2} \sum_{k,l} [J_{k,l} - K_{k,l}] . \tag{2.20}$$

Since the $\hat{J}$ and $\hat{K}$ operators depend on the molecular orbitals, the solution of the HF equation is obtained self-consistently as outlined below, starting with an initial guess for the molecular orbitals (see Section 2.2.1). Notably, HF is still a costly method as O(N^4) with the number of basis functions [67] due to the computation of direct and exchange Coulomb integrals in Eq. 2.17 and Eq. 2.18.

Computational procedure

The wavefunction for the i^{th} electron can be represented as a linear combination of atomic basis function Φ_W as shown in the equation,

$$\psi_i(\mathbf{r}) = \sum_W C_{Wi}\Phi_W(\mathbf{r}), \tag{2.21}$$

where there are multiple options for the choice of atomic basis set functions as presented in the following Subsection 2.2.1. From the Fock operator, with a given basis set, and an initial guess for the molecular orbitals (*e.g.*, atomic orbitals) due to the dependence of the MO in the Coulomb and exchange operators (see Eqs. 2.17 and 2.18) , a self-consistent field (SCF) procedure is employed to find the converged molecular orbitals and their energies. The SCF procedure involves iterative solving of the Roothaan equations - a generalized eigenvalue problem, due to the nonorthogonality of the basis set

$$\mathbf{FC} = \mathbf{SC}\epsilon, \tag{2.22}$$

where ϵ is the diagonal matrix of orbital energies and $\mathbf{S}$ is the overlap matrix quantifying the overlap of AOs. Thus the variational converged MOs and their energies are obtained

from the final solution to the diagonalised Fock matrix (representation of the Fock operator in a specific basis set), whose elements are presented as

$$F_{WV} = h_{WV} + \sum_{fg} P_{fg}(WV|fg) - \frac{1}{2}(Wf|Vg), \tag{2.23}$$

where χ_W and χ_V are the basis functions, f and g are the summation indices for building the electron density matrix P_{fg}; h_{WV} are the matrix elements of the $\hat{h}(i)$ operator; $(WV|fg)$ is the two-electron Coulomb integral, describing the classical electrostatic repulsion between electron densities; and $(Wf|Vg)$ is the exchange integral. Generally, Coulomb and exchange matrices creation is the most costly computational step of the SCF procedure, while for large systems, the diagonalization of the Fock matrix can also pose a notable computational overhead. The 1/2 term shown in Eq. 2.23 corresponds to the base form of HF, *i.e.*, restricted HF applicable to closed shell systems. The unrestricted HF addresses also open shell systems with unpaired electrons, but requires two separate Fock matrices, for alpha-spin and beta-spin electrons but is beyond the scope of this work.

The electron density matrix elements are given by P_{fg} as a sum over the occupied molecular orbitals (MOs):

$$P_{WV} = 2\sum_{i}^{\text{occ}} C_{Wi}C_{Vi}, \tag{2.24}$$

where C_{Wi} and C_{Vi} are the MO coefficients giving the relationship between the i^{th} molecular orbital ψ_i and the μ^{th} atomic basis function Φ_W.

2.2.1 Atomic Orbitals and Basis sets

The atomic orbitals from the basis sets function are approximations on the Hydrogen orbitals from quantum chemistry are given by

$$\Phi_{n\ell m}(r,\theta,\phi) = R_{n\ell}(r)\cdot Y_\ell^m(\theta,\phi), \tag{2.25}$$

where $R_{n\ell}(r)$ is probability of electron density as a function of the distance from the nucleus and $Y_\ell^m(\theta,\phi)$ is the spherical harmonics $Y_\ell^m(\theta,\phi)$ function of the angular part of the orbital as a function of the polar angle θ and azimuthal angle ϕ. The spherical harmonics are defined as

$$Y_\ell^m(\theta,\phi) = \sqrt{\frac{(2\ell+1)}{4\pi}\cdot\frac{(\ell-m)!}{(\ell+m)!}}\cdot P_\ell^m(\cos\theta)\cdot e^{im\phi}, \tag{2.26}$$

with P_l^m as the associated Legendre polynomial. Due to the computational overheads associated with integral evaluation, approximations on an atomic basis set are made for a finite discrete set of functions with three popular types: Slater-type orbitals (STOs), Gaussian-type orbitals (GTOs), and on numerically tabulated atom-centered orbitals (NAOs) , the latter two being more popular due to the computational expense of the Slater-type orbitals [78, 79].

GTOs have two widely used forms: primitive and contracted. The original, or 'primitive', GTOs are given by

$$\Phi(\mathbf{r}) = N \cdot x^a y^b z^c \cdot e^{-\alpha r^2}, \tag{2.27}$$

where $x^a y^b z^c$ is a Cartesian polynomial giving the angular momentum of the orbital dependent on a, b, and c, whose sum is the total angular momentum quantum number; α is the Gaussian exponent defining the spread of the orbital with inverse proportionality to localisation; and $e^{-\alpha r^2}$ is the radial part of the Gaussian function dependent on the radial distance r. The improved contracted GTOs are a linear combination of primitive GTOs with corresponding coefficients d_i.

$$\Phi(\mathbf{r}) = \sum_{i=1}^{N} d_i \cdot \chi_i(\mathbf{r}) = \sum_{i=1}^{N} d_i \cdot x^a y^b z^c \cdot e^{-\alpha_i r^2} \tag{2.28}$$

The coefficients of the GTOs have been fitted empirically or by a set of rules depending on the basis set type, *e.g.*, to match Slater-type orbitals for atoms (such as STO-nG []), to reproduce the energy, to capture core and diffuse valence orbitals behaviour, or to converge post-Hartree–Fock calculations systematically to the complete basis set (CBS) limit using empirical extrapolation techniques (the Dunning basis sets [80]). Most highly accurate basis sets in current use combine multiple features, *e.g.*, aug-cc-pVTZ [81], and aug-cc-pV(X+d)Z [82] from the aug- (augmented correlation-consistent basis sets) focus on the convergence to the complete basis sets but also include diffuse valence orbital behaviour.

The NAOs used in the FHI-aims electronic structure code [83], are a type of basis function of the form

$$\Phi(\mathbf{r}) = \frac{\mu(r)}{r} Y_\ell^m(\theta, \phi). \tag{2.29}$$

Unlike the commonly used GTOs or STOs, NAOs are not defined by a simple analytical function. Instead, their radial part $\mu(r)$ is calculated numerically for a free atom and then tabulated, providing a highly accurate and flexible representation of the orbitals, especially near the nucleus where the electron density changes rapidly. This all-electron approach, which avoids pseudopotentials, allows for a precise description of all electrons from light to heavy elements while remaining computationally efficient due to the strictly localized nature of the basis functions. The NAOs are hierarchically organized into "tiers" (e.g., light, tight, really_tight) to allow for a systematic convergence of calculations from fast, qualitative results to meV-level accuracy.

2.3 Density Functional Theory

Density Functional Theory (DFT) was developed around the core idea of simulating chemical systems at a QM level as functions of their electronic density ρ rather than by their wavefunctions, thus reducing the degrees of freedom from $3n$ to 3. This approach was first proposed by Hohenberg and Kohn [84], and generalised into a practically applicable

self-consistent method by Kohn and Sham [85], which has gone onto becoming one of the most widely used theoretical methods to describe molecular and solid-state systems at the quantum level.

The foundations of DFT are based on two theorems by Hohenberg and Kohn [84]. The first theorem states that for any system of interacting electrons in an external potential $v_{\text{Ne}}(\mathbf{r})$, that contains the interactions between the nuclei and the electrons, the potential energy is uniquely determined (up to a constant) by the electron density $\rho(\mathbf{r})$:

$$V_{Ne}[\rho] = \int \rho(\mathbf{r}) v_{Ne}(\mathbf{r})\, d\mathbf{r}. \tag{2.30}$$

The second Hohenberg-Kohn theorem [84] states that exists a universal functional $F[\rho]$ such that the minimum of the total energy functional in ground state $E_0[\rho]$ and the ground state electron density $\rho_0(\mathbf{r})$ can be obtained via variational minimisation of the equation

$$E_0[\rho] = F[\rho] + V_{ne}[\rho], \tag{2.31}$$

given that $\int_V \rho(\mathbf{r}) d\mathbf{r} = n$.

Based on these proven assumptions, Kohn and Sham [85] further developed this approach for practical calculations, forming the Kohn–Sham DFT (KS-DFT) approach by rewriting the density as a sum over the square modulus of single-particle molecular orbitals $\psi_k^{KS}(\mathbf{r})$ times their occupation number f_k

$$\rho(\mathbf{r}) = \sum_{k=1}^{\text{occ}} f_k |\psi_k^{KS}(\mathbf{r})|^2 \tag{2.32}$$

and obtaining the set of equations

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_{\text{Ne}}(\mathbf{r}) + V_{\text{H}}(\mathbf{r}) + V_{\text{xc}}(\mathbf{r}) \right] \psi_k^{KS}(\mathbf{r}) = \varepsilon_k \psi_k^{KS}(\mathbf{r}) \tag{2.33}$$

that can be solved self-consistently, where ε_k are the energies associated to the $\psi_k^{KS}(\mathbf{r})$ Kohn–Sham orbitals.

The external potential $V_{\text{Ne}}(\mathbf{r})$ represents the interaction of the electrons with the nuclei, the Hartree potential V_{H} accounts for the classical electrostatic repulsion between electrons, and the exchange–correlation potential $V_{\text{xc}}(\mathbf{r})$ incorporates all many-body effects. The explicit form of the Kohn-Sham equations are written as follows

$$\left[-\frac{\hbar^2}{2m_e}\nabla^2 - \sum_A \frac{Z_A e^2}{4\pi\varepsilon_0 |\mathbf{r} - \mathbf{R}_A|} + \int \frac{\rho(\mathbf{r}')}{4\pi\varepsilon_0 |\mathbf{r} - \mathbf{r}'|}\, d\mathbf{r}' + \frac{\delta E_{\text{XC}}[\rho]}{\delta\rho(\mathbf{r})} \right] \psi_k(\mathbf{r}) = \varepsilon_k \psi_k(\mathbf{r}) \quad \forall \quad k, \tag{2.34}$$

where E_{xc} is the exchange–correlation energy, and as above Z_A is the atomic number and $\mathbf{R}_A$ the position of nucleus A, $\mathbf{r}$ is the position of the electron. Ultimately, the overall energy is split into the sum of contributions

$$E[\rho] = T_s[\rho] + V_{\text{ext}}[\rho] + E_{\text{Hartree}}[\rho] + E_{\text{xc}}[\rho], \tag{2.35}$$

where in particular, $T_s[\rho]$ is the non-interacting kinetic energy and $E_{\text{xc}}[\rho]$ is the unknown exchange-correlation functional that contains the many-body effects due to exchange and electron-electron correlations. Due to the unknown exact form of the exchange-correlation function, many approximated forms have been introduced in the past decades, and their development is still an active and important field of research.

A popular categorisation of for the ensemble of functionals, dating back to 2001, is Jacob's ladder [86], in which they are grouped in four main categories: Local Density Approximation (LDA), Generalized Gradient Approximation (GGA), Meta-GGA, Hybrid, Double hybrid (including range-separated double hybrid) functionals, according to their accuracy, moving from the 'Hartree hell' towards the 'chemical accuracy heaven', *i.e.*, within 1 kcal/mol deviation in predictions [86, 87].

In the following sections, the main characteristics of the categories of XC functionals used in this thesis are discussed.

2.3.1 Exchange-Correlation Functionals

LDA [85] is the simplest approximation used to construct XC functionals and is based on the exchange between electrons in the uniform electron gas, which works reasonably well in solid state physics. Mathematically, it is defined as follows

$$E_x^{\text{LDA}}[\rho] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} \int \rho^{4/3}\, d\mathbf{r}. \tag{2.36}$$

Since the correlation functional could not be derived from first principles, its parameters were determined by fitting to results from Monte Carlo simulations of a uniform electron gas. This approach allowed for accurate interpolation between the known high- and low-density behaviours. A more advanced approximation is that of the GGA class of functionals, which also takes into account the gradient of the electron density, improving the estimation of the interaction energies in molecular systems of about one order of magnitude [87]. Some of the most successful and now commonly used GGA functionals are the PBE (Perdew–Burke–Ernzerhof) [88] and BLYP (Becke–Lee–Yang–Parr) [89, 90] functionals, which also serve as the basis for the more advanced class of XC functionals and are defined as follows

$$E_x^{GGA}[\rho] = \int \rho^{4/3} F(x)\, dr, \tag{2.37}$$

where F_{x} is an enhancement factor depending on the reduced density gradient s and constants, with F_{x} chosen to obey gradient expansion for uniform electron gas. Some functionals, like PBE, have a non-empirically derived form

$$E_{\text{x}}^{\text{PBE}}[\rho] = \int \rho(\mathbf{r})\, \varepsilon_{\text{x}}^{\text{LDA}}(\rho(\mathbf{r}))\, F_{\text{x}}(s(\mathbf{r}))\, d\mathbf{r}, \tag{2.38}$$

where $F_{\text{x}}(s)$ is a function of the reduced density gradient s and constants, whose values are derived from theoretical constraints of the behaviour of the exchange-correlation energy.

The reduced density gradient s is defined as a function of the local Fermi wavevector from the uniform electron gas model k_{F} as

$$s(\mathbf{r}) = \frac{|\nabla\rho(\mathbf{r})|}{2k_{\mathrm{F}}(\mathbf{r})\rho(\mathbf{r})}, \tag{2.39}$$

where $k_{\mathrm{F}}(\mathbf{r}) = (3\pi^2\rho(\mathbf{r}))^{1/3}$ is defined as the maximum momentum of electrons at 0K. The theoretical constrains informing the constants in F_{x} include, among others, the Lieb-Oxford inequality [91, 92] for a lower bound on the exchange-correlation energy, the bounds on behaviour at very large or very small $s(\mathbf{r})$ values. Two well-established GGA correlation functionals are PBE and LYP. The PBE functional is non-empirical and designed to satisfy fundamental constraints, while the LYP functional is based on the Colle-Salvetti correlation energy expression, derived from a correlated wave function for a two-electron system.

The next step in the Jacob's ladder is the meta-GGA class of functionals build upon GGA ones by the addition of the kinetic energy density term incorporating information about the local orbital structure and thus providing enough information to distinguish different type of bonding, however the meta-GGA functionals are beyond the scope of this thesis.

Hybrid functionals

Hybrid functionals have also the addition in the functional of a term that contributes with HF exchange, as represented in

$$E_x^{\mathrm{HF}} = -\frac{1}{2}\sum_{kj\sigma}\iint \frac{\psi^*_{k\sigma}(\mathbf{r})\psi_{j\sigma}(\mathbf{r})\psi^*_{j\sigma}(\mathbf{r}')\psi_{k\sigma}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,d\mathbf{r}\,d\mathbf{r}'. \tag{2.40}$$

The idea to add HF exchange into the DFT functional originated with the Brecke linear model [93] that mixes in local DFA exchange and correlation functionals, resulting in the creation of the BHLYP class of functionals *e.g.*, BH&HLYP

$$E_{xc}^{BHH} = \frac{1}{2}E_x^{HF} + \frac{1}{2}W_1^{LDA}. \tag{2.41}$$

There are different ways in which hybrid functionals have been designed, and in this work focus on two relevant creation models for hybrid functionals - fitting to data and range-separation. One of the historically most popular functionals in computational Chemistry, B3LYP, which was devised by fitting to the experimental dataset 'Gaussian' [94, 95] and is given by the formula

$$E_{xc}^{\mathrm{B3LYP}} = 0.2E_x^{\mathrm{HF}} + 0.8E_x^{\mathrm{LDA}} + 0.72\Delta E_x^{\mathrm{B88}} + 0.81E_c^{\mathrm{LYP}} + 0.19E_c^{\mathrm{VWN}}, \tag{2.42}$$

where contributions with empirically obtained weights from the aforementioned HF exchange, LYP and LDA functionals, as well as such from the VWN (Vosko-Wilk-Nusair [96]) functional and the gradient correlation to exchange from Becke 1988 [97] are summed. Attempts to create better functionals based on better constraint satisfaction and better

parametrisation on more extensive datasets have given rise to the Minnesota functionals. Among them, M06-2X [98] with a high proportion of HF exchange (over 50%) is the best from the 06 family of Minnesota functionals [15] for organic molecules thermochemistry, kinetics and non-covalent interactions.

The PBE0 hybrid functional obtained from the correlation of the non-empirical PBE GGA functional, and its exchange term in 3 to 1 ratio (from experimental data fit [99]) to the HF exchange

$$E_{\text{xc}}^{\text{PBE0}} = \frac{1}{4}E_{\text{x}}^{\text{HF}} + \frac{3}{4}E_{\text{x}}^{\text{PBE}} + E_{\text{c}}^{\text{PBE}}. \tag{2.43}$$

The second relevant general approach to building global hybrids to be discussed here is range separation. All the functionals presented above are considered 'global hybrids' for using globally the exchange contribution from HF as given in Eq. 2.40. The range-separation idea relies on the split of electron-electron interactions in two parts, with the split governed by a single parameter μ

$$\frac{1}{r_{ij}} \equiv \underbrace{\frac{\text{erf}(\mu r_{ij})}{r_{ij}}}_{\text{long-range } U^{\text{lr}}(r_{ij},\mu)} + \underbrace{\frac{\text{erfc}(\mu r_{ij})}{r_{ij}}}_{\text{short-range } U^{\text{sr}}(r_{ij},\mu)}, \tag{2.44}$$

where the long-range part is given by

$$E_x^{\text{lr-HF}} = -\frac{1}{2}\sum_{ij\sigma}\iint \frac{\psi_{i\sigma}^*(\mathbf{r}_1)\psi_{j\sigma}(\mathbf{r}_1)\,\text{erf}(\mu r_{12})\,\psi_{j\sigma}^*(\mathbf{r}_2)\psi_{i\sigma}(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|}\, d\mathbf{r}_1\, d\mathbf{r}_2. \tag{2.45}$$

Examples of range-separated functionals include ωB97X [100] built on GGA contribution for the exchange, with range-separation at $0.3 Bohr^{-1}$ and 10 adjustable parameters fitted on datasets focused on thermochemistry, kinetics, and non-covalent interactions; the more computationally demanding ωB97M [101] built on meta-GGA for the exchange, with 12 adjustable parameters, range-separation optimised during the fitting on the GMTKN55 dataset [55], and includes built-in dispersion correction VV10 nonlocal [102]; and CAM-B3LYP [103] built on B3LYP with 2 empirical parameters fitted on datasets focused on excitation energies and charge transfer processes.

On a computational note, Resolution-of-identity is a popular computational scheme using auxiliary atomic orbitals or numeric atom-centered orbitals basis to reduce expensive four-center two-electron integrals to more efficient three- and two-center integrals, and works for DFT, HF, MP2, etc. [104]

Double Hybrid functionals

Double Hybrid functionals aim to improve on hybrid functionals through the combination of contributions from both DFT and wavefunction-based methods *e.g.*, second-order perturbation theory (MP2-like) correlation. The form the double hybrids is

$$E_{XC} = a_x E_x^{\text{HF}} + (1 - a_x)E_x^{\text{DFT}} + (1 - a_c)E_c^{\text{DFT}} + a_c E_c^{\text{MP2}}. \tag{2.46}$$

The Møller–Plesset perturbation theory of second order (MP2) [105], is a perturbative post-HF quantum chemistry approach that builds on HF by adding electron correlation contributions ($E^{(2)}$ term), which arise from including the correlation potential as a perturbation, corresponding to the difference between the full interacting Hamiltonian and Fock operator. This results in higher computational cost with scaling of $O(N^5)$ with the number of basis functions. MP2 method uses HF orbitals without electron correlation and misses higher order correlation effects like those in Coupled Cluster (see Appendix A), which results in a tendency to overcorrect and overestimate the dispersion interactions - particularly relevant in non-covalent complexes calculations [106]. Mathematically, the MP2 energy can be expressed as

$$E_{\text{MP2}} = E_{\text{HF}} + E^{(2)} = E_{\text{HF}} + \sum_{k<j}^{\text{occ}} \sum_{a<b}^{\text{virt}} \frac{|\langle kj||ab\rangle|^2}{\varepsilon_k + \varepsilon_j - \varepsilon_a - \varepsilon_b}, \tag{2.47}$$

where k and j are the occupied MOs indices, whereas a and b are those of the unoccupied MOs; ε_k, ε_j, ε_a, and ε_b are the energies of the corresponding orbitals, and $\langle kj||ab\rangle$ gives the antisymmetrised two-electron integral ($\langle kj|ab\rangle - \langle kj|ba\rangle$, where $\langle kj|ab\rangle$ is the standard Coulomb integral for the electron repulsion between orbitals k, j and a, b).

An example of a double hybrid functional is PBE-QIDH [107], for which the Eq. 2.46 has as DFT choice the PBE functional, and the coefficients for HF and MP2 correspondingly 0.7 and 0.33.

2.3.2 van der Waals dispersion models

Despite the great success of KS-DFT, due to the incredible accuracy reached in the construction of increasingly more accurate exchange-correlation functionals, these still struggle in describing correctly vdW interactions due to the fact that they essentially are local or semi-local operators, despite the inclusion of HF exchange in Hybrid functionals. VdW interactions arise from the collaboration of electrostatic, polarisation (or induction), dispersion, and exchange-repulsion contributions [108] and are responsible for the long-range electronic correlation effects, whose description is crucial for describing binding energies. Therefore, in order to improve the description of these energy contributions in KS-DFT, many different formulations have been introduced as additive energetic corrections to KS-DFT functionals without those effects, including exchange-hole dipole moment (XDM) [109, 110], Tkatchenko-Scheffler (TS) [111], Grimme's D3 and D4 corrections [112, 113], and Many-Body Dispersion (MBD) [59, 60], including MBD-Non-Local (MBD-NL) [114], among many others. In the next subsections, we will briefly describe the main approaches compared in this thesis, highlighting the main characteristics and differences.

.

Casimir-Polder formalism for van der Waals interactions

The descriptions of the dispersion vdW interactions of many, though not all, dispersion correction methods in the following subsections rely on the conceptual framework of the Casimir-Polder formalism [115]. It can be seen as a perturbative method to compute the correlation energy between two (generally non-overlapping) charge distributions, *i.e.*, the energy shift in the ground state energy due to the multipolar electrostatic coupling between two quantum charge distributions w.r.t the non-interacting case. The approach employs second-order perturbation theory applied to the multipolar coupling. This results in the generation of instantaneous virtual excitations of the two systems, such that a transition multipolar moment of the charge distribution is generated in each system. Consequently, the dispersion energy can be expressed in terms of a function of the distance between the charge distributions and of the so-called dispersion coefficient C_n, which depends on the integral of the imaginary part of the multipolar dynamic polarisabilities of the two systems in the non-interacting ground state.

The interaction energy of the interacting A-B system is obtained as a function of the dispersion coefficients as follows

$$E(R) = -\frac{C_6}{R^6} - \frac{C_8}{R^8} - \frac{C_{10}}{R^{10}} - \cdots , \tag{2.48}$$

where R is the distance between the systems A and B.

In general, the dispersion coefficients can be expressed as a functional of the multipole dynamic polarizabilities: the ability of an atomic electron cloud to respond to the instantaneous dipole created by a neighbouring atom. The C_6 coefficient depends, for instance, on the dynamical dipole polarizability of both interacting systems. Here, we must recall that the dynamic dipole polarizability $\alpha_{A,ij}(\boldsymbol{\omega})$ of the A-th quantum charge distribution is defined via its dipole operator $\hat{\boldsymbol{d}}_A$. Namely, for a quantum system in equilibrium state, being ρ_0 its unperturbed density, the induced dipole due to the presence of an external stationary electric field $\delta\boldsymbol{E}(\omega)$, that perturbs ρ_0, is given by $\delta\langle\boldsymbol{d}(\omega)\rangle_{\rho_0} = \langle\boldsymbol{d}(\omega)\rangle_{\rho} - \langle\boldsymbol{d}(\omega)\rangle_{\rho_0}$. The matrix elements of the dynamic polarizability tensor $\boldsymbol{\alpha}(\omega)$ of this system are defined as

$$\alpha_{ij}(\omega) = \lim_{|\delta E|\to 0} \frac{\delta\langle d_i(\omega)\rangle_{\rho_0}}{\delta E_j(\omega)}. \tag{2.49}$$

For spherically symmetric charge distributions, the relevant quantity is the isotropic polarisability $\alpha(\omega)$ as a measure of the degree of malleability of the electron distribution of the system to an external potential, obtained as an average over the diagonal terms of the dynamic polarizability tensor as

$$\alpha_A(\omega) = \frac{1}{3}\,\mathrm{Tr}[\boldsymbol{\alpha}_A(\omega)] = \frac{1}{3}\sum_{i=x,y,z} \alpha_{A,ii}(\omega)\,. \tag{2.50}$$

Given these definitions, the dispersion coefficient C_6 is computed from Casimir-Polder integrals over the frequencies of products of the imaginary part of the dynamic dipole polar-

izabilities α of the atoms that are computed, *i.e.*,

$$C_6^{AB} = \frac{3}{\pi} \int_0^\infty \alpha_A(i\omega)\, \alpha_B(i\omega)\, d\omega. \tag{2.51}$$

These relations can be generalized for higher-order dispersion coefficients that will depend on the higher-order multipole polarizabilities.

Exchange Dipole Model

The eXchange Dipole Model (XDM), also known as eXchange-hole Dipole Moment [109, 116] dispersion correction, starts from the consideration of the idea of holes existing with each electron due to the Pauli principle. XDM derives an attractive dispersion energy non-empirically from the interaction between instantaneous dipole moments from the exchange-holes obtained from the partition of the density in atomic fragments. XDM is thus obtained as an electron density-dependent correction derived as a function of the dispersion coefficient C_6 [109]. For a molecular system, the XDM dispersion correction energy is given by

$$E_{\text{XDM}} = \sum_{A>B} E_{AB} = -\sum_{A>B} \frac{C_{6,AB}}{R_{AB}^6 + \kappa C_{6,AB}/(E_A^C + E_B^C)}, \tag{2.52}$$

where R_{AB} is the distance between two atoms A and B; the second term in the denominator is the damping factor that avoids divergences as the two atoms come close ($1/R^6$); E_X^C is the correlation energy of the free atom $X = A, B$; and κ is a dimensionless universal empirical parameter. A and B are atomic indices. It is notable, however, that in practice, the XDM dispersion coefficients are obtained using a different approach to the Casimir-Polder integral, *i.e.*, it is built on a practical atom-in-molecule approach for the polarisability. The atomic polarisability is given by α_i, with the sum of the atomic contributions recovering the isotropic molecular polarisability $\alpha(\omega)$. A Hirshfeld charge partitioning scheme is employed, where $\omega_i(r)$ is the Hirshfeld partition of atom i and $V_i = \int \omega_i(\mathbf{r})\, \rho(\mathbf{r})\, d\mathbf{r}$ is the Hirshfeld effective atomic volume in the molecule (correspondingly V_i^{free} denotes the atomic volume of the free atom - both defined as functions of $\rho(\mathbf{r})$, the molecular electron density and ρ_i the spherically symmetric reference density of free atom i, respectively). Then, α_i is obtained by rescaling free-atom reference polarisabilities of the free atoms α_i^{free} by the Hirshfeld atomic volumes as follows

$$\alpha_i = \alpha_i^{\text{free}} \left(\frac{V_i}{V_i^{\text{free}}} \right). \tag{2.53}$$

The second ingredient in the XDM formalism is the multipole moment M_l of the exchange hole, where l=1 corresponds to dipole moment, n=2 to quadrupole moment, etc. defined as a sum of integrals for spin up and down depending on the spin σ. The expectation value of the multipole moment M_l is used as a measure of the strength of the multipole exchange-hole couplings and is defined as follows

$$\langle M_\ell^2 \rangle_i = \sum_\sigma \int \omega_i(\mathbf{r})\, \rho_\sigma(\mathbf{r}) \left| \mathbf{r}_i^\ell - \left(\mathbf{r}_i - \mathbf{d}_{X,\sigma}(\mathbf{r})\right)^\ell \right|^2 d\mathbf{r}, \tag{2.54}$$

where the exchange-hole dipole vector $\mathbf{d}_{X,\sigma}(\mathbf{r})\big)^{\ell}$ is the result of the displacement of center of the hole charge from $\mathbf{r}$. The dispersion coefficients are obtained as follows

$$C_{6,ij} = \frac{\alpha_i \alpha_j \langle M_1^2 \rangle_i \langle M_1^2 \rangle_j}{\langle M_1^2 \rangle_i \alpha_j + \langle M_1^2 \rangle_j \alpha_i}. \tag{2.55}$$

A specialised version of XDM dependent on C_6, C_8, and C_{10} is beyond the scope of this thesis.

Semiempirical DFT-D3 and DFT-D4 long-range dispersion corrections

Let us now consider the DFT-D [113, 117, 118] dispersion correction scheme, which started with D2 and has continued with the currently relevant D3 and D4 to include 2 and 3-body terms. A distinguishing feature of this dispersion correction model is the use of an approximation for the dispersion as a sum of additive effective atom–atom interactions within their chemical environment, and the use of empirical functions and parameterisation on reference data. In particular, the D3 correction [112] is written as the sum of two-body $E^{(2)}$ and three-body $E^{(3)}$ contributions

$$E_{\text{D3}} = E^{(2)} + E^{(3)}. \tag{2.56}$$

The 2-body term is dependent on the C_6 and C_8 coefficients via the equation

$$E^{(2)} = -\sum_{A}^{\text{atoms}} \sum_{B<A}^{\text{atoms}} \left[s_6 \left(\frac{C_{6,AB}}{R_{AB}^6} \right) f_{\text{damp},6}(R_{AB}) + s_8 \left(\frac{C_{8,AB}}{R_{AB}^8} \right) f_{\text{damp},8}(R_{AB}), \right] \tag{2.57}$$

where the parameters s_n are global scaling factors specific to the chosen XC functional, to which they are applied, and the $f_{\text{damp,X}}$ damping functions are chosen to smoothly reduce the dispersion energy at short interatomic distances avoiding unphysical divergences. A typical choice for these functions is the form proposed by Brecke and Johnson [119] given as follows

$$f_{damp,n}(R_{AB}) = \frac{s_n \, R_{AB}^n}{R_{AB}^n + (a_1 R_{0,AB} + a_2)^n}, \tag{2.58}$$

where $R_{0,AB}$ is the reference distance between van der Waals radii and a_1 and a_2 are constants fitted to the DFT functionals. Finally, the three-body $E^{(3)}$ term corresponds to the Axilrod–Teller–Muto (ATM) [120] three-body interaction

$$E^{(3)} = \frac{C_9^{ABC}(3\cos\theta_a \cos\theta_b \cos\theta_c + 1)}{(r_{AB} r_{BC} r_{CA})^3}, \tag{2.59}$$

where we consider A, B, and C are three interacting atoms forming a triangle with θ_a, θ_b, and θ_c being their relative angles and r_{AB}, r_{BC} and r_{CA} their pairwise distances. The dispersion coefficient C_9^{ABC} for three bodies, differs from that of the two-body terms and is approximated as

$$C_9^{ABC} = \frac{3}{\pi} \int_0^{\infty} \alpha^A(i\omega)\alpha^B(i\omega)\alpha^C(i\omega) d\omega \approx -\sqrt{C_6^{AB} C_6^{AC} C_6^{BC}} \tag{2.60}$$

for practical computation, and

$$C_{6,\text{ref}}^{AB} = \frac{3}{\pi}\int_0^\infty \frac{1}{m}\left[\alpha^{A_m,H_n}(i\omega) - \frac{n}{2}\alpha^{H_2}(i\omega)\right] \times \frac{1}{k}\left[\alpha^{B_k,H_1}(i\omega) - \frac{l}{2}\alpha^{H_2}(i\omega)\right] d\omega. \quad (2.61)$$

Here, A_mH_n and B_kH_l are the reference compounds with atoms A and B having a specific coordination number. In practice, the above integral is estimated numerically over a fixed number of points in terms of effective atomic polarizabilities, and then C_6^{AB} is finally estimated by rescaling those pre-stored $C_{6,\text{ref}}^{AB}$ based on the atomic coordination numbers in the target system to effectively account for the chemical environment. The last generation in the DFT-D framework, the D4 dispersion correction is obtained by adding to the dispersion energy the four-body contribution as follows

$$E_{\text{disp}}^{\text{D4}} = -\sum_{A<B}\sum_{n=6,8} s_n \frac{C_n^{AB}(q)}{R_{AB}^n + f_{\text{damp},n}(R_{AB})} - s_9 \sum_{A<B<C} \frac{C_9^{ABC}(q)}{(R_{AB}R_{BC}R_{CA})^3} f_{\text{damp},9}, \quad (2.62)$$

in which the dispersion coefficients are dependent on partial charges, used to better capture the local electronic environment further improving with respect to the D3 method, especially for densely packed systems which becomes significant in densely packed systems in densely packed systems [113, 118].

Tkatchenko-Scheffler and Many-Body Dispersion method

The MBD correction captures, as the name would suggest, collective many-body correlation effects in non-covalently interacting systems. Like its pairwise predecessor, the TS pairwise method, MBD is also based on Hirshfeld partitioning of the atomic density to compute effective atomic polarisability in molecules. Thus, let us first focus briefly on describing the latter.

The Hirshfeld effective volume V_A of each atom A in a molecular system is defined as

$$V_A = \int \frac{\rho_A^0(\mathbf{r})}{\sum_B \rho_B^0(\mathbf{r})} \cdot \rho(\mathbf{r})\, d\mathbf{r}, \quad (2.63)$$

where $\rho_A^0(\mathbf{r})$ is the spherically averaged electron density of atom A in its free state (weighted over a sum of the spherically averaged electron densities of the other atoms, which quantifies the fractional contribution of atom A to the total reference density at each point in space). Using the factor $\chi_A = \frac{V_A}{V_A^0}$ to rescale the atomic polarisabilities $\alpha_A = \chi_A \cdot \alpha_A^0$ and $C_6^{AA} = \chi_A^2 \cdot C_6^{AA,0}$ coefficients, the dispersion correction is written as

$$E_{\text{disp}}^{\text{TS}} = -\sum_{A<B}\left[\frac{2\chi_A^2 C_6^{AA,0}\cdot\chi_B^2 C_6^{BB,0}}{\left(\frac{\chi_B\alpha_B^0}{\chi_A\alpha_A^0}\cdot\chi_A^2 C_6^{AA,0} + \frac{\chi_A\alpha_A^0}{\chi_B\alpha_B^0}\cdot\chi_B^2 C_6^{BB,0}\right)}\right] \cdot \frac{1}{R_{AB}^6} \cdot f_{\text{damp}}(R_{AB}), \quad (2.64)$$

including only pairwise interactions between atomic fragments.

In MBD, on the other hand, the many-body correlation effects are modelled via a physics-motivated framework based on quantum harmonic oscillators (QHOs). Specifically, in MBD the electric response property of every atom-in-molecule is modelled as a QHO with rescalable parameters q_A, m_A, and ω_A, representing the charge, mass, and frequency of the QHO, respectively. The imaginary part of the dynamic QHO polarisability $\alpha(i\nu)$ is directly used in the QHO Hamiltonian to compute the dispersion interaction, obtained as follows

$$\alpha_A^{QHO}(i\nu) = \frac{q_A^2}{m_A(\omega_A^2 + \nu^2)}, \tag{2.65}$$

where ν is a frequency variable. The MBD Hamiltonian for the system of N coupled QHOs, in the quantum Drude oscillators (QDOs) displacements representation, is then given as follows

$$\hat{H}_{\text{MBD}} = \frac{1}{2}\sum_{A=1}^{N} \nabla_{\boldsymbol{\xi}_A}^2 + \frac{1}{2}\sum_{A=1}^{N} \omega_A^2 |\boldsymbol{\xi}_A|^2 + \frac{1}{2}\sum_{A=1}^{N}\sum_{B=1}^{N} \omega_A\omega_B (\alpha_A(0)\alpha_B(0))^{1/2}\, \boldsymbol{\xi}_A^T T_{AB} \boldsymbol{\xi}_B, \tag{2.66}$$

where $\boldsymbol{\xi}_X = \sqrt{m_X}\,(\mathbf{r}_X - \mathbf{R}_X)$ is the mass-weighted offset of the Xth QHO from its center $\mathbf{R}_X$; $\alpha_X(0)$ is its static isotropic polarisability; and $T_{AB}(\boldsymbol{R}_{AB})$ is the 3×3 dipole-dipole effective interaction matrix between the QHOs A and B. The dipole-dipole coupling referred to here is an effective one as the $T_{AB}(\boldsymbol{R}_{AB})$ tensor generally accounts for the local screening effects and is thus represented as a damped dipole interaction tensor parametrised through single-oscillator properties. The functional form of the damping is tied to a specific flavour of the MBD model, while variations in its parameters are typically introduced to better adapt a given MBD Hamiltonian to the reference exchange–correlation DFT functional.

The MBD Hamiltonian can be re-written with all displacement vectors compacted into a $3N$-dimensional vector $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N)$. The second and third term in the MBD Hamltonian, which represent the potential energy, can be defined as $\frac{1}{2}\,\boldsymbol{\xi}^T V \boldsymbol{\xi}$, where V is the effective potential matrix of size $3N \times 3N$, which can be separated into two distinct contributions as follows

$$V = \Omega^2 + C, \tag{2.67}$$

where $\Omega^2 = \text{diag}(\omega_1^2 I_3,\ \omega_2^2 I_3,\ \ldots,\ \omega_N^2 I_3)$ is a block-diagonal matrix containing the bare restoring forces of the uncoupled QHOs and thus encoding the intrinsic response of each QHO, while C is a matrix of the dipole-dipole couplings between different QHOs and thus accounting for the intermolecular couplings. The off-diagonal 3×3 blocks of C are given by a dispersion coefficient defined as

$$C_{AB} = \omega_A\omega_B\,(\alpha_A(0)\alpha_B(0))^{1/2}\, T_{AB}(\boldsymbol{R}_{AB}). \tag{2.68}$$

V can be diagonalised by a rotation into a new coordinate system of the MBD normal modes, where the couplings go to zero. To that end, an orthogonal matrix U is found that diagonalises V, such that $U^T V U = \text{diag}(\tilde{\omega}_1^2, \ldots, \tilde{\omega}_{3N}^2)$. Thus, the displacement coordinates $\boldsymbol{\xi}$ are transformed into into new coordinates $\boldsymbol{\zeta} = U^T\boldsymbol{\xi}$, for which the potential energy can

be written as the sum of the potential energies of $3N$ independent QHOs (the MBD normal modes), each with characteristic frequency $\tilde{\omega}_p = \tilde{\omega}_p(\boldsymbol{R}_1, ..., \boldsymbol{R}_N)$ [121].

Finally, the MBD dispersion energy E_{MBD}, providing an estimate of the correlation energy E_c is obtained as the difference between the zero-point energy of the interacting and non-interacting system of QDOs, *i.e.*, in atomic units. It is expressed as

$$E_{\mathrm{MBD}} = \frac{1}{2}\left(\sum_{p=1}^{3N}\tilde{\omega}_p - 3\sum_{A=1}^{N}\omega_A\right) = \frac{1}{2}\left(\mathrm{Tr}[\sqrt{V}] - \mathrm{Tr}[\Omega]\right). \tag{2.69}$$

Notably, the dispersion energy $E_{\mathrm{MBD}} = E_{\mathrm{MBD}}(\boldsymbol{R}_1, ..., \boldsymbol{R}_N)$ is a highly non-linear function of atomic center coordinates that is generally not separable as a sum of only two- or three-body contributions. In fact, the MBD contribution is highly non-local and depends on the global geometry QDOs arrangement: the force acting on the A-th atomic nucleus can depend on the relative configuration of any set of $M \leq N-1$ atomic nuclei not containing A and coupled to them. From a computational point of view, the MBD model allows for a reduction of the highly complex problem of the many-body correlation energy estimate. It does so via a diagonalisation of a dense $3N \times 3N$ matrix, once a post-DFT procedure has been defined to fix the parameter of the MBD Hamiltonian.

A key factor in the success of the MBD method for quantum chemistry applications is the approach of fixing the QDO parameters, $\alpha_X(0)$ and ω_X, together with the specification of the short-range damping functional form of the dipole–dipole tensor T_{AB} for a given MBD scheme. The primary version employed throughout this work is the MBD@rsSCS approach, where rsSCS denotes the range-separated Self-Consistent Sreening procedure to fix the QDOs parameters [60], combined with a Fermi damping of the dipole–dipole interaction. A single range-separation parameter β controls the damping: it sets how rapidly (and at what distance scale) the short-range part of T_{AB} is suppressed and the long-range MBD coupling is recovered; generally β is fitted to a DFT functional [122] or to a DFTB set of parameters like 3ob [123](*vide infra*) on a benchmark set such as the S66×8 [54].

Another relevant MBD variant is MBD-NL [114], which was designed to be broadly applicable to more molecular and materials systems, including ionic and metallic compounds, and hybrid metal-organic interfaces. It starts from a VV-style (Vydrov–Van Voorhis [102]) semi-local, electron density-based functional, and its gradient, to model local dynamic polarisability, instead of using the MBD effective atomic volumes. Then, the Hirshfeld partitioning is applied to the continuous polarisability field to obtain atom-centred dynamic polarisabilities that define a set of QDOs. Similarly to MBD, the oscillators are coupled through a long-range dipole interaction (with smooth Fermi-type damping at short range), and diagonalising the coupled-oscillator Hamiltonian yields the MBD-NL energy contribution. To prevent double counting with semi-local exchange–correlation in regions of slowly-varying electron density and gradient. Firstly, it renormalises the atom-resolved polarisabilities and pair coefficients to free-atom reference values [124], guaranteeing correct asymptotic behaviour of the dispersion interactions. Secondly, it smoothly suppresses

vdW energy contributions from near-uniform, metal-like, electron gas density regions identified by local descriptors such as a density-based ionisation potential and the iso-orbital indicator, so that only truly non-local correlations are retained. All the above lead to a more consistent and robust approach for describing vdW interactions in a broader range of systems than previously possible.

2.4 Semi-empirical approaches

The semi-empirical approaches are a class of methods using different approximations to combine the inclusion of some QM effects with the computational efficiency needed for large-scale molecular simulations on the order of a few thousand atoms [125] where more accurate quantum Chemistry methods become prohibitively expensive. The semi-empirical methods simplify the electronic structure problem by incorporating empirical parameters fitted to experimental data or data from high-fidelity computational methods. Among the most popular semi-empirical approaches are Density Functional Tight Binding (DFTB) [64, 65] and extended Tight-Binding (xTB) [126] families of methods, which this work will focus on.

2.4.1 Density Functional Tight Binding

The DFTB family of methods uses a Taylor expansion of the Kohn-Sham total energy in DFT (Eq. 2.34) as a function of the electronic density $\rho(\mathbf{r})$ around a non-interacting reference ρ_0 typically chosen as a superposition of neutral atomic densities.

The version of DFTB parametrised for and geared towards organic molecules is the third-order self-consistent charge DFTB (*i.e.*, DFTB3) [127–129] used in this work, for which only the ground-state DFTB is relevant and the primary source used for this Subsection is reference 130. Let us now consider the DFTB mathematical formalism for the energy and then individually each of its terms.

The DFTB formalism involves a Taylor expansion of the Kohn–Sham energy for the electron density of the interacting system $n(\mathbf{r})$ around a reference density ρ_0, a superposition of neutral atomic valence densities, where $\delta n(\mathbf{r})$ encodes the redistribution of the electron charge due to bonding obtained as $\delta n(\mathbf{r}) = n(\mathbf{r}) - \rho_0(\mathbf{r})$. In practice, $\delta n(\mathbf{r})$ is replaced by atomic charge deviations $\Delta q_A = q_A - Z_A^{\text{val}}$, computed via Mulliken population analysis [131], where Z_A^{val} is the count of valence electrons for atom A and q_A denotes the Mulliken electron population on atom A. The Z_A^{val} in DFTB is approximated with the use of a minimal basis set of valence atomic orbitals (*i.e.*, only the valence atomic orbitals necessary to describe bonding). The Mulliken electron population q_A is calculated as follows

$$q_A = \sum_{i \in A} \sum_j P_{ij} S_{ij}, \tag{2.70}$$

where P_{ii} is the density matrix over the occupied basis set of orbitals similarly to the Eq. 2.24, and $S_{ij} = \int \phi_i(\mathbf{r})\, \phi_j(\mathbf{r})\, d^3\mathbf{r}$ is the overlap matrix for orbitals i and j.

The DFTB3 energy, truncated at the third-order term in the Taylor expansion around a reference electron density ρ_0, is given by

$$
\begin{aligned}
E_{\text{DFTB3}} = \sum_a f_a \left\langle \psi_a \middle| \overbrace{-\frac{1}{2}\nabla^2 + V_{\text{ext}}(\mathbf{r}) + V_H[\rho_0](\mathbf{r}) + \left.\frac{\delta E_{\text{XC}}[\rho]}{\delta\rho(\mathbf{r})}\right|_{\rho_0}}^{\hat{H}_{\text{eff}}} \middle| \psi_a \right\rangle \\
+ \frac{1}{2}\iint \left(\frac{1}{\|\mathbf{r}-\mathbf{r}'\|} + \left.\frac{\delta^2 E_{\text{XC}}[\rho]}{\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')}\right|_{\rho_0} \right) \delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')\, d\mathbf{r}\, d\mathbf{r}' \\
+ \frac{1}{6}\iiint \left.\frac{\delta^3 E_{\text{XC}}[\rho]}{\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')\delta\rho(\mathbf{r}'')}\right|_{\rho_0} \delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')\delta\rho(\mathbf{r}'')\, d\mathbf{r}\, d\mathbf{r}'\, d\mathbf{r}'' \\
+ \sum_{A<B} \frac{Z_A Z_B}{\|\mathbf{R}_A - \mathbf{R}_B\|} - \frac{1}{2}\iint \frac{\rho_0(\mathbf{r})\rho_0(\mathbf{r}')}{\|\mathbf{r}-\mathbf{r}'\|}\, d\mathbf{r}\, d\mathbf{r}' \\
+ E_{\text{XC}}[\rho_0] - \int \left.\frac{\delta E_{\text{XC}}[\rho]}{\delta\rho(\mathbf{r})}\right|_{\rho_0} \rho_0(\mathbf{r})\, d\mathbf{r} + \mathcal{O}((\delta\rho)^4),
\end{aligned}
\tag{2.71}
$$

where the Hartree potential is defined as $V_H[\rho](\boldsymbol{r}) = \int_{\mathbb{R}^3} \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\, \mathrm{d}^3\boldsymbol{r}'$. The term $\mathcal{O}((\delta\rho)^4)$ represents higher-order corrections, while the subtraction of the linear exchange–correlation and Hartree contributions avoids double counting, as these are already included in the band structure and second-order terms. f_a is the occupation number of the single-particle state ψ_a, hence the first term denotes the band structure energy as the expectation value of the effective single-particle Hamiltonian over the occupied orbitals. The tight-binding approximation is applied, so the molecular orbital ψ_a is written as a linear combination of atomic orbitals, $\psi_a = \sum_i c_{ia}\phi_i$, with coefficients c_{ia}.

For efficiency, another approximation also applied is the two-center approximation, which restricts Hamiltonian and overlap integrals to pairs of atoms. Thus, the values for the Hamiltonian matrix elements $H^0_{\mu\nu} = \langle\phi_\mu|\hat{H}_{\text{eff}}|\phi_\nu\rangle$ and $S_{\mu\nu} = \langle\phi_\mu|\phi_\nu\rangle$ (with μ on atom A and ν on atom B) are pre-tabulated in Slater–Koster calculated as functions of interatomic distance and direction cosines $\langle\phi^{(A)}_\mu|\hat{H}_{\text{eff}}|\phi^{(B)}_\nu\rangle \longrightarrow \beta^{AB}_{\mu\nu}(R_{AB})$. The tables are distributed in parameter sets (*e.g.*, , the `3ob` set for organic/biomolecular systems [132, 133]). Given that the first energy term of Eq. 2.71 can be computed as follows

$$
E_{\text{band structure}} = \sum_a f_a \sum_{A,B} \sum_{\mu\in A} \sum_{\nu\in B} c^*_{\mu a} c_{\nu a}\, \beta^{AB}_{\mu\nu}(R_{AB}),
\tag{2.72}
$$

where the pre-tabulated β values are called the hopping integrals and represent the off-diagonal elements of the Hamiltonian matrix elements H_{ij} between atomic orbitals located on different atoms.

The second term in Eq. 2.71 provides the energy contribution due the first-order perturbation of the electron density, *i.e.*, the redistribution of the charge in the system resulting

from electrostatic interactions - the Coulomb interaction and some exchange-correlation contributions. The term is approximated using the aforementioned Mulliken population analysis and the δ softening (screening) parameter that models electrons overlap and includes exchange effects. It is derived for atoms by the reciprocal of the Hubbard parameter and inversely proportional to it for pairs of atoms fitted to DFT data and stored in the Slater-Koster files. The Hubbard parameter used comes from the Hubbard model - a simplified model for electron-electron interaction in a lattice, considering two competing effects: electron hopping between neighbouring atoms via a hopping parameter to represent delocalisation and electron repulsion by the on-site Hubbard parameters U_A. Thus, the term γ_{AB} is obtained as a model for the interaction energy between partial charges $\Delta q = q - q^0$ on two atoms A and B due to the charge redistribution.

$$\gamma_{AB}(R_{AB}) = \frac{1}{\sqrt{R_{AB}^2 + \delta^2}}, \tag{2.73}$$

which combined with the charge approximation above gives as an approximated form of the second energy term in Eq. 2.71

$$E \approx \frac{1}{2} \sum_{A,B} \gamma_{AB} \Delta q_A \Delta q_B. \tag{2.74}$$

The third term in the DFTB energy (Eq. 2.71), *i.e.*, the third order correction, is similarly approximated as

$$E^{(3)} \approx \frac{1}{6} \sum_{A,B} (\Delta q_A^2 \Delta q_B \Gamma_{AB} + \Delta q_A \Delta q_B^2 \Gamma_{BA}) \quad \text{with} \quad \Gamma_{BA} = \left. \frac{d\gamma_{AB}}{dq_B} \right|_{q_A^0}, \tag{2.75}$$

and it accounts for the non-linear change of the exchange-correlation energy w.r.t the electron density at different spatial points - especially important in the presence of strong polarisation *e.g.*, Hydrogen bonds. The fourth term, besides the already discussed subtractions, includes the classical Colomb electrostatics and can be approximated as follows

$$\begin{aligned} E_{\text{rep}} = \sum_{A<B} \frac{Z_A Z_B}{\|\mathbf{R}_A - \mathbf{R}_B\|} - \frac{1}{2} \int V_H[\rho_0'] \rho_0' \, d\mathbf{r} + E_{\text{xc}}[\rho_0] - \left. \int \frac{\delta E_{\text{xc}}[\rho]}{\delta \rho} \right|_{\rho_0} d\mathbf{r} + \mathcal{O}((\delta\rho)^4) \\ \approx \sum_{A<B} v_{\text{rep}}^{(AB)}(\|\mathbf{R}_A - \mathbf{R}_B\|), \end{aligned} \tag{2.76}$$

where $v_{\text{rep}}^{(AB)}$ is a repulsive potential chosen such that it reproduces the (semi-)local DFT or *ab initio* reference and thus also accounts for the many-body effects not explicitly introduced otherwise.

Despite its computational efficiency, the DFTB approach inherits limitations from its underlying approximations: the minimal LCAO basis can lead to inaccuracies in capturing electron delocalisation and charge transfer [134]; the truncation of the energy expansion

around the reference density—typically at second or third order—can introduce significant errors in systems exhibiting strong charge redistribution or high reactivity [134, 135]. Furthermore, while the use of parameterised Slater–Koster integrals and pairwise repulsive potentials enables rapid evaluation of inter and intramolecular interactions, it also introduces empirical biases. These can impair the transferability of the model, particularly when parameters optimised for small molecules are applied to larger or more complex (bio)molecular systems without additional reparameterisation [135, 136].

2.4.2 Generalized Frequency Non-covalent extended Tight-Binding

Generalized Frequency Non-covalent eXtended Tight-Binding (GFN-xTB) [126] is one of the most popular semi-empirical methods, parametrised for organic molecules and elements in the Periodic table up to Radon. The particular version of GFN-xTB relevant to this thesis is GFN2-xTB, which includes the dispersion correction D4 [118] (see section 2.3.2), multipole moments for electrostatics, and third-order density fluctuation terms similar to DFTB3. It also uses a minimal basis set like DFTB3 and has a self-consistent charge treatment, and uses pre-tabulated parameters *e.g.*, Slater-Koster values, atomic parameters, D4 values, etc. for fast calculations. An interesting difference with DFTB is the global parametrisation per element fitted on a large dataset in xTB as opposed to focus on element-pairs parameters in environments giving specific basis sets for different types of molecular systems. Also notable is the presence of dispersion correction in GFN2-xTB as opposed to DFTB3, which is why the latter is often used for organic systems with an MBD correction as will be applied in this work, namely DFTB3+MBD.

2.5 Classical Force Fields

Classical FFs are a set of computational models used to describe how atoms and molecules interact with each other, offering a pragmatic alternative to computationally intensive quantum mechanical calculations. This approach is based on classical mechanics treatment of molecules as a collection of atoms treated as point charges and masses, connected by springs to model their bonds, meaning they can elongate and compress around an equilibrium distance. Each atoms has a parametrisation *e.g.*, mass, partial charge, van der Waals parameters, bonding information based on distance to other atoms. Classical FFs, also called molecular mechanics (MM) FFs have historically been the only computational model able to access and produce molecular dynamics (MD) simulations of large molecular systems *e.g.*, of the size of biomolecules like proteins. Thanks to their constantly improving parametrisations fitted to experimental and to high-fidelity computational data these are still an indispensable tool for biochemical simulations. FFs have limited transferability and different FFs have been developed to capture specific areas of chemical systems, *e.g.*, an FF developed for proteins would not be suitable in parametrisation for small organic molecules. A further limitation is the additive pairwise model of contributions neglecting many-body effect as is in many cases the lack of response to the environment due to lack of charge redistribution

approximation. While with the increase in calculation power available, the MMFFs have also improved by also incorporating more accurate contributions, *e.g.*, polarisation in the polarisable FFs like AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications) [137, 138]. Polarisable FFs continue to develop rapidly [139], however, they have still not reached the general applicability of the most common FFs like CHARMM (Chemistry at HARvard Macromolecular Mechanics) [140], GAFF2 [141] or CGenFF [142]. In this Section I aim to provide a quick overview of the theoretical framework [143] and focus on a few relevant FFs to this thesis, GAFF2 [141] and CGenFF [142].

A force field consists of an interatomic potential energy, $U(\mathbf{r})$ reliant on a set of parameters, which for the classical FFs is typically obtained from such an expression

$$U = \sum_{\text{bonds}} \frac{1}{2} k_b (\mathbf{r} - \mathbf{r_0})^2 + \sum_{\text{angles}} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{\text{torsions}} \frac{V_n}{2} [1 + \cos(n\phi - \delta)] + \sum_{\text{improper}} V_{\text{imp}} + \sum_{i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{\text{elec}} \frac{q_i q_j}{r_{ij}}, \quad (2.77)$$

where the first term, $\frac{1}{2} k_b (r - r_0)^2$, accounts for the bond stretching and models the energy associated with deviations from the equilibrium bond length $\mathbf{r_0}$, with k_b as the bond force constant and $\mathbf{r}$ the current bond length. The second term, $\frac{1}{2} k_a (\theta - \theta_0)^2$, the angle bending term, provides the energy due to deviations from the equilibrium bond angle θ_0, with correspondingly k_a as the angle force constant and θ the current angle. In the third term, the torsional interactions are described by $\frac{V_n}{2}[1 + \cos(n\phi - \delta)]$, where V_n is the barrier height, n is the periodicity, ϕ is the dihedral angle, and δ is the phase shift. Planarity or chirality are maintained thanks to the fourth term, denoted V_{imp} and finally the last two terms capture the non-bonded interactions. The energy contributions of the van der Waals interactions are given by the Lennard-Jones potential, $4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$, with ϵ_{ij} as the potential well depth and σ_{ij} the distance at zero potential energy, while the electrostatic interaction term, $\frac{q_i q_j}{r_{ij}}$ represents Coulombic interactions between partial charges q_i and q_j apart at r_{ij} distance. Obtaining each of the aforementioned parameters is a laborious undertaking due to the necessary data for fitting. For example, for $\mathbf{r_0}$ X-ray diffraction references are used, while the spring constant estimates can rest on infrared or Raman spectra [143].

GAFF2 (General AMBER Force Field 2) is based on the RESP (Restrained ElectroStatic Potential) charge model using the semi-empirical AM1-BCC [144] charges augmented with bond charge corrections based on data fitting to 442 neutral organic solutes with known experimental hydration free energies [145] and suitable for organic molecules in drug discovery. CGenFF (CHARMM General Force Field), is also designed particularly to cover a range of diverse organic molecules relevant to drug discovery. CGenFF uses a bond-charge increment scheme, where atomic charges are derived by summing predefined charge increments for each bond based on chemical environment - thus effects up to three bonds away can be included. The charges are automatically assigned by using a fragment-based analogy

approach, which matches by similarity parametrised molecular fragments parametrised in the CHARMM FF database. A known limitation of standard MMFFs is their inability to capture bond creation and breaking in MDs due to the underlying framework of fixed spring connectivity between atoms as in Eq. 2.77 - a particularly relevant effect for cases of ligand-protein interactions with covalent bond creation and breaking. The classical FF have been used in Chapters 3 and 6.

Nudged elastic banding

Given the limitations of classical FFs but their low cost and speed, they are often integrated in pipelines as initial steps for configuration sampling. For example, classical modelling of molecules has applications directly for creation of molecular structures between two states with Nudged Elastic Banding (NEB) [146], with the structures then evaluated at higher level of accuracy. In NEB, the minimum energy path between a given starting and final state is found by creating a series of intermediate structures called images, connected by virtual springs mimicking an elastic band. In NEB, the forces acting on these images to are minimised to find the optimal path with the key difference to other methods being the projection of the forces ("nudging") along the path: only the perpendicular component of the true force $\mathbf{F}_i^{\perp}$ and the parallel component of the spring force $\mathbf{F}_i^{\parallel}$ are used so the virtual springs push along the minimum energy path. The total force acting on 'image' i along the path is thus $\mathbf{F}_i$

$$\mathbf{F}_i = \mathbf{F}_i^{\perp} + \mathbf{F}_i^{\parallel}, \tag{2.78}$$

where the perpendicular force $\mathbf{F}_i^{\perp}$ is given by

$$\mathbf{F}_i^{\perp} = -\nabla V(\mathbf{P}_i) + \left[(-\nabla V(\mathbf{P}_i) \cdot \hat{\tau})\,\hat{\tau}\right], \tag{2.79}$$

where $\nabla V(\mathbf{P}_i)$ is the gradient of the potential energy at the position $\mathbf{P}_i$, representing the true physical force acting on the system, $\hat{\tau}$ is the unit tangent vector to the path at image i, used to decompose forces. The parallel force $\mathbf{F}_i^{\parallel}$ component is given by

$$\mathbf{F}_i^{\parallel} = \left[(k_{i+1}(\mathbf{P}_{i+1} - \mathbf{P}_i) - k_i(\mathbf{P}_i - \mathbf{P}_{i-1})) \cdot \hat{\tau}\right]\hat{\tau}, \tag{2.80}$$

where k_i is the spring constant between images $i-1$ and i, which can then uniformly sample the minimum energy path between two states. Thus, the NEB can be used to describe transitions, *e.g.*, for torsional transitions to model rotational profiles in molecules [147], and it can also be combined with higher-accuracy methods for energy calculations *e.g.*, DFT [63, 148] as seen in Chapter 5.

Molecular dynamics

In organic systems, there is an additional consideration for real-life applications, namely the effect of the environment and *in vivo* temperature for performing molecular dynamics simulations (MDs). Thus, (bio)molecular simulations need to be performed at 300 K. In classical FFs, the temperature is not explicitly included in the potential energy functions

themselves but is accounted for via the velocities of the classically treated atoms. Mechanistically, this is implemented with a thermostat that maintains the system at constant temperature and thus samples configurations of the canonical ensemble *i.e.*, at NVT conditions - constant number of particles, volume, and temperature. A commonly used example is the Langevin thermostat [149, 150]. It modifies Newton's 2nd law by the addition of two extra forces in addition to the interatomic potential $U(\mathbf{r})$: the γ friction coefficient, controlling how strongly the system is coupled to the heat bath, and $\sqrt{2\gamma k_B T}\,\eta(t)$ accounting for thermal noise as a stochastic force based on a Gaussian white noise distribution given by $\eta(t)$. The Langevin thermostat is then given by

$$m\frac{d^2\mathbf{r}}{dt^2} = -\nabla U(\mathbf{r}) - \gamma m\frac{d\mathbf{r}(t)}{dt} + \sqrt{2\gamma k_B T}\,\eta(t) \tag{2.81}$$

and can be used beyond classical FFs in ML-integrated pipelines. For example, using MLFFs instead of MMFFs to compute interatomic forces, which are then passed onto an MD calculation package *e.g.*, Atomic Simulation Environment (ASE) [151] in which the Langevin thermostat is implemented.

Another role of the temperature in the molecular dynamics simulations is as a tool to perform simulations out of local equilibria to further sample configurational space. At increased temperature, the molecular systems can evolve through CCS regions that would have otherwise at lower temperatures have taken much longer time to visit. Thus increased temperature MDs, for temperatures where the systems are expected to remain physically stable, are an excellent tool to reduce the timescales of MDs needed and associated costs.

2.6 Machine Learning

Machine Learning (ML) models offer a powerful paradigm towards capturing complex patterns and behaviour in molecular systems and modelling them with high accuracy and efficiency. There are two main types of ML models in use - those based on kernel methods and those based on neural networks (NN). While they can have different use cases, the general aim of ML in physical chemistry has been to create a bridge between the accuracy of *ab initio* or DFT methods and the efficiency of classical molecular mechanics. In particular, the complex non-linearity inherent in the structure-property relationship is well-suited as a task for ML, a tool for capturing non-linear patterns and thus offering a shortcut to solving the explicit equations that quantum chemistry and semi-empirical methods are based on. Two types of molecular properties are predicted - those that mostly depend on molecular size, *i.e.*, extensive properties (*e.g.*, atomisation energy, molecular polarisability) and those that largely depend on the electronic structure of molecules, *i.e.*, intensive properties (*e.g.*, HOMO-LUMO gap, dipole moment), the latter of which remain more challenging. Moreover, building reliable ML models also requires comprehensive and accurate training data, reflecting a representative sample of the chemical compound space (CCS) to learn the relevant inter- and intra-molecular effects. Overall, the goal of using ML here is to capture the underlying physicochemical effects sufficiently to produce a model that is scalable and

transferable to different and large systems, thus substituting the need for (prohibitively) expensive reference calculations.

2.6.1 Machine Learning Force Fields

A particular research line in ML is the creation of MLFFs, which aim to offer greater accuracy with respect to classical FFs, by incorporating quantum effects 'learned' by synthetic data, usually from DFT, maintaining high efficiency and a computational cost comparable to that of classical FFs. Most commonly, MLFFs are applied for MDs, where the forces along the MDs are derived from the PES as the negative gradient to the potential energy U w.r.t. the atomic positions $\mathbf{R}$ [45]. The performance of the MLFFs is typically evaluated on a test system or dataset, which was not included in the training or validation data and their predictions are assessed with common regression metrics such as mean absolute error (MAE), root mean square error (deviation) RMSE (RMSD). An important further test to MLFF's viability as a better substitute for classical FFs is also their ability to produce stable and accurate MDs on larger or sufficiently different molecular systems from those in the training data. The performance of the models is judged over a variety of measures, most commonly the goal is defined as the "chemical accuracy" at 1 kcal/mol MAE, but other appropriate metrics at different level of aggregation are also employed - see further measures in Chapter 5. A sufficient accuracy would indicate that the extrapolation between the structure and properties is captured well by the ML potential, enabling it to produce scalable and transferable models. To that end, multiple strategies have been developed, but before diving into specific examples, let us take a look at the creation of ML models for MLFFs.

In general terms, the workflow for the creation and application of ML models to obtain molecular properties can be described as

$$\text{Input } (x, y) \xrightarrow[\text{process}]{\text{ML training}} f_\theta \xrightarrow{x_{\text{new}}} y_{\text{new}},$$

where (x, y) denote the training inputs and the known associated properties respectively. The training process adjusts the parameters θ of the ML model f_θ to produce the trained model, which can be applied to new input x_{new} to predict the properties y_{new} associated with it. In each model, f_θ contains both chosen fixed parameters called hyperparameters and parameters optimised during the training. While this general schematic applies to MLFFs obtained both from kernels *e.g.*, sGDML (Symmetric Gradient Domain Machine Learning) [152] and NNs *e.g.*, Allegro [35], MACE [153], SO3krates [34], they qualitatively differ in how the training is implemented. Both ML training procedures minimise a loss function during the training (*e.g.*, a squared loss in kernel ridge regression and the chosen loss function in NNs). Kernel methods apply a one-shot implicit non-linear mapping of input to output features via a kernel function, whereas NNs require iterative gradient-based updates to adjust their numerous learnable parameters. This links to the concept of 'depth' in ML models that refers to the number of sequential non-linear transformations in the

input-to-output mapping. Hence, the kernel methods are generally shallow models, and thus expressivity is largely determined by the encoding of the inputs via a hand-crafted descriptor and the choice of non-linear function. Consequently, a lot of research has been conducted to create appropriate bespoke descriptors to capture molecular features as well as on the best ways to utilise non-linear functions to capture structure-property relationships. Meanwhile, NNs learn task-specific latent representations (embeddings) end-to-end based on unprocessed or minimally processed features (*e.g.*, respectively, atomic coordinates or neighbour lists, to be expanded on later). To that end, NNs leverage the sequential non-linear transformations to learn the relationships between the inputs and thus the impetus is on building blocks of the NN architecture to facilitate this. Let us now consider the individual building blocks of the ML methods and discuss specific examples relevant to this thesis. First, the concepts of invariance and equivariance will be introduced below as they pertain to the description of molecules and molecular environments for ML models, elaborated on here.

Invariance and Equivariance

By Noether's theorem [154], each continuous symmetry of the action of a physical system is associated to a corresponding conservation law for a closed physical system, such as the molecular systems considered here. Thus, a key ingredient in modern ML architectures is accounting for such symmetries, initially by including invariance, and more recently by the addition of equivariance in building such models. Invariance is mathematically defined as follows

$$f(g \cdot x) = f(x) \quad \forall g \in G,\ x \in X, \tag{2.82}$$

where the function f is invariant under the action of a group G, if applying any group element $g \in G$ to the input $x \in X$ does not change the output of the function. On the other hand, equivariance is defined as follows

$$f(g \cdot x) = g \cdot f(x) \quad \forall g \in G,\ x \in X, \tag{2.83}$$

where $g \cdot x$ denotes the group action of $g \in G$ on $x \in X$. The function f is said to be equivariant if applying f after the group action is equivalent to applying the group action after f.

The idea of equivariance as applied to NNs for molecular systems is relatively new in the field and firmly established itself since its first use in an NN for chemical systems in 2020 [38, 155]. However, it is inherently linked to the well-established question of how molecules can be represented, with multiple existing strategies, such as through nuclear densities or as a discrete set of atomic positions. This relates back to the fact that the energy is invariant under rotation, while the force vector, defined as the negative gradient of the energy, is equivariant. The analytical inclusion of invariance and equivariance prevents the model from re-learning the same physical law for the same molecule from different perspectives [156]. The choice of representation comes with a corresponding need for symmetry preservation as it determines the relevant symmetry group, w.r.t. which the model

(representations and operations) requires invariance or equivariance - for continuous spatial symmetries, this is typically the special orthogonal group SO(3) *i.e.*, the group of all the 3D rotations in space that do not involve reflection or scaling. Meanwhile, discrete atomic representations require invariance under permutation groups and although they are formally exact, they also come with challenging overheads when the associated permutation groups become large and complex. This is why many methods represent molecular structures as continuous densities, which are more naturally compatible with Lie groups like SO(3). These groups admit well-defined irreducible representations (irreps), which can be used to efficiently encode rotational symmetries in ML models. This encodes a fundamental property of molecules as physical objects and allows for changes in the vectors output (*e.g.*, a rotation) corresponding to the changes in the input.

2.6.2 Descriptors and embeddings

The role of the ML descriptor is to represent the molecule by capturing a vector of informative features needed to learn and predict molecular properties. A descriptor x typically involves atomic positions $\mathbf{R}$, atomic numbers Z, and, depending on the kernel, angles between atoms (*i.e.*, 3-body terms) and further informative encodings. While initially two-body descriptors like Colomb matrix [157] and Bag-of-bonds [158] were used for their lower expense, three-body contributions have also been developed to capture more complex chemical moieties, improving the accuracy of the property predictions [159, 160] and are thus now the norm in kernel approaches. A typical example among them is the SLATM (Spectrum of London and Axilrod–Teller–Muto potentials) [161] descriptor x with 2- and 3-body terms as follows

$$x = [\sum_i \delta(Z_i - Z),\ \sum_{i<j} \frac{C_6(Z_i, Z_j)}{r_{ij}^6} \cdot \exp\left(-\frac{(r_{ij} - \mu_k)^2}{2\sigma^2}\right),$$
$$\sum_{i<j<k} \frac{C_9(Z_i, Z_j, Z_k)}{r_{ij}^3 r_{ik}^3 r_{jk}^3} \cdot (1 + 3\cos\theta_{ijk}\cos\theta_{jik}\cos\theta_{kij}) \cdot \exp\left(-\frac{(\theta_{ijk} - \nu_l)^2}{2\sigma_\theta^2}\right)], \quad (2.84)$$

where r_{ij} is the interatomic distance between atoms i and j and θ_{ijk} is an angle formed at atom i by atoms j and k. The 1-body term $\delta(Z_i - Z)$ encodes the atomic identity, the 2-body term $\frac{C_6(Z_i,Z_j)}{r_{ij}^6} \cdot \exp\left(-\frac{(r_{ij}-\mu_k)^2}{2\sigma^2}\right)$ models pairwise interactions using a London dispersion-like potential, and the last terms is the 3-body term captures angular correlations using the Axilrod–Teller–Muto potential. The μ_k parameter in the 2-body term and the ν_l parameter in the 3-body term are the grid centers for discretisation for interatomic and angular distances, respectively. A Gaussian basis function is centered at each grid center with the overlap between functions evaluated and stored in a vector.

Another classic descriptor, SOAP (Smooth Overlap of Atomic Positions) [162] is based on encoding the local atomic environments by a scalar field centred on an atom i and composed of Gaussian functions. The electron density is then expanded in orthonormal basis functions combining radial basis functions and spherical harmonics (Eq. 2.26), and the overlap

integrals of atomic neighbour densities are computed. The SOAP descriptor is introduced as a power spectrum as follows

$$p^{(i)}_{nn'l} = \sum_m c^{(i)}_{nlm} \left(c^{(i)}_{n'lm} \right)^* , \tag{2.85}$$

where $c^{(i)}_{nlm}$ are the coefficients from the expansion of the atomic neighbour density in a basis of radial functions and spherical harmonics as given by

$$c^i_{nlm}(\mathbf{R}) = \iiint_{R^3} dV \; g_n(\mathbf{R}) Y_{lm}(\theta, \phi) \rho^i(\mathbf{R}), \tag{2.86}$$

where $g_n(\mathbf{R})$ is the radial basis function, and $\rho^i(\mathbf{R})$ is the Gaussian smeared atomic density for atom i.

Another current descriptor is FCHL19 [163], particularly created with efficiency in mind, which represents similarly to SLATM atomic environments using a combination of 1-, 2-, and 3-body terms, each discretised using Gaussian basis functions. The key difference is that the 2- and 3-body terms are defined as follows

$$\begin{aligned} x^{(2)}_{ij} &= f(Z_i, Z_j) \cdot \exp\left(-\frac{(r_{ij} - \mu_k)^2}{2\sigma^2} \right) \\ x^{(3)}_{ijk} &= f(Z_i, Z_j, Z_k) \cdot \exp\left(-\frac{(r_{ij} - \mu_k)^2}{2\sigma^2} \right) \cdot \exp\left(-\frac{(\theta_{ijk} - \nu_l)^2}{2\sigma^2_\theta} \right), \end{aligned} \tag{2.87}$$

where the $f(Z_i, Z_j)$ and $f(Z_i, Z_j, Z_k)$ functions are element-dependent scaling functions defined for pairs and triplets of elements. A universal set of hyperparameters is obtained from a Monte Carlo optimisation, which then does not undergo re-optimisation in application.

Finally, let us discuss the sGDML (symmetry-adapted Gradient-Domain Machine Learning) descriptor x, which is built upon all unique pairs of Euclidean distances between atoms i and j, for all indices $i < j$ as follows

$$x = \left(\|\mathbf{r}_i - \mathbf{r}_j\| \right)_{i<j} . \tag{2.88}$$

For sGDML the symmetry adaptation augments the training set by finding all possible permutations in a molecule (*i.e.*, for identical atoms swapping) and concatenating those to the descriptor. All descriptors described thus far are invariant to translation, rotation, and permutation of atoms.

In NNs, besides direct descriptors, an alternative approach is learned descriptors, also called embeddings, which become internal representation to the model. Early NNs, in this field, dating back to the 90s [164], were trained directly on unprocessed Cartesian coordinates to model PES and thus lacked invariance in the build leading to an inherent limit to their

transferability. The Behler-Parrinello NNs, from 2007 [165] were the first systematic scalable class of NNs that incorporated invariance in the molecular representation via fixed hand-crafted symmetry functions of interatomic distances and bond angles, *i.e.*, symmetry-aware descriptors. This idea is conceptually related to developments in kernel descriptors such as the aforementioned FCHL19 and SOAP, and is an idea that evolved in NNs with more intricate introductions of symmetry, *e.g.*, via the incorporation of spherical harmonic expansions. Notably, however, modern NNs have progressed increasingly beyond set descriptors and rely overwhelmingly on embeddings, based on which information the NNs learn the latent molecular representations. An embedding usually includes atomic numbers and interatomic position *e.g.*, for the SchNet [39] and Allegro [35] NNs. The interatomic distances can then further be expanded using radial basis functions (RBF) [34, 39] for smooth local features *e.g.*, the NNs SchNet and SO3krates. Further, the relative position vectors can be projected onto a set of spherical harmonics functions (see Eq. 2.26) to encode the angular geometry between neighbouring atoms *e.g.*, for the NNs MACE [153] and SO3krates. The embeddings can further include more involved and physically-motivated features such as charge and spin, as in the case of the SpookyNet [31] NN.

To respect rotational symmetry, the embeddings are grouped into l (angular momentum) channels that each correspond to a particular irreducible representation (irrep) of the rotation group SO(3), *i.e.*, a channel for scalars ($\ell = 0$) unchanged under rotation, a channel for vectors ($\ell = 1$) rotating as bond directions or forces, and higher-order channels ($\ell \geq 2$) transforming like tensors to capture angular correlations. This organisation of multipole features ($\ell = 0, 1, 2, ..$) ensures that every component of the latent representation transforms correctly under physical laws, namely under rotation, scalar features remain unchanged while vector and tensor features rotate consistently with the geometry. The updates of the latent representation with the training of equivariant NNs also ensure that equivariance is preserved. This is done in a physics-inspired fashion by tensor products of the different channels, with the result decomposed into irreducible representations using Clebsch–Gordan coefficients.

The Clebsch-Gordan coefficients are typically precomputed for ML and stored in lookup tables or generated using libraries like e3nn [166], which handle the Clebsch–Gordan tensor products efficiently on GPU, speeding up the process. Clebsch–Gordan coefficients arise when adding angular momentum eigenstates, such as orbital or spin angular momentum. When two angular momenta $\vec{J}_1$ and $\vec{J}_2$ are combined, the total angular momentum states $|j, m\rangle$ are expressed as linear combinations of product states $|j_1, m_1\rangle \otimes |j_2, m_2\rangle$ using Clebsch–Gordan coefficients $C^{jm}_{j_1 m_1 j_2 m_2}$, such that

$$|j, m\rangle = \sum_{m_1, m_2} C^{jm}_{j_1 m_1 j_2 m_2} |j_1, m_1\rangle \otimes |j_2, m_2\rangle.$$

These coefficients encode how rotational symmetries (*e.g.*, SO(3)) are preserved in the coupling of quantum states. In the equivariant NNs, the Clebsch–Gordan coefficients are used to combine spherical tensor features (*e.g.*, vectors, dipoles, quadrupoles) in a way that ensures the learned representations are equivariant under 3D rotations.

Loss functions

The training of the ML model depends on a loss function, which is used to quantify the discrepancy between the model's predictions versus the target reference data. The aim is to minimise the loss function and thus is a key step in optimisation processes, *i.e.*, model training. The aim of the loss function, depending on how it is constructed, can be also to prioritise target properties - for example, energies and forces when aiming to obtain MD simulations. The loss function $\mathcal{L}$ can be a sum of performance estimate terms (usually Mean Square Error, MSE) over different properties, each multiplied by a coefficient λ_i assigning the importance of each term. Mathematically, the loss function $\mathcal{L}$ can be expressed as a weighted sum of error terms,

$$\mathcal{L} = \sum_i \lambda_i \, \mathrm{MSE}\left(y_i^{\mathrm{pred}}, y_i^{\mathrm{ref}}\right),$$

where λ_i gives the relative importance of property y_i. In KRR, a quadratic loss is used to obtain an analytic solution, while for NNs the same principle is applied, but the loss function is used in an iterative gradient-based optimisation. For modern NNs in the field, the loss function can also frequently be physically motivated *e.g.*, for SpookyNet [31] it includes the total energy E, atomic forces $\mathbf{F}$, dipole moments $\boldsymbol{\mu}$, atomic charges q, and polarisabilities $\boldsymbol{\alpha}$, additional physical quantities beyond those of a typical MLFF, as follows

$$\mathcal{L} = \lambda_E \cdot \mathrm{MSE}(E) + \lambda_F \cdot \mathrm{MSE}(\mathbf{F}) + \lambda_\mu \cdot \mathrm{MSE}(\boldsymbol{\mu}) + \lambda_q \cdot \mathrm{MSE}(q) + \lambda_\alpha \cdot \mathrm{MSE}(\boldsymbol{\alpha}), \quad (2.89)$$

while for SO3krates the loss function includes energies and forces, for its next incarnation SO3LR, built on SO3krates, partial charges and Hirshfeld ratios [167] are also included in its loss function. Notably, for some NNs like SpookyNet, SchNet [39], and SO3krates [34] the forces are obtained as gradients of the energies w.r.t. positions, and for some like MACE the forces can be obtained can also be learned directly. In some NNs there is greater focus on forces only learned directly, *e.g.*, for Allegro [35] and SO3LR [168].

2.6.3 Kernel Ridge Regression

While many NN architectures in recent years have been devised to predict molecular features [169], they generally rely on large training sets, which are expensive to generate with accurate QM methods, and have become more available in large quantities only in recent years, especially for chemically diverse out-of-equilibrium molecular systems. As an advantage to NNs, the KRR approach is a more data-efficient method, enabling the possibility of working with sparse datasets. However, the two approaches are naturally related. In the context of NNs, kernel models can be viewed as single-layer architectures, for which the non-linearity is defined by a kernel function rather than a learned activation function, and the input is mapped using a fixed instead of a learned descriptor. Unlike typical deep NNs, kernel models do not learn internal representations and instead operate in a fixed feature space, where the model parameters are optimised. kernel models minimise a loss function conceptually similarly to NNs, but with the advantage of certain formulations—such

as kernel ridge regression (KRR)—to admit a closed-form solution. This allows for efficient solvers, while the hyperparameters are typically optimised in a separate outer loop. KRR also provides a less complex mathematical approach with a reduced number of tunable parameters (typically less than 10) compared to possibly hundreds of thousands for NNs. Furthermore, KRR offers high interpretability of results in contrast to the "black box" model developed with deep NN. Let us now consider the building blocks of KRR.

Mathematically, a ridge regression is built on the concept of linear regression by adding a regularisation parameter λ to avoid overfitting and handle highly correlated features (of size m) in the input matrix X of size n samples times m features [170]. The weights w are learned as per the normal equation of the ridge regression as follows: $\mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$, where y is the vector of target property and w are the weights to be learned. In KRR, the role of X is taken by the kernel K which encodes all pairwise similarities between training points mapped to a high-dimensional Hilbert space [171]. The KRR equation for obtaining the target properties y_{test} is

$$y_{\text{test}} = K_{\text{test}}^\top (K_{\text{train}} + \lambda I)^{-1} y_{\text{train}}, \quad \text{(output in } \mathbb{R}^{n_{\text{test}} \times 1}\text{)}. \tag{2.90}$$

When using KRR for molecular property prediction, besides the hyperparameter optimisation procedure, another crucial point is the determination of the optimal set of KRR components, such as kernel functions, molecular descriptors [172], and distance metrics [173].

Distance metric

The distance metric is a function that quantifies the level of similarity between two data points in the input space. The Minkowski metric (or L_p norm) is the general form of some of the most commonly used distance metrics in KRR method such as Manhattan L_1 and Euclidean L_2. The Minkowski metric is expressed as follows

$$L_p = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} = ||x_i - y_i||, \tag{2.91}$$

with parameter $p \in \mathbb{R}^+$, and molecular representation vectors x_i and y_i. Setting $p = 1$ recovers the Manhattan metric, while $p = 2$ corresponds to the Euclidean distance metric.

Kernels

The kernel function k computes the inner product between the images of x and x' under the mapping ϕ in the feature space $\mathcal{H}$ as follows

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}. \tag{2.92}$$

Specific prominent examples of kernel functions are the Laplacian and Gaussian kernels, mathematically defined as follows

$$K_{Laplacian} = e^{-\frac{||x_i - y_i||}{\sigma}}, \qquad K_{Gaussian} = e^{-\frac{||x_i - y_i||^2}{2\sigma^2}}, \tag{2.93}$$

where σ is the length-scale hyperparameter, and the second hyperparameter is a small and mathematically necessary regularisation parameter λ, which secures the invertibility of the kernel matrix. It is multiplied by the identity matrix in the equation used to obtain the target molecular property, y_{test}, as a function of the training and test set kernels, as well as the target property of the training set, y_{train}. The Gaussian kernel is also known as the radial basis function (RBF) kernel and is notable for the smoothness with which it models the reference function due to its infinite differentiability. Using the Gaussian kernel, more involved kernel models like FCHL19 [163].

More involved kernels like the Matérn one [174, 175] have been used for more sophisticated encoding of the molecules, but usually come at a higher cost. The Matérn kernel can be seen as a generalised form of the Gaussian kernel and is given in mathematical terms by the expression

$$k_{\text{Matérn}}(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|x - y\|}{\sigma} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu}\|x - y\|}{\sigma} \right), \tag{2.94}$$

where $\Gamma(\nu)$ is the gamma function and K_{ν} is the modified Bessel function of the second kind. The parameter $\nu > 0$ controls the smoothness of the resulting function: for example, $\nu = 1/2$ yields an exponential kernel (non-differentiable), $\nu = 3/2$ gives a once-differentiable function, and $\nu = 5/2$ results in a twice-differentiable function. As $\nu \to \infty$, the Matérn kernel converges to the Gaussian kernel, making it the better choice for less smooth underlying functions. The key idea behind the sGDML method is not only the kernel but also learning directly the shape of the PES, *i.e.*, the forces, rather than differentiating the learned energies to obtain the forces. Limiting the aggregated errors in differentiation makes this method particularly stable for MD simulations, where accurate forces are critical to the system's evolution with time. The forces $\mathbf{F}$ are obtained from the Hessian of the kernel $\mathbf{K}_{\text{Hess}}$ (*i.e.*, 2nd derivative w.r.t input coordinates),

$$\mathbf{K}_{\text{Hess}} + \lambda \mathbf{I}\alpha = \nabla E = -\mathbf{F}, \tag{2.95}$$

which acts as a covariance matrix in the regression, encoding how different force components are related. α is the set of learnable parameters, through which this approach explicitly enforces that the forces are a gradient of a function, *i.e.*, that there exists an energy function that is conserved as per the law of conservation - the energy E can be recovered by integration.

The SOAP kernel, meanwhile, is a similarity measure between two atomic environments, computed as the dot product of two SOAP descriptors as obtained from Eq. 2.85 raised to the power of ζ.

$$k(\mathcal{X}, \mathcal{X}') = \left(\mathbf{p}(\mathcal{X}) \cdot \mathbf{p}(\mathcal{X}')\right)^{\zeta}, \tag{2.96}$$

where $\mathbf{p}$ the SOAP descriptor for two atomic environments $\mathcal{X}$ and $\mathcal{X}'$ and ζ is a hyperparameter controlling the sharpness of the similarity [162]. The normalised version of the kernel,

$k(\mathcal{X}, \mathcal{X}') = \left(\frac{\mathbf{p}(\mathcal{X}) \cdot \mathbf{p}(\mathcal{X}')}{\sqrt{(\mathbf{p}(\mathcal{X}) \cdot \mathbf{p}(\mathcal{X}))(\mathbf{p}(\mathcal{X}') \cdot \mathbf{p}(\mathcal{X}'))}} \right)^{\zeta}$ is also scale-invariant and suitable for chemically similar environments. The SOAP descriptor (and SOAP kernel) can be combined with the Gaussian approximation potential (GAP) [162, 176] model to produce the SOAP/GAP model. There, GAP is a framework for energy and forces predictions in molecular systems, which uses the Gaussian process regression (GPR) as a non-parametric Bayesian method to learn a function that maps atomic environments to energy contributions. In that case, GPR assumes that the function values, *i.e.*, energies, follow a multivariate Gaussian distribution over a given kernel, which is then trained on data to obtain a posterior distribution.

2.6.4 Neural Networks

Artificial neural networks are a ML type of model comprised of a computational graph made up of layers of interconnected nodes, which take an input, to which a linear transformation is applied, and then the result is passed through a non-linear function called the activation function. A neural network typically has a layered structure with many nodes per layer, whose specific organisation of the input layer, hidden layers, and output layer is called architecture. Consequently, the deep neural networks include two or more hidden layers [177], where the output from a node a_j in a hidden layer can be defined as the result of the activation function ϕ applied to the sum of the contributions of the connected nodes from the previous layer $a_i^{(\text{prev})}$, scaled by learnable parameters called weights w_{ji} and the learnable bias term b_j of the node as follows

$$a_j = \phi \left(\sum_i w_{ji} \cdot a_i^{(\text{prev})} + b_j \right). \tag{2.97}$$

Examples of activation functions include shifted softplus (*e.g.*, in SchNet and SpookyNet) $\phi(x) = \ln(0.5e^x + 0.5)$ and SiLU, also called Swish, $\phi(x) = x \cdot \sigma(x) = \frac{x}{1+e^{-x}}$ (*e.g.*, in Allegro, MACE, and SO3krates). The hyperparameters of the NN are learned, like the weights and biases, in the process of backpropagation. After a (possibly random) initialisation of the NN, the output of the output layer is compared to the reference training data and the information is propagated back through the NN (using the chain rule). The backpropagation follows the computed gradients of a loss function w.r.t. the NN parameters, where the loss function quantifies the discrepancies between target and prediction values (further details on the loss function choice for chemical systems will be provided in Subsection 2.6.2). An optimiser utilises these gradients to update parameters iteratively, applying a learning rate to control the step size of each update and the training is performed on batches (*i.e.*, subsets of the training data). A pass through the NN is called an epoch. The training stops when a pre-defined number of epochs elapses, a loss on the training data reaches a desired threshold, or if the performance on a separate validation dataset converges for a set number of consecutive epochs (called patience).

Message-Passing Graph Neural Networks

Graph Neural Networks (GNN) are particularly relevant and well-suited to modelling chemical systems due to the intrinsic match between molecular structure and a graph. The molecules are mapped to an undirected connected graph $G = (V, E)$ of nodes (vertices V) with each node an atom and edges E connecting the nodes representing bonds - both atoms and bonds can have features (*e.g.*, , atom type, hybridisation, bond order, etc).

A particularly prominent type of GNN is message passing NNs (MPNNs) - all the aforementioned NNs are message-passing graph NNs. In them, each node is updated by gathering information from its neighbouring nodes through the so-called message passing. The neighbour list of atoms in these NNs is based on a spatial definition of the interatomic distance. The atomic features are updated iteratively across the molecular graph, achieving a collection of atomic features based on the local environment ('locality' based on a pre-defined cutoff radius *e.g.*, 5-15Å).

Attention

The seminal paper 'Attention is all you need' [27], introduced the concept of attention in ML, which led to a series of innovations such as transformers, with wider impact including in the field of ML for chemical systems. The attention mechanism is a way to dynamically weight the importance of different elements of a system and shift focus based on importance. It builds on previous concepts of ML in learning all dependencies between atoms dynamically. This is conceptually similar to the aforementioned kernel ridge regression process based on the similarity measure (*e.g.*, distance metric) to globally compute how similar two atomic objects are. Furthermore, message passing neural networks (MPNNs), which build on convolutional NNs, are conceptually linked to the idea of the attention mechanism. MPNNs extend the idea of application of the convolutional kernel to graph representation - the convolutional kernel (not to be confused with the aforementioned kernels) generalises to a learned message function that aggregates features from neighbouring nodes and acts as a local filter repeatedly applied (*i.e.*, shared weights for the NN) throughout the input to extract patterns. For the traditional convolutional NNs in signal processing, that would correspond to applying it across, for example, pixels to extract spatial features, while for the MPNNs in this field, it is applied across neighbourhoods of atoms to extract information about structure-property relationships. However, the attention mechanism removes the inherent constraints of the convolutional kernel filter on local dependencies between atoms and allows for non-local interactions.

Formally, it is given by the following function

$$\text{Attention}(Q, K, V) = \textit{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{2.98}$$

where Q, K, and V are the 'Query', 'Key', and 'Value' matrices, respectively. The dot product $QK^\top$ gives the similarity between Q and K, indicating how much *attention* each atom

should pay to the others. The result is scaled by $\sqrt{d_k}$ to prevent large values that could lead to vanishing gradients and given as inputs to the *softmax* function, which converts them into probabilities used to compute a weighted sum of the value vectors V and represents the aggregated information each atom gathers from the other atoms within the cutoff (aka neighbours list of atoms). The *softmax* function has the mathematical form of a normalised probability distribution as follows

$$\textit{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}, \tag{2.99}$$

where z_i is the input score from the i^{th} element.

Attention has two different types used in different contexts - cross and self-attention. In cross attention, the queries are obtained from one set of data, while the keys are values stem from another and obtaining the relationship between the two *e.g.*, between different molecular representations or between structures and properties [155]. In self-attention, the source of all Q, K, and V is the same, which in the context of chemical systems provides a relationship between the atoms and their relative importance to each other based on their interactions. Self-attention is also a key ingredient in the transformer architecture [27], which underpins some state-of-the-art (SOTA) ML models *e.g.*, SO3krates [34]. Self-attention enables each atom to dynamically weigh and aggregate information from other atoms based on learned interaction strengths. It can be applied, as *multi-head self-attention*, to multiple targets simultaneously allowing the model to simultaneously consider different types of atomic relationships, *e.g.*, interaction patterns for different angular momentum $l = 0, 1, 2$.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \tag{2.100}$$

where concatenated results from different heads are passed through a final linear projection with weights W^O, mixing the information from different heads into a unified representation, where each head is obtained as

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{2.101}$$

with W_i^Q, W_i^K, W_i^V, and W^O as learnable parameters.

Physics-inspired modules

The Ziegler–Biersack–Littmark (ZBL) [178] potential describes the short-range repulsive interaction between two nuclei and is especially useful in high-energy collision simulations. The potential is given by

$$V_{\text{ZBL}}(r) = \frac{Z_1 Z_2 e^2}{4\pi\varepsilon_0 r} \cdot \phi\left(\frac{r}{a}\right), \tag{2.102}$$

where Z_1 and Z_2 are the atomic numbers of the interacting nuclei, e is the elementary charge, ε_0 is the vacuum permittivity, and r is the interatomic distance. The screening

function $\phi(x)$ accounts for the partial shielding of nuclear charges by the electron clouds and is defined as

$$\phi(x) = 0.1818e^{-3.2x} + 0.5099e^{-0.9423x} + 0.2802e^{-0.4029x} + 0.02817e^{-0.2016x}. \quad (2.103)$$

The reduced interatomic distance $x = r/a$ uses a screening length a given by

$$a = 0.8854a_0 \left(Z_1^{0.23} + Z_2^{0.23}\right)^{-1}, \quad (2.104)$$

where a_0 is the Bohr radius. The ZBL potential smoothly interpolates the screened Coulomb repulsion between atoms at very short distances, and while the origins of its empirical parameters were in materials science, it has proven as a useful physical term in NNs for molecular systems.

Specific architectures

Let us now consider specific architectures of NNs relevant in this work and also representative of the trends in the field as a whole. *Allegro* builds on the NequIP (Neural Equivariant Interatomic Potentials) [36] framework and is notable for being fully local and linear-scaling. There, a combination of radial basis functions (RBFs) and spherical harmonics form a symmetry-aware representation of the atomic neighbourhood. These features are processed through equivariant tensor layers that ensure the model's outputs, such as energies and forces, and transforms consistently under 3D rotations, translations, and permutations. Notably, Allegro avoids the typical iterative message-passing instead, and instead computes all atomic interactions in a single forward pass, which significantly boosts computational efficiency and parallelism.

MACE , on the other hand, also utilises rotationally equivariant tensors constructed from RBFs but also directional information via spherical harmonics. The equivariant feature of an atom i with angular momentum ℓ, denoted $\mathbf{f}_i^{(\ell)}$, is computed as

$$\mathbf{f}_i^{(\ell)} = \sum_{j \in \mathcal{N}(i)} \sum_n R_n(d_{ij}) \cdot \left[Y_\ell(\hat{r}_{ij}) \otimes \mathbf{c}_{n\ell}\right], \quad (2.105)$$

where $R_n(d_{ij})$ are RBFs of the pairwise distance d_{ij}, $Y_\ell(\hat{r}_{ij})$ are spherical harmonics of order ℓ, and $\mathbf{c}_{n\ell}$ are learnable weight tensors. To build higher-order features, MACE uses multiplicative interactions between equivariant features from previous layers. These are combined via tensor products and projected onto irreducible representations using Clebsch–Gordan coefficients $C^{\ell,m}_{\ell_1,m_1;\ell_2,m_2}$ as follows

$$\mathbf{f}_i^{(\ell)} = \sum_{\ell_1,\ell_2} \sum_{m_1,m_2} C^{\ell,m}_{\ell_1,m_1;\ell_2,m_2} \left(\mathbf{f}_i^{(\ell_1)} \otimes \mathbf{f}_i^{(\ell_2)}\right)^{m_1,m_2}, \quad (2.106)$$

where the sum runs over all valid combinations of angular momentum indices. This operation ensures that the resulting features $\mathbf{f}_i^{(\ell)}$ transform correctly under rotations. This

multiplicative combination over angular channels and neighbours aggregates directional information and forms an implicit atomic cluster expansion, giving rise to the name MACE: Multi Atomic Cluster Expansion.

In a Deep tensor NN (DTNN) [179] like *SchNet* [39], for example, the message from atom j to atom i, $\mathbf{v}_{ij}$, can be encoded in a tensor $\mathbf{W}$ that is built as a tensor product between the transposed embedding of atom j, $\mathbf{c}_j$, and a vector ϕ that encodes the interatomic distance d_{ij} via a Gaussian RBFs (see Subsection 2.6.3). Mathematically, this is expressed as follows

$$\mathbf{v}_{ij} = \mathbf{c}_j^\top \mathbf{W}[\phi(d_{ij})]. \tag{2.107}$$

While the aforementioned architectures use tensor products and physics-inspired design, let us now focus on NNs architectures that leverage also the attention mechanism for molecular property prediction. *SpookyNet* [31] builds on the idea of physics-informed message-passing NNs with the addition of a global attention mechanism and specific explicit physically-motivated energy terms. Its atom-centered features are updated through two complementary modules: a local interaction module and a non-local attention module. In the local module, atom-wise feature vectors are refined via message-passing neural networks, where information is aggregated from neighbouring atoms within a fixed cutoff radius (*e.g.*, 5 Å). In the non-local module, SpookyNet leverages a global multi-head self-attention mechanism, allowing each atom to attend to all others in the system, regardless of spatial distance. Attention weights α_{ij} are computed from a learned compatibility function based on atomic features and their relative positions, and used to update the atomic features as follows

$$\mathbf{f}_i' = \mathbf{f}_i + \sum_j \alpha_{ij} \cdot \mathbf{W}\mathbf{f}_j, \tag{2.108}$$

where $\mathbf{W}$ is a learnable projection matrix. This mechanism enables the network to model long-range quantum interactions—such as polarisation and dispersion—hence the NN's name alluding to "spooky action at a distance" [180]. Additionally, SpookyNet maintains latent electronic variables per atom, which are updated jointly with the atomic features and influence both local and non-local updates.

To further incorporate physical mechanisms, SpookyNet has optional explicit energy contributions through differentiable physics-based modules: a short-range ZBL repulsion term to prevent unphysical overlaps (*vide retro*), a Coulombic electrostatics term based on predicted partial atomic charges q_i,

$$E_{\text{Coul}} = \frac{1}{4\pi\varepsilon_0} \sum_{i<j} \frac{q_i q_j}{r_{ij}}; \tag{2.109}$$

and a D4-style van der Waals dispersion correction capturing long-range interactions via damped pairwise potentials as follows

$$E_{\text{vdW}} = -\sum_{i<j} \sum_{n\in\{6,8\}} s_n \frac{C_n^{ij}}{r_{ij}^n} f_{\text{damp}}^{(n)}(r_{ij}). \tag{2.110}$$

After successive updates through these modules, the final atomic features are used to predict molecular properties *e.g.*, per-atom energy contributions, which are summed to yield the total molecular energy or atomic forces obtained as energy gradients.

SO3krates is an $\mathrm{SO}(3)$-equivariant NN, which constructs atom-centered features transforming as irreducible representations (irreps) of the rotation group. Given atomic positions $\{\mathbf{r}_i\}$, relative position vectors $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ are expanded using learnable RBFs $R_n(d_{ij})$ and spherical harmonics $Y_\ell(\hat{\mathbf{r}}_{ij})$, where the invariant and equivariant features are separate representations used to form geometric basis tensors $\mathbf{B}_{ij}^{(\ell)} = \sum_n R_n(d_{ij}) Y_\ell(\hat{\mathbf{r}}_{ij})$ - components, which bear similarity with the MACE NN. However, for SO3krates, each atom i has feature tensors $\mathbf{f}_i^{(\ell)}$ of angular momentum order ℓ, initialised from atomic embeddings and updated via equivariant message passing. Neighbouring features are combined via Clebsch–Gordan tensor products $\mathbf{B}_{ij}^{(\ell_1)} \otimes \mathbf{f}_j^{(\ell_2)}$, which are projected onto irreps ℓ using pre-computed Clebsch–Gordan coefficients to ensure equivariance, yielding messages $\mathbf{m}_i^{(\ell)}$. Features are updated by applying equivariant non-linearities $\phi^{(\ell)}$, such as gated or norm-based activations, to $\mathbf{f}_i^{(\ell)} + \mathbf{m}_i^{(\ell)}$. This process is repeated over multiple layers to aggregate information from larger neighbourhoods while preserving rotational equivariance. Finally, scalar or tensorial molecular properties are predicted by linearly mapping the appropriate irreps of the final atom features, typically using scalar ($\ell = 0$) components for scalar targets.

The *SO3LR* [168] NN builds on the ideas of So3krates and SpookyNet as it augments the SO3krates architecture with physics-inspired modules including a ZBL repulsion term and dispersion and electrostatic terms. Unlike SpookyNet, however, the dispersion term in SO3LR is based on MBD, and the electrostatic one is a fixed pairwise Coulomb-like term, which are implemented separately.

Chapter 3

Extending quantum-mechanical benchmark accuracy to biological ligand-pocket interactions

---

Parts of this chapter have been published in this or similar form in:

M. Puleva, L. Medrano Sandonas, B. Lőrincz, J. Charry, D. M. Rogers, P. R. Nagy, A. Tkatchenko "Extending quantum-mechanical benchmark accuracy to biological ligand-pocket interactions", *Nature Communications* **16**, 8583 (2025),

and have been produced in collaboration with the above authors. I generated the QUID dimers and performed DFT optimisation; I performed also DFT, semiempirical, empirical, and SAPT calculations, organised, analysed and visualised the results, and contributed with conceptualisation and methodology.

Accurate computational modelling of physicochemical phenomena in protein-ligand systems is vital for accelerating the early stages of the drug development pipeline [181–184]. In fact, reliable and "clean room" experimental measurements of the binding affinity can be expensive due to multiple factors, *e.g.*, dissolved electrolytes, solvent concentration, and target protein misfolding [185], making robust computational methods crucial for improving efficiency and gaining detailed mechanistic understanding into the ligand-protein systems [186]. Within the computational frameworks that are at our disposal, understanding and consequently controlling non-covalent interactions (NCIs) in ligand-protein systems in particular, can aid the compound design process to achieve optimal target selection [187]. For this purpose, accurate calculations are indeed critically important, as even errors of 1 kcal/mol can lead to erroneous conclusions about relative binding affinities [188] (see also Chapter 4). MMFFs, along with docking and free-energy methods, have been widely used due to their affordable computational cost for estimating structural and thermodynamical properties of complex (bio)molecular systems[189, 190].

Despite significant progress in the development of accurate and polarisable MMFFs[141, 142], most of them treat ubiquitous non-covalent polarisation and dispersion interactions using effective pairwise approximations, often resulting in inaccuracies or lack of transferability between different chemical subspaces, both being prerequisites for accurate novel drug predictions [191, 192]. Meanwhile, a broad range of QM methods have become available with different trade-offs between accuracy and size spanning from the less expensive but more approximate semi-empirical [126, 129] approaches, passing through the DFT [58–60, 98, 100, 111, 113, 114, 118] ones to the "gold standard" CC [193] and QMC [194–196] methods. However, achieving a reproducible and sufficiently reliable for drug design purposes QM description of NCIs remains computationally prohibitive for realistic ligand-pocket systems, preventing further development of accurate and efficient free-energy simulation methods as well as enhanced mechanistic models for ligand-protein design.

Building accurate DFT, SE, MMFF or ML models, furthermore, depends on reliable reference data, which remains a challenge for large and more complex non-covalent molecular systems. Due to the typically small strength of the NCIs (∼0.5-5 kcal/mol) compared to covalent bonds (∼50-200 kcal/mol) and the possible long-range effect of the NCIs, they are particularly challenging for most computational methods. Thus, ensuring the reliability of benchmark data for protein-ligand interactions is a stepping stone towards very accurate and reliable computational approaches for large (bio)molecular systems, whether DFT, classical FF, or ML methods. Pre-existing datasets aiming to model protein-ligand interactions, such as the SPICE [197] and Splinter [198] datasets, are typically generated as structural data and optimised at a FF level. This gap in general has also been recently targeted by the later release of Meta dataset OMol25[199] (in June 2025) as an aggregation of many types of chemical systems.

Aiming to look at the systematic steps needed to close this gap and improve our understanding of ligand-pocket binding, the "QUantum Interacting Dimer" (QUID) benchmarking framework was developed here. QUID is a dataset of larger molecular dimers up to 65 atoms, that contains complex reference systems that are toy models for protein-ligand interactions. The purpose of the data set is to create a reliable benchmark for equilibrium and out-of-equilibrium interactions along non-covalent bond dissociations. The quality of the dataset calculations is compared against two *ab initio* approaches, namely Coupled Cluster (CC) and Quantum Monte Carlo (QMC). Building on these references, an in-depth investigation into state-of-the-art DFT, SE, and classical FF approaches is conducted, also including the analysis of other physicochemical properties such as van der Waals force components at equilibrium, polarisabilities, and dipole moment.

The QUID dataset, introduced in the next sections, provides a blueprint on robust dataset creation for complex systems; as well as a challenging benchmark for classical, physical chemistry, and MLFF models; and an informative probe into current methods to shed light into their performances and limitations and inform the shortcomings of reference data that the latter are built on.

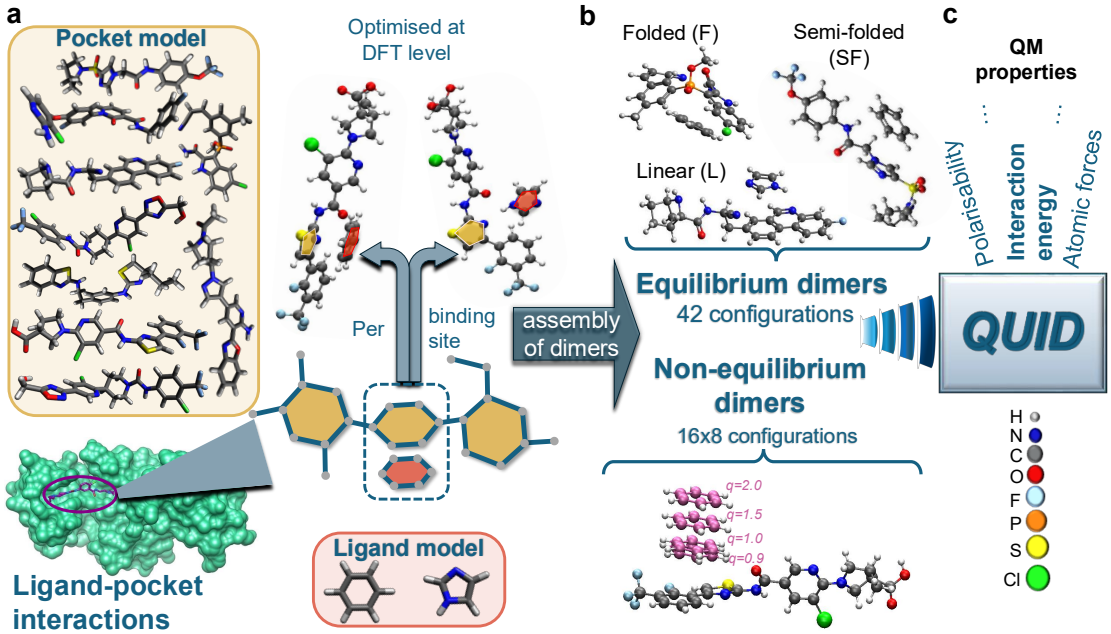

Figure 3.1: Figure from [57]. In panel **a**, the molecules forming a QUID structure modelling a protein pocket and those modelling a ligand are presented at the top and right side of a protein-ligand complex [200] visualised with ChimeraX [201]. Each dimer is composed of one of nine big monomers containing multiple potential binding sites and a small monomer binding to one of them. The resultant dimer arrangement is optimised at DFT level PBE0+MBD. The resulting conformer geometries are shown in panel **b**, categorised as Linear, Semi-Folded, and Folded. For 16 equilibrium dimers, eight non-equilibrium conformations are designed along the dissociation of the non-covalent bond as illustrated by an example. q is a multiplicative dimensionless factor in the range of 0.9 to 2, which denotes the ratio of the inter-monomer distance to that of the equilibrium dimer. In panel **c**, a graphic summary of the QUID dataset with its chemical composition and some available QM molecular properties are shown.

3.1 Generation

3.1.1 Creation rationale

In order to represent a variety of motifs of ligand-pocket systems, different binding sites of nine large flexible chain-like drug molecules from the Aquamarine dataset[56] were exhaustively explored to create the QUID structures. They were obtained via systematic probing of these pocket-like binding sites with a small monomer, once with benzene (C_6H_6) and once with imidazole ($C_3H_4N_2$), representing the ligand motif.

For a selection of 16 equilibrium dimers, eight non-equilibrium conformations are also generated per each, sampling along the non-covalent bond dissociation direction. Given the structural and chemical diversity of the resulting conformations, a single dimer can exhibit

multiple types of steric effects and NCIs simultaneously, including, but not limited to, polarisation, π–π stacking, hydrogen and halogen bonds - to be elaborated upon in Section 3.3.

The design process of QUID was inspired by several landmark QM datasets. Most of them are focused on DFT-based physicochemical properties of single molecules up to a few hundred atoms[56, 202–205]. Only a limited number investigate NCIs, via the interaction energies (E_{int}), in molecular systems at the benchmark *ab initio* level of CCSD(T)/CBS. Among them, one can find the well-established S22 [206, 207] and S66(x8) [53, 54] datasets; L7 with a few specific larger systems [208], as well as the newer NENCI[209], DES370K[210], and SAPT10K[211] ones with improved chemical diversity. Specifically for modelling ligand-pocket interactions, a recent dataset Splinter [198] has been developed – in it two distinct small molecules represent common fragments in proteins and small-molecule ligands. While Splinter features charged monomers and good chemical diversity, its compounds are all of similar size, up to ≈40 atoms, thus offering limited venue for reproducing size-dependent NCIs or geometric arrangements typical of ligands in a pocket. Furthermore, large QM datasets of energies and atomic forces have been generated for non-covalent systems by combining structural data from MMFF-based simulations with DFT methods[23, 168, 197, 212] to develop more robust machine-learned FFs for (bio)molecular simulations. A recent development to address this has been the newly released in June 2025 OMol25 dataset [199] featuring a plethora of organic molecules, including some molecular dimers optimised at DFT level, but designed to create diverse systems to cover CCS for high throughput pipelines rather than for systematic in-depth analysis and benchmarking for a specific target type.

The QUID framework aims to redefine the state-of-the-art in benchmarking NCIs in complex molecular systems. Firstly, a tight agreement between two completely different gold standard methods for solving the Schrödinger equation is established: LNO-CCSD(T) and FN-DMC, thereby largely reducing the uncertainty in highest-level QM calculations. Secondly, the analysis of interaction components allows for the description of a wide range of NCIs of relevance to ligand-protein systems. Thirdly, a comprehensive analysis of approximate empirical, semi-empirical, and first-principles calculations paves the way to pinpoint improvements required in each of these methodologies to move towards trustworthy free-energy simulation methods. Hence, only a comprehensive combination of such benchmark analyses enables an unbiased understanding of NCIs in realistic molecular complexes as illustrated in this Chapter for model ligand-pocket systems.

3.1.2 Generation procedure

Modelling the NCIs of a ligand in a protein pocket is essential for determining the structural arrangements of natural enzyme substrates and drug candidates. This is why the QUID model systems, which represent the most frequent ligand-pocket interaction types, were created to investigate the impact of adequately describing NCIs on binding features and the influence of structural binding conformations on electronic properties. Each QUID dimer comprises a large monomer as a host and a small monomer representing a ligand motif.

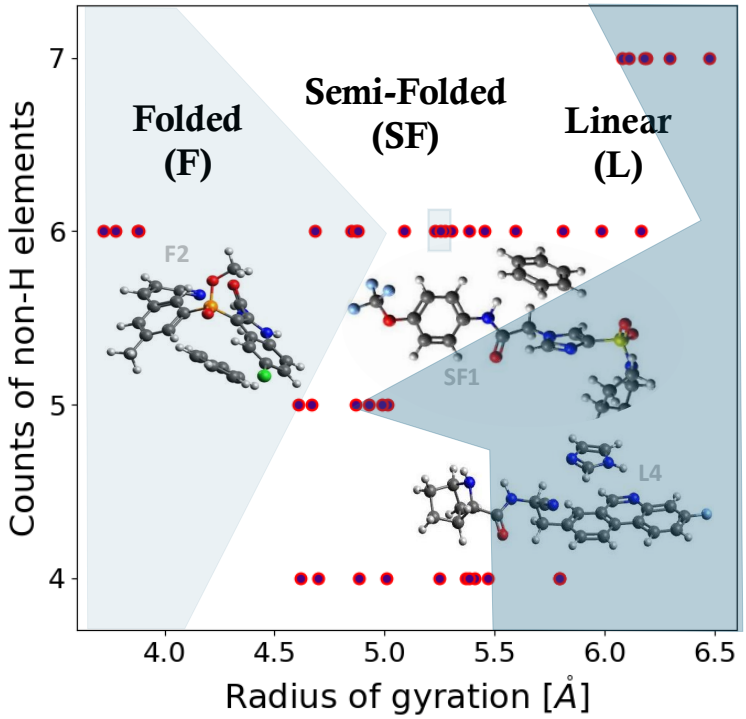

Figure 3.2: Figure from [57]. The number of different heavy (non-Hydrogen) elements in a dimer versus the radius of gyration for all equilibrium QUID dimers. The 'Folded', 'Semi-Folded', and 'Linear' geometry categories are shown on the scatter plot as well.

To achieve a proxy model representation of the interactions on the pocket-ligand surface, the large monomers are chosen to meet stringent criteria from chemically diverse drug-like molecules of ∼50 atoms, with flexible chain-like geometry allowing for folding and multiple accessible binding sites (aromatic rings) (see Fig. 3.1). In doing so, nine molecules (including C, N, O, H, F, P, S, and Cl atoms) that met the criteria were extracted from the Aquamarine dataset [56]. Two small monomers were selected to represent the ligand interactions: benzene, the quintessential aromatic compound present in the phenylalanine side-chain, and imidazole, present in histidine, a more reactive and also a commonly used drug motif [213]. The resulting complexes represent the three most frequent interaction types appearing on the pocket-ligand surface, that is aliphatic-aromatic, H-bonding, and π-stacking, which are found in more than 100,000 interactions within PDB structures [214]. The QUID dimers are comprised of monomers interacting in one or more of the aforementioned ways, with many presenting non-covalent effects of mixed character, *e.g.*, combining π-stacking and H-bonding, which will be explored in-depth in Section 3.3. In each initial dimer conformation, the aromatic ring of the small monomer was aligned with that of the binding site at a distance of 3.55±0.05 Å (similar to S66 dimers [53]), and the dimer was then optimised at the PBE0+MBD level of theory to produce the equilibrium dimer structures in QUID.

Post-optimisation, 42 QUID equilibrium dimers were obtained and then split into three

categories based on the structural shape of the corresponding large monomer: 'Linear', in which the original chain-like geometry is mainly retained; 'Semi-Folded', in which parts of the large monomer are bent while other sections remain linear, and 'Folded', in which the big monomer encapsulates the smaller one. Thus, a variety of pockets with different packing densities are represented by dimers included in the QUID dimers, *e.g.*, the folded F1I3 mimicking a more crowded binding pocket [215], or the linear L2B1 representing a toy model of a more open surface pocket [216]. This classification is shown in terms of the radius of gyration and chemical diversity in Fig. 3.2. As a result, a wide range of interaction energies E_{int} between the monomers is produced, ranging from -24.3 to -5.5 kcal/mol (at PBE0+MBD level single point calculations), with imidazole usually resulting in stronger non-covalent bonds. The prediction of the E_{int} values will be investigated in Section 3.4 with diverse computational methods, and further properties of these systems are explored in Section 3.6.

Next, a representative selection of 16 dimers is used to construct non-equilibrium conformations along the dissociation pathway of the non-covalent bond (along π-π or H-bond vector, see Section 3.2), modelling snapshots of a ligand binding to a pocket. These conformations were generated at eight distances, characterised by a multiplicative dimensionless factor q, defined as the ratio of the inter-monomer distance to that of the equilibrium dimer. The chosen values of q are 0.90, 0.95, 1.00, 1.05, 1.10, 1.25, 1.50, 1.75, and 2.00, where $q = 1.00$ denotes the equilibrium dimer. The structure of these non-equilibrium dimers was also optimised at PBE0+MBD level with the heavy atoms of the small monomer and the respective binding site kept frozen. The resulting systems demonstrate the varied E_{int} spectrum for different pocket types via the structure categories in equilibrium and along the dissociation paths. The generation protocol for the 42 equilibrium and 128 (16x8) non-equilibrium dimers is schematically outlined in Fig. 3.1 and detailed in this Section. Their investigation with different methods is presented in Section 3.5. These model systems represent a significant step forward in accurately investigating ligand-pocket interactions, characterised by robustly optimised molecular dimers that exhibit chemical diversity, larger size, and complex binding conformations.

3.2 Computational details

In this work the used basis sets are def2-QZVPPD, aug-cc-pVTZ, 'tight' in the `FHI-aims` software [217], a highly accurate numerical atom-centered orbitals including polarisation and diffuse functions and devised to reach 'chemical accuracy' *i.e.*, 1 kcal/mol [78]. General information about the basis sets can be found in the theoretical background in Chapter 2 and exact details about the ones employed here in Appendix B.

Counterpoise corrections were applied to PBE0+MBD, PBE-QIDH+D3, and CCSD(T) single-point calculations. The basis set superposition error was negligibly small (under 1.5%) for DFT and ca. 4% on the average for CCSD(T) when extrapolated to the complete basis set (CBS) limit (see results in Fig. B.3 in Appendix B).

The interaction energies E_{int} of QUID dimers were calculated using the supramolecular approach,

$$E_{\text{int}} = E_{dimer} - (E_{L_{monomer}} + E_{S_{monomer}}). \tag{3.1}$$

To investigate the level of agreement among QM methods for calculating E_{int} of QUID dimers, a selection of well-performing hybrid, double hybrid, range-separated hybrids DFT functionals is considered here, including M06-2X [98], ωB97X+D3 [100], ωB97M-V [101] ωB97X-V [218], PBE+MBD [219], PBE-QIDH+D3 [107], B3LYP+D3 [220, 221], CAM-B3LYP+XDM [103] and BH&HLYP+XDM [93]. Additionally, the PBE0 functional was combined with multiple two-body or many-body corrections: MBD [59, 60] (range-separated self-consistent screening (MBD@rsSCS) approach), MBD-NL [114], XDM [109], TS-vdW [111], D4 [113, 118], ωB97X+D3 [100], ωB97M-V [101]. These calculations were performed using either the `FHI-aims`, `Psi4` [222, 223] or the `QChem` software [224] - please find exact details on the used software versions, basis sets, and for XDM parametrisation in Appendix B. SAPT energy decomposition calculations were carried out at the sSAPT0/jaDZ level [225] employing the `Psi4` software [222] (version 1.9.1). At the semiempirical level, E_{int} was calculated via single-point calculations using DFTB3+MBD [129] with DFTB+ software [226] and GFN2-xTB with the xTB software [227].

Regarding MM methods, the results for AMBER [141] were obtained using Openbabel [228] for molecular format conversion. The parametrisation with AmberTools and GAFF2 [141] required manual assignment and adjustment of bonds for more complex cases, such as ring interactions, as well as modification of the self-consistent loop limits for the F2I1 dimer. The CHARMM-CGenFF [142] calculations were conducted using OpenMM [229] following a CGenFF2 [230] parametrisation. For these calculations, manual inclusion of the dihedral angles for the flexible side chains, such as the 'C-C-N' type, was necessary. An example is the L4B1 dimer, which was assumed to exhibit relatively low flexibility due to the nature of its bonds and chemical environment.

Additionally, the optimised structures of equilibrium and non-equilibrium QUID dimers were also utilised for more accurate QM single-point calculations using PBE0+MBD level of theory to compute other physicochemical properties (as detailed in Table B.1 in Appendix B). For these calculations, the `FHI-aims` code [231] was used together with "tight" settings for basis functions and integration grids (exact details in Appendix B). The MBD energies and MBD atomic forces were here computed using the range-separated self-consistent screening (rsSCS) approach [60], while the atomic C_6 coefficients, isotropic atomic polarisabilities, molecular C_6 coefficients and molecular polarisabilities (both isotropic and tensor) were obtained *via* the SCS approach [59]. Here, also computed were the van der Waals forces using D4 and XDM methods. Hirshfeld ratios correspond to the Hirshfeld volumes divided by the free atom volumes. In the TS dispersion energy the vdW radii were obtained using the SCS approach *via* $R_{\text{vdW}} = \left(\alpha^{\text{SCS}}/\alpha^{\text{TS}}\right)^{1/3} R^{\text{TS}}_{\text{vdW}}$, where α^{TS} and $R^{\text{TS}}_{\text{vdW}}$ are the atomic polarisability and vdW radius computed according to the TS scheme, respectively. Atomisation energies were obtained by subtracting the atomic PBE0 energies from the PBE0 total energy of each molecular conformation.

3.3 Analysis of non-covalent interaction components

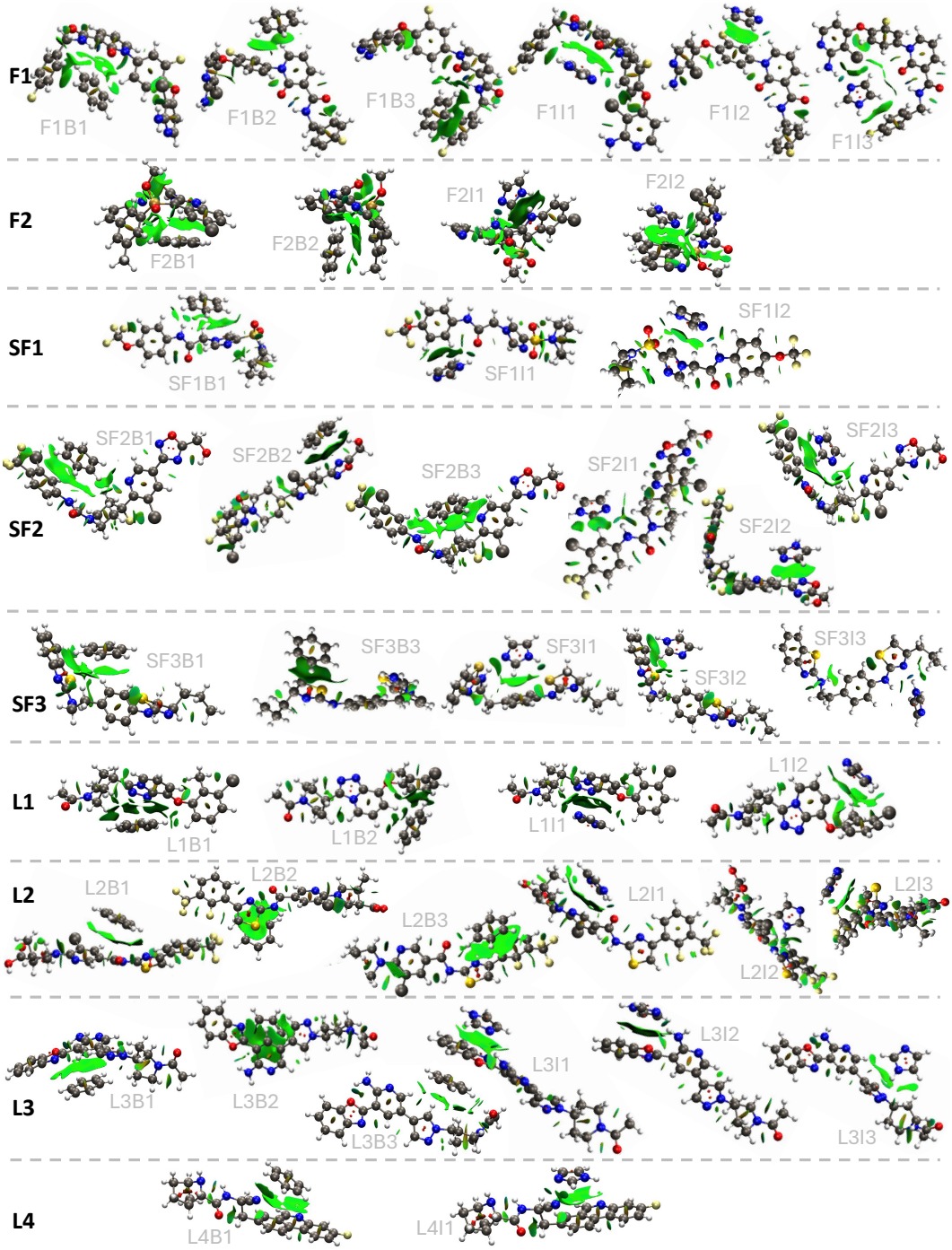

Figure 3.3: Figure from [57]. Analysis of the character of the non-covalent bonds in all equilibrium QUID dimers with NCI-plot software [232], where the green surfaces denotes van der Waals-type interactions between the dimers or between atoms on the same dimer.

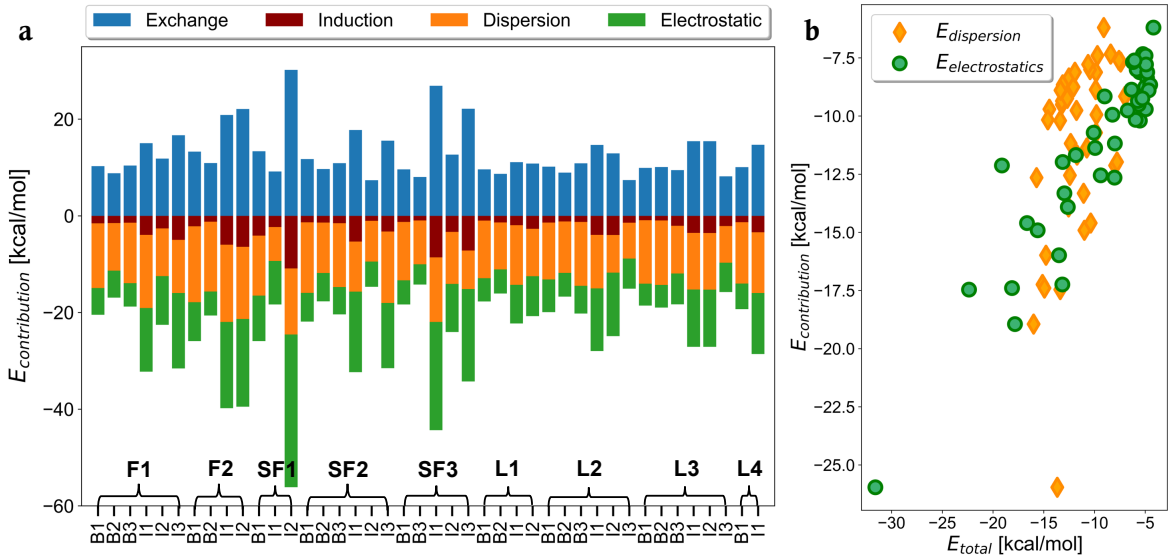

Figure 3.4: Figure taken from [57]. a) The Exchange, Induction, Dispersion, and Electrostatic contributions to the interaction energies E_{int} of all equilibrium QUID dimers are depicted for the results obtained with sSAPT0/jaDZ. b) Scatter plot of the Dispersion and Electrostatics contributions compared to the total interaction energies is shown for all equilibrium QUID dimers.

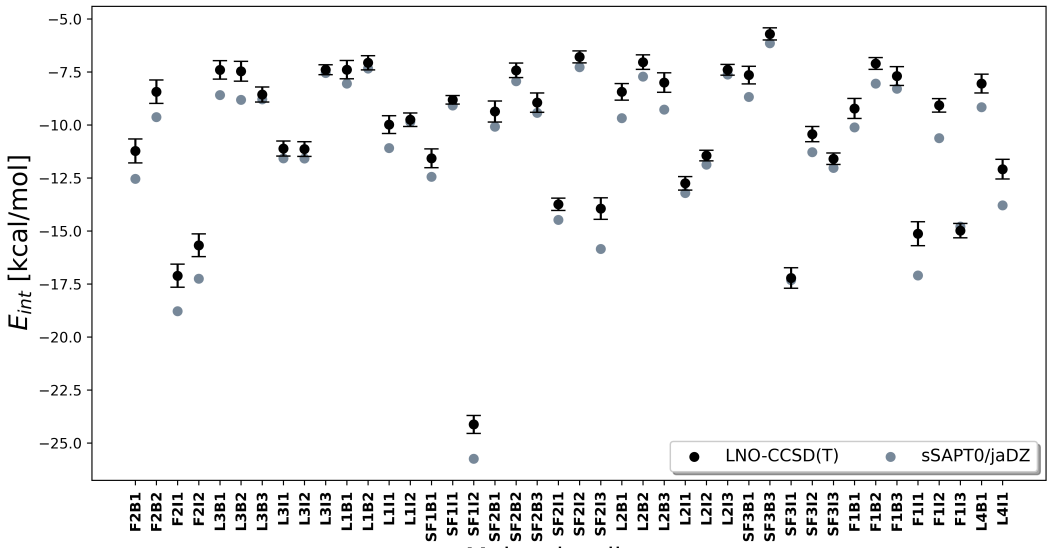

Figure 3.5: Figure from [57]. QUID molecular dimers dataset analysis of non-covalent bonds in all equilibrium dimers with sSAPT0 method implemented in Psi4 for all equilibrium dimers. QUID molecular dimers dataset analysis of non-covalent bonds in all equilibrium dimers with sSAPT0 method compared to the AMBER-GAFF2 results with the interaction energies w.r.t LNO-CCSD(T) results.

A detailed characterisation of the physical interactions in ligand-pocket systems is needed to both aid understanding and ensuring diversity in the coverage of interactions. Within that context, a quantitative SAPT analysis provides insight into the energy components of the NCIs, namely induction, dispersion, electrostatic, and exchange contributions, which elucidate the balance of intermolecular interactions. A decomposition of E_{int} of QUID systems with SAPT analysis was performed, specifically with the sSAPT0 version due to its good balance between accuracy and computational cost [225], results are shown in Fig. 3.4. sSAPT0 is a slightly modified recipe that involves a rescaling of $E^{(20)}_{exch-disp}$ and $E^{(20)}_{exch-ind}$ terms based on an empirically adjusted proportion between $E^{(10)}_{exch}$ and $E^{(10)}_{exch}(S^2)$ - further details on the SAPT framework can be found in Appendix D. The sSAPT0 E_{int} predictions for the equilibrium dimers were found qualitatively consistent with those computed at the LNO-CCSD(T) level (MAE of 0.85 kcal/mol). The largest discrepancies are found for dimers with imidazole as the small monomer and with both π–π stacking and H-bonding contributions to the non-covalent interaction. The most pronounced discrepancy occurs for the folded F1I1 dimer, with a value of 1.97 kcal/mol. Case-by-case results for all equilibrium QUID dimers are shown in Fig. 3.5. This SAPT measures of dispersion and electrostatic components has been used before to gain insight into specific protein-ligand interactions [211, 233] and will provide a solid basis for interpreting the different predictions of E_{int} from different QM methods. Qualitatively, each QUID system's interaction landscape is shown in Fig. 3.3, where the different type of NCIs contribute to the isocountour surfaces. Among them, of particular interest in this work are the H-bonds and $\pi - \pi$ interactions giving rise to the observed dominant contributions of the electrostatics and dispersion terms, correspondingly.

3.4 Equilibrium QUID dimers results

3.4.1 *Ab initio* reference

The calculations were performed by Peter Nagy and Balázs D. Lőrincz for Coupled Cluster and prepared by Jorge Charry for Quantum Monte Carlo.

Reliable models for pocket-ligand systems rest on robust methods performing consistently and accurately in such systems. A prerequisite towards understanding and estimating the performance of current methods is the existence of reliable data, which can be challenging when results in literature obtained at "gold standard" level of computation have been found to disagree [234, 235], and comprehensive and computationally expensive studies are needed to explore sources of the discrepancy [236]. Hence, to establish a thoroughly dependable reference for the interaction of ligands with a protein pocket, E_{int} for the QUID proxy systems has been obtained and compared for the two gold standard methods LNO-CCSD(T) [237–240] and FN-DMC [194–196] to produce a "platinum standard" (not to be confused with the platinum standard as used in the quantum chemistry community). Within our methodology, the thus defined "platinum standard" was used as a

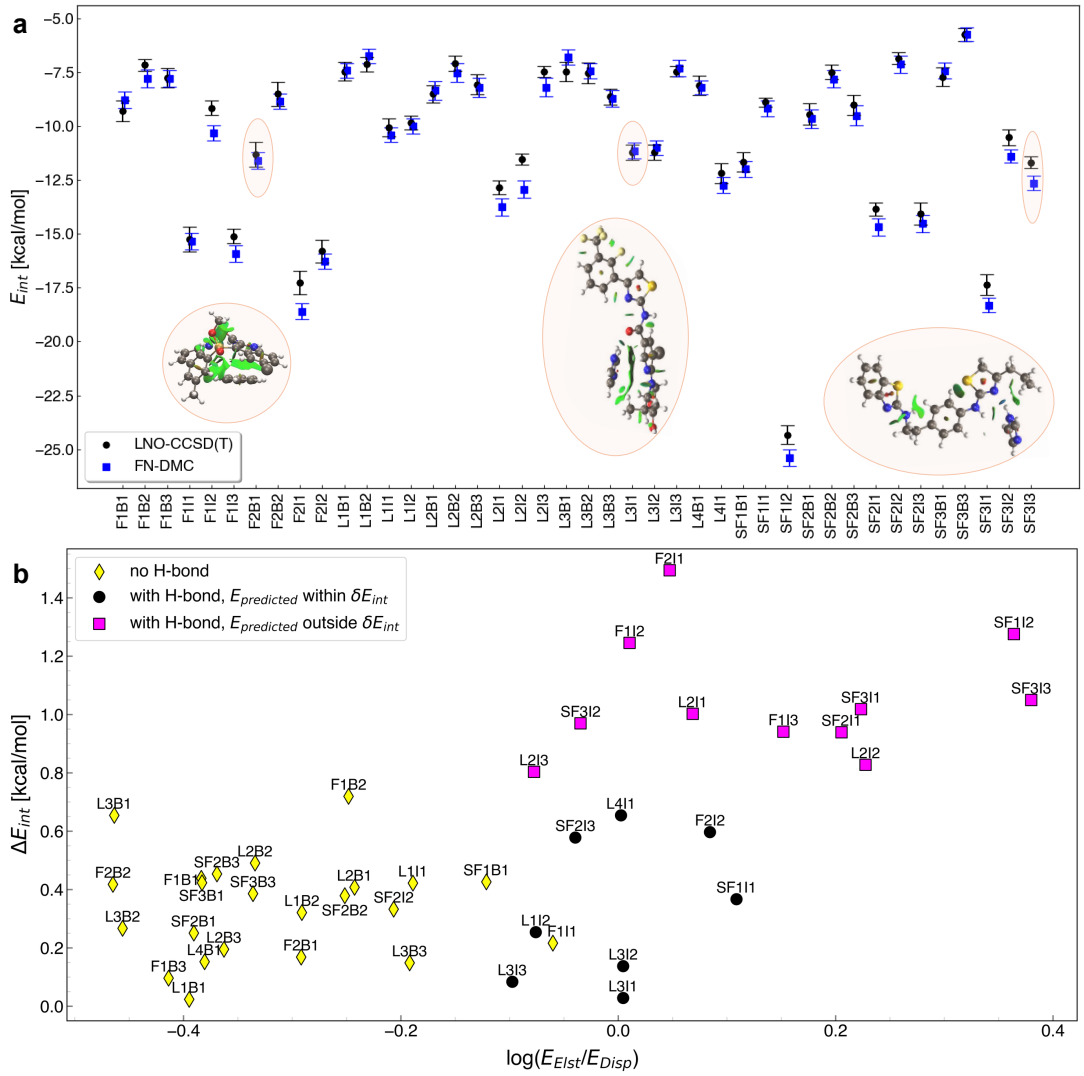

Figure 3.6: Figure from [57]. **a** Comparison of the interaction energies (E_{int}) computed using FN-DMC (0.015 or 0.025 time step) and LNO-CCSD(T) (extrapolated to Complete Basis Set (CBS) and Local Approximation-Free (LAF) limit) for the 42 equilibrium QUID structures. For FN-DMC, error bars represent estimated one-σ statistical error for 4×10^8 configurations. For LNO-CCSD(T), the error bars correspond to the estimated uncertainty from the best CBS and LAF extrapolations. Three cases are highlighted: L3I1 for which the methods are in perfect agreement, F2B1 for which the methods agree within their uncertainty estimates, and SF3I3 as the one case for which they are in slight disagreement. The NCI plots illustrating the non-covalent interactions in those molecular dimers are also shown [232]. **b** Scatter plot of the absolute differences in the prediction of the interaction energies between LNO-CCSD(T) and FN-DMC, ΔE_{int}, versus the log of the ratio between the electrostatic (Elst) and dispersion (Disp) SAPT components from sSAPT0 [225]. The equilibrium QUID dimers are divided into three subsets: yellow (no H-bond in non-covalent interaction) and black (H-bond in non-covalent interaction) symbols indicate cases where the E_{int} predictions agree within their uncertainty estimates. Pink symbols denote dimers for which the predictions do not agree within uncertainty estimates, all of these feature an H-bond between the monomers.

basis for further exploration, and to ensure its robustness both approaches were employed with particular care to achieve convergence in accordance to best practices, see the "Methods" section for more details. The results were compared for the equilibrium set of QUID dimers and found in agreement within the uncertainty estimates of the two reference quantum methods in 31 of the 42 cases (*i.e.*, 74%) as seen in Fig. 3.6. The MAE between the two methods is 0.47 kcal/mol compared to 0.38 kcal/mol mean absolute value of the uncertainty estimate for both FN-DMC and LNO-CCSD(T), respectively. The benchmark *ab initio* methods are in good agreement for the QUID systems, a result previously found unattainable for larger non-covalent systems with dispersion-dominated interactions, *e.g.*, in the L7 dataset [234, 241]. In QUID, let us take as an example one case (SF3I3) out of the studied ones, where the LNO-CCSD(T) prediction with its uncertainty estimate lies outside one-sigma agreement with the FN-DMC result but remains within two sigmas. This still means statistical consistency considering 68% and 95% assigned to one and two sigma intervals, respectively.

Analysis of the discrepancy patterns between FN-DMC and LNO-CCSD(T) was performed by assessing the character of the non-covalent bonding from sSAPT0/jaDZ. The results were found to be consistent with a recent study performed on the S66 dimers dataset [235] - in both cases the results indicate that the dominant electrostatic component of the interaction energy correlates with the disagreements between the gold benchmark methods. Details are given in Fig. 3.6b, presenting a plot of the difference in the interaction energy predictions versus a log of the ratio of the electrostatic and dispersion sSAPT0 components. For the QUID systems, all the cases of disagreement involve a H-bond between the monomers, although some dimers with H-bonds in the interaction with higher dispersion contribution were found to be in agreement. These results are in line with the 0.9 kcal/mol deviation found by Shi et al. [235] for the acetic acid dimer. From this perspective, the QUID systems differ from the supramolecular complexes with extended π–π interactions, where some considerable disagreements between CCSD(T) and FN-DMC were uncovered [234]. As noted in Ref. 234 and recent studies, [235, 242–244] the FN approximation, time-step discretization, pseudopotential, post-CCSD(T) terms, basis set extrapolation used in CCSD(T), and other high-order effects could be notable for extended π–π interactions. However, beyond CCSD(T) corrections are deemed to be very small for our purposes in H-bonded dimers [242, 244]. Hence, based on these studies and the comparison between "gold standard" methods, LNO-CCSD(T) is taken as a practical and reliable reference for E_{int} of ligand-pocket NCIs in the complex QUID dimers. LNO-CCSD(T) results were subsequently obtained for all 42 equilibrium dimers and the full dissociation curves of a representative selection of 6 dimers (details in Subsection 3.5).

3.4.2 Methods exploration

Given the robust E_{int} reference, next a comprehensive and reliable examination of its prediction has been conducted and corresponding measures obtained from different approximate computational methods for capturing NCIs in QUID equilibrium systems. This is done

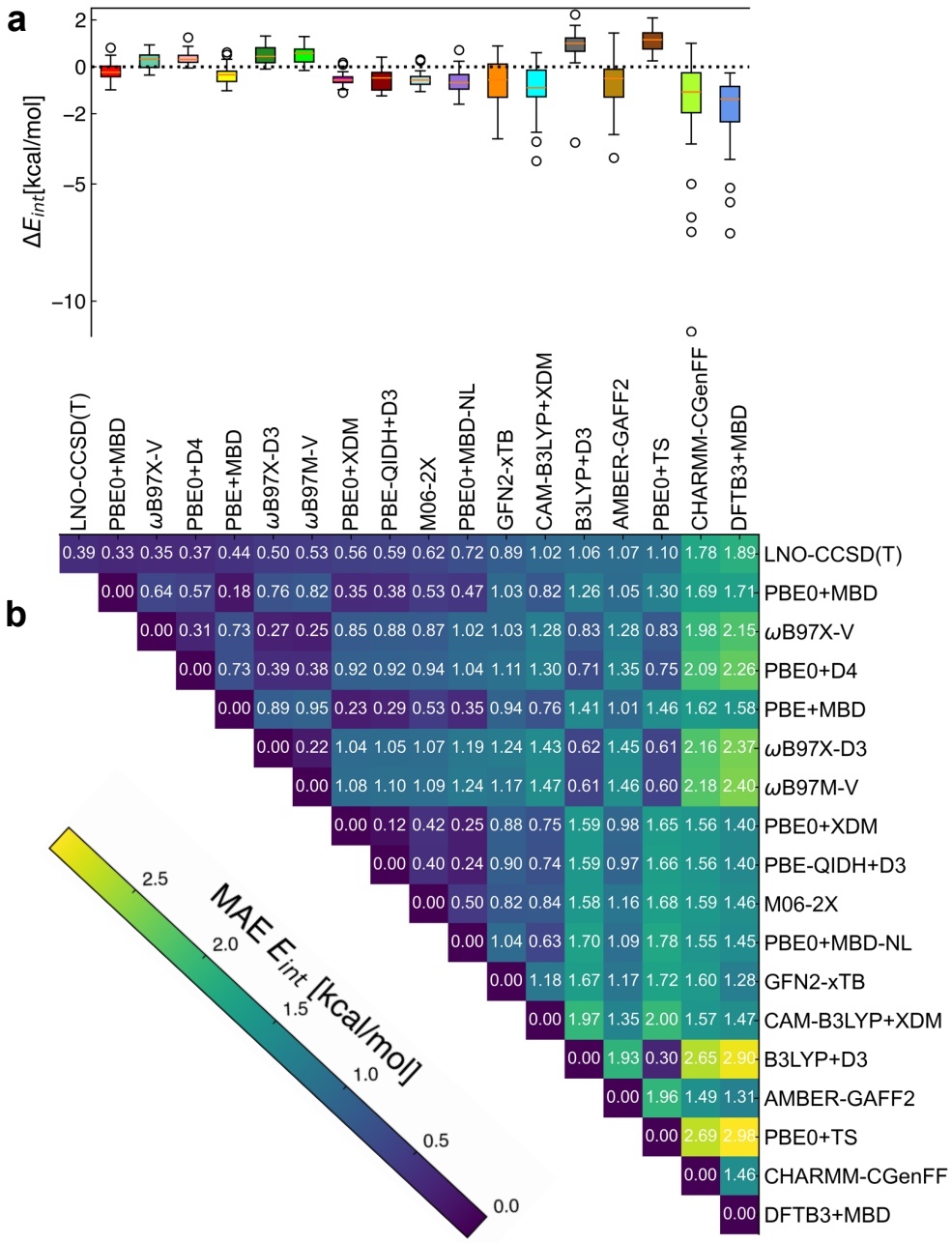

Figure 3.7: Figure from [57]. a) Distributions of interaction energy predictions w.r.t. LNO-CCSD(T), ΔE_{int}, showed via box plots, for a selection of computational methods - DFT methods: PBE0+MBD, ωB97X-V, PBE0+D4, PBE+MBD, ωB97X+D3, ωB97M-V, PBE0+XDM, PBE-QIDH+D3, PBE0+MBD-NL, CAM-B3LYP+XDM, B3LYP+D3, PBE0+TS; SE methods: DFTB3+MBD, GFN2-xTB; and classical FFs: AMBER-GAFF2 and CHARMM-CGenFF. The negative ΔE_{int} values signify underbinding, while the positive ones overbinding. b) A heatmap of MAE values of predicted E_{int} w.r.t LNO-CCSD(T) for the QUID equilibrium dimers in the first column, and the MAE of all methods w.r.t each other in subsequent columns. The computational methods to predict E_{int} were for the same methods as in a. *For the LNO-CCSD(T) method, the value shown with asterisk is the mean absolute of the uncertainty estimates for E_{int}.

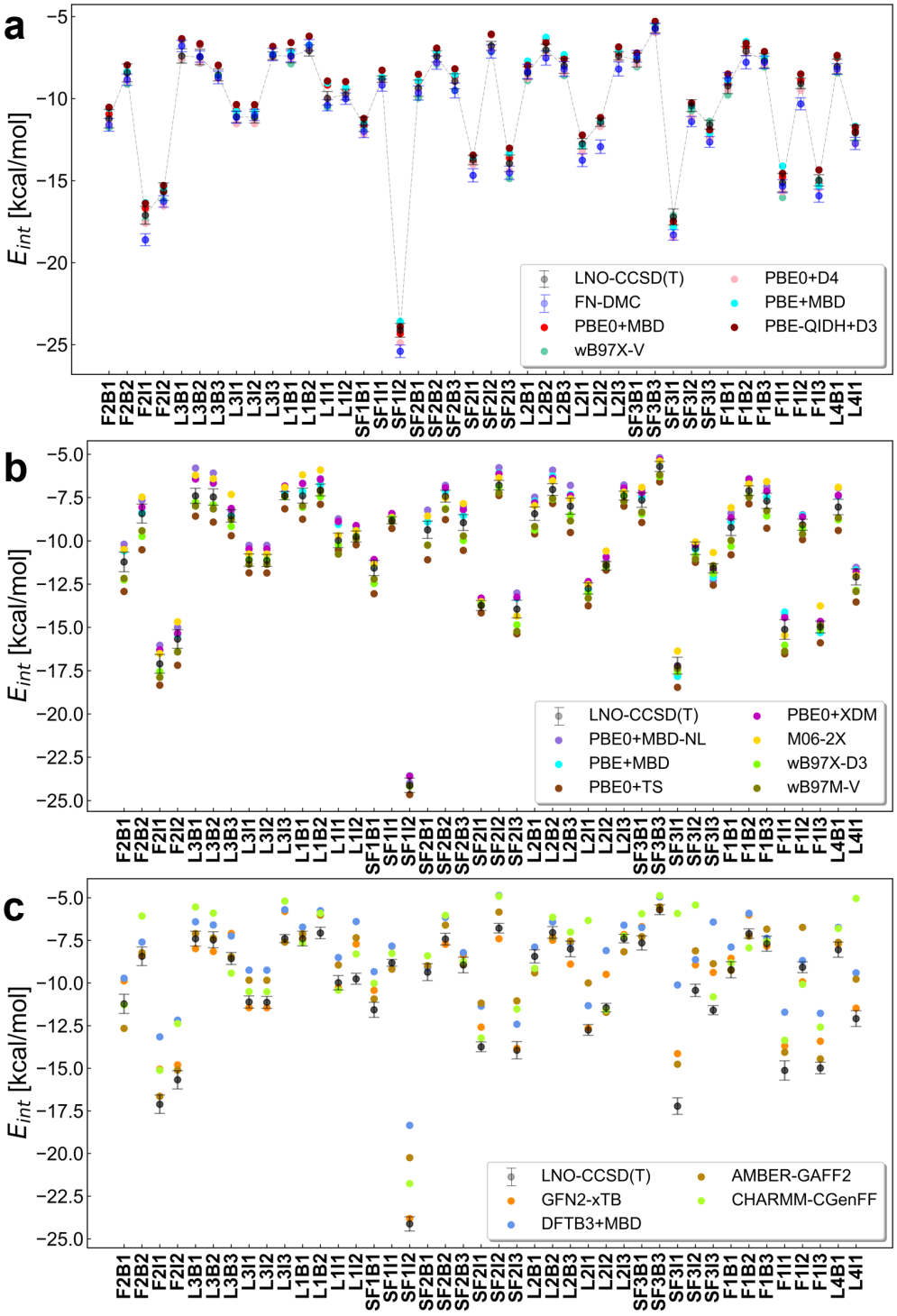

Figure 3.8: Figure from [57]. QUID 42 equilibrium structures: interaction energy E_{int} calculated with LNO-CCSD(T) as reference and FN-DMC - both with uncertainty estimates, and the methods a) at DFT level, PBE0+MBD, PBE0+D4, ωB97X-V, PBE+MBD, and PBE-QIDH+D3. b) PBE0+MBD-NL, PBE0+TS, PBE0+XDM, ωB97X+D3, ωB97M-V, M06-2X, PBE+MBD. c) the semiempirical GFN2-xTB and DFTB3+MBD, and the classical FFs AMBER-GAFF2 and CHARMM-CGenFF.

with the goal of identifying approximate methods that can be used in eventually building a trustworthy pipeline for calculating binding affinities.

With the aim of providing a systematic investigation of QM and MM approximations, a wide selection of methods is included here. Firstly, a variety of DFT functionals is studied (*e.g.*, global, range-separated, and double hybrids) with dispersion interactions selected from previous benchmark studies [245, 246], namely PBE0+MBD, PBE0+D4, ωB97X-V, ωB97X+D3, ωB97M-V, PBE+MBD, PBE0+XDM, PBE-QIDH+D3, CAM-B3LYP+XDM, B3LYP+D3, M06-2X, PBE0+MBD-NL, and PBE0+TS. Secondly, among the SE methods the DFTB3+MBD [129] and GFN2-xTB [126] were studied. From the available empirical classical FFs, GAFF2 (computed with AMBER) [141] and CHARMM-CGenFF [142] (computed with OpenMM) [229] are included. The results are presented in Table 3.1 and as a spread of E_{int} predictions obtained with these methods with respect to the LNO-CCSD(T) reference values, ΔE_{int}, shown in Fig. 3.7a. They are presented in ascending order of MAE, whose values can be found in the first column of Fig. 3.7b. The performance must be analysed in the context of the intrinsic uncertainty estimate of the LNO-CCSD(T) method on the QUID dataset (mean value 0.39 kcal/mol). From this perspective, PBE0+MBD, ωB97X-V, and PBE0+D4 are within the uncertainty of LNO-CCSD(T), although a case-by-case analysis reveals deviations beyond the uncertainty of the benchmark data.

Overall, it is encouraging to note that all recent **DFT** approximations yield rather accurate results. Even if some of them underestimate (PBE0+XDM, M06-2X, PBE0+MBD-NL, PBE-QIDH+D3) or overestimate (B3LYP+D3, PBE0+TS) the reference interaction energies on average, the spread of deviations is rather narrow for *all* DFT methods (with the only exception B3LYP+D3 despite the high-level def2-QZVPPD basis set used). On the other hand, both **empirical and semiempirical methods** show a tendency to underbind, producing larger spreads and exhibiting large outliers. The most prominent outliers are found for the methods CHARMM-CGenFF and DFTB3+MBD, with errors ranging from -12.5 kcal/mol to -5 kcal/mol (details in Fig. 3.8) and -7.5 kcal/mol to -4.5 kcal/mol, correspondingly. Examining the DFTB3+MBD outliers–SF1I2, SF3I1, and SF3I3 dimers–reveals that for SF1I2 the strongest interactions are driven by electrostatics, including contributions from a sulfonyl group at the binding site, while the SF3Ix dimers exhibit reactive thiazole groups. It is noticeable that the error distributions of GFN2-xTB (a SE method) and AMBER-GAFF2 (an empirical FF) are quite similar, although GFN2-xTB has a slightly lower average deviation from LNO-CCSD(T). Since GFN2-xTB was partially fitted to CCSD(T) data while AMBER-GAFF was not, this is not surprising.

For AMBER-GAFF2, dimers with high electrostatic contributions result in larger errors (see Fig. 3.9) pointing to a limitation of fixed partial charges. Contrarily, for GFN2-xTB, the higher errors appear to be associated with the local chemical environment. For example, the presence of P (in all F2Ix dimers), S (in all SF3Ix dimers) or Cl atoms (in both Folded F1I1, F1I3 and Linear structures L1I2, L2I2) as well as H-bonds (in L3I3) affects the bonding of imidazole ligands, presenting greater challenges for the method.

To evaluate the level and areas of agreement between the investigated DFT, SE, and FF

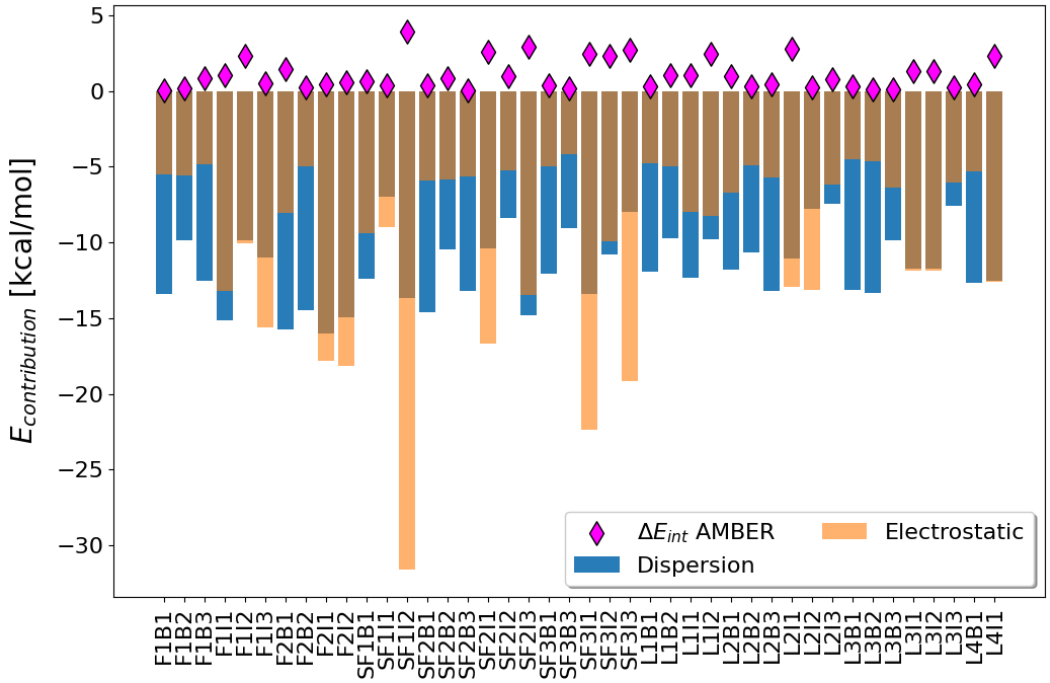

Figure 3.9: Figure from [57]. QUID molecular dimers dataset analysis of non-covalent bonds in all equilibrium dimers with sSAPT0 method implemented in Psi4 for all equilibrium dimers. QUID molecular dimers dataset analysis of non-covalent bonds in all equilibrium dimers with sSAPT0 method compared to the AMBER-GAFF2 results with the interaction energies w.r.t results from LNO-CCSD(T).

methods, also computed are the MAE values for E_{int} of QUID equilibrium dimers relative to each other (see Fig. 3.8b). The **DFT** functionals ωB97X+D3 and ωB97X-V show excellent agreement with each other (MAE = 0.25 kcal/mol) despite the distinct incorporation of dispersion terms, D3 and non-local correlation VV10, in ωB97X+D3 and ωB97X-V, respectively [100, 218]. This indicates that the critical similarity between ωB97X-V and ωB97X+D3, which sets them apart from the Minnesota functional M06-2X or PBE0+MBD, is the range-separation treatment of the DFT functional [15]. Particularly important for QUID systems appears to be the long-range handling of the electron-electron interactions, as the short-range ones differ for the GGA and meta-GGA functionals [15] (see theoretical background in Chapter 2). In the same vein, considered are the related PBE0+MBD and PBE0+MBD-NL methods (MAE of 0.47 kcal/mol) - it is noticeable that the MBD-NL method increases the deviation compared to MBD in almost all cases except for the dimer with two S atoms and imidazole ligand, SF3I1-3 (see Fig. 3.8). Also worth noting is that the MBD-NL functional was designed to achieve broad applicability to inorganic solids and molecular systems, while the original MBD method was developed for molecular systems. Interestingly, while the PBE0+MBD and the PBE+MBD functionals produce comparable results, with MAEs of 0.33 kcal/mol and 0.44 kcal/mol respectively, the double-hybrid PBE-QIDH+D3 functional, based on the same PBE method, achieves an MAE of 0.59 kcal/mol (2.63 kcal/mol without the D3 correction). This may be a result of the shortcomings found in MP2 contributions for large flexible π-π dispersion-dominated molecular systems [247] and the role of van der Waals parametrisations for double hybrid functionals [248].

The next investigation step is delving more in-depth in the performance of the three best performing methods PBE0+MBD, ωB97X-V, and PBE0+D4, all of which obtain E_{int} within the LNO-CCSD(T) uncertainty estimate (0.39 kcal/mol) at 0.33 kcal/mol, 0.35 kcal/mol, and 0.37 kcal/mol, respectively. As the chemical environment and energetic balance in the NCIs proved to be a more distinguishing factor for the method than the structure types, the focus is on a consideration of dispersion versus electrostatics contributions to E_{int}. Overall, the MAE value of the 14 electrostatics-dominated dimers for PBE0+MBD is 0.26 kcal/mol, notably better than the PBE0+D4 results with an MAE of 0.56 kcal/mol. On the other hand, for the 28 dispersion-dominated dimers, PBE0+D4 yields 0.27 kcal/mol, while PBE0+MBD obtains a close MAE of 0.35 kcal/mol. This suggests that systems with stronger electrostatic interactions pose a greater challenge for the D4 dispersion correction. This could stem from the different underlying mechanisms of the two approaches for modelling long-range correlation effects [59, 60, 118, 249, 250]. ωB97X-V achieves correspondingly 0.25 kcal/mol, outperforming for the electrostatics-dominated dimers but has higher MAE of 0.40 kcal/mol for the dispersion-dominated ones.

In summary, empirical and semi-empirical methods have demonstrated variable performance for ligand-pocket model systems in QUID, yielding a MAE of about 1 kcal/mol or higher and exhibiting a tendency to underbind. In contrast, among the many DFT methods examined, PBE0+MBD, ωB97X-V, and PBE0+D4 proved most effective in capturing the complex QM effects contributing to E_{int} calculations, while PBE0+XDM also showed excellent performance as a pairwise dispersion method. These findings enhance the understanding

ab initio	MA(ΔE)[kcal/mol]	
LNO-CCSD(T)	0.39	-
DFT functional + dispersion	MAE [kcal/mol]	RMSE [kcal/mol]
GGA hybrids		
PBE+MBD	0.44	0.51
Global hybrids		
PBE0+MBD	0.33	0.41
PBE0+D4	0.37	0.45
PBE0+XDM	0.56	0.60
M06-2X	0.62	0.73
PBE0+MBD-NL	0.72	0.81
B3LYP+D3	1.06	1.20
PBE0+TS	1.10	1.18
BH&HLYP+XDM	1.34	1.98
Range-separated hybrids		
ωB97X-V	0.35	0.41
ωB97M-V	0.53	0.63
ωB97X+D3	0.50	0.62
CAM-B3LYP+XDM	1.02	1.34
Double hybrids		
PBE-QIDH+D3	0.59	0.63
Semiempirical methods	MAE [kcal/mol]	RMSE [kcal/mol]
DFTB3+MBD	1.90	2.43
GFN2-xTB	0.89	1.13

Table 3.1: Table from [57]. MAE and RMSE values for a variety of DFT and semiempirical methods w.r.t LNO-CCSD(T) in kcal/mol for the equilibrium QUID dimers, where the DFT functionals are listed with the added dispersion method, and ordered based on their type, and per type in ascending order of MAE values for the equilibrium QUID dimers. The mean absolute value of the uncertainty estimate for the LNO-CCSD(T) results is also provided as a reference at the top.

of the applicability and limitations of the various investigated computational methods – however, the choice of method for simulating a ligand binding to a protein pocket should

also account for non-equilibrium conformations.

3.5 QUID dimers dissociation results

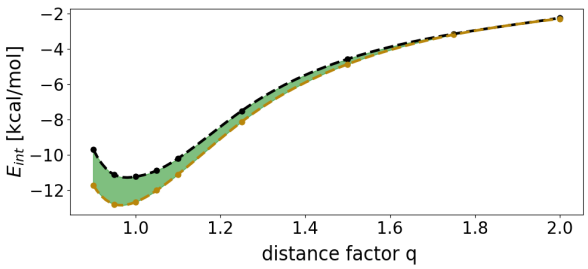

Figure 3.10: Figure from [57] and based on an image in [251]. An example of the delta metric (as employed in literature [251, 252]) for the F2B1 dimer, where the method shown in AMBER-GAFF2.

Dimer	Δ[kcal/mol]						
	PBE0+MBD	PBE0+D4	ωB97X-V	PBE0+XDM	GAFF2	GFN2-xTB	DFTB3+MBD
F2B1	0.11	0.23	0.33	0.32	0.55	0.96	0.86
F2I1	0.23	0.20	0.33	0.44	0.36	1.67	2.59
SF2B2	0.16	0.21	0.21	0.20	0.39	0.36	0.92
SF2I2	0.23	0.22	0.24	0.23	0.31	0.36	0.98
L2B3	0.28	0.22	0.23	0.29	0.35	0.45	0.52
L2I3	0.10	0.22	0.16	0.14	0.70	0.10	0.48

Table 3.2: Table from [57]. Delta metic (Δ) results for the dissociation curves w.r.t LNO-CCSD(T) of the selection of 6 dimers F2B1, F2I1, SF2B2, SF2I2, L2B3, and L2I3 for the best DFT, semiempirical, and classical FF methods investigated.

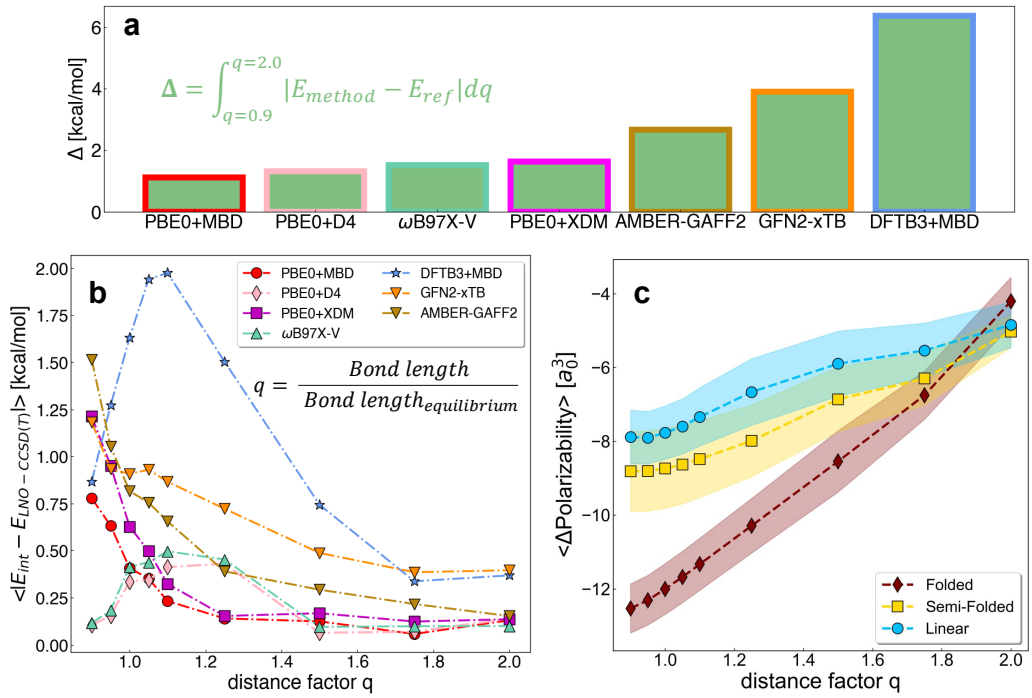

Figure 3.11: Figure from [57]. **a** The delta metric (Δ, see formula on the plot) results are shown for four DFT methods: PBE0 including MBD, D4, and XDM, ωB97X-V; two SE methods: DFTB3+MBD, GFN2-xTB; and a classical FF method: AMBER-GAFF2. **b** Average of the absolute difference of predicted interaction energy with the LNO-CCSD(T) reference along the dissociation of the non-covalent bond of a selection of dimers. The average is calculated at each multiplicative distance factor q (ranging from 0.9 to 2.0), defined as the ratio between the bond length and the equilibrium non-covalent bond length for the corresponding dimer. The average is shown for six selected molecular dimers: F2B1, F2I1, SF2B2, SF2I2, L2B3, and L2I3, using the same methods as in plot **a**. **c** Average of the difference in the molecular polarisability of the dimer and the sum of isotropic polarisabilities of its corresponding monomers at each distance factor q. The results are shown as an average over all non-equilibrium dimers in the QUID dataset, split by structural type in Linear, Semi-Folded, and Folded.

Non-covalent bond dissociation pathways: non-equilibrium dimers. A key factor in modelling the dynamics of ligand-pocket systems is the capability of a physical model to investigate systems out of equilibrium accurately. To that end, six representative dissociation curves are considered (*i.e.*, F2B1, F2I1, L2B3, L2I3, SF2B2, and SF2I2) and conducted an in-depth analysis of the performance of selected computational methods: PBE0+MBD, PBE0+D4, PBE0+XDM, GFN2-xTB, DFTB3+MBD, and AMBER-GAFF2. The choice of DFT functional was based on the long-range effects in the elongated non-covalent bond regime as seen in Table 3.3 and allowed for direct comparison within the same functional PBE0. In Fig. 3.11a are presented the averaged results (over six dimers) for the 'Delta metric' Δ that measures the agreement between the dissociation curves of a given computational method

Method	MA($\Delta E_{compressed}$)	MA($\Delta E_{elongated}$)
LNO-CCSD(T)	0.45	0.26
DFT functional + dispersion	MAE $E_{compressed}$	MAE $E_{elongated}$
GGA functionals		
PBE+MBD	0.84	0.20
Global hybrids		
PBE0+MBD	0.61	0.17
PBE0+D4	0.20	0.24
PBE0+XDM	0.93	0.23
PBE0+TS	0.83	0.59
PBE0+MBD-NL	1.10	0.44
M06-2X	2.12	0.60
Range-separated hybrids		
ωB97X-V	0.24	0.28
ωB97X+D3	0.35	0.37
ωB97M-V	0.70	0.27
Double hybrids		
PBE-QIDH+D3	1.13	0.41
Semiempirical and classical FF methods		
AMBER-GAFF2	1.13	0.41
GFN2-xTB	1.01	0.63
DFTB3+MBD	1.26	1.14

Table 3.3: Table from [57]. Interaction energy MAE in kcal/mol for non-equilibrium dimers w.r.t. LNO-CCSD(T) for a variety of methods: DFT, semiempirical, and classical FF, where the DFT functionals are listed with the added dispersion method, and ordered based on their type. The non-equilibrium dimers are considered in two regimes: 'compressed', for which the non-covalent bond length is smaller or equal to the equilibrium one, and 'elongated', for which the non-covalent bond length is longer than in equilibrium. As a reference, also the mean of the LNO-CCSD(T) uncertainty estimates in the two regimes are provided at the top.

and the LNO-CCSD(T) reference (see Fig. 3.10). To assess the efficiency of the methods in different interaction regimes, two ranges have been defined: *compressed* for $q \leq 1.0$ and *elongated* for $q > 1.0$. The MAE values for E_{int} in the *compressed* and *elongated* regimes obtained using all methods are provided in Table 3.3. Indeed, it is evident that SE and classical FF methods, which are the tools of choice for biomolecular modelling, perform notably worse than DFT methods. The Δ values per dimer are listed in Table 3.2, with the best performing equilibrium methods PBE0+MBD and PBE0+D4 achieving smaller Δ values. These findings are confirmed by analysing the average error of E_{int} w.r.t. LNO-CCSD(T) at each q, see Fig. 3.11b (corresponding six individual plots are available in Fig. 3.12-Fig. 3.17). Notably, the performance of each method shows a strong dependence on the intermolecular distance. To elucidate the results, the dissociation curve profiles for all methods are presented individually in Fig. 3.12-3.17. Interestingly, unlike DFT methods (*vide supra*), AMBER-GAFF2 either underestimates or overestimates E_{int}, depending on the dimer configuration. The discrepancies are more pronounced in configurations where dispersion components dominate the NCI, and for those dimers particularly at distances with factor $q < 1.0$, where dispersion interactions are stronger. On the other hand, both SE methods predominantly underestimate E_{int} and fail to accurately capture the position of the minimum on the dissociation curve or its overall shape. This behaviour changes only for GFN2-xTB at $q < 1.0$, where it overestimates E_{int} in most cases.

Concerning DFT methods, the best performance across the dissociation curves is consistently displayed by PBE0+MBD (underestimation) and PBE0+D4 (overestimation), with errors remaining within the uncertainty estimates of LNO-CCSD(T). As expected from previous results, PBE0+D4, ωB97X-V, ωB97X-D3, PBE0+D4, and PBE0+MBD yield the best results in the *compressed* regime, while PBE0+MBD, PBE+MBD, PBE0+XDM, PBE0+D4, ωB97M-V, and ωB97X-V perform best in the *elongated* regime, indicating the consistently good performance of the PBE0 and ωB97X functionals. The SF2B2 and SF2I2 dimers proved to be the most challenging among those examined, likely due to the interaction of a 5-membered oxadiazole ring (C_2N_2O) via π–π stacking with the small monomer. The presence of two N and one O atoms in the aromatic ring contributes to an increase in the dipole moment and polarisability of the monomer, thereby enhancing both electrostatic and dispersion interactions. Interestingly, F2B1 and F2I1 are the easiest dimers to predict among the examined methods, as the molecular environment contributing to NCI is located within a few Å of the molecule.

While the analysis of the dissociation curves has confirmed the performance of the methods for computing E_{int} (*vide supra*), it has also revealed that the accuracy of SE methods and classical FF strongly depends on the distance range and the dimer configuration. This is a critical result, as both approaches are widely used to investigate intermolecular interactions in biomolecular simulations, raising questions about the reliability of the results obtained in molecular dynamics simulations carried out with empirical methods.

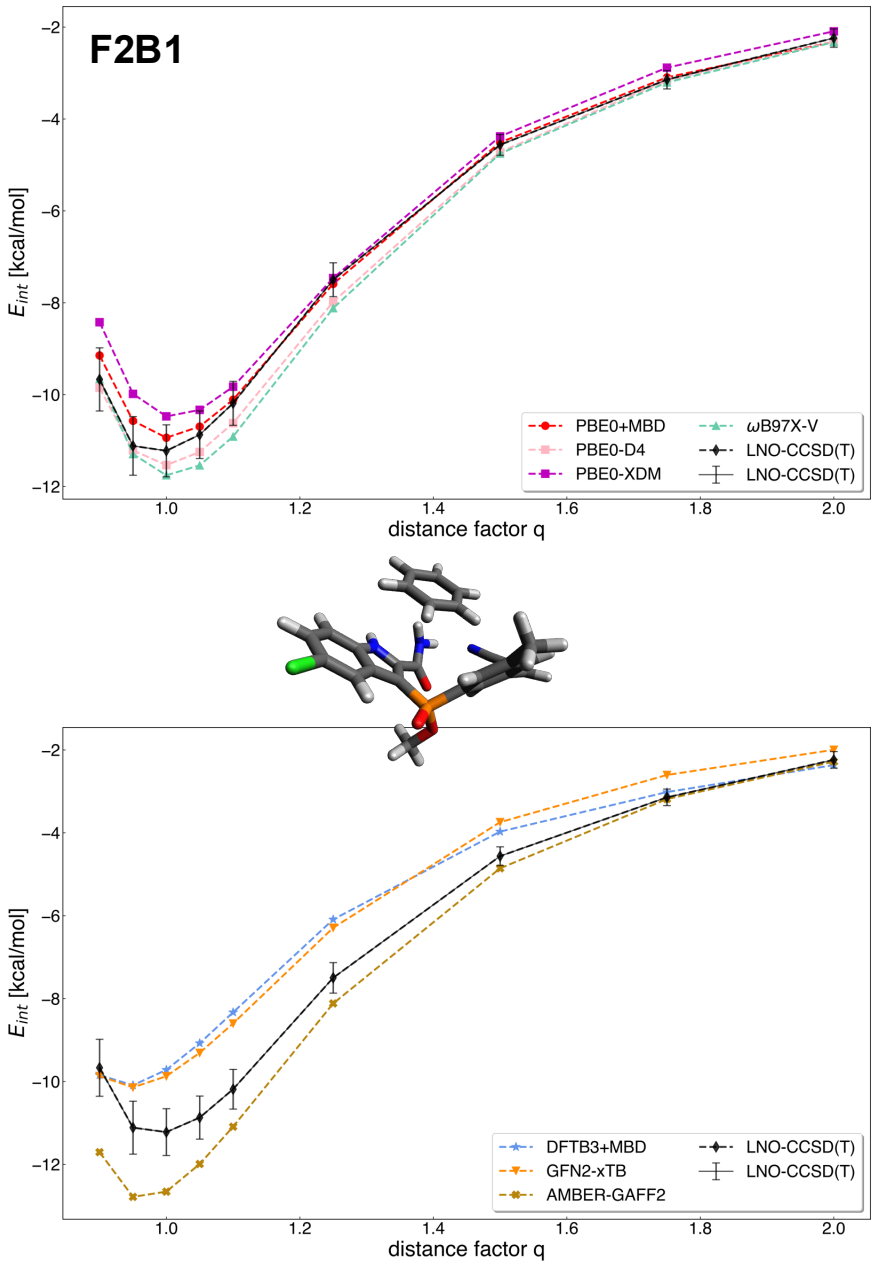

Figure 3.12: Figure from [57]. Dissociation curves for the the non-equilibrium dimer F2B1, for which LNO-CCSD(T) level reference is shown. The DFT methods PBE0+MBD, PBE0-D4, PBE0+XDM, ωB97X-V in the top plot, and semi-empirical methods GFN2-xTB, and DFTB3+MBD, as well as the classical force field AMBER-GAFF2 in the bottom plot, are presented.

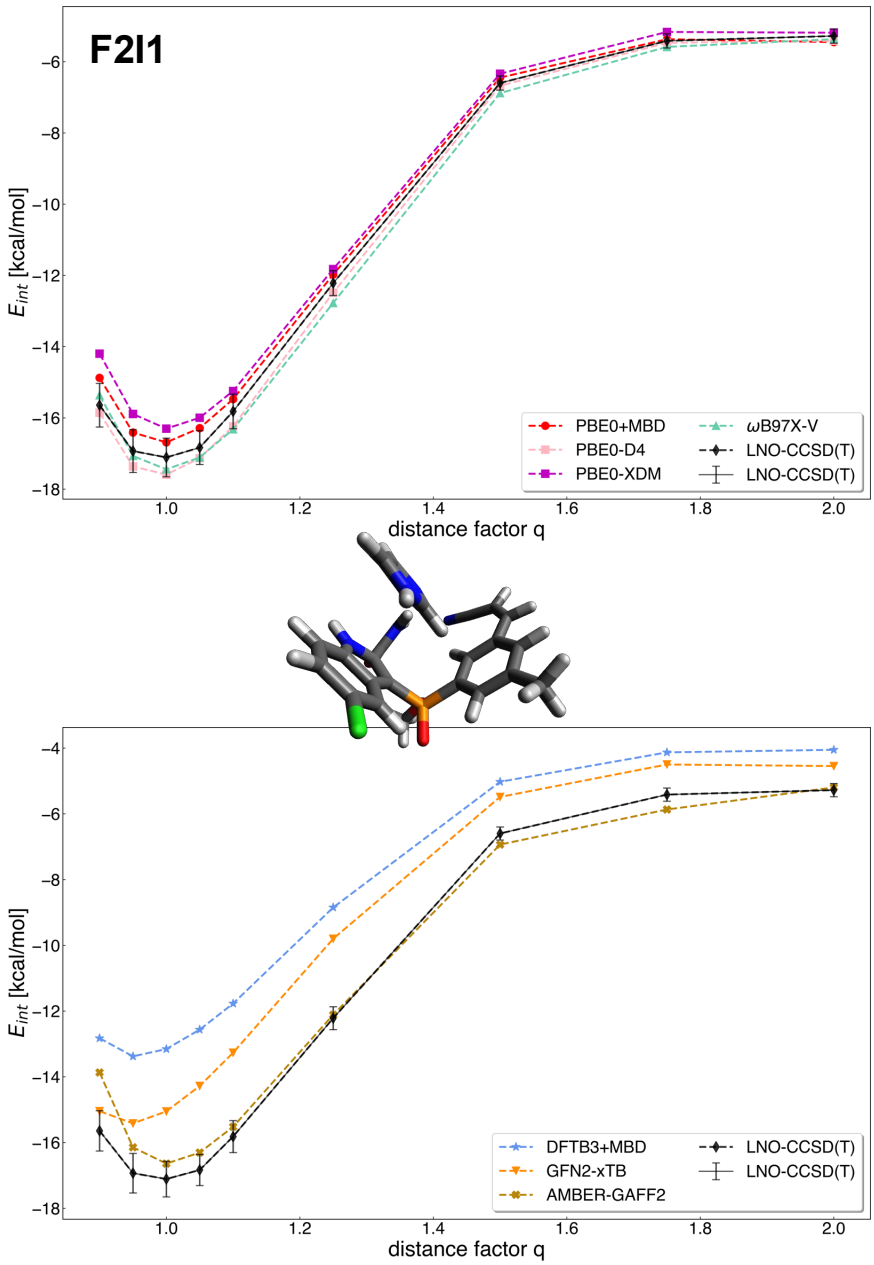

Figure 3.13: Figure from [57]. Dissociation curves for the the non-equilibrium dimer F2I1, for which LNO-CCSD(T) level reference is shown. The DFT methods PBE0+MBD, PBE0+D4, PBE0+XDM, ωB97X-V in the top plot, and semi-empirical methods GFN2-xTB, and DFTB3+MBD, as well as the classical force field AMBER-GAFF2 in the bottom plot, are presented.

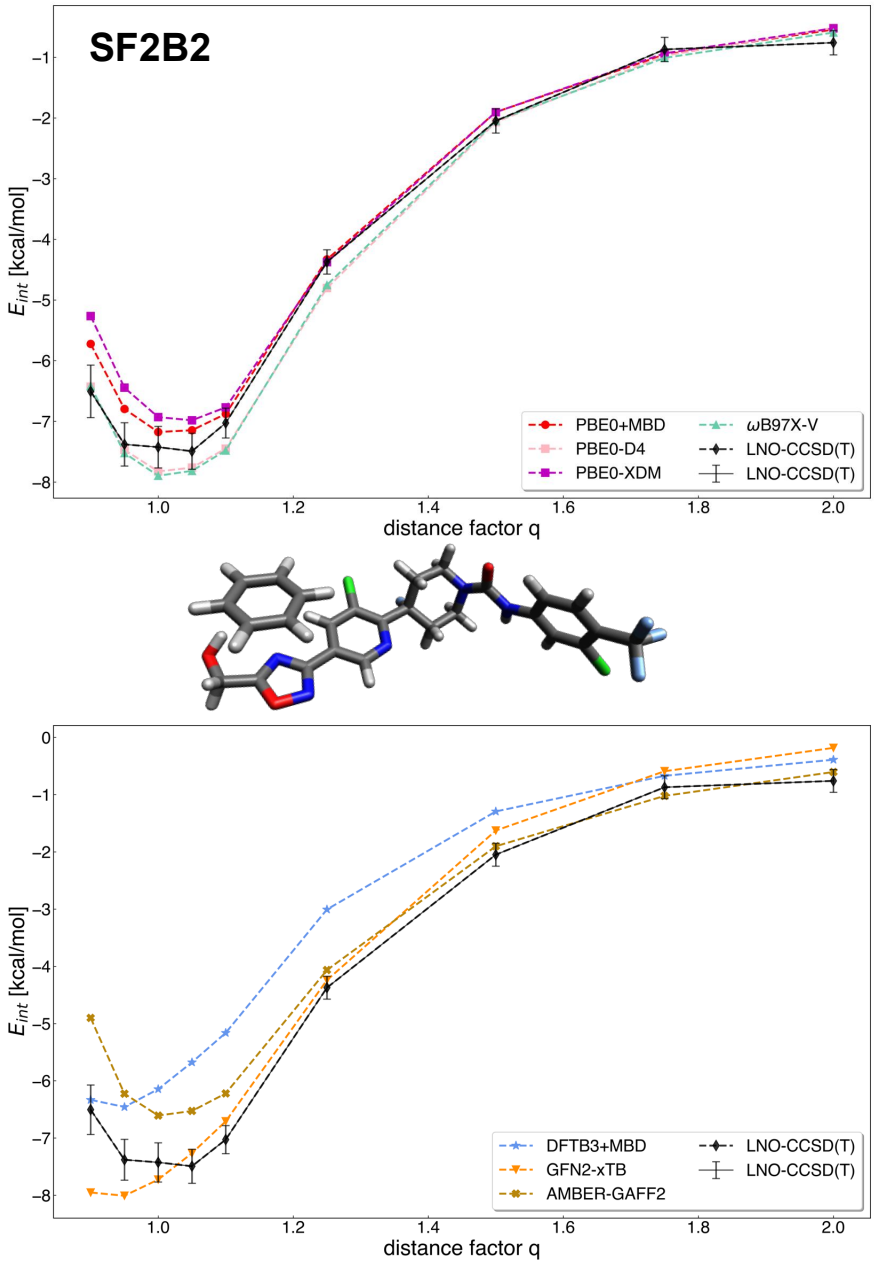

Figure 3.14: Figure from [57]. Dissociation curves for the the non-equilibrium dimer SF2B2, for which LNO-CCSD(T) level reference is shown. The DFT methods PBE0+MBD, PBE0+D4, PBE0+XDM, ωB97X-V in the top plot, and semi-empirical methods GFN2-xTB, and DFTB3+MBD, as well as the classical force field AMBER-GAFF2 in the bottom plot, are presented.

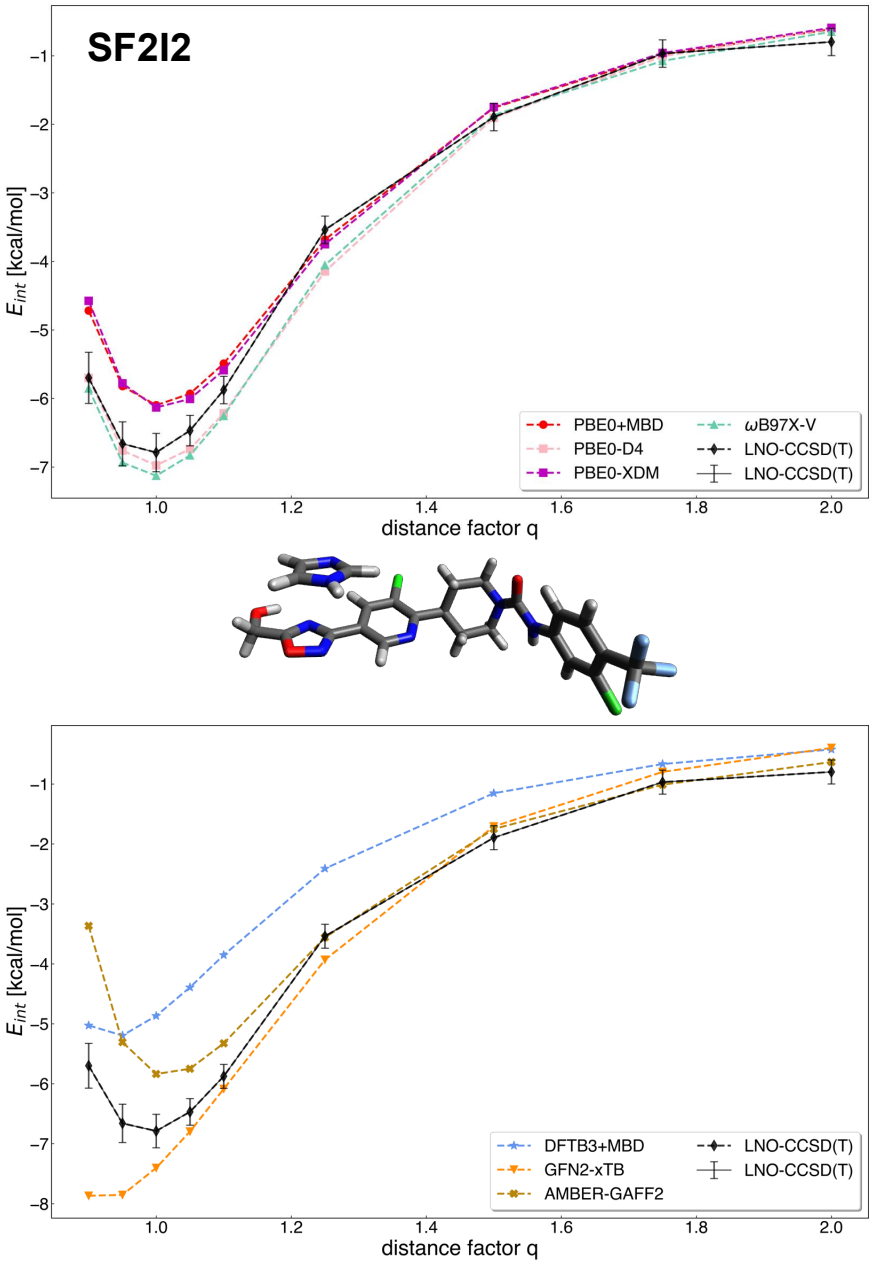

Figure 3.15: Figure from [57]. Dissociation curves for the the non-equilibrium dimer SF2I2, for which LNO-CCSD(T) level reference is shown. The DFT methods PBE0+MBD, PBE0+D4, PBE0+XDM, ωB97X-V in the top plot, and semi-empirical methods GFN2-xTB, and DFTB3+MBD, as well as the classical force field AMBER-GAFF2 in the bottom plot, are presented.

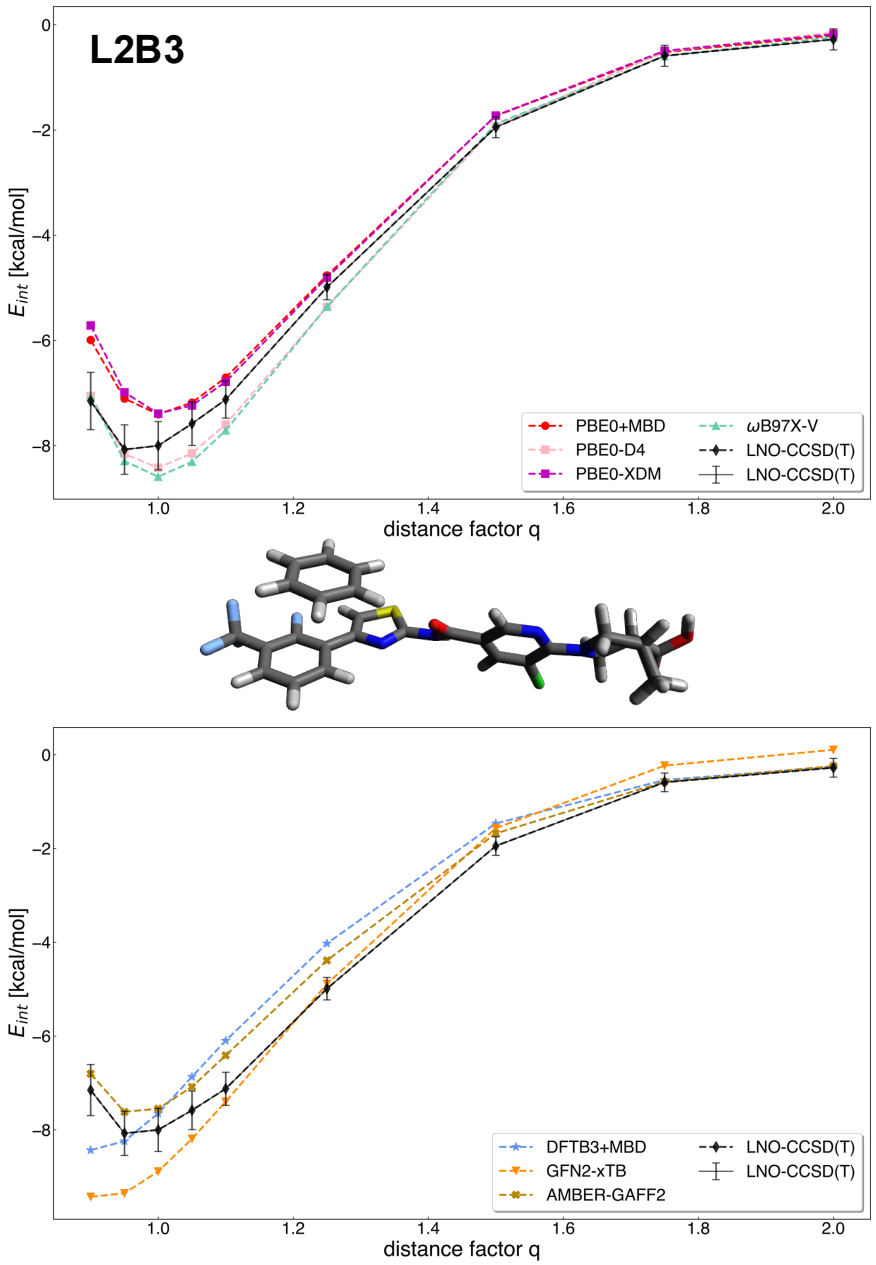

Figure 3.16: Figure from [57]. Dissociation curves for the the non-equilibrium dimer L2B3, for which LNO-CCSD(T) level reference is shown. The DFT methods PBE0+MBD, PBE0+D4, PBE0+XDM, ωB97X-V in the top plot, and semi-empirical methods GFN2-xTB, and DFTB3+MBD, as well as the classical force field AMBER-GAFF2 in the bottom plot, are presented.

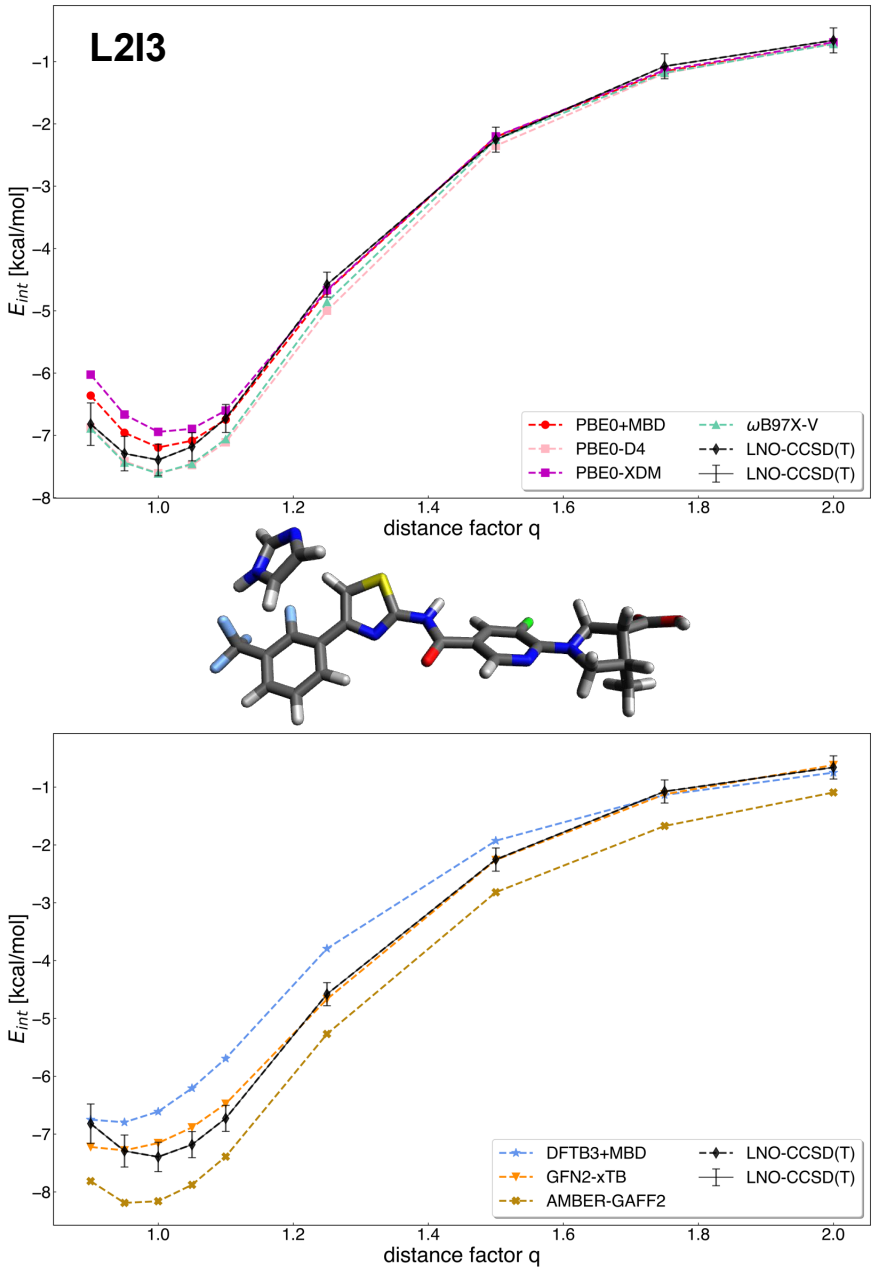

Figure 3.17: Figure from [57]. Dissociation curves for the the non-equilibrium dimer L2I3, for which LNO-CCSD(T) level reference is shown. The DFT methods PBE0+MBD, PBE0+D4, PBE0+XDM, ωB97X-V in the top plot, and semi-empirical methods GFN2-xTB, and DFTB3+MBD, as well as the classical force field AMBER-GAFF2 in the bottom plot, are presented.

3.6 QUID dimers properties

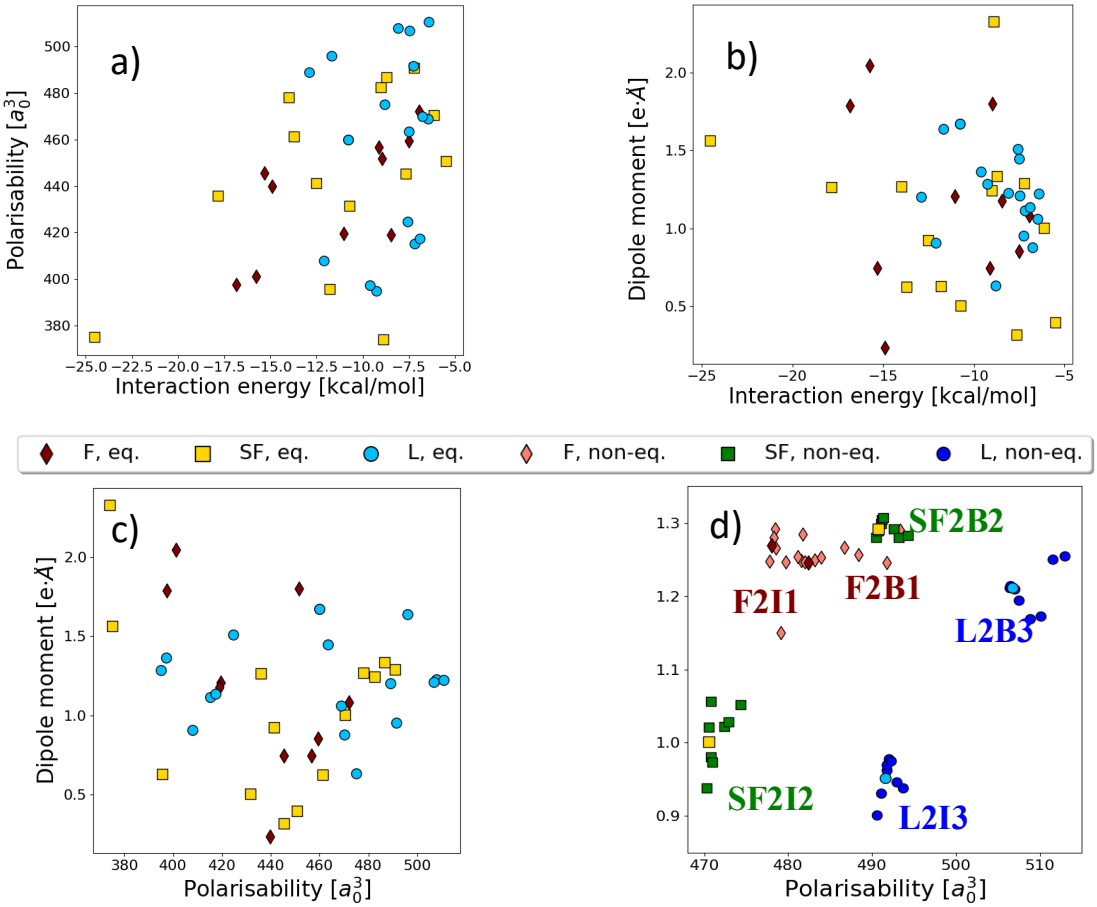

Figure 3.18: Figure from [57]. a) and b): Molecular properties correlation plot for all 42 QUID equilibrium dimers colour-coded by geometry category for a)total dipole moment versus interaction energy and b) molecular polarisability versus interaction energy. c) and d): Molecular properties correlation plot of the total dipole moment versus molecular polarisability for a) all equilibrium QUID dimers, folded configurations in red, semi-folded in yellow, and linear in blue; and b) for the equilibrium and non-equilibrium configurations of six configurations F2B1 and F2I1 in pink, SF2B1 and SF2I1 in green, L2B3 and L2I3 in dark blue.

Quantum-mechanical property space of QUID systems. To enhance the understanding of the effects of dimer configuration and intermolecular distances on the properties of pocket-ligand systems, several global and local physicochemical properties, in addition to E_{int}, were computed for all QUID dimers with PBE0+MBD (details in Subsection 3.4). A full list of properties, similar to those in the Aquamarine [56] dataset of large monomers, is provided in Table B.1. Further, quantities describing the Hirshfeld partitioning, *i.e.*, Hirsh-

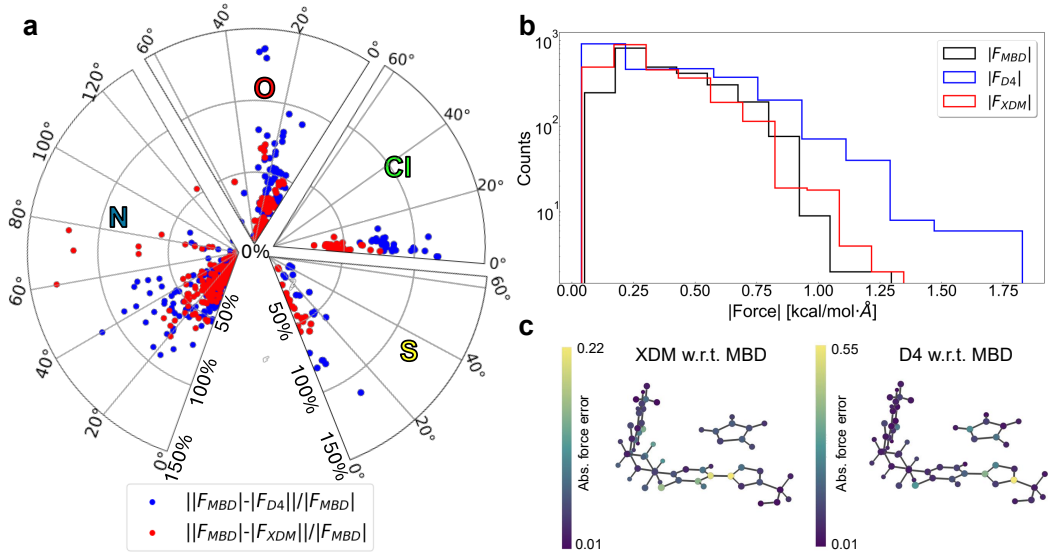

Figure 3.19: Figure from [57]. a) A polar plot, where the radius is the difference in magnitude of the vdW forces (given as a percentage) and the angle is the arccosine between the dot product of the force vectors. Four such polar plots depict the distributions of the forces acting on all the N, O, Cl, and S atoms, respectively, in the equilibrium QUID dimers. b) Overall distribution of the magnitudes of the atomic vdW forces on all atoms for equilibrium QUID dimers using the MBD, D4, and XDM methods. c) An example of the deviation in atomic van der Waals (vdW) forces between different methods is visualised for the SF2I2 dimer (produced by the FFAST software [253]). The range is represented by colours varying along a viridis colour bar from purple to yellow.

feld volumes, ratios, charges, (scalar) dipole moments provides information for the electron response of an atom-in-molecule environment. Overall the molecular properties can also be useful as ML descriptors [254]. Here, the initial focus is on the molecular polarisability, α, as an additional measure of the NCIs (theoretical background details in Section 2.3.2).

To that end, analogous to E_{int}, the difference in α between each dimer and the sum of its corresponding large and small monomers, $\Delta\alpha$, was calculated for all 128 non-equilibrium conformations. The average $\Delta\alpha$ values for each structure type as a function of the distance factor q are plotted in Fig. 3.11c). Overall, these plots for the three structure types exhibit an almost linear behaviour, with slight deviations near the equilibrium distance for the Linear and Semi-Folded structures. According to a recent concept of chemical bonding based on α, proposed by D. Hait and M. Head-Gordon [255], this linear behaviour suggests no significant modification in the covalent bond arrangements along the dissociation curve. The variation can thus be attributed to the self-consistent screening effect between the monomers. Indeed, in both Linear and Semi-Folded structures, the small monomer affects fewer atoms of the large monomer. In contrast, in Folded structures, the small monomer remains within 5 Å of a significant number of atoms of the large monomer, substantially influencing the electrostatics and dispersion in the pocket site, resulting in a steeper curve.

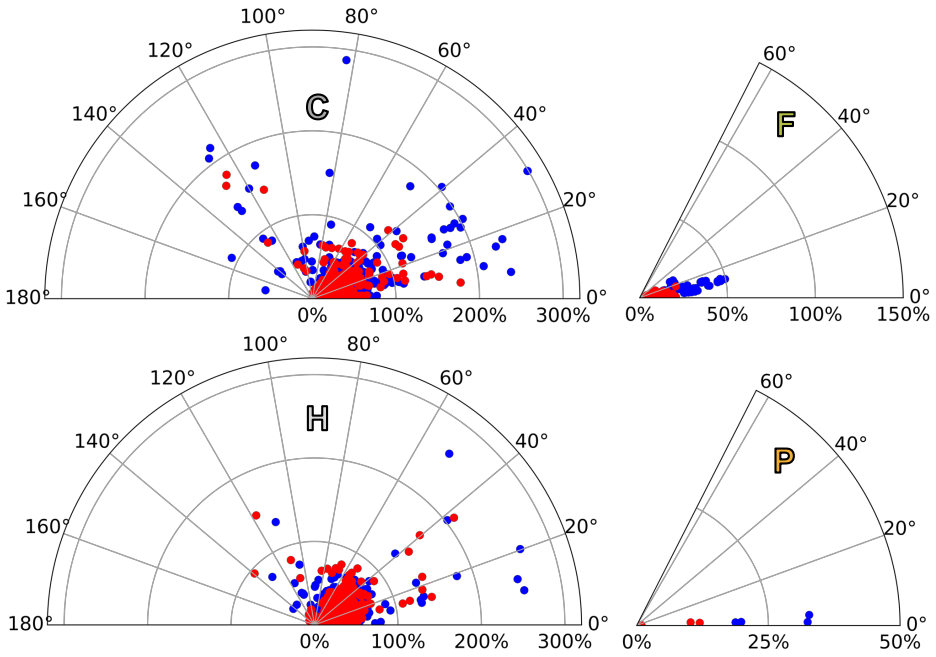

Figure 3.20: Figure from [57]. Atomic forces radial plot of the differences between the forces of PBE0+D4(blue) and PBE0+XDM (red) w.r.t PBE0+MBD plotted as an arccos of the forces vectors versus a radius of the differences of the magnitudes scaled by the PBE0+MBD magnitude and provided in percentages. The plots are for the atomic forces for H, C, F, and P atoms in all equilibrium molecules.

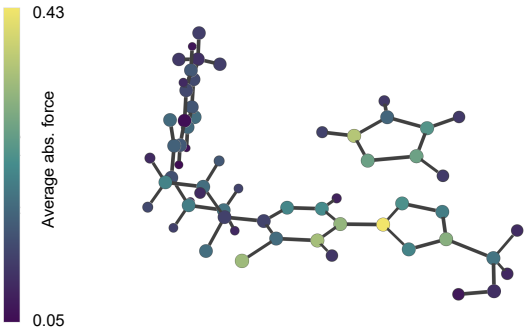

Figure 3.21: Figure from [57]. An image of the force vectors corresponding to the MBD force for the SF2I2 dimer, produced with FFAST software.

Moreover, no correlation between α with E_{int} and μ emerges from the exploration QUID dimers (see Fig. 3.18). Hence, the interplay between electrostatic and dispersion interactions in a structurally and chemically complex local environment in the QUID pocket-ligand proxies requires exact QM models beyond the capture of a single key global property in the high-dimensional chemical compound space.

Another relevant property for understanding NCIs in QUID dimers is the atomic forces,

which are widely used to parameterise MLFFs for (bio)molecular systems. Indeed, the molecular conformational sampling at a given temperature strongly depends on the accuracy of the chosen computational method in adequately describing the forces acting on the atoms in the molecule. In MLFF methods molecular data is routinely optimised at one level of theory, method or functional or dispersion correction, and computed at a different one to serve as input or reference for energies and forces. Unfortunately, while for E_{int} highly accurate LNO-CCSD(T) reference data was calculated and even compared against another highly accurate method, FN-DMC, obtaining forces at benchmark *ab initio* level is prohibitively expensive. Current research in developing gradients for LNO-CCSD(T) [256] and FN-DMC [257] could facilitate a reference benchmark for the vdW forces in the future and provide a clearer picture for the accuracy of the DFT methods for MD of ligands binding to a pocket. Accordingly, the atomic force distributions have been analysed using a selection of DFT methods: the best performing ones PBE0+MBD and PBE0+D4, and the well-performing pairwise PBE0+XDM method. Since these methods share the DFT functional and the primary interest for QUID lies in NCIs, the focus will be on the vdW components of the atomic forces. The MBD method is taken as the baseline for comparison because geometry optimisations have been carried out with the PBE0+MBD level of theory. In addition, MBD is the only method that includes non-local screening effects in the polarisability, which were shown to be important in Fig. 3.11c).

The differences in the vdW atomic forces are examined in terms of their magnitude and direction, treated as two distinct yet interrelated driving factors in MD simulations. The results are presented in Fig. 3.19a as a radial plot illustrating the difference in forces of D4 and XDM with respect to MBD. In this plot, the angular part represents the arccosine between the dot product of two force vectors, while the radius is the difference in force magnitudes, scaled by the MBD force. This analysis focuses on the differences observed per atom type, with the results for N, O, Cl, and S atoms shown in Fig. 3.19a (see the other atom types in Fig. 3.20).

The overall force analysis is provided by the plot in Fig. 3.19b, which shows that MBD yields the smallest forces on average. The vdW force magnitudes are generally not much higher than 1 kcal/mol/Å; this is expected given that equilibrium geometries are involved. Higher vdW forces would be expected for non-equilibrium geometries. Further, largest discrepancies in force magnitude were found in D4 compared to XDM. This is corroborated by the distribution of the atomic force magnitudes for all atoms. The most significant outliers in force magnitude discrepancies are associated with D4 and Fig. 3.19a) demonstrates that those outliers of up to 3-4 times the magnitude of the MBD force are found on C atoms in descending order for L2I3, L2B3, L3I1, and L3I2 (Fig. 3.20). The `FFAST` software [253] allows for visualisation of the discrepancies between the vdW atomic forces to confirm that the C atoms were found at the binding site, an example of such a visualisation can be seen in Fig. 3.19c) depicting SF2I2. The difference in force magnitudes is visible not only on the atoms of the binding site as expected but also in proximity to it showing the differing impact of the ligand interactions in the different methods. For the SF2I2, there is a notable difference between the comparisons of XDM and D4 w.r.t. MBD (the MBD depiction is

available in Fig. 3.21).

This also holds true in general for the vdW forces on the Cl atoms in the QUID dimers. As seen on Fig. 3.19a) XDM is in better agreement with MBD than D4 for Cl atoms, and the same is true for the other halogen element F, as well as the few P atoms (Fig. 3.20). Hence, the presence of more electronegative atoms can hint at a systematic difference between the different vdW atomic forces. In that vein, a particular outlier is seen for the S element, with a 113% gap between the magnitude of the MBD and D4 forces for the SF1I1 dimer, in the sulfonyl group of the binding site of the imidazole ligand (Fig. 3.3). Interestingly, there is a split in the vdW force directions between 'heavier' atoms in the QUID dimers, *i.e.*, O, F, Cl, S, and P and the 'lighter' ones, *i.e.*, H, C, and N. The 'lighter' atoms represented more in organic molecules demonstrate a larger spread of angle difference between the forces up to 120°-180°. By construction, for the pairwise D4 and XDM methods, the vdW force is a simple vector sum where all force vectors are aligned along the vector connecting pairs of atoms. In contrast, many-body effects that are treated to infinite order in MBD can thus substantially alter the force directions, and this difference is much more pronounced than for force magnitudes. This could have a visible effect on MD trajectories, although these implications remain to be assessed in the future with the observed difference in force directions having potential repercussions for MLFFs. The MLFF analysis of MD performance with particular attention to atomic force prediction will be discussed as particularly relevant for the results in the next Chapter 4.

3.7 Conclusions

The QUID benchmarking framework was developed as a state of the art in QM-based modelling of motifs representative for ligand-protein pocket interactions. Currently, QUID contains 170 structurally and chemically diverse large molecular dimers (42 equilibrium and 128 non-equilibrium) of up to 64 atoms, including the H, N, C, O, F, P, S, and Cl chemical elements, encompassing most atom types of interest for drug discovery purposes. This diversity enables a single dimer to exhibit multiple types of steric effects and non-covalent interactions simultaneously, including, but not limited to, $\pi-\pi$ stacking, hydrogen, and halogen bonds. Accordingly, the SAPT energy decomposition of the interaction energy revealed that ligand-pocket interactions are predominantly governed by dispersion and electrostatics—types of interactions often inadequately represented by molecular mechanics (MM) methods. To create a truly robust basis for the improvements of MM, ML, and quantum chemistry approaches, the interaction energy accuracy of the ligand-pocket interactions was obtained at the "gold standard" level of Coupled Cluster and Quantum Monte Carlo methods (specifically LNO-CCSD(T) and FN-DMC), contrasting the results. Notably, the previously reported disagreement between LNO-CCSD(T) and FN-DMC for large non-covalent systems [234] was not observed at such scale for the QUID dimers. Here only a small discrepancy of approximately 0.5 kcal/mol is observed, driven predominantly by electrostatic interactions in accordance with a recent study on the S66 dataset [235]. Among all stud-

ied MM and QM approaches, DFT methods such as PBE0+MBD, ωB97X-V, and PBE0+D4 achieve excellent agreement with the computationally expensive LNO-CCSD(T) reference for both equilibrium and non-equilibrium interaction energies. Additionally, certain limitations in the widely used semiempirical (*e.g.*, GFN2-xTB and DFTB3+MBD) and MM methods (*e.g.*, AMBER-GAFF2 and CHARMM-CGenFF) were identified for the investigated complex non-covalent motifs, which raises questions about their reliability in binding affinity simulations. These intriguing results highlight the relevance of determining the appropriate level of theory to accurately characterise protein-ligand systems, particularly in the development of extensive QM datasets utilised in physical method benchmarking and ML-based investigations.

Furthermore, QUID provides access to a diverse set of extensive and intensive (global and local) properties obtained at high level DFT – going beyond the interaction energy enables the electronic characterisation of chemical environments within ligand-pocket motifs—a limitation of current benchmark frameworks, which primarily focus on structural and energetic features. The additional electronic properties for training ML on structure-property relations, as well as data that can be used for refining the candidate selection when targeting specific pocket sites with molecules with a desired set of QM properties. A further relevant physicochemical property for molecular systems, especially in the context of robust data used for the creation of FFs, is the atomic forces. The observed discrepancies in the vdW component of the atomic forces using MBD, D4, and XDM methods also show the importance of further investigations of existing methods in non-covalent complexes. An inaccurate description of vdW interactions can strongly impact the reaction pathway and the resulting binding position, when simulating the interaction of ligands with protein pockets.

While the insights gained from this work highlight the importance of an appropriate QM description for inter- and intramolecular properties of ligand-pocket motifs, further efforts are needed and should incorporate more flexible and charged pocket structures, as well as solvation effects [56, 258]. Thus, this project can pave the way for a more informed use and refinement of physical and chemical models for simulating ligand-pocket interactions, offering particular value in fine-tuning MLFFs and ML-augmented semi-empirical models, which are increasingly integrated into screening pipelines for drug discovery.

Chapter 4

Crash Testing Machine Learning Force Fields for Organic Molecules in TEA Challenge 2023

Parts of this chapter have been published in this or similar form in:

I. Poltavsky, A. Charkin-Gorbulin, **M. Puleva**, G. Fonseca, I. Batatia, N. J. Browning, S. Chmiela, M. Cui, J. T. Frank, S. Heinen, B. Huang, S. Käser, A. Kabylda, D. Khan, C. Müller, A. J. A. Price, K. Riedmiller, K. Töpfer, T. W. Ko, M. Meuwly, M. Rupp, G. Csányi, O. A. von Lilienfeld, J. T. Margraf,u K.-R. Müller, and A. Tkatchenko "Crash testing machine learning force fields for molecules, materials, and interfaces: model analysis in the TEA Challenge 2023" *Chemical Science* **16**, 3720-3737, 2025

and

I. Poltavsky, **M. Puleva**, A. Charkin-Gorbulin, G. Fonseca, I. Batatia, N. J. Browning, S. Chmiela, M. Cui, J. T. Frank, S. Heinen, B. Huang, S. Käser, A. Kabylda, D. Khan, C. Müller, A. J. A. Price, K. Riedmiller, K. Töpfer, T. W. Ko, M. Meuwly, M. Rupp, G. Csányi, O. A. von Lilienfeld, J. T. Margraf,u K.-R. Müller, and A. Tkatchenko "Crash testing machine learning force fields for molecules, materials, and interfaces: molecular dynamics in the TEA challenge 2023" *Chemical Science* **16**, 3738-3754, 2025

and have been produced in collaboration with the above authors. The data can be found in the Zenodo archive 10.5281/zenodo.13832724. My contribution involved data curation, formal analysis, investigation, project administration, software, validation, and visualisation - predominantly for Challenges I and II.

While the previous Chapter 3 investigated the performance of a plethora of physical methods at different levels of QM approximation when applied to toy models of a protein pocket-

ligand systems, the current chapter examines the capabilities of ML methods to model organic systems.

The “Crash Testing Machine Learning Force Fields for Molecules, Materials, and Interfaces" (TEA) Challenge 2023 aimed to evaluate a snapshot of the field of ML models in 2023-24, and while the work of the QUID dimers focused on the non-covalent interactions predictions of configurational snapshots in terms of energetics and touching upon atomic forces, here the ML models capabilities are benchmarked directly on the dynamic behaviour, *i.e.*, the ability produce stable and accurate simulations of molecular dynamics in different chemical systems.

The TEA Challenge 2023 captured a representative sample of the field of ML Force Fields (MLFFs) by involving a diverse set of models – from lighter kernel models like sGDML, FCHL19*, and SOAP/GAP, the lightest of which with ∼120k trainable parameters to equivariant NNs, MACE and SO3krates, whose heaviest model involved ∼3M parameters. The project was carried out in collaboration with their developers, who provided the trained models on the data sent to them for the challenges. Not only were different types of ML methods tested, but also their ability to reproduce potential energy surfaces (PES), to extrapolate within different training data regimes from incomplete data, and their efficiency was considered. The results and their in-depth analysis were presented in two papers focusing on the analysis of the models' performances [61] , and the molecular dynamics simulations produced with them [62]. The examined ML architectures included in The TEA Challenge 2023 involved four specific challenges, each one on a different molecular system - Challenges I and II as the explorations of tests on organic molecules, and Challenges III and IV on periodic systems, with my work focused, beyond the general project, on Challenges I and II. Thus, the results for the biochemically relevant alanine tetrapeptide system of Challenge I and the N-acetylphenylalanyl-pentaalanyl-lysin system of Challenge II will be presented in the following chapter.

4.1 Molecular Dynamics Datasets

The organic molecular systems for Challenges I and II are presented in Fig. 4.1. The alanine tetrapeptide dataset (Ac-Ala3-NHMe, 42 atoms) in Challenge I is a part of the MD22 benchmark dataset [259]. It was generated as a single MD trajectory under constant NVT conditions sampled at a temperature of 500 K with a time step of 1 fs. The dataset contains a total of 85,109 structures, whose corresponding potential energy and atomic forces were calculated at PBE [88]+MBD [59, 60] level. Meanwhile, the N-acetylphenylalanyl-pentaalanyl-lysine dataset (Ac-Phe-Ala5-Lys, 100 atoms) in Challenge II was specifically generated for the TEA Challenge 2023. It contains a limited sampling near the 200 lowest energy conformers identified using the CREST software package [260]. The sampling was also performed by running NVT MD simulations at a temperature of 500 K with a time step of 0.5 fs starting from each conformer's equilibrium structure. All the trajectories are 250 fs long and contain 500 configurations, resulting in a total of 100,000 reference structures,

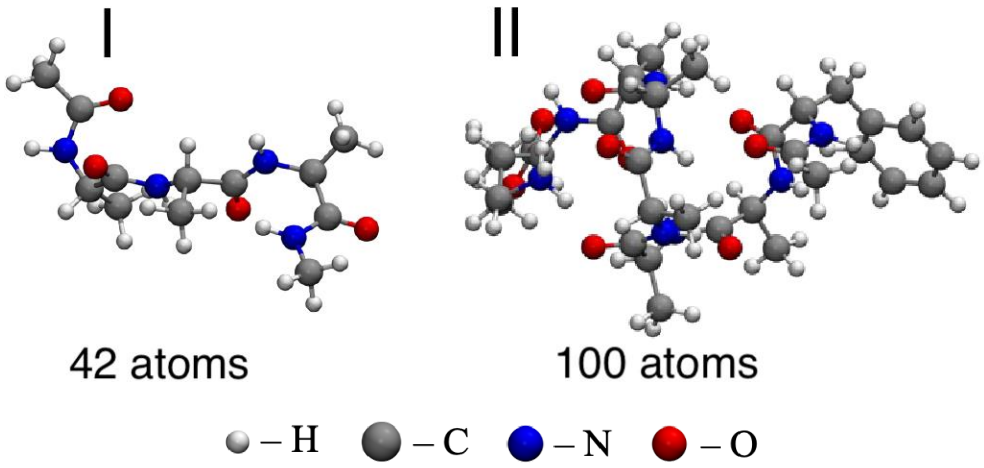

Figure 4.1: Figure excerpt from [61]. Snapshots of system geometries for the molecules in the datasets for Challenge I Ac-Ala3-NHMe and Challenge II Ac-Phe-Ala5-Lys.

whose energies and forces were computed using the PBE0 [58]+MBD-NL [114]. All datasets have been generated with the `FHI-Aims` software [104, 231, 261]. The Ac-Phe-Ala5-Lys is protonated in all calculations. Notably, only the alanine tetrapeptide dataset was published before the TEA Challenge 2023, while the N-acetylphenylalanyl-pentaalanyl-lysine dataset was generated by collaborators to the TEA Challenge 2023. Therefore, the MLFF developers had limited information about three out of four datasets.

4.2 Models analysis

The analysis of the MLFFs is carried out by evaluating their performance on test datasets and examining the outcomes of the MD simulations. Specifically, the TEA Challenge 2023 results are presented not only in terms of MAEs and RMSEs but also as maximum errors (denoted MAX) in energy and force predictions. The MAX error for forces is computed as follows

$$\Delta F_{max} = \max_{n} \left| \vec{F}_n^{true} - \vec{F}_n^{ml} \right| , \tag{4.1}$$

where $\vec{F}_n^{true}$ is the reference force acting on an atom n, $\vec{F}_n^{ml}$ is the corresponding force predicted by an ML model, $|\ldots|$ is the norm of a vector, and $\max$ is the maximum over atomic forces for all system configurations.

To gain more detailed insight into MLFFs performance, a comprehensive analysis is carried out for atomistic force predictions beyond the aggregate measures using colour coding of the MAE for each individual atom in the system with the FFAST software [253]. Finally, to recognise practical considerations about the use of the MLFF models, the stability and computational speed of the MD simulations are also assessed. Detailed analyses of the MD results are presented in the Sections 4.3 and 4.4 .

4.2.1 Aggregated Accuracy

The first step in assessing the performance of the models in the TEA 2023 challenges is by considering the measures of aggregated accuracy as presented in Table 4.1. A comprehensive set of energy and forces MAE, RMSE, and MAX errors across all four challenges can be found there, with the MAE values below 1 kcal/mol or 1 kcal/(mol·Å) highlighted in bold. To simplify the comparative analyses of the MLFFs performance, the MAE and MAX energy and force errors are also plotted in Fig. 4.2, with the models' naming schemes detailed in the List of abbreviations. Tables 4.2 and 4.3 contain relative prediction errors normalized by the mean absolute values and by the standard deviations of the energies and forces in the reference datasets, respectively.

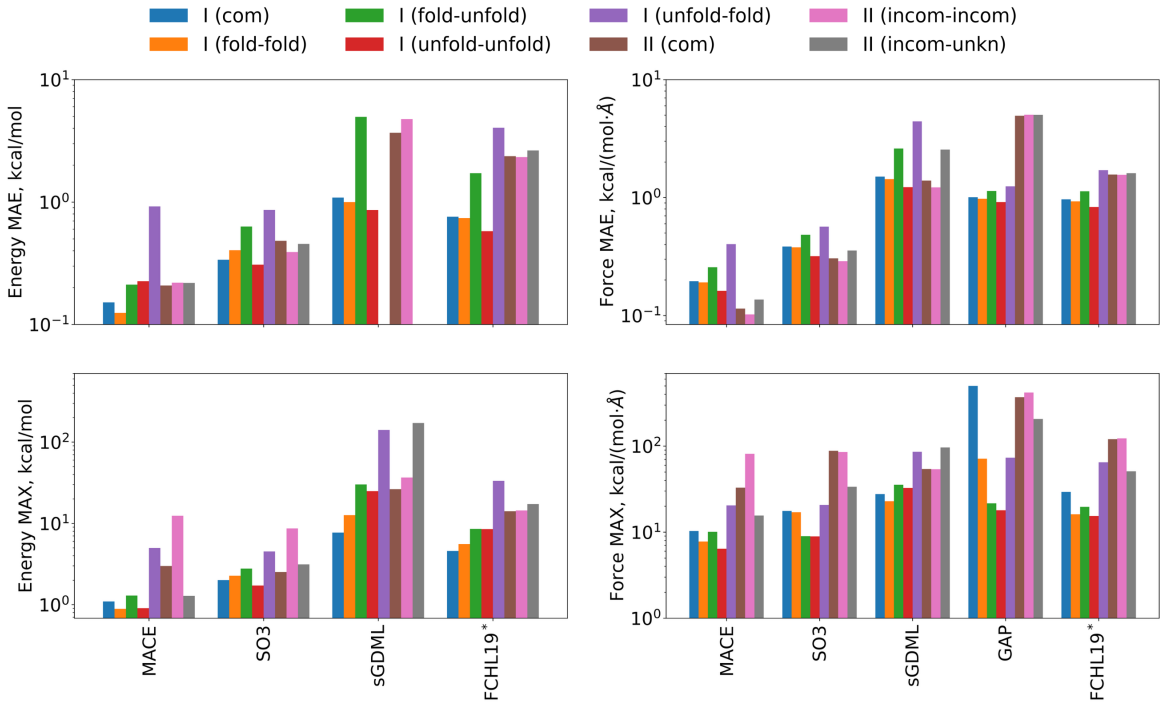

Figure 4.2: Figure excerpt from [61]. Graphical representation of MAE and MAX errors in energy and forces for the Challenges I and II as listed in Table 4.1.

Considering Table 4.1, a significant reduction in MAE and RMSE values for energy and forces can be observed throughout the tested systems for the equivariant NN architectures in MACE and SO3krates models compared to the kernel-based models sGDML, SOAP/GAP, and FCHL19*. While the traditional benchmark for MLFF, 'chemical accuracy' 1 kcal/mol MAE for energy and 1 kcal/(mol·Å) MAE for forces, the equivariant architectures can surpass the latter and reduce typical MAEs by a factor of 2 to 5, depending on the number of parameters within the NN. This reduction represents a substantial leap in predictive accuracy, establishing a new standards for force predictions in the range of 0.2 – 0.5 kcal/(mol·Å). Conversely, the standards for energy predictions have not seen a similar reduction. The

increased size and complexity of the simulated systems complicate the reconstruction of global quantities, such as energy. While the MAE and RMSE for forces and energies are fairly similar for both molecules, the energy errors for the solids in the other Challenges increases, however this is beyond the scope of this thesis.

The second finding concerns the MAX errors. Despite the advancements in MLFF architectures, MAX errors remain substantial even for the investigated state-of-the-art MLFF models. When trained on comprehensive (complete) datasets, the MACE and SO3krates models exhibit significant prediction discrepancies in the forces, with maximum MAX errors of 33 and 88 kcal/(mol·Å) for Challenge II, respectively. Moreover, the maximum MAX force error for the sGDML model is comparable at 54 kcal/(mol·Å). Still, a significant improvement is observed compared to two other kernel-based models, SOAP/GAP and FCHL19*, which display considerably higher maximum MAX force errors of 370 and 120 kcal/(mol·Å), respectively. On the other hand, sGDML generally displays smaller MAX energy errors than the other kernel models, indicating that the balance between energies and forces in the respective loss functions also plays a role here. The issue of large MAX errors can be further exacerbated when the MLFFs are trained on partial datasets, *i.e.*, missing data in the coverage of the relevant chemical compound space. For the MACE model, the MAX error increases from 33 to 81 kcal/(mol·Å), while for the sGDML model, it escalates from 54 to 140 kcal/(mol·Å). Overall for the organic systems of Challenges I and II the equivariant NN architectures showed larger improvements compared to their kernel counterparts.

4.2.2 In-depth Accuracy Analysis

The previous analyses, *vide supra*, follow the typical approach based on aggregated accuracy metrics. While the well-established method is reliable for small and chemically relatively simple systems, it becomes insufficient as system size and composition diversity increase. To address this gap, the FFAST software [253] is used to provide detailed performance quantifiers on the atomistic level.

Fig. 4.6 shows the MAEs of atomic forces for the Ac-Phe-Ala5-Lys molecule on the test set for all five MLFFs participating in Challenge II. In Figs. 4.6b through 4.6f, atomic colours encode the force MAE for each individual atom, with corresponding colour bars indicating numeric values. A similar Fig. 4.3 can be found for the Ac-Ala3-NHMe molecule. Notably, individual colour scales are used for each model and the absolute scale of the errors shown varies considerably between the models. In both figures 4.6 and 4.3, a significant heterogeneity in the atomic force prediction accuracy across the molecule can be seen. The MAE ratio between the worst and best-predicted atoms ranges from 6 for the NNs to 4-5 for the kernel-based methods. In particular, even among atoms of the same type, such as Carbon, the best-predicted atoms can have MAEs that are 3 times smaller than the worst-predicted carbons. This force predictions heterogeneity indicates that there is still room for improvement for MLFF architectures to achieve consistent accuracy across individual atoms. Interestingly, in Fig. 4.6 the pattern of force prediction heterogeneity is very similar across all five MLFFs. The colour bar values non-withstanding, one could notice minimal differences

Task		MACE			SO3			sGDML			SOAP/GAP			FCHL19*		
I (com)	E	**0.15**	0.19	1	**0.34**	0.43	2	1.09	1.42	8	1.16	1.50	9	**0.76**	0.98	5
	F	**0.20**	0.30	10	**0.38**	0.57	18	1.50	2.14	28	1.01	1.67	500	**0.97**	1.36	29
	F(H)	**0.11**	0.17	9	**0.23**	0.33	12	**0.99**	1.32	22	**0.65**	0.87	23	**0.65**	0.88	29
	F(C)	**0.29**	0.40	9	**0.56**	0.76	18	2.13	2.81	26	1.41	2.21	500	1.36	1.80	22
	F(N)	**0.32**	0.44	10	**0.62**	0.83	10	2.57	3.39	28	1.46	2.77	490	1.42	1.85	17
	F(O)	**0.23**	0.32	9	**0.45**	0.60	14	1.39	1.87	22	1.31	1.70	19	1.07	1.41	20
I (f-f)	E	**0.12**	0.16	0.9	**0.40**	0.50	2	1.00	1.33	13	1.11	1.44	7	**0.74**	0.95	6
	F	**0.19**	0.29	8	**0.38**	0.55	17	1.43	2.05	23	**0.98**	1.39	71	**0.93**	1.30	16
	F(H)	**0.11**	0.17	7	**0.24**	0.34	17	**0.95**	1.27	16	**0.64**	0.86	20	**0.63**	0.85	16
	F(C)	**0.28**	0.39	6	**0.54**	0.73	8	2.04	2.72	23	1.37	1.84	71	1.31	1.72	15
	F(N)	**0.31**	0.43	7	**0.61**	0.81	11	2.39	3.18	23	1.38	1.83	71	1.37	1.78	16
	F(O)	**0.23**	0.32	8	**0.43**	0.58	11	1.29	1.76	20	1.26	1.65	25	1.01	1.33	11
I (f-u)	E	**0.21**	0.28	1	**0.63**	0.75	3	4.96	7.01	30	1.44	1.80	7	1.72	2.24	9
	F	**0.26**	0.40	10	**0.48**	0.73	9	2.60	3.79	36	1.14	1.63	22	1.13	1.63	20
	F(H)	**0.14**	0.20	10	**0.28**	0.41	9	1.65	2.26	30	**0.71**	0.98	22	**0.73**	1.02	16
	F(C)	**0.39**	0.55	7	**0.70**	0.95	7	3.78	5.10	32	1.59	2.13	16	1.63	2.19	20
	F(N)	**0.42**	0.57	6	**0.8**2	1.11	8	4.36	5.71	36	1.69	2.22	21	1.72	2.27	16
	F(O)	**0.34**	0.46	5	**0.61**	0.83	7	2.55	3.49	25	1.54	2.01	20	1.28	1.70	15
I (u-u)	E	**0.23**	0.26	0.9	**0.31**	0.39	2	**0.86**	1.43	25	**0.90**	1.17	4	**0.58**	0.81	9
	F	**0.16**	0.26	6	**0.32**	0.47	9	1.23	1.80	33	**0.92**	1.28	18	**0.83**	1.16	15
	F(H)	**0.10**	0.15	6	**0.20**	0.28	6	**0.83**	1.15	33	**0.61**	0.81	14	**0.58**	0.80	13
	F(C)	**0.23**	0.34	5	**0.45**	0.62	9	1.71	2.35	21	1.29	1.70	13	1.14	1.51	10
	F(N)	**0.26**	0.38	6	**0.51**	0.69	7	2.05	2.80	23	1.28	1.67	18	1.22	1.60	15
	F(O)	**0.19**	0.27	5	**0.36**	0.49	6	1.11	1.55	15	1.14	1.48	10	**0.90**	1.18	9
I (u-f)	E	**0.92**	1.19	5	**0.86**	1.04	5	23.4	31.2	140	1.94	2.48	10	4.04	5.67	33
	F	**0.40**	0.74	21	**0.57**	0.87	21	4.43	6.66	86	1.25	1.85	73	1.71	2.62	65
	F(H)	**0.20**	0.35	20	**0.33**	0.52	21	2.95	4.62	86	**0.80**	1.22	73	1.18	1.93	65
	F(C)	**0.59**	0.99	14	**0.81**	1.13	18	6.00	8.22	75	1.65	2.25	48	2.11	2.93	36
	F(N)	**0.75**	1.07	21	**0.96**	1.28	12	7.34	9.74	58	1.96	2.63	50	3.03	4.17	30
	F(O)	**0.63**	0.99	12	**0.74**	1.02	19	4.96	7.12	53	1.75	2.40	25	2.04	2.91	49
II (com)	E	**0.21**	0.27	3	**0.48**	0.60	3	3.68	5.05	26	7.79	9.71	41	2.37	3.00	14
	F	**0.11**	0.17	33	**0.30**	0.46	88	1.39	2.06	54	4.93	6.80	370	1.57	2.18	120
	F(H)	**0.07**	0.10	33	**0.19**	0.28	57	**0.92**	1.30	54	3.41	4.63	230	1.05	1.41	120
	F(C)	**0.16**	0.22	10	**0.42**	0.59	16	1.90	2.62	36	7.04	9.10	110	2.22	2.90	81
	F(N)	**0.18**	0.25	24	**0.49**	0.67	88	2.40	3.28	47	6.25	8.03	230	2.14	2.75	59
	F(O)	**0.14**	0.20	22	**0.36**	0.50	43	1.34	1.89	37	4.91	6.43	370	1.69	2.22	35
II (i-i)	E	**0.22**	0.31	12	**0.39**	0.50	9	4.76	6.79	37	3.17	3.94	22	2.33	2.94	14
	F	**0.10**	0.16	81	**0.29**	0.44	85	1.22	1.85	54	5.04	6.93	420	1.56	2.17	120
	F(H)	**0.07**	0.12	81	**0.19**	0.29	85	**0.81**	1.17	54	3.47	4.70	340	1.04	1.41	120
	F(C)	**0.14**	0.20	17	**0.40**	0.56	36	1.67	2.35	31	7.29	9.35	100	2.20	2.87	84
	F(N)	**0.16**	0.24	34	**0.44**	0.63	52	2.10	2.93	32	6.35	8.23	420	2.13	2.75	59
	F(O)	**0.12**	0.18	35	**0.34**	0.49	79	1.18	1.71	49	4.75	6.14	93	1.69	2.23	110
II (i-un)	E	**0.22**	0.28	1	**0.45**	0.58	3	30.2	35.4	170	2.81	3.47	14	2.64	3.37	17
	F	**0.14**	0.21	16	**0.36**	0.55	34	2.55	3.59	96	5.03	6.92	210	1.61	2.25	51
	F(H)	**0.08**	0.12	10	**0.21**	0.32	34	1.88	2.74	96	3.46	4.68	160	1.07	1.46	40
	F(C)	**0.19**	0.27	7	**0.51**	0.71	17	3.21	4.21	41	7.28	9.35	110	2.26	2.97	42
	F(N)	**0.23**	0.30	9	**0.58**	0.79	21	4.11	5.32	45	6.37	8.22	210	2.25	2.90	47
	F(O)	**0.18**	0.25	16	**0.45**	0.62	24	2.60	3.49	39	4.75	6.14	78	1.76	2.32	51

Table 4.1: Table from [62]. MAE, RMSE, and MAX errors, in that order, for energy and forces are reported (in bold for values <1) in units of kcal/mol and kcal/(mol·Å), respectively.

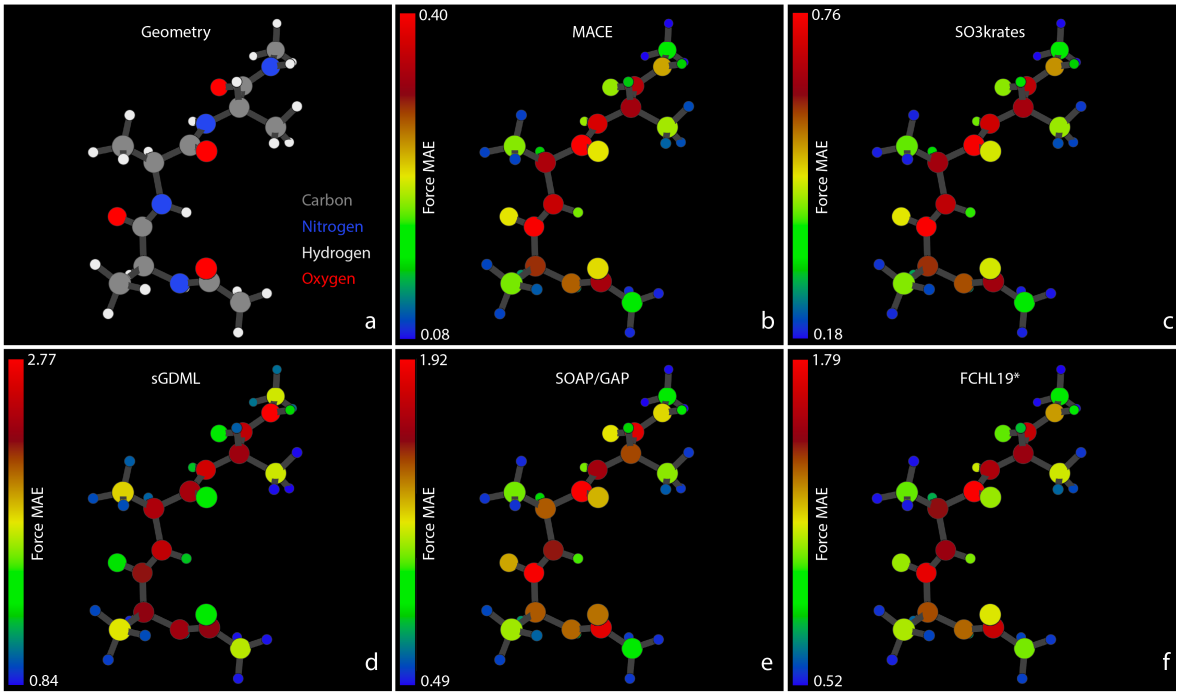

Figure 4.3: Figure from [61]. Atomic Force MAEs for Ac-Ala3-NHMe. Fig. (a) shows a snapshot of the system geometry and atom types: Carbons – grey, Nitrogens – blue, Hydrogens – white, and Oxygens – red. Figures (b) through (f) display the MAEs for forces, measured in kcal/(mol·Å), acting on individual atoms within the Ac-Ala3-NHMe. The MAEs correspond to the MLFFs predictions on the test set and are represented with different colours according to the colour bars shown with the corresponding scaling numbers, different for different MLFFs. Note the rather different absolute scale of the colour bars, ranging from 0.40 for MACE to 2.77 for sGDML.

between the relative errors of atom pairs, especially in the middle part of the molecule. Notable differences in prediction patterns appear mainly at the molecule's extremities. For example, examining the aromatic ring in the upper right corner of the figures, one can observe variations in accuracy for the Carbon atoms, particularly those connecting the ring to the rest of the molecule, going from MACE or SO3krates MLFFs to sGDML, and finally to SOAP/GAP and FCHL19* models. Also the observed MAX error patterns in the MACE, SO3krates, and SOAP/GAP models show a similarity in the relative error between different atoms for these MLFFs. Fig. 4.9 illustrates the atomistic force errors for configurations from the test set where these models exhibit the most significant deviations from the reference data (Figs. 4.4 and 4.5 contain more examples from Challenges I and II, respectively). Although the three geometries in Fig. 4.9 are not identical, they represent consecutive points along one of the MD trajectories used to form the reference dataset. Notably, the employed visualisation software, which determines chemical bonds based on interatomic distances, erroneously depicts the Hydrogen atom as being bonded to Oxygen, whereas, it is bonded to the Nitrogen atom, as illustrated in Fig. 4.6 (see the NH_3 tail in the lower central part).

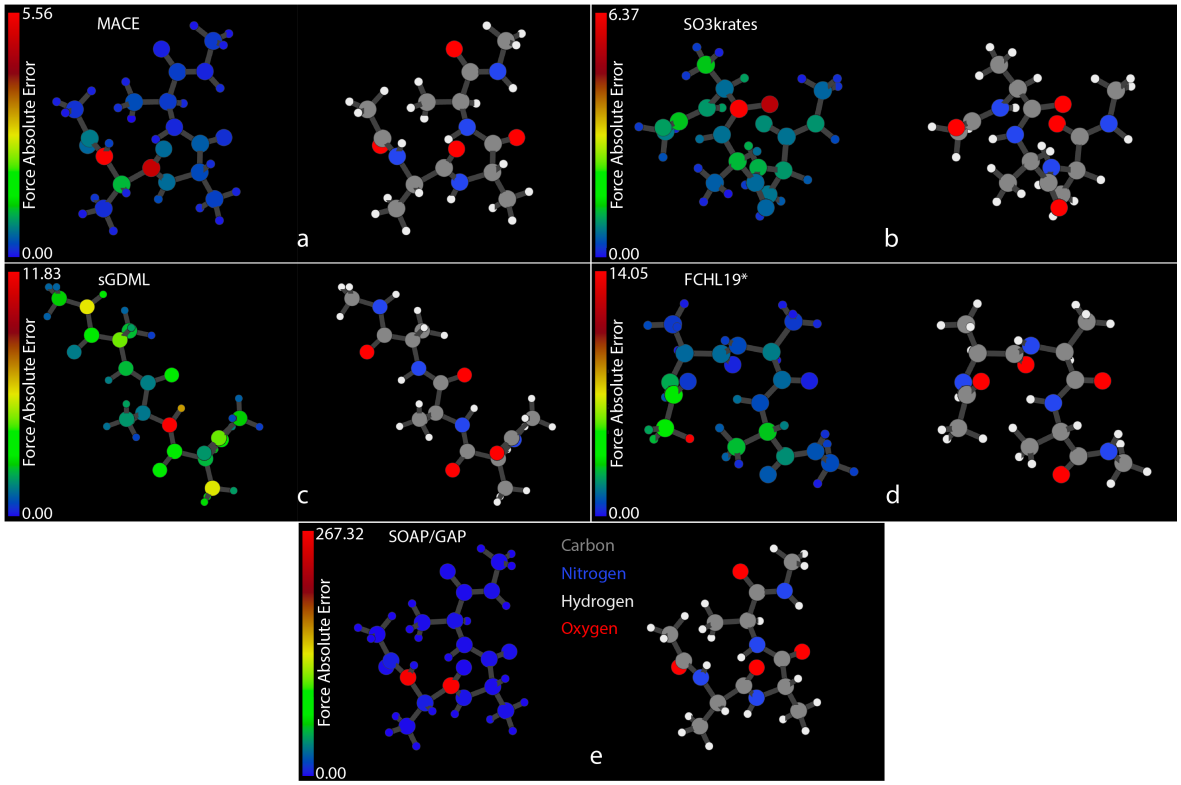

Figure 4.4: Figure from [61]. Maximum Atomic Force Errors for Ac-Ala3-NHMe. Figures (a) through (e) display the absolute atomic force errors, measured in kcal/(mol·Å), acting on individual atoms within the Ac-Ala3-NHMe system for the geometries where these errors are maximum for the corresponding ML model. The MAX values are represented with different colours according to the colour bars shown with the scaling numbers, different for different MLFFs.

4.2.3 Stability and Speed

One of the critical requirements for any MLFF is the ability to run efficient and reliable simulations. Beyond achieving low MAE, RMSE, or MAX errors on test sets, an MLFF must also ensure stable and computationally efficient MD trajectories. To that end, an assessment is carried out with 12 independent MD simulations, each for 1 M steps for each of the MLFF architectures participating in the TEA Challenge 2023. Let us first focus on the stability analyses and the averaged computational times required to generate 1 ns of dynamics (*i.e.*, 1 m energy and force evaluations).

Table 4.6 provides a summary of the test results of the stability, which was assessed using a broken bond criterion, terminating the simulation if any covalent chemical bond in the system exceeded a length of 2 Å. Under the given simulation conditions, the test systems were not expected to undergo proton transport or other processes that would alter their bonding patterns. The MACE architecture demonstrated the highest stability in the tests, success-

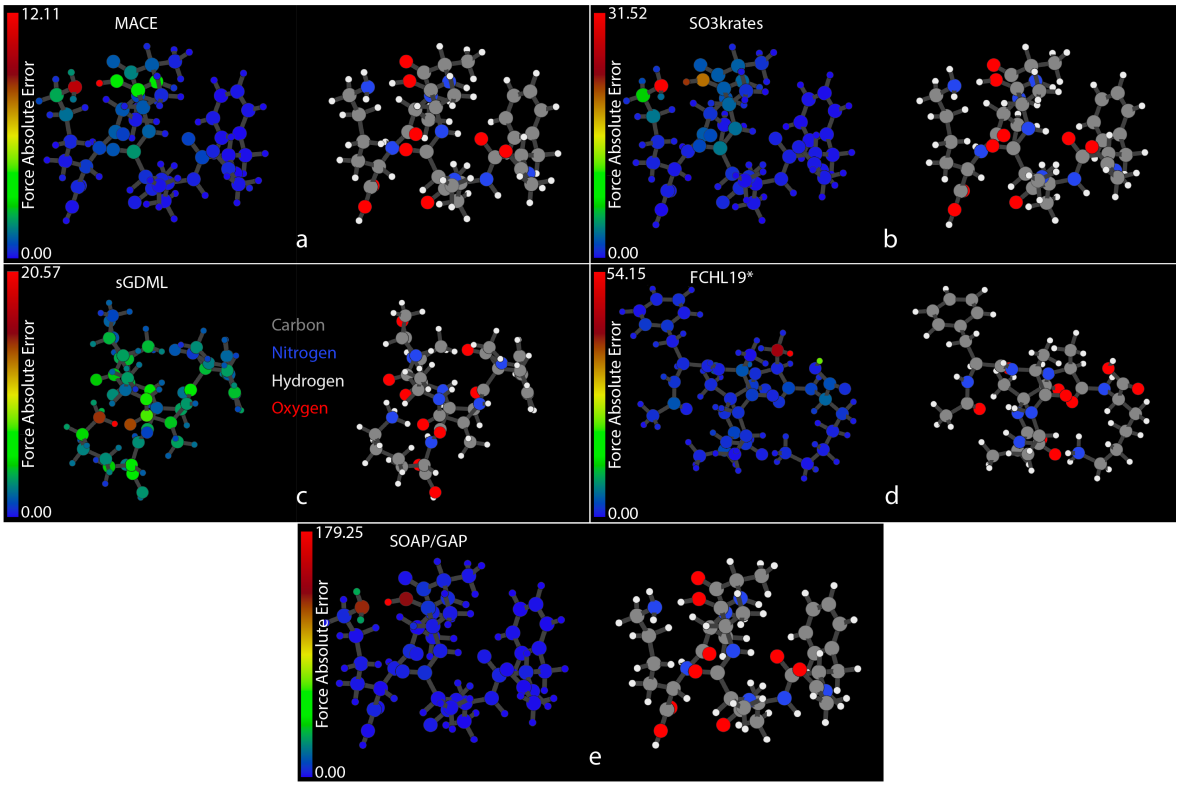

Figure 4.5: Figure from [61]. Maximum Atomic Force Errors for Ac-Phe-Ala5-Lys. Figures (a) through (e) display the absolute atomic force errors, measured in kcal/(mol·Å), acting on individual atoms within the Ac-Phe-Ala5-Lys system for the geometries where these errors are maximum for the corresponding ML model. The MAX values are represented with different colours according to the colour bars shown with the scaling numbers, different for different MLFFs.

fully completing most MD trajectories. The second most stable architecture, SO3krates, faced stability issues in some challenges when the temperature increased to 500 K and above. The causes of the difference in stability between the equivariant NN architectures is currently speculative and requires further investigation. Among the kernel-based models, sGDML exhibited the highest stability. It performed well at 300 K across all challenges except Challenge I, which had a significantly incomplete training set. It also maintained stability in Challenges II at 500 K. Nevertheless, the limited sampling of the PES in the reference dataset for Challenge I led to instability in the MD simulations of the alanine tetrapeptide molecule at 500 K. The SOAP/GAP and FCHL19* models, however, struggled to provide stable dynamics across all challenges, with some simulations failing within just a few thousand steps. The results here are consistent with the expected for such kernel methods due to their recommended training in active learning workflows to achieve stable dynamics [176, 262–264]. Nevertheless, achieving simulation durations ranging from dozens to even hundreds of picoseconds is feasible using the SOAP/GAP and FCHL19*

models, particularly in well-sampled near-equilibrium PES regions.

Overall, the results highlight the varying degrees of stability among different MLFF architectures and emphasise the need for continued improvements to ensure reliable MD simulations, cf. also [265, 266]. Additionally, the MACE and SO3krates models, when trained on the unfolded alanine tetrapeptide dataset underscore that despite the much lower mean and maximum force errors of these models compared to the kernel ones, this did not necessarily translate to better stability of MD simulations at 500 K, for the case of SO3krates. This suggests a weak correlation between the accuracy of MLFFs on the test set and their reliability in actual simulations and demonstrates the unreliability of high aggregate accuracy on reference data as a guarantee for stable and reliable performance in practical MD simulations. However, in relative terms, the models with the most accurate aggregate statistics (MACE and SO3krates) also consistently exhibited the highest stability, so overall accuracy is necessary for stability but not sufficient on its own. Building on this, incorporating uncertainty estimations into the workflow would represent a crucial step in enhancing MLFF modelling reliability. For kernel-based models, Bayesian methods provide a relatively straightforward way to achieve this. However, NN models such as MACE and SO3krates currently lack intrinsic mechanisms for estimating prediction uncertainty. Meanwhile, uncertainty estimation in NNs has been a widely researched topic in other fields [267]. Adapting and integrating the developed methodologies into MLFF codes should become a key research focus, especially as the complexity and scale of target systems grow, making reference calculations computationally prohibitive.

Table 4.5 summarises the average computational time required for each ML model to produce 1 ns of MD dynamics (1 million steps) on a single NVIDIA A100-40 GPU. The reported times are averaged across all MD trajectories within each Challenge. All run conditions and scripts were identical for all models except for the ASE calculator and the specific ML architecture provided by the participant [151]. It is important to note that the SOAP/GAP models cannot be executed on a GPU, so these models were run on CPU nodes consisting of 2 AMD Rome CPUs, each with 64 cores at 2.6 GHz, for a total of 128 cores and 512 GB of RAM. This hardware configuration was used to ensure that SOAP/GAP models could complete the MD simulations within a reasonable timeframe.

The SO3krates model, with its current settings, proved to be the fastest in production, outperforming its NN competitor, MACE, by a factor of $\tilde{2}0$. Surprisingly, for the MD simulations for Challenges I, it was even faster than the sGDML model, known for its minimalistic architecture. The third fastest ML model in production was FCHL19*, but this came at the cost of the largest instability in MD simulations. The MACE architecture, with its current settings, ranked fourth. For Challenges I and II, MACE showed only $\tilde{2}.5$ times increase in prediction speed compared to the SOAP/GAP model, even though SOAP/GAP ran on CPU rather than GPU. As the organic molecular systems in these challenges were smaller than the typical target of the MLFFs and the given the findings of also Challenges III and IV, no straightforward relationship between the simulation time and system size could be established firmly by the TEA Challenge 2023.

4.2.4 Guidelines for MLFF development and use

Defining precise guidelines, pitfalls, building a vision for the future is important and as efforts in this direction are already recognised in literature [268, 269].

When utilising any modern MLFF, it is important to recognise that much of the software is developed by scientists with varying degrees of coding experience, and not professional software engineers. Below, we outline some of the issues the TEA team encountered in the process of installation, integration within common frameworks, and during MD runs and tests with various MLFF architectures:

1. **Multi-Model Validation**: Even the most advanced single MLFF architecture could result in inaccuracies. Cross-checking results between different MLFF models can help increase the reliability of simulations, particularly where reference data (computational or experimental) is sparse or unavailable.

2. **Performance Analysis Beyond Aggregate Measures**: The aggregate measures of MLFFs' performances *e.g.*, MAE or RMSE could be misleading and do not tell the whole story in complex systems. In such cases, a detailed analysis of MLFF performance (per atom, per chemical element, per environment type) is crucial. Reducing the heterogeneity of atomistic MAEs while maintaining acceptable overall accuracy leads to more reliable MLFFs than those trained solely to minimise aggregated errors.

3. **Training Dataset Quality**: The completeness, composition, and accuracy of training datasets significantly impacts the performance of MLFFs. Related to the previous point, using datasets that over-represent certain types of states can decrease overall MAE and RMSE but might lead to incorrect simulation results.

4. **Appropriate Accuracy Levels**: Depending on the application, MLFFs with MAEs of, for instance, 0.5 or 0.1 kcal/(mol·Å) may produce the same results in MD simulations. A more accurate model requires more computationally demanding reference data and is slower in production and training, without providing any significant practical benefits. Even within the same MLFF architecture, modellers should explore the trade-off between model size, accuracy, and computational efficiency.

5. **Saving Training Information**: While some publishers may now enforce this, it is important in general use to document the complete training settings (hyperparameters), MLFF software version, and details of the training and validation datasets to ensure future applicability and potential retraining of an ML model. Ideally, this information should be automatically embedded in the MLFF model files, enabling the exact reproduction of the training process if the initial dataset is available.

6. **Transparency**: Developers of MLFFs should provide comprehensive details about modifications between different software and ML model versions, optimal preprocessing of training data beyond the intrinsic MLFF procedures, the units used, and any related offsets (*e.g.*, for energy) present in the outputs.

Several such pitfalls were corrected in the course of the challenge, highlighting the benefits such collaborative efforts bring to the entire community. By adhering to these guidelines, the development and application of MLFFs can achieve greater reliability and reproducibility of results, ensuring more robust and trustworthy simulations.

4.3 MDs Challenge I: Alanine tetrapeptide

Challenge I aims to verify the ability of MLFF models to reproduce the PES for flexible organic molecules, where the training dataset is chosen to include only folded or alternatively only unfolded configurations. The alanine tetrapeptide reference dataset comprises of 85,109 molecular geometries representing an *ab initio* NVT MD trajectory generated at 500 K [259]. For Challenge I, it was split into training, validation, and test sets employing three different strategies for the task. As a benchmark, the initial test involved MLFFs trained on representative samples of all possible extended and compact Ac-Ala3-NHMe structures with training/validation datasets of 200, 400, 600, 800, and 1,000 size to validate the convergence of the models. Next, the entire alanine tetrapeptide dataset was divided into two subsets based on the distance between the farthest non-Hydrogen atoms as a measure of the compactness of the molecule: 70% of the most compact structures form the folded dataset and the remaining 30% form the unfolded one. The classification is based solely on whether the corresponding interatomic distance is larger or smaller than the empirically defined threshold of 10.06 Å found suitable for this dataset. The challenge posed to the participants was to train MLFFs using only one of the proposed subsets of configurations (provided to all participants) and to predict the unseen data from the other subset. The evaluation criteria were the accuracy and stability of three types of MLFF models. Hence, the ability of MLFFs to extrapolate in the configurational space was investigated. All analyses on accuracy, stability, and performance of the provided MLFF models were conducted based on the most accurate models trained on 1000 reference geometries. Different training/validation test sizes were used solely to ensure the convergence of the ML models during the training process.

For the stability and performance tests, the trained models provided by participants were used by the TEA 2023 evaluation team to run 12 independent NVT MD simulations employing a Langevin thermostat with a friction coefficient of 10^{-3} fs^{-1} and a time step of 1 fs to measure the trajectory's average length before the system enters a non-physical state (*i.e.*, breaking chemical bonds). The 12 MD runs were started at a temperature of 300 K from 12 fixed non-identical system configurations. The 12 starting configurations were always chosen from the training datasets. For the models, which completed all 12 trajectories reached 1 ns length without entering an unphysical state, the next test was carried out at an increased temperature of 500 K. For the most robust models, 12 MDs were run also at 700 K. The starting conditions and the MD settings were identical for all MLFFs participating in the challenge.

For a decade, organic molecules of a few dozen atoms in the gas phase have been routinely

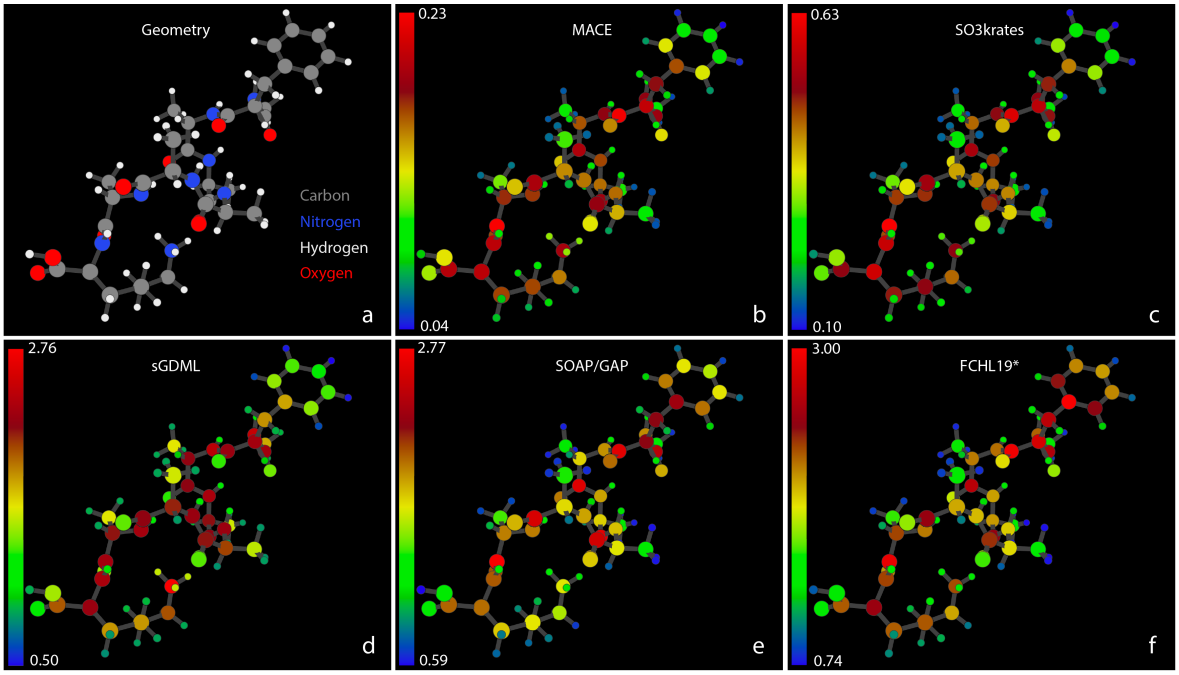

Figure 4.6: Figure from [61] Atomic Force MAEs for Ac-Phe-Ala5-Lys. Fig. **(a)** shows a snapshot of the system geometry and atom types: Carbons – grey, Nitrogens – blue, Hydrogens – white, and Oxygens – red. Figs. **(b)** through **(f)** display the MAEs for forces, measured in kcal/(mol·Å), acting on individual atoms within the Ac-Phe-Ala5-Lys system. The MAEs correspond to the MLFFs predictions on the test set and are represented with different colours according to the colour bars shown with the corresponding scaling numbers, different for different MLFFs. Note the rather different absolute scale of the colour bars, ranging from 0.23 for MACE to 3.00 for FCHL19*.

handled by different MLFF architectures. Additionally, such molecules are well within the capabilities of DFT codes, which can routinely compute hundreds of thousands of geometries without excessive computational effort. Analysis of the MD trajectories and a comparison between the outputs for the different MLFFs are presented here via Ramachandran plots, as shown in Fig. 4.7, to visualise the allowed conformational space of the peptide backbone [270]. For the Ac-Ala3-NHMe tetrapeptide, two selected pairs of dihedral angles in Fig. 4.7a, A and B, are conventionally referred to as ϕ_2/ψ_2 and ϕ_1/ψ_1, respectively [270–272]. The initial points for MD simulations are depicted in Fig. 4.7b. In order to obtain an informative picture of the results, analysis involving a clustering algorithm was carried out to identify the high density regions of population during the MD and filter out noise and unrepresentative low density areas [273, 274]. The full step-by-step algorithm description is available in Appendix C.

Our analysis algorithm identifies different (meta)stable domains in the Ramachandran plots, illustrated in various colors, Fig. 4.7c. The transitions between these (meta)stable domains obtained with SO3krates folded/unfolded and sGDML complete/folded models are repre-

sented in the graph form in Fig. 4.7d. It is important to note that the benchmark reference trajectory was obtained at 500 K to allow for more extensive sampling of the conformational space and bond lengths. To ensure statistically significant results, only the Ramachandran plots for ML models that produced stable 1 ns dynamics are presented. Consequently, the analyses are based on MLFF MD trajectories obtained at 300 K, as most MLFF models failed to produce stable 1 ns dynamics at 500 K. Nevertheless, the Ramachandran plots for the reference *ab inito* MD at 500 K still provide a qualitative guideline for the 300 K MLFF MD results.

The MACE models trained on both complete and folded datasets exhibit excellent mutual agreement. However, they undersample the upper left corner of the Ramachandran plot for A and the upper part of the dihedral cluster for B compared to the reference data. Both of these areas correspond to the highly unfolded conformations of the tetrapeptide. Training the MACE model on the unfolded dataset results in better qualitative agreement with the reference MD trajectory. Quantitative agreement estimation is challenging due to the short length of the *ab initio* MD trajectory and the difference in temperatures.

The SO3krates models display distinct Ramachandran profiles depending on whether they are trained on complete, folded, or unfolded datasets. Firstly, when trained on both complete and folded datasets, SO3krates model also undersample regions of highly unfolded conformations, similar to MACE. This indicates that this might be due to lower simulation temperature compared to the reference. Notably, the SO3krates model trained on the folded dataset exhibits additional metastable states with low populations (3% for A and 1% for B) and low transition probabilities. For dihedral B, the SO3krates model trained on the unfolded dataset identifies an extra metastable region with a 9% cluster population, though the transition probability into this state is low, with only one transition observed in 12 ns of total dynamics (12x1 ns). The seeds of this cluster also appear as two small clusters with a 1% population and relatively high mutual transition rates in the SO3krates model trained on the folded dataset. These molecular geometries were not observed in MD simulations using the SO3krates model trained on the complete dataset or by any MACE or sGDML models. However, exploration to these regions have been studied previously in the original SO3krates article [275]. It is worth noting that the two small clusters in the *SO3 fold* plot, dihedrals B, and the two clusters in the *SO3 unfold* plot, dihedrals A, can be merged due to their relatively high transition rates. The clustering algorithm employed here uses a predefined fixed number of transitions to merge clusters chosen to suit the data in general. It does not account for their population, providing suboptimal results when at least one of the clusters is small.

The sGDML model trained on a complete dataset also demonstrates acceptable results. The total number of transitions between the two clusters identified for B is 86, slightly below the manually selected threshold of 100 for merging the clusters into one. However, *removing parts of the reference geometries in folded or unfolded datasets leads to significant differences in the Ramachandran profiles or even MD instability.* This sensitivity is attributed to the sGDML model's interpolation in the space of a global system descriptor of inverse distances,

making the model highly dependent on the quality and completeness of the training dataset compared to MLFFs employing local descriptors.

In SI, a comprehensive Table 4.7 lists the chemical bonds responsible for the instability of all MLFF architectures trained on different datasets at 300, 500, and 700 K. Most broken bonds involve carbon atoms connected to other elements. Additionally, there are notable differences in bond-breaking patterns between kernel-based and equivariant NN-based ML models. For sGDML, SOAP/GAP, and FCHL19*, the specific bond causing molecular instability was readily identifiable. By contrast, for SO3krates, bond breaking exhibited an explosion-like behavior, with a large part of the molecule decomposing into atoms within a few dozen steps, making it challenging to pinpoint the exact bond responsible for the instability.

In summary, when trained on a complete dataset, we observed strong mutual agreement in the molecular dynamics of the alanine tetrapeptide generated by MACE and SO3krates models. Discrepancies arise when the training datasets lack either folded or unfolded geometries. Both models reasonably explored the unfolded PES basin when trained on the unfolded dataset. The MACE folded model is consistent with the complete one. The SO3krates folded and unfolded models were mostly consistent, while the sGDML MLFFs demonstrated increased sensitivity to variations in the training data.

4.4 MDs Challenge II: N-acetylphenylalanyl-pentaalanyl-lysine

Challenge II verifies the ability of MLFFs to deal with a different type of reference data incompleteness and is based on the Ac-Phe-Ala5-Lys dataset. As a baseline, the “complete" ML models are trained on a fixed number (10 and 20) of randomly selected molecule configurations from each of the 200 MD trajectories. Additionally, the ML models were trained on a separate dataset to simulate “incomplete" training data by sampling a fixed number (16 and 32) of randomly selected configurations from 125 out of the 200 MD trajectories. The remaining 75 trajectories were excluded from training and used as an “unseen" set. The accuracy of the MLFFs was compared for complete and incomplete ML models tested on both seen and unseen test datasets. Notably, the training and validation dataset sizes are 2000 and 4000 for both types of models, and the datasets are the same for all MLFF architectures, while for the analysis, only the models trained on 4000 reference geometries were used. The stability and performance of the MLFFs were estimated in the same manner as for Challenge I.

Challenge II in the TEA Challenge 2023 was performed on a larger organic system, namely the protonated Ac-Phe-Ala5-Lys peptide with 100,000 data points. The main aim of the challenge was to assess the ability of MLFFs to handle different types of reference data incompleteness. To this end, two training datasets were created for the Ac-Phe-Ala5-Lys peptide system based on MD simulations around the 200 lowest energy conformers. The

“complete" dataset consisted of 20 randomly selected molecular configurations from each of these trajectories, while the “incomplete" dataset contained 32 randomly selected configurations from only 125 trajectories. The clustering algorithm used here is the same one as for Challenge I (see Appendix C). Fig. 4.10 presents the Ramachandran plots for three selected pairs of neighbouring dihedral angles along the peptide. The dihedral angle pairs, labelled as A, B, and C (see Fig. 4.10a), represent key cases for comparing different MLFFs. The initial points for MD simulations are shown in Fig. 4.10b. Only the Ramachandran plots for ML models that produced stable 1 ns long dynamics are displayed. The FCHL19* and SOAP/GAP MLFFs did not achieve the required 1 ns to perform statistically-significant analysis. The conducted analysis identifies distinct (meta)stable domains in the Ramachandran plots, represented in different colours, Fig. 4.10c. Since the original dataset was not derived from an MD trajectory, reference density plots are unavailable. Therefore, transition graphs illustrate the transitions between (meta)stable domains for dihedral C across different ML models, Fig. 4.10d.

The initial comparison of the MD results of the MACE and SO3krates models, each trained on complete (com) and incomplete (incom) datasets, Fig. 4.10c. The most noticeable difference is the increased number of clusters in the MDs generated by the models trained on the incomplete dataset, primarily reflecting variations in the starting points for the MD simulations. Additional red clusters appear for the A and B dihedrals due to distinct starting points in the incomplete trajectories (highlighted by red circles in Fig. 4.10b). These starting points significantly differ from other initial configurations, leading to divergences in the Ramachandran plots. Both models, however, accurately reproduce the close-to-equilibrium regions of the PES for A and B, providing similar positions of the minima and shapes of the probability distributions around them (within 3% agreement in population distributions between clusters). For the C dihedrals, a block of four starting points (highlighted by an orange circle in Fig. 4.10b) appears like an orange cluster. Additionally, a shift in the red cluster position from 0 to $-\pi/2$ along the Y-axis is attributed to differences in starting configurations between the complete and incomplete datasets (emphasized by a red circle in Fig. 4.10b). Differences in the MLFFs from the MACE and SO3krates architectures significantly influence the distribution of the C dihedrals, as supported by transition graphs in Fig. 4.10d, differing especially in the case of training on an incomplete dataset. The MACE MLFF shows a 54% relative population for the complete C dihedral case for the largest cluster $(-\pi/2, \pi/2)$ compared to 60% with the SO3krates MLFF. This difference stems from a higher transition probability within the MACE PES from this state to the green cluster $(-\pi, \pi/2)$. Conversely, the green cluster population is higher with the MACE PES at 40%, compared to 32% for the SO3krates PES, due to a nearly twice higher transition probability between the green and red $(\pi, 0)$ metastable states with the SO3krates MLFF trained on the complete dataset. These results suggest that the primary difference between MACE and SO3krates MLFFs lies in the description of out-of-equilibrium regions of the PES, which are responsible for rare transitions or large geometry fluctuations. Further comparison of the MACE and SO3krates models trained on the incomplete dataset for the C dihedrals supports this. Noticeably different transition patterns emerge, as well as differing cluster

populations and even the appearance of an extra metastable state (blue cluster) within the PES reconstructed with the MACE MLFF (incomplete dataset). It is also likely that such differences would emerge between different MLFFs trained on the same data with the same architecture, just using a different set of initial weights. However, it is worth noting that the statistically converged analyses of the transition patterns would require more extended MD simulations or the employment of enhanced sampling techniques, which is beyond the scope of the current work. Simultaneously, the MACE and SO3krates architectures, when trained on the complete dataset, demonstrate remarkable mutual agreement in modelling the MDs of the Ac-Phe-Ala5-Lys peptide. This consistency highlights the capability of modern equivariant NNs with different architectures to handle relatively large and complex organic molecules and model their molecular dynamics and transitions.

For the sGDML architecture, however, only qualitative agreement is found with the results from both NNs in Challenge II. The shape and population of clusters differ significantly due to the global nature of the sGDML model, which requires a reliable and comprehensive set of representative query configurations for effective interpolation between system states. Achieving such a representative set is challenging for large molecules undergoing complex structural transformations, especially with a relatively small training dataset of 4,000 configurations. Consequently, the sGDML model operates in a low-data regime, leading to a notable depreciation in performance.

Table 4.8 details the broken chemical bonds responsible for the instability of MD simulations across all MLFF architectures, where similar behaviour to that reported for Challenge I is observed. Notably, the kernel-based MLFFs utilising local descriptors, namely SOAP/GAP and FCHL19*, could not sustain stable dynamics over a 1 ns duration for both peptides evaluated in this study. This finding indicates that MD simulations employing global sGDML models or NN architectures, incorporating non-locality through message-passing elements, exhibit greater stability than MLFFs based solely on local ML models. Overall, the observations based on the MD simulations Challenge II indicate that the results for MACE, SO3krates, and sGDML MLFFs in this test were consistent with those from Challenge I. Hence, the equivariant NN MLFFs can be reliably trusted to reproduce the MDs of organic molecules (at least up to 100 atoms). The main discrepancies between MD results occur when the training datasets lack representative reference geometries. This suggests that active learning or similar iterative approaches [276, 277] should become an integral part of MLFF training procedures to ensure the completeness and representativeness of training datasets.

4.5 Conclusions

The TEA Challenge 2023 provided a comprehensive evaluation of a representative sample of modern MLFF models, capturing the current state of the art, and identifying strengths and weaknesses of existing architectures. ML models were tested on their ability to predict potential energy surfaces and forces with high accuracy and reliability. The selected models

included both kernel and NN architectures, namely MACE, SO3krates, sGDML, SOAP/GAP, and FCHL19*, whose MLFF developers participated in TEA Challenge 2023. The evaluation comprised four distinct tasks designed to test the limits of MLFFs under various conditions and system complexities, the first two of which were for large flexible organic molecules obtained at highly accurate reference DFT level of PBE0+MBD(-NL). The results demonstrate significant advancements in the accuracy of MLFFs, particularly with novel equivariant neural network architectures such as MACE and SO3krates, which showed marked improvements in MAE and RMSE for energy and force predictions compared to kernel-based models like sGDML, SOAP/GAP, and FCHL19* in the context of TEA 2023 Challenge I and II.

However, maximum errors remain a critical challenge, highlighting the necessity for further refinement to reduce rare but substantial prediction discrepancies. Additionally, all ML models demonstrated significant heterogeneity in force error across the different atoms in a given system, with a worst-to-best atomistic MAE ratio up to $\sim$3 times. Interestingly, the heterogeneity patterns were consistent across all MLFF architectures. While all of them offer opportunities for trade-offs between model accuracy and computational efficiency, the balance in this study depends on the expert intuition of the participating developers of these methods. With this individual's choices in mind, the SO3krates model here outperformed its competitors in computational speed by a significant margin, demonstrating the potential for highly efficient MD simulations. Overall, MLFF performances may vary by two orders of magnitude for the same task and even within the same class of ML models, *e.g.*, MACE and SO3krates, prediction speed can differ by a factor of 25. In addition to potential code optimisation, the speed is related to model size, so a rigorous study of the MLFF's performance speed is needed to explore the aforementioned trade-off between model size, corresponding speed, and prediction accuracy.

Stability assessments in the context of MD simulations also revealed that the MACE and SO3krates models generally exhibited better stability in MD simulations, completing most 1 ns trajectories under varying conditions, while kernel-based models like SOAP/GAP and FCHL19* showed substantial instability, particularly at higher temperatures, though performing adequately in well-sampled, near-equilibrium regions of the potential energy surface. The comprehensive comparative analysis of MD simulations conducted under identical conditions aimed to identify reliable application areas in the current models and those requiring further improvement. For the organic molecules in Challenge I and II, excellent agreement was observed between the MD results of the MACE and SO3krates MLFFs when trained on comprehensive datasets. Discrepancies were primarily in the transition regions between (meta)stable states or large atomic fluctuations, likely due to the incompleteness of the training dataset rather than the ML architecture itself. The sGDML model also provided reliable MD trajectories for the smaller peptide. In contrast, the other two investigated kernel-based models, SOAP/GAP and FCHL19*, exhibited insufficient stability for extended MD simulations. Despite the success of MLFFs trained on comprehensive datasets, the dynamics of the molecules in Challenges I and II revealed noticeable artifacts when MLFFs were trained on incomplete datasets. This issue affected both kernel-based

models and NNs, underscoring the importance of reliable, high-quality, and comprehensive training data as a major bottleneck in developing effective MLFFs for organic molecules.

Several technical pitfalls were encountered during the TEA Challenge 2023. While direct communication with MLFF architecture developers allowed for the resolution of most issues, this may not be available for ordinary users to such an extent. Hence, the reliable integration of such models into drug development pipelines also necessitates a stable, clear, and robust version of the ML code release.

Overall, the TEA Challenge 2023 demonstrated notable progress in MLFFs over the past decade, achieving new standards for accuracy and stability. However, continued innovation and optimisation are essential to fully realise the potential of MLFFs in practical applications, ensuring reliable and efficient simulations across a broad spectrum of chemical systems. Incorporating active learning or similar iterative approaches into MLFF training procedures is crucial to ensure thorough and representative datasets. Novel equivariant MLFF architectures such as MACE and SO3krates significantly enhance the faithful reconstruction of molecular PES. However, a major challenge remains that of eliminating rare but substantial errors in force predictions. The observed maximum force errors are comparable to those of previous MLFF generations, such as sGDML. Force reconstruction demonstrates significant heterogeneity between different atoms of a given system that is consistent across all tested models, including both kernel-based and NN approaches. The consistent error patterns are primarily due to the Euclidean L2 nature of the loss function applied to the forces vector across all models. The observed heterogeneity, which does not always align with the magnitude of the force acting on an atom, suggests a direction for further development of MLFFs. Furthermore, moving beyond the aggregate accuracy measurements like the MAE or RMSE of the system can uncover potential pitfalls of MLFFs.

Challenge		MACE			SO3			sGDML			SOAP/GAP			FCHL19*		
I (com)	E	0.54	0.68	3.9	1.20	1.52	7.1	3.87	5.06	27	4.14	5.34	32	2.70	3.48	16
	F	1.01	1.54	53	1.99	2.93	91	7.77	11.0	140	5.20	8.65	2580	4.99	7.03	152
	F(H)	0.80	1.18	64	1.67	2.36	83	7.07	9.40	155	4.61	6.22	160	4.61	6.30	209
	F(C)	1.07	1.49	34	2.07	2.81	65	7.88	10.4	95	5.23	8.18	1850	5.04	6.67	82
	F(N)	1.24	1.69	40	2.38	3.17	40	9.85	13.0	106	5.61	10.6	1890	5.46	7.09	64
	F(O)	1.23	1.71	50	2.37	3.18	76	7.33	9.87	118	6.88	8.96	98	5.62	7.43	106
I (fold-fold)	E	0.28	0.36	2.0	0.91	1.14	5.1	2.26	3.00	29	2.52	3.25	16	1.67	2.15	12
	F	0.99	1.51	49	1.96	2.87	113	7.41	10.6	148	5.05	7.20	435	4.80	6.74	88
	F(H)	0.80	1.18	67	1.69	2.40	154	6.74	8.99	133	4.50	6.09	175	4.45	6.04	121
	F(C)	1.04	1.45	30	2.02	2.72	37	7.58	10.1	105	5.10	6.85	309	4.85	6.39	60
	F(N)	1.22	1.68	35	2.35	3.13	49	9.26	12.3	111	5.34	7.10	326	5.32	6.91	63
	F(O)	1.22	1.70	46	2.30	3.09	57	6.86	9.35	116	6.70	8.78	139	5.36	7.06	69
I (fold-unfold)	E	0.79	1.02	4.8	2.34	2.79	10	18.4	26.1	112	5.34	6.67	26	6.39	8.33	32
	F	1.33	2.08	57	2.52	3.79	61	13.6	19.7	207	5.91	8.47	114	5.88	8.51	108
	F(H)	1.00	1.47	80	2.07	2.99	85	12.0	16.5	244	5.19	7.17	159	5.30	7.40	121
	F(C)	1.45	2.05	30	2.59	3.54	30	14.1	18.9	140	5.92	7.90	70	6.04	8.13	77
	F(N)	1.60	2.16	26	3.10	4.21	35	16.6	21.7	151	6.44	8.43	81	6.54	8.62	68
	F(O)	1.77	2.39	29	3.18	4.32	36	13.2	18.1	147	7.96	10.4	111	6.63	8.80	83
I (unfold-unfold)	E	0.84	0.98	3.4	1.14	1.43	6.4	3.20	5.31	93	3.35	4.33	16	2.14	3.01	32
	F	0.84	1.33	46	1.66	2.45	61	6.38	9.37	231	4.76	6.67	109	4.33	6.06	98
	F(H)	0.72	1.09	56	1.48	2.07	62	6.07	8.36	323	4.42	5.92	121	4.24	5.80	132
	F(C)	0.87	1.26	20	1.68	2.32	43	6.35	8.74	109	4.78	6.31	56	4.24	5.59	47
	F(N)	1.00	1.45	33	1.93	2.63	37	7.79	10.6	119	4.88	6.35	80	4.62	6.08	71
	F(O)	0.98	1.40	28	1.86	2.52	34	5.76	8.02	89	5.90	7.66	57	4.66	6.12	67
I (unfold-fold)	E	2.08	2.70	11	1.95	2.36	10	53.1	70.7	319	4.40	5.62	23	9.17	12.9	75
	F	2.08	3.82	131	2.93	4.52	113	22.9	34.5	468	6.44	9.59	503	8.83	13.6	353
	F(H)	1.39	2.45	180	2.32	3.67	154	20.9	32.7	641	5.67	8.67	689	8.39	13.7	484
	F(C)	2.19	3.69	53	3.03	4.21	73	22.3	30.5	278	6.14	8.34	194	7.85	10.9	162
	F(N)	2.91	4.15	96	3.73	4.96	56	28.5	37.8	236	7.62	10.2	221	11.8	16.2	151
	F(O)	3.37	5.27	69	3.92	5.42	104	26.4	37.8	314	9.33	12.8	180	10.9	15.5	294
II (comp)	E	0.35	0.45	5.0	0.80	1.00	4.2	6.12	8.40	44	12.8	16.0	70	3.94	4.99	23
	F	0.59	0.88	170	1.57	2.37	455	7.19	10.6	280	25.5	35.1	1910	8.09	11.3	621
	F(H)	0.52	0.75	238	1.40	2.04	409	6.69	9.45	393	24.7	33.5	1630	7.59	10.2	871
	F(C)	0.60	0.83	38	1.59	2.22	62	7.19	9.88	137	26.6	34.4	404	8.37	11.0	305
	F(N)	0.68	0.93	90	1.84	2.54	333	9.08	12.4	176	23.6	30.4	879	8.10	10.4	222
	F(O)	0.74	1.03	115	1.89	2.61	221	6.97	9.85	192	25.5	33.4	1930	8.77	11.5	183
II (incom-incom)	E	0.36	0.51	21	0.65	0.83	14	7.90	11.3	61	13.8	17	60	3.87	4.88	24
	F	0.53	0.85	420	1.49	2.28	442	6.31	9.55	279	26.1	35.8	2170	8.06	11.2	636
	F(H)	0.48	0.86	587	1.35	2.06	617	5.85	8.42	390	25.1	33.9	2420	7.54	10.2	889
	F(C)	0.53	0.74	66	1.50	2.11	137	6.31	8.89	119	27.6	35.4	389	8.31	10.9	317
	F(N)	0.61	0.90	129	1.69	2.40	196	7.96	11.1	120	24.1	31.2	1590	8.10	10.4	224
	F(O)	0.65	0.97	182	1.79	2.57	412	6.17	8.98	256	25.0	32.2	490	8.88	11.7	555
II (incom-unkn)	E	0.44	0.56	2.6	0.92	1.17	6.3	60.9	71.6	346	16.2	20.3	82	5.33	6.81	35
	F	0.70	1.07	80	1.83	2.82	173	13.1	18.5	497	25.9	35.7	1060	8.29	11.6	263
	F(H)	0.57	0.85	75	1.53	2.33	245	13.7	19.9	702	25.2	34.1	1130	7.8	10.6	293
	F(C)	0.73	1.01	28	1.90	2.69	65	12.1	15.9	154	27.4	35.2	395	8.52	11.2	156
	F(N)	0.85	1.12	34	2.19	2.95	79	15.4	20.0	167	23.9	30.9	774	8.44	10.9	175
	F(O)	0.92	1.28	80	2.29	3.18	123	13.3	17.8	201	24.3	31.4	397	9.00	11.9	261

Table 4.2: Table from [62]. Normalised MAE, RMSE, and MAX errors for relative energy and forces in %, w.r.t. the mean absolute energies and forces in the DFT reference.

Challenge		MACE			SO3			sGDML			SOAP/GAP			FCHL19*		
I (com)	E	1.86	2.35	13.5	4.16	5.26	24.6	13.4	17.5	94.7	14.3	18.5	111	9.35	12.0	56.3
	F	0.75	1.14	39.5	1.47	2.17	67.5	5.77	8.22	106	3.86	6.41	1920	3.70	5.21	112
	F(H)	0.61	0.89	48.5	1.27	1.79	62.7	5.36	7.12	117	3.49	4.71	121	3.49	4.77	158
	F(C)	0.85	1.18	26.6	1.64	2.22	51.6	6.24	8.25	75.3	4.14	6.48	1460	3.99	5.28	65.0
	F(N)	0.97	1.31	30.8	1.85	2.47	31.0	7.67	10.1	82.7	4.36	8.28	1470	4.25	5.52	50.3
	F(O)	0.92	1.27	37.2	1.77	2.37	56.7	5.47	7.36	87.9	5.13	6.68	72.7	4.19	5.54	79.4
I (fold-fold)	E	1.56	2.00	11.1	5.06	6.29	28.4	12.5	16.6	158	14.0	18.0	88.3	9.26	11.9	69.4
	F	0.73	1.12	36.0	1.45	2.13	83.7	5.50	7.88	110	3.75	5.34	323	3.56	5.00	65.7
	F(H)	0.60	0.89	50.2	1.27	1.81	117	5.09	6.79	101	3.40	4.60	132	3.36	4.56	91.5
	F(C)	0.82	1.15	23.8	1.60	2.16	29.2	6.00	8.00	83.5	4.04	5.42	245	3.84	5.06	48.2
	F(N)	0.95	1.30	27.1	1.82	2.43	38.3	7.20	9.58	86.3	4.15	5.52	253	4.13	5.37	49.4
	F(O)	0.91	1.27	34.5	1.72	2.31	42.7	5.12	6.99	86.6	5.00	6.56	104	4.00	5.28	51.5
I (fold-unfold)	E	2.58	3.37	15.8	7.70	9.19	33.9	60.6	85.7	368	17.5	21.9	85.0	21.0	27.4	104
	F	0.99	1.54	42.5	1.87	2.81	45.0	10.1	14.6	154	4.38	6.28	84.3	4.36	6.31	80.1
	F(H)	0.77	1.12	61.3	1.58	2.28	64.9	9.19	12.6	186	3.96	5.47	122	4.04	5.65	92.4
	F(C)	1.15	1.62	23.3	2.05	2.80	23.8	11.1	15.0	111	4.69	6.26	55.1	4.78	6.44	61.1
	F(N)	1.25	1.68	20.3	2.42	3.28	27.2	12.9	16.9	118	5.02	6.57	62.9	5.10	6.72	53.1
	F(O)	1.32	1.78	21.4	2.37	3.21	26.6	9.83	13.4	109	5.92	7.75	82.4	4.93	6.54	62.4
I (unfold-unfold)	E	2.75	3.22	11.1	3.76	4.71	21.0	10.5	17.5	305	11.0	14.2	53.1	7.05	9.89	104
	F	0.62	0.98	33.9	1.23	1.82	45.1	4.73	6.95	171	3.53	4.95	80.7	3.21	4.49	73.0
	F(H)	0.55	0.83	42.3	1.13	1.58	47.0	4.63	6.38	247	3.37	4.51	92.0	3.23	4.42	101
	F(C)	0.69	0.99	15.9	1.33	1.84	34.4	5.03	6.92	86.5	3.79	5.00	44.4	3.36	4.43	37.5
	F(N)	0.78	1.13	26.0	1.50	2.05	28.5	6.07	8.29	92.6	3.80	4.95	62.0	3.60	4.74	56.0
	F(O)	0.73	1.04	21.0	1.38	1.88	25.4	4.28	5.96	65.8	4.39	5.70	42.0	3.46	4.55	50.5
I (unfold-fold)	E	11.5	15.0	62.5	10.8	13.1	56.6	294	392	1770	11.0	14.2	53.1	50.8	71.2	418
	F	1.55	2.84	97.4	2.17	3.35	83.7	17.0	25.6	348	4.78	7.12	373	6.55	10.1	262
	F(H)	1.05	1.85	136	1.75	2.77	117	15.8	24.7	484	4.29	6.55	520	6.34	10.3	365
	F(C)	1.73	2.92	42.2	2.40	3.33	57.7	17.6	24.2	220	4.86	6.61	154	6.21	8.61	128
	F(N)	2.26	3.23	74.9	2.90	3.86	43.6	22.1	29.4	184	5.92	7.94	172	9.14	12.6	117
	F(O)	2.52	3.93	51.7	2.93	4.05	77.8	19.7	28.3	234	6.97	9.54	134	8.11	11.6	220
II (com)	E	1.51	1.95	21.7	3.51	4.37	18.3	26.8	36.8	192	55.9	69.9	305	17.2	21.8	103
	F	0.43	0.65	125	1.15	1.74	333	5.28	7.80	205	18.7	25.8	1400	5.93	8.26	456
	F(H)	0.38	0.56	176	1.04	1.51	302	4.93	6.97	290	18.2	24.7	1200	5.60	7.56	643
	F(C)	0.47	0.65	29.6	1.24	1.74	48.6	5.64	7.75	107	20.9	27.0	317	6.57	8.59	239
	F(N)	0.53	0.72	69.8	1.42	1.96	257	7.02	9.57	136	18.3	23.5	680	6.26	8.05	171
	F(O)	0.55	0.77	85.6	1.40	1.94	164	5.18	7.32	142	19.0	24.8	1430	6.52	8.56	136
II (incom-incom)	E	1.60	2.24	90.7	2.84	3.67	63.1	34.7	49.5	267	60.7	75.5	265	17.0	21.5	105
	F	0.39	0.63	309	1.10	1.68	324	4.63	7.01	205	19.1	26.3	1590	5.91	8.23	467
	F(H)	0.35	0.64	434	1.00	1.52	456	4.32	6.22	288	18.5	25.1	1787	5.57	7.52	656
	F(C)	0.41	0.58	51.5	1.18	1.66	108	4.96	6.98	93.2	21.7	27.8	305	6.52	8.51	249
	F(N)	0.47	0.70	100	1.31	1.86	152	6.15	8.60	92.4	18.6	24.2	1230	6.26	8.06	173
	F(O)	0.48	0.72	136	1.33	1.92	307	4.60	6.69	191	18.6	24.0	365	6.61	8.72	414
II (incom-unkn)	E	1.57	2.01	9.28	3.28	4.17	22.6	218	256	1240	57.8	72.7	294	19.1	24.3	125
	F	0.51	0.78	58.8	1.34	2.07	127	9.62	13.5	364	19.0	26.1	7790	6.07	8.50	192
	F(H)	0.42	0.63	55.2	1.12	1.72	180	10.1	14.7	517	18.5	25.1	832	5.73	7.80	216
	F(C)	0.57	0.79	22.0	1.49	2.11	51.1	9.48	12.5	120	21.5	27.6	310	6.68	8.79	123
	F(N)	0.65	0.87	26.4	1.69	2.28	60.7	11.9	15.4	129	18.5	23.8	598	6.52	8.40	135
	F(O)	0.68	0.95	59.0	1.69	2.35	91.0	9.87	13.2	149	18.0	23.2	294	6.66	8.79	193

Table 4.3: Table from [62]. MAE, RMSE, and MAX errors for relative energy and forces in %, w.r.t. the standard deviations of the energies and forces in the DFT reference.

Simulation		Challenge				
Model	Temperature, K	I			II	
		com	fold	unfold	com	incom
MACE	300	12/0	12/0	12/0	12/0	12/0
	500	12/0	12/0	12/0	12/0	12/0
	700	12/0	12/0	12/0	12/0	12/0 §
SO3krates	300	12/0	12/0	11/1	12/0	12/0
	500	12/0	11/1	5/7	12/0	12/0
	700	4/8	7/5	0/12	5/7	8/4
sGDML	300	12/0	12/0	0/12	12/0	12/0
	500	7/5	7/5	0/12	12/0	12/0
	700	–	–	–	12/0	12/0
SOAP/GAP	300	2/10	0/12	8/4	0/12	0/12
FCHL19*	300	0/12	0/12	0/12	0/12	0/12

Table 4.4: Table from [61]. Molecular dynamics stability in A/B ratio. Here, A is the number of completed trajectories (1 ns), and B is the number of failed trajectories (broken chemical bonds).

Simulation	Challenge	
Model	I	II
MACE	34.4	43.6
SO3krates	1.4	2.0
sGDML	2.4	2.7
SOAP/GAP	83.6	111.0
FCHL19*	11.9	12.4

Table 4.5: Table from [61]. Average simulation time in hours per 1 ns of molecular dynamics. The red colour indicates models that failed to generate a single 1 ns MD trajectory.

Simulation		Challenge				
Model	Temperature [K]	I			II	
		com	fold	unfold	com	incom
MACE	300	12/0	12/0	12/0	12/0	12/0
	500	12/0	12/0	12/0	12/0	12/0
	700	12/0	12/0	12/0	12/0	12/0
SO3krates	300	12/0	12/0	11/1	12/0	12/0
	500	12/0	11/1	5/7	12/0	12/0
	700	4/8	7/5	0/12	5/7	8/4
sGDML	300	12/0	12/0	0/12	12/0	12/0
	500	7/5	7/5	0/12	12/0	12/0
	700	–	–	–	12/0	12/0
SOAP/GAP	300	2/10	0/12	8/4	0/12	0/12
FCHL19*	300	0/12	0/12	0/12	0/12	0/12

Table 4.6: Table excerpt from [62]. Molecular dynamics stability in A/B ratio. Here, A is the number of completed trajectories (1 ns), and B is the number of failed trajectories (broken chemical bonds).

Task		MACE	SO3	sGDML	SOAP/GAP	FCHL19*
I (com) @300K	C-H	0	0	0	2	14
	N-H	0	0	0	1	0
	C-C	0	0	0	6	0
	C-N	0	0	0	1	0
	C-O	0	0	0	0	1
I (com) @500K	C-H	0	0	1	9	-
	C-C	0	0	4	9	-
	C-N	0	0	1	1	-
I (com) @700K	C-H	0	67	1	-	-
	N-H	0	20	0	-	-
	C-C	0	26	4	-	-
	C-N	0	32	0	-	-
	C-O	0	15	0	-	-
I (fold) @300K	C-H	0	0	0	2	17
	N-H	0	0	0	1	0
	C-C	0	0	0	6	0
	C-N	0	0	0	1	0
	C-O	0	0	0	0	1
I (fold) @500K	C-H	0	0	0	9	-
	C-C	0	0	4	9	-
	C-N	0	0	1	1	-
I (fold) @700K	C-H	0	66	1	-	-
	N-H	0	19	0	-	-
	C-C	0	26	4	-	-
	C-N	0	32	0	-	-
	C-O	0	14	0	-	-
I (unfold) @300K	C-H	0	1	7	0	8
	N-H	0	0	1	2	11
	C-C	0	7	0	0	1
	C-N	0	4	0	2	1
I (unfold) @500K	C-H	0	2	7	0	-
	N-H	0	11	3	5	-
	C-C	0	0	5	5	-
	C-N	0	0	1	2	-
I (unfold) @700K	C-H	0	3	-	-	-
	N-H	0	19	-	-	-
	C-N	0	1	-	-	-
	C-O	0	2	-	-	-

Table 4.7: Table excerpt from [61]. Chemical bonds broken by separation or fusing of atoms in Challenge I. The same MD test of 1M steps is run 12 separate times with each model. The possible bond types that can be broken are: 'C-H', 'N-H', 'O-H', 'C-C', 'C-N', and 'C-O'; '-' indicates that no MD was run for the specific model at this temperature. If a bond type is not listed for a specific tasks, all models have performed well for that bond, which was not broken in the termination of the MD run.

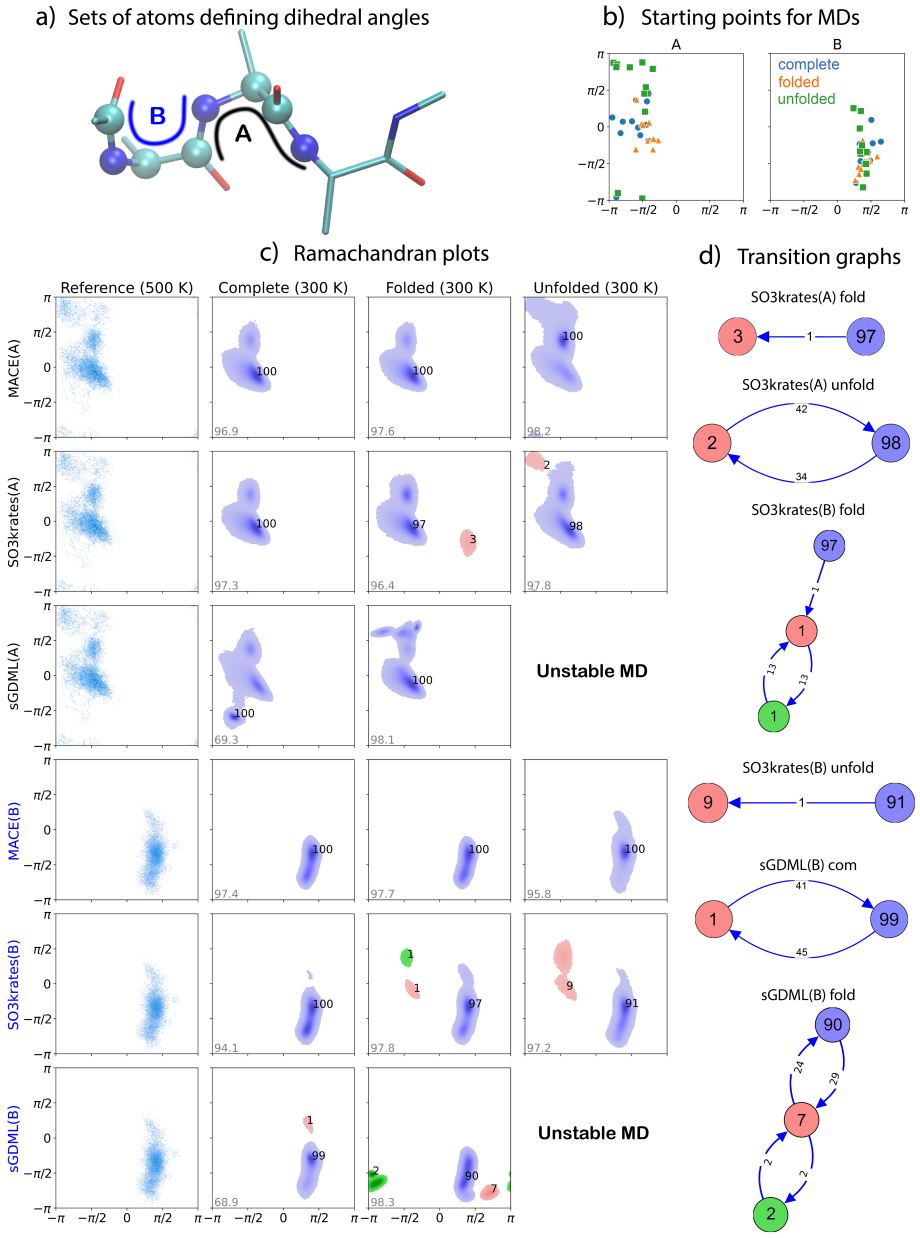

Figure 4.7: Figure from [62]. Challenge I, Alanine tetrapeptide: Analysis and Ramachandran plots for MD simulations at 300 K. **a)** Diagram of the peptide depicts the peptide with atom sets A and B forming two consecutive pairs of dihedral angles (left four atoms - x-axis, right four atoms - y-axis). **b)** Initial points for 12 MD trajectories. **c)** Ramachandran plots for the reference systems at 500 K and for MD simulations at 300 K using MACE, SO3krates, and sGDML MLFFs trained on com (complete), fold (folded), and unfold (unfolded) datasets. The numbers near the clusters indicate their relative population (in percent), while the grey number in the lower left corner of each plot shows the percentage of configurations from the MD trajectories identified as belonging to one of the clusters. **d)** Graphical representation of the transitions between different (meta)stable domains. The values on the arrows show the number of transitions identified in the dynamics.

Figure 4.8: TEA Challenge I: ML predictions of dihedral angles

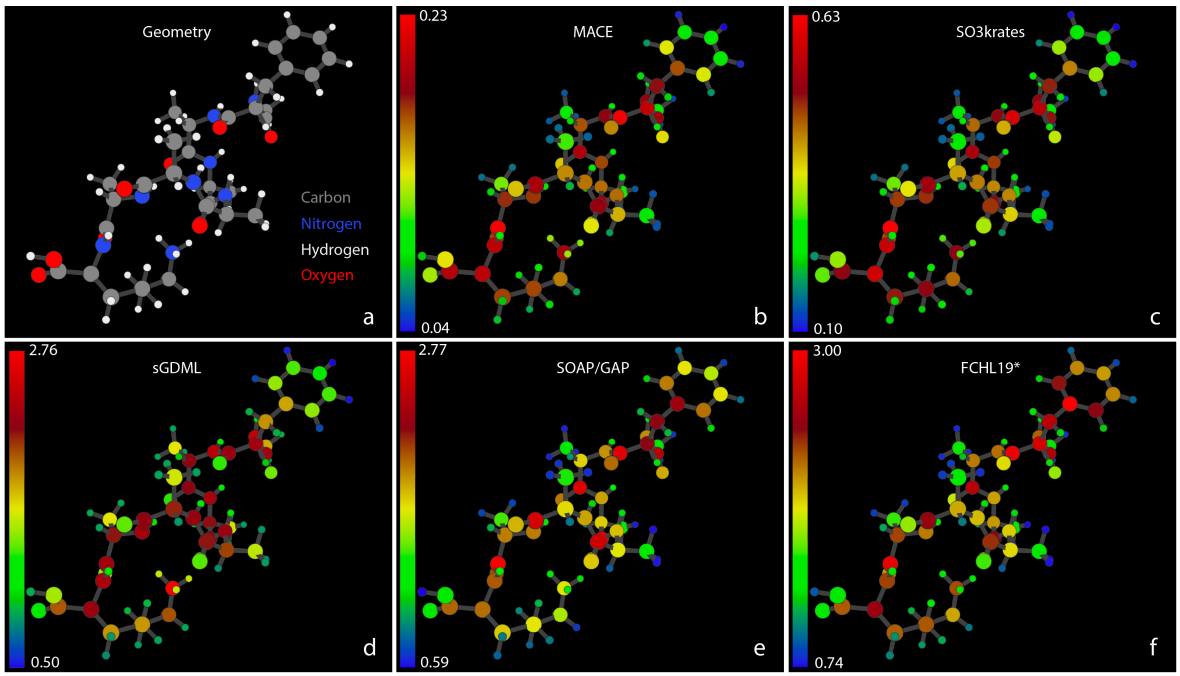

Figure 4.9: Figure from [61]. Maximum Atomic Force Errors for Ac-Phe-Ala5-Lys. Fig. **(a)** shows a snapshot of the system geometry and atom types: Carbons – grey, Nitrogens – blue, Hydrogens – white, and Oxygens – red. Figs. **(b)** through **(d)** show the absolute atomic force errors in kcal/(mol·Å), as indicated by the corresponding colour. Note the different scales for the colour bars, ranging from 12.11 for MACE, 31.52 for SO3krates, to 179.25 for SOAP/GAP. Fig. **(e)** presents the projection of the force acting on hydrogen atoms along the vectors connecting the nitrogen and hydrogen atoms in the R-NH_3 (with maximum prediction errors). These projections are shown as a function of the nitrogen–hydrogen distance. The solid lines represent the force predictions made by the MACE and SO3krates MLFFs, while the dotted lines correspond to the reference results obtained from DFT calculations. The red line depicts the hydrogen atom with the largest force prediction error. At the top of the figure, we also plot the distance between this hydrogen atom and the neighbouring oxygen atom, which is only relevant for the red curve. Fig. **(f)** illustrates the work required to move the hydrogen atom along bond vectors, with the energy minima used as reference zero points. Vertical dashed lines on **(e)** and **(f)** represent the longest NH bond distance found in the training set (grey) and the NH bond distance corresponding to the configuration with the maximum error (black). Blue and green solid curves show the PES scan for the two hydrogen atoms in the R-NH_3 that move away from the molecule as the bond length increases (for comparison purposes).

Task		MACE	SO3	sGDML	SOAP/GAP	FCHL19*
I (com) @300K	C-H	0	0	0	9	22
	N-H	0	0	0	4	0
	O-H	0	0	0	1	0
I (com) @500K	None	12	12	12	-	-
I (com) @700K	C-H	0	3	0	-	-
	N-H	0	2	0	-	-
	O-H	0	2	0	-	-
I (incom) @300K	C-H	0	0	0	7	20
	N-H	0	0	0	2	5
	O-H	0	0	0	0	0
	C-C	0	0	0	5	2
	C-N	0	0	0	0	1
I (incom) @500K	C-H	0	0	0	3	-
	N-H	0	0	0	9	-
	O-H	0	0	0	1	-
	C-N	0	0	0	1	-
I (incom) @700K	C-H	0	21	0	-	-
	N-H	0	9	0	-	-
	C-C	0	15	0	-	-
	C-N	0	7	0	-	-
	C-O	0	2	0	-	-

Table 4.8: Table excerpt from [62]. Chemical bonds broken by stretching or atoms fusing in Challenge II. The same MD test of 1M steps is run 12 separate times with each model. The possible bond types that can be broken are: 'C-H', 'N-H', 'O-H', 'C-C', 'C-N', and 'C-O'; 'None' indicates that no bonds were broken and the MD run successfully completed the set 1M steps; '-' indicates that no MD was run for the specific model at this temperature. If a bond type is not listed for a specific tasks, all models have performed well for that bond, which was not broken in the termination of the MD run. The SOAP/GAP model's results are for the best-performing 'heavy' model for this task.

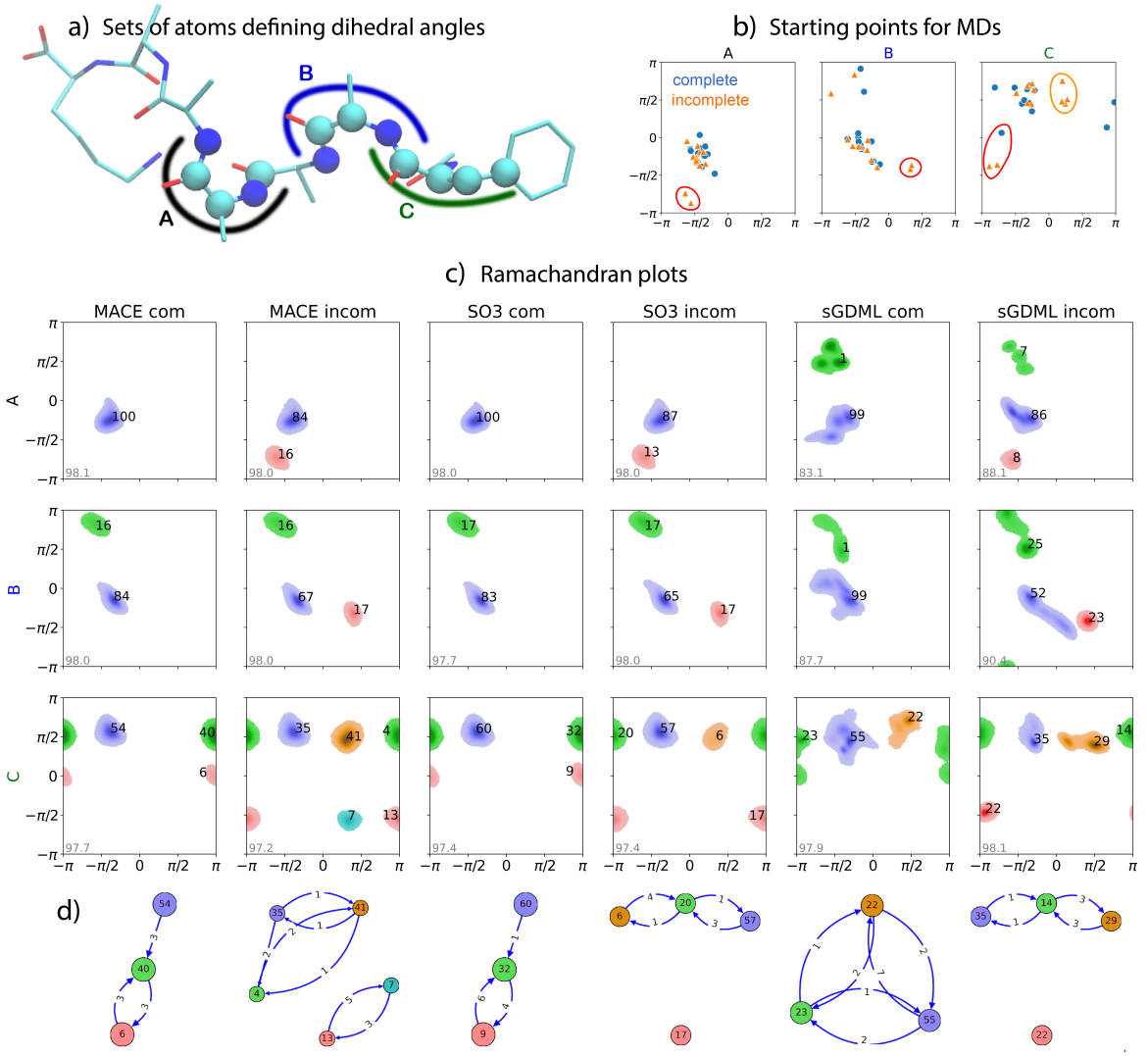

Figure 4.10: Figure from [62]. Challenge II, N-acetylphenylalanyl-pentaalanyl-lysine peptide: Analysis and Ramachandran plots for MD simulations at 300 K. **a)** Diagram of the peptide with sets of atoms A, B, and C forming three consecutive pairs of dihedral angles (right four atoms - x-axis, left four atoms - y-axis). **b)** Initial points for 12 MD trajectories. **c)** Ramachandran plots for MD simulations employed different MLFF models: MACE com (complete), MACE incom (incomplete), SO3 com, SO3 incom, sGDML com, and sGDML incom. The numbers near the clusters indicate their relative population (in percent). The grey number in the bottom left corner of each plot shows the relative number of configurations (in percent) from the MD trajectories identified as belonging to one of the clusters. **d)** Graphical representation of transitions between different (meta)stable domains for dihedrals C. The values on the arrows show the number of transitions identified in the dynamics.

Chapter 5

Advancing Density Functional Tight-Binding method for Organic Molecular Systems through Equivariant Neural Networks

---

Parts of this chapter have been published in this or similar form in:

L. Medrano Sandonas, **M. Puleva**, R. Parra Payano, M. Stöhr, G. Cuniberti, and A. Tkatchenko, "Advancing Density Functional Tight-Binding method for Large Organic Molecules through Equivariant Neural Networks", ChemRxiv, 10.26434/chemrxiv-2025-z3mhh, 2025

and have been produced in collaboration with the above authors. I contributed with data curation, performing computational experiments, methodology, formal analysis, and visualisation.

The applicability domains of different computational methods in terms of the size of the (bio)molecular systems they can model is ever expanding with improvements in the current hardware capabilities. However, as outlined in the introductory Chapter 1, the computational approaches currently capable of accessing the size domain of proteins and large biomolecular complexes are classical FF, semi-empirical approaches, and increasingly so, MLFFs. In physical models, the ability to describe larger systems generally comes at the expense of reduced accuracy – an example of this is shown in Chapter 3 with the prediction of the interaction energy of the QUID dimers, as proxy models for a ligand interacting with a protein pocket. For two representative models for both classical FFs and semi-empirical methods, including DFTB3+MBD [64, 65], an underestimation of the strength of the non-covalent interactions is found for the QUID systems due to the lack of, or a significant approximation of, QM effects On the other hand, the TEA Challenge presented a snapshot of the capabilities of the MLFFs in 2023-2024 in the previous Chapter 5, which already

display excellent results in modelling organic molecules like the Alanine tetrapeptide and N-acetylphenylalanyl-pentaalanyl-lysine peptide with equivariant NNs, further improved in more recent studies [168, 278]. Due to the proven advantages of ML models, the idea of improved parametrisation of semi-empirical methods through better optimisation [279] or augmentation through ML methods has been explored in the past few years [280–294].

Exploring on these ideas, the research presented in this chapter is based on the desire to improve upon the investigations reported in a previous work work, J. Phys. Chem. Lett. 11, 16 (2021) [289], which outlined the use of ML models to improve upon the most complex part of the semi-empirical DFTB method, ie. the repulsive component to the energy. In this work, the repulsive components of DFTB method were parameterised with a many-body Δ_{TB} potentials instead of the standard pairwise repulsive potentials $\mathrm{pw}_{\mathrm{rep}}$, where the many-body Δ_{TB} potentials were learned using an invariant NN SchNet [39] to match to a hybrid DFT-PBE0 functional reference. However, while those results are promising, they were contained to the properties prediction of single small organic molecules.

The prowess of the novel physics-inspired equivariant NNs, *e.g.*, SpookyNet [31], Allegro [35], and MACE [33, 295, 296], and their promising methodology in the earlier work motivated the creation of the "Equivariant networks for Delta Tight-Binding" (EquiDTB) framework with the goal of finding optimal synergies between the semi-empirical and ML methods and thus achieving scalable and transferable ML-corrected DFTB models. Indeed, encouraging results were obtained here from tests on larger organic molecules as well as non-covalent molecular dimers beyond the scope of the CCS covered by the reference data. The ML-corrected DFTB models were tested extensively on different experiments informing their scope in modelling organic systems *e.g.*, predicting the interaction energies and atomic forces of equilibrium and non-equilibrium molecular dimers, determining the minimum energy path between isomers and analysing structural transitions during dynamical simulations. Indeed, the EquiDTB-produced models were compared against both unaltered tight-binding approaches DFTB and GFN2-xTB [126] and reference ML potentials to demonstrate the utility of the proposed methodology for the aforementioned molecular properties.

5.1 EquiDFTB methodology

The flowchart for the EquiDTB method used in this work is shown in Figure 5.1.

5.1.1 EquiDTB: A hybrid ML/DFTB framework

To improve the accuracy and transferability of the DFTB method, the standard pairwise repulsive potential ($\mathrm{pw}_{\mathrm{rep}}$) is replaced by a ML-based many-body potential (denoted as Δ_{TB} potentials) for the prediction of the target energies ΔE_{TB} and atomic forces $\Delta\mathbf{F}_{\mathrm{TB}}$ as follows

$$\Delta E_{\mathrm{TB}} = E^{(\mathrm{at})}_{\mathrm{DFT}} - E^{(\mathrm{el,at})}_{\mathrm{DFTB}} \qquad \text{and} \qquad \Delta\mathbf{F}_{\mathrm{TB}} = \mathbf{F}_{\mathrm{DFT}} - \mathbf{F}^{(\mathrm{el})}_{\mathrm{DFTB}} \ , \tag{5.1}$$

where $E^{(\mathrm{at})}_{\mathrm{DFT}}$ and $\mathbf{F}_{\mathrm{DFT}}$ are the DFT atomisation energy and the DFT atomic forces of a given molecular system, whereas $E^{(\mathrm{el,at})}_{\mathrm{DFTB}}$ and $\mathbf{F}^{(\mathrm{el})}_{\mathrm{DFTB}}$ are the DFTB electronic atomisation energy and atomic forces, respectively. The electronic components are obtained from the DFTB3 [127–129] method using `3ob` parameters [132, 133]. The parameterisation of Δ_{TB} potentials with ML methods represents a complex multidimensional fitting problem to reproduce DFT reference results, which renders it the most challenging step in the development of our methodology. Accordingly, building upon the previous work [289], equivariant NNs are used here to improve the performance of Δ_{TB} potentials to meet the so-called EAST requirements: enhancing the efficiency, accuracy, scalability, and transferability [297]. These NNs ensure that quantities like energies and forces transform correctly under 3D rotations, translations, and permutations, thereby preserving physical consistency. The EquiDTB framework currently involves the equivariant message-passing NNs `SpookyNet` [31] (SP), `Allegro` (AG) [298], and `MACE` (MC) [24] (see details on their architectures in Chapter 2), and can be extended to consider others. The models were chosen as they differ in their approaches to chemical environment representation and interactions modelling. The design diversity is crucial to determine the most reliable EquiDTB model for accurate prediction of ΔE_{TB} and $\Delta\mathbf{F}_{\mathrm{TB}}$. After constructing the Δ_{TB} potentials, the predicted ΔE_{TB} and $\Delta\mathbf{F}_{\mathrm{TB}}$ values are added to the DFTB electronic components computed with the `DFTB+` code [299], yielding the final ML-corrected energies and forces (see Figure E.2(a)). Both contributions are integrated into a single calculator instance *via* a locally modified QM/ML calculator within the ASE package [300], employed to perform all the computational tasks in this work. The EquiDTB results are compared against the pure tight binding methods of two of the most popular semi-empirical choices for organic molecules - the DFTB3+MBD and GFN2-xTB (including the D4 [301] dispersion contribution directly) methods. Additionally, to further highlight the importance of developing Δ_{TB} potentials, we compare their performance with that of a reference ML potential (referred to as rMLP potential throughout the text), which was trained using `MACE` architecture on absolute PBE0+MBD energies and forces for approximately 500 k conformations extracted from the QM7-X and DES15K datasets.

In the choice of training data, a driving question was how the inclusion of non-covalent systems influences the development of the Δ_{TB} potential. In doing so, equilibrium and non-equilibrium conformations of both small single molecules and molecular dimers (only including C, N, O, and H atoms) were considered, as extracted from QM7-X [202] and DES15k [302] datasets, respectively. QM7-X dataset contains property data of approximately 4 M small drug-like molecules of up to 7 heavy atoms computed through PBE0 hybrid functional [99, 303] in conjunction with the tightly converged numeric atom-centered basis sets [304]. The DES15K dimer dataset involved property data for $7,565$ small molecular dimers calculated using a different QM method, so new single-point calculations were carried out here at the PBE0 level to homogenise the reference data. Initially, the developed Δ_{TB} potentials were obtained using only QM7-X molecules, for which we add the label '1' after the name of the NN model, *e.g.*, `SP1` for the `SpookyNet` model. For Δ_{TB} potentials trained on both datasets, we will add the label '2', *e.g.*, for the `SpookyNet` model, it would

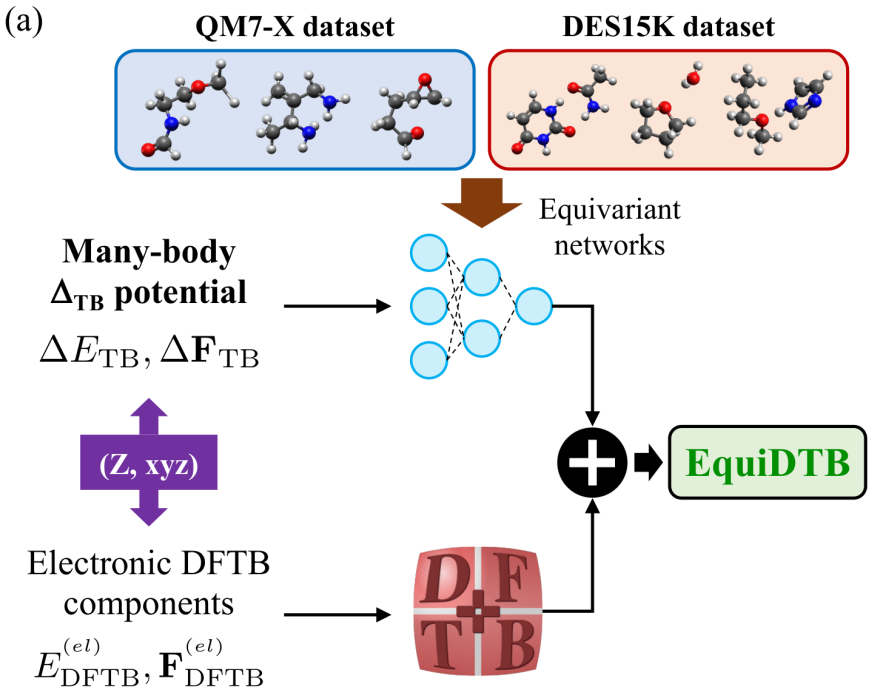

Figure 5.1: Figure from [63]. Schematic representation of the EquiDTB framework (which stands for "Equivariant networks for Delta Tight-Binding") used in this work. Many-body Δ_{TB} potentials $(\Delta E_{TB}, \Delta \mathbf{F}_{TB})$ are developed using SE(3)-equivariant neural networks (ENN) to replace the standard pairwise DFTB repulsive potentials (pw_{rep}). Quantum-mechanical datasets of small single molecules (QM7-X) and molecular dimers (DES15k) were used to train these potentials. The reference level of theory for these datasets is PBE0 hybrid functional.

be `SP2`. The notations for the NNs used are `SpookyNet` (SP), `MACE` (MC), and `Allegro` (AG).

5.1.2 Examination of the Δ_{TB} potentials

The performance of the Δ_{TB} potentials is initially assessed by predicting the corrections in energy ΔE_{TB} and atomic force $\Delta \mathbf{F}_{TB}$ of the equilibrium and non-equilibrium conformations of both QM7-X single molecules and DES15K molecular dimers. Figure 5.2 depicts the boxplots of the errors for ΔE_{TB} and $\Delta \mathbf{F}_{TB}$ computed using the Δ_{TB} potentials trained with SpookyNet, MACE, and Allegro for QM7-X and DES15K in addition to QM7-X datasets. For comparison, the results obtained with our previously developed NN_{rep} model are also presented. One can see that the inclusion of equivariant NNs for property prediction of QM7-X molecules–independent of the training datasets–results in a median of the error distribution of ΔE_{TB} that is closer to zero, along with a reduced data spread. This finding is verified by calculating the MAEs, see Table 5.1. For the prediction of $\Delta \mathbf{F}_{TB}$, the `Allegro` and `MACE` models displayed the best performance on QM7-X molecules, with `AG1` and `AG2` yielding the

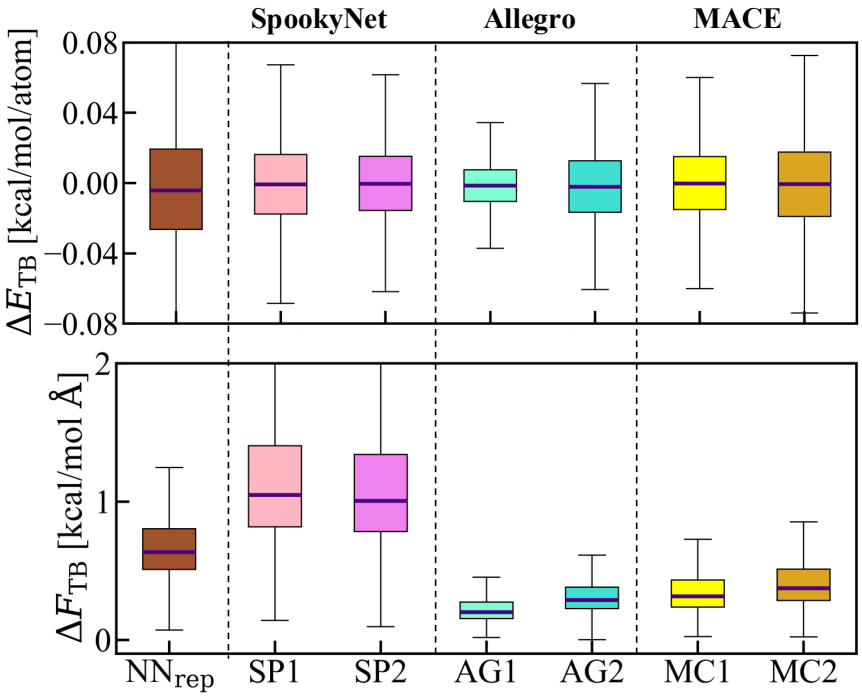

Figure 5.2: Figure from [63]. Boxplot of the error in predicting the correction for energies (ΔE_{TB}) and atomic forces ($\Delta \mathbf{F}_{\mathrm{TB}}$) in QM7-X molecules, computed using Δ_{TB} potentials trained with the SpookyNet (SP), Allegro (AG), and MACE (MC) architectures. Results are shown for models trained on QM7-X (label '1') and both datasets (label '2').

lowest MAE values. Interestingly, the models trained with these equivariant NNs on both reference datasets exhibit slightly higher MAE values than those trained solely on QM7-X. Moreover, although training on both datasets reduces the errors in predicting ΔE_{TB} and $\Delta \mathbf{F}_{\mathrm{TB}}$ for small molecular dimers (see Table 5.1), the models still present only moderate performance, as evidenced by large MAE values and considerable data spread. It is worth noting that the `Allegro` models are the most accurate on both datasets, a result that can be attributed to the strict locality feature in their design. These results highlight the challenge that equivariant NNs pose for simultaneously learning ΔE_{TB} and $\Delta \mathbf{F}_{\mathrm{TB}}$ of covalently- and non-covalently bonded systems. Accordingly, `SP2`, `AG2`, and `MC1` models were selected as representative equivariant Δ_{TB} potentials for further investigation.

The comparison in performance of different equivariant NN architectures for the parameterisation of Δ_{TB} potentials is not a trivial task. To conduct a fair examination, we have used the training set size of 500 k conformations as it was used in our previous work where we developed the many-body repulsive potential, $\mathrm{NN_{rep}}$, using SchNet architecture [289] and QM7-X molecules. The validation set comprised 50 k molecular structures, while the remaining structures were used as the test set. Training samples were randomly selected. Moreover, the default hyperparameters (or alternatives suggested by developers) for each ENN were considered. We only set the number of features to 128 and the cut-off radius to

Model	Δ_{TB} potential	QM7-X		DES15K	
		ΔE_{TB}	$\Delta \mathbf{F}_{\mathrm{TB}}$	ΔE_{TB}	$\Delta \mathbf{F}_{\mathrm{TB}}$
$\mathrm{NN_{rep}}$	SN1	0.031	0.70	0.874	5.76
EquiDTB	SP1	0.025	1.24	0.863	6.78
	AG1	0.013	0.24	0.863	5.31
	MC1	0.023	0.38	0.852	5.12
	SP2	0.022	1.17	0.017	6.12
	AG2	0.020	0.34	0.094	0.71
	MC2	0.028	0.44	0.198	2.18

Table 5.1: Table from [63]. Performance of Δ_{TB} potentials in predicting the corrections in energies ΔE_{TB} and in atomic forces $\Delta \mathbf{F}_{\mathrm{TB}}$ for molecules from QM7-X and DES15K datasets. The MAEs are reported for our previously developed model $\mathrm{NN_{rep}}$[289], trained using SchNet (SN), as well as for the new EquiDTB models trained using equivariant NNs such as SpookyNet (SP), Allegro (AG), and MACE (MC). The number following the name of each NN architecture indicates whether the model was trained on QM7-X only (label '1') or on both datasets (label '2'). Errors for energies and atomic forces are given in kcal/mol/atom and kcal/mol·Å, respectively.

5 Å. More details about the hyperparameters employed in the training procedure can be found in Section

To thoroughly evaluate the equivariant Δ_{TB} potentials, we designed a series of rigorous experiments to assess their scalability and transferability in investigating the properties of larger, more flexible systems, as well as non-covalent complexes (*vide infra*). These experiments require the inclusion of a many-body treatment of vdW/dispersion interactions (MBD) due to its relevance in describing long-range effects that are not properly captured by the PBE0 hybrid functional [121, 122]. To this end, the energy and forces resulting from the MBD formalism have been added to the ML-corrected DFTB energies and forces using `libMBD` package, already implemented in the DFTB+ code. The results obtained using the Δ_{TB} potentials will be compared to those produced by two well-established semi-empirical methods: the DFTB3+MBD method with pairwise repulsive potentials and the GFN2-xTB, which already includes the D4 [301] correction in its design. Additionally, to further highlight the importance of developing Δ_{TB} potentials, we compare their performance with that of a reference ML potential (referred to as rMLP potential throughout the text), which was trained using the `MACE` architecture on absolute PBE0+MBD energies and forces for approximately 500 k conformations extracted from the QM7-X and DES15K datasets.

5.2 Results: single molecules

Let us first outline the experiments for the evaluation of the performance of Δ_{TB} potentials models on single molecules and their rationale. The first one involves the rotational energy

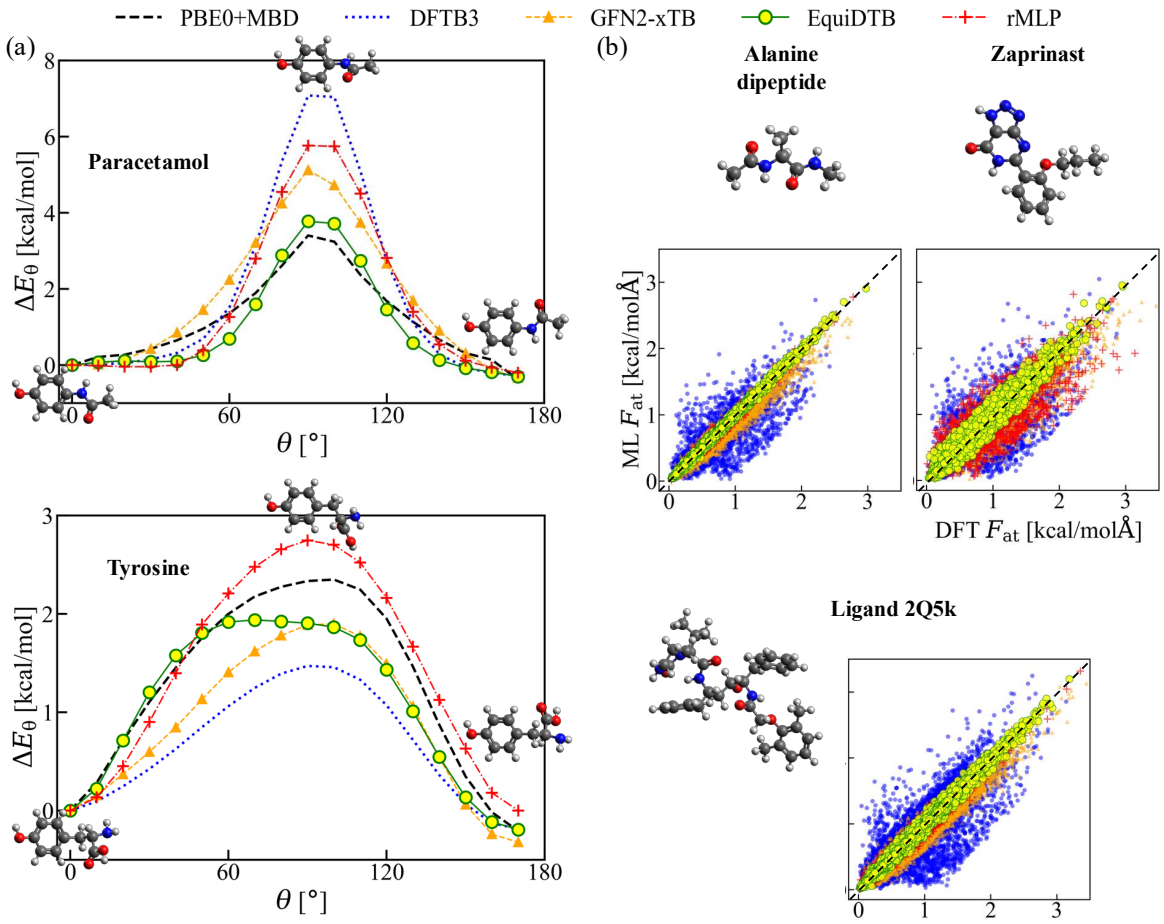

Figure 5.3: Figure from [63]. Assessing predictions of non-equilibrium properties of flexible molecules. (a) Minimum energy path for the rotation of the dihedral connecting the aromatic ring and the linear-type structure in paracetamol and tyrosine. The rotational profiles were computed performing Nudged Elastic Band (NEB) calculations. (b) Analysis of atomic forces F_{at} during an MD trajectory of 100 ps at 300 K for alanine dipeptide, zaprinast, and ligand 2Q5k. We present correlation plots between DFT and ML values of F_{at} for 200 randomly selected conformations sampled during the last 40 ps of the trajectory. In all graphs, we present the results obtained by the widely used TB methods (DFTB3 and GFN2-xTB), EquiDTB model, and rMLP potential. Reference values for energies, forces, and vibrational frequencies were calculated using PBE0+MBD. All calculations include the MBD correction, except for xTB, which considers D4 correction.

profiles of paracetamol ($C_8H_9NO_2$) and tyrosine ($C_9H_{11}NO_3$), chosen as a test in the prediction of properties of molecules beyond the scope of the training set (see Figure 5.3(a)). These systems are good candidates for analysing transferability in chemical space and scalability with system size, as they exhibit greater flexibility and complexity compared to QM7-X molecules. Paracetamol and tyrosine contain 11 and 13 heavy atoms, respectively, and their

	system	metric	DFTB3	GFN2-xTB	EquiDTB	rMLP
Rotational profile (ΔE_θ)	Paracetamol	ρ	0.98	0.99	**0.97**	**0.97**
		MAE	1.05	0.67	**0.38**	0.84
	Tyrosine	ρ	0.99	0.99	**0.97**	0.98
		MAE	0.64	0.41	**0.23**	0.24
MD trajectory (F_{at})	Alanine dipeptide	ρ	0.81	0.98	**1.00**	**1.00**
		MAE	4.00	1.98	**0.31**	0.43
	Zaprinast	ρ	0.76	0.94	**0.98**	0.93
		MAE	4.71	3.15	**1.06**	2.16
	Ligand 2Q5k	ρ	0.87	0.98	**1.00**	0.99
		MAE	3.24	2.00	**0.43**	0.63

Table 5.2: Table from [63]. Performance of EquiDTB model in computing the relative energies (ΔE_θ) along the minimum energy path between isomers and the atomic forces (F_{at}) of conformations extracted from MD trajectories of unseen flexible molecules. The Pearson correlation factor (ρ) and MAE for ΔE_θ and F_{at} are shown as obtained by the widely used tight-binding methods (DFTB3 and GFN2-xTB), the EquiDTB model, and the rMLP potential. Reference values for relative energies and atomic forces were calculated using PBE0+MBD. All calculations include a MBD treatment, except for xTB, which considers D4 correction. The error values for energies and forces are given in kcal/mol and kcal/Å·mol, respectively.

molecular motifs are absent in the training set. To build the reference data for rotational profiles, NEB calculations (see details on NEB in Chapter 2) at the PBE0+MBD level were carried out to determine the minimum energy path for the rotation of the dihedral connecting the aromatic ring and the linear-type structure on the molecules, *i.e.*, C-C-N-C and C-C-C-C dihedrals for paracetamol and tyrosine, respectively. These calculations were performed using `FHI-aims` calculator implemented in the ASE package. The relative energy of the rotated structures with respect to the initial structure is given by $\Delta E_{\theta_i} = E_{\theta_i} - E_{\theta_0}$, with E_{θ_i} as the energy of the structure at the i^{th} NEB interpolation step corresponding to a dihedral angle θ.

The capabilities of the Δ_{TB} potentials are also examined by exploring the PES of larger molecules with biological relevance. To that end, MD simulations were performed at constant temperature for three molecules with biological relevance of increasing size and flexibility, namely alanine dipeptide (22 atoms), zaprinast (33 atoms), and ligand 2Q5k (94 atoms) (see Figure 5.3(b)). The MD simulations were conducted at 300 K for 100 ps using the Langevin thermostat, as implemented in the ASE package. The simulation timestep was set to 0.5 fs, with a friction coefficient of 2×10^{-3}. All simulations using the DFTB method were supplemented with an MBD treatment. The MD simulations with GFN2-xTB were also performed at 300 K but using a Berendsen thermostat with a timestep of 0.5 fs, available in the current and last version of xTB code. Before starting the MD run, all geometries were initially optimised with the corresponding approach. The optimised geometries ob-

tained with EquiDTB model are shown in Figure 5.3(b). To better understand the efficiency of each approach, the atomic forces $\mathbf{F}_{\text{at}}$ and structures of 200 randomly selected conformations extracted from the last 40 ps of each MD trajectory were considered. Accordingly, single-point calculations at the PBE0+MBD level were carried out using the `FHI-aims` software for each conformation to obtain the reference atomic force data.

5.2.1 Exploring potential energy surfaces of flexible molecules

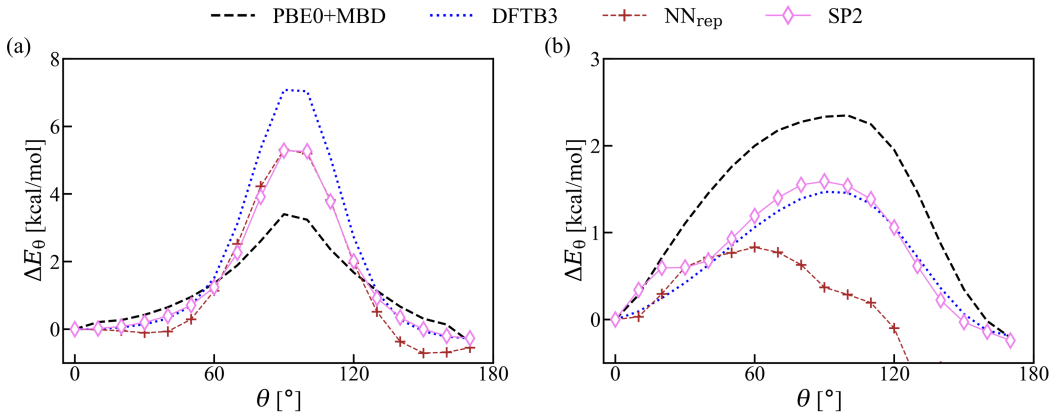

Figure 5.4: Figure from [63]. Assessing predictions of non-equilibrium properties of flexible molecules. Minimum energy path for the rotation of the dihedral connecting the aromatic ring and the linear-type structure in (a) paracetamol and (b) tyrosine. The rotational profiles were computed performing NEB calculations. The results were obtained by the DFTB3 method, NN_{rep} model, and SP2 model. Reference values for energies were calculated using PBE0+MBD. All calculations include a MBD treatment.

The atomic configurations of the selected molecular rotational profiles systems, paracetamol ($C_8H_9NO_2$) and tyrosine ($C_9H_{11}NO_3$), are embedded in the graphs in Figure 5.3(a). In the case of paracetamol, the rotational energy profile is qualitatively well described by all approaches, *i.e.*, the largest ΔE_θ is located around $\theta = 90°$. However, some discrepancies exist in defining this barrier height, with $\Delta E_{90°} = 7.08$ kcal/mol from the DFTB3 method and $\Delta E_{90°} = 3.77$ kcal/mol from the EquiDTB model being the farthest and the closest to the PBE0+MBD reference value (3.4 kcal/mol), respectively. This is also reflected in the evaluation of the Pearson correlation coefficient (ρ) and the MAE for ΔE_θ, see Table 5.2. Overall, ρ values are close to 1.0, indicating an almost perfect linear correlation between predicted and reference ΔE_θ values. The key difference lies in the MAE values, where the EquiDTB model outperforms the other approaches with an error of 0.38 kcal/mol. Similarly, ρ values are close to 1.0 for predicting the rotational energy profile of a more flexible molecule like tyrosine (compared to paracetamol, see additional carboxyl group in its linear molecular building block), with the best performance achieved by the EquiDTB model and the rMLP potential, yielding MAE values of 0.23 kcal/mol and 0.24 kcal/mol, respectively. Notably, for both molecules, GFN2-xTB exhibits higher accuracy than DFTB3, which uses pairwise

repulsive potentials. It is also worth noting that the previously developed NN_{rep} model has a moderate prediction for ΔE_θ values of paracetamol with an MAE of 0.83 kcal/mol, however, it predicts poorly the rotational energy profile for tyrosine with $\rho = 0.64$ and a MAE of 1.26 kcal/mol (see Figure 5.4).

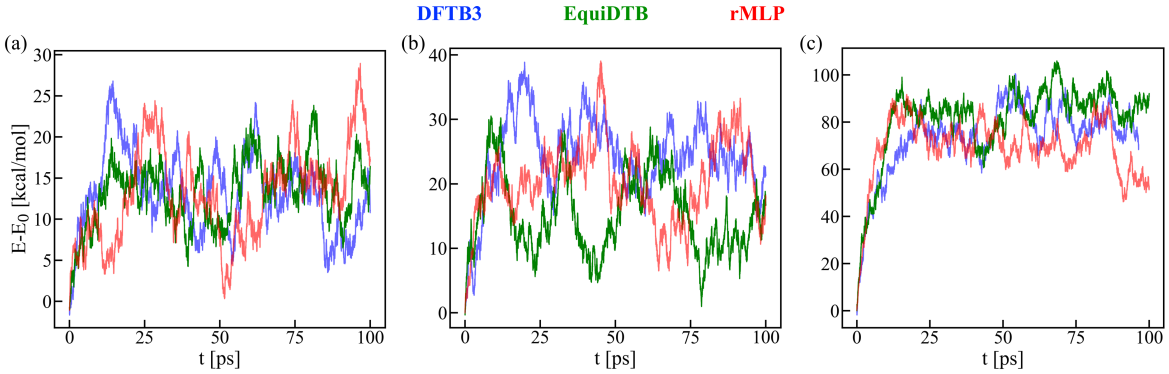

Figure 5.5: Figure from [63]. Variation of the total energy of (a) alanine dipeptide, (b) zaprinast, and (c) ligand 2Q5k as a function of simulation time. Molecular dynamics simulations were performed at 300 K for 100 ps. Results are shown for the DFTB3 method, the EquiDTB model, and the rMLP potential. All calculations include MBD treatment.

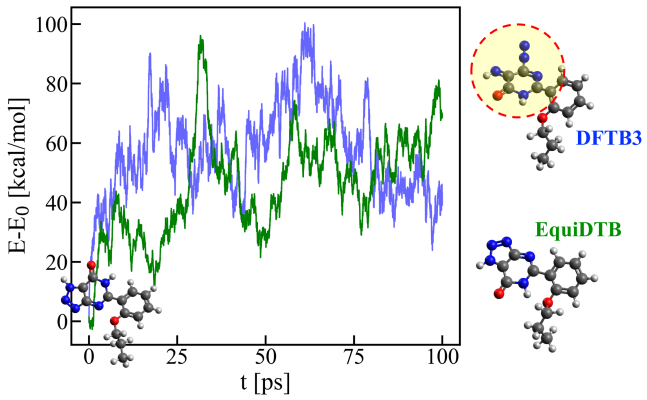

Figure 5.6: Figure from [63]. Variation of the total energy of zaprinast as a function of simulation time at 600 K. Results are shown for the DFTB3 method and the EquiDTB model. Atomistic representations of the molecular structure of zaprinast at 80 ps of simulation are also provided to highlight bond-breaking in the triazole ring observed with the DFTB3 method.

The second experiment of PES exploration of larger molecules of the alanine dipeptide, zaprinast and ligand 2Q5k, is carried out with MD trajectories at 300 K. The analysis of the time evolution of their total energy shows that the MD trajectories are stable for all approaches, with no abrupt energy change due to bond breaking or formation (see Figure 5.5). Based on the correlation plots between true and predicted F_{at} for the 200 selected conformations (see Figure 5.3(b)), it can be determined that the EquiDTB model provides a more

accurate probing of the PES for these molecules. As expected, the rMLP potential shows moderate performance, which is better than that of standard tight binding-based methods. Similar to the analysis performed for ΔE_θ, these observations are quantified by computing ρ and MAE for F_{at} predictions (values listed in Table 5.2). Specifically, we obtained an average ρ of 0.993 over the three molecules for EquiDTB model, followed by 0.973 for rMLP potential and 0.967 for GFN2-xTB. However, when analysing the MAE values, a clear performance difference between EquiDTB and rMLP is found, with the EquiDTB model exhibiting the lowest MAE values for all alanine dipeptide, zaprinast, and ligand 2Q5k. These findings highlight the superior scalability and transferability of the ML-corrected DFTB method compared to an ML potential trained on absolute energies and forces of molecules from the QM7-X and DES15K datasets. Among the studied molecules, zaprinast is the most challenging system for the DFTB method, primarily due to the presence of the aromatic heterocycle triazole. DFTB is known to have certain shortcomings in describing aromaticity and delocalized π-systems, especially in molecules containing N≡N interactions. Accordingly, the present work demonstrated that replacing the pairwise repulsive potential with an equivariant Δ_{TB} potential can considerably improve the description of these electronic configurations, as reflected in the reduction of MAE values from 4.71 kcal/Å·mol with DFTB3 to 1.06 kcal/Å·mol with the EquiDTB model. Additionally, the triazole ring in zaprinast was found to break apart after 80 ps of MD simulation at 600 K using DFTB3 (see Figure 5.6), whereas the equivariant Δ_{TB} potentials help preserve the structure for longer simulation time—another compelling example of the robustness of the proposed methodology.

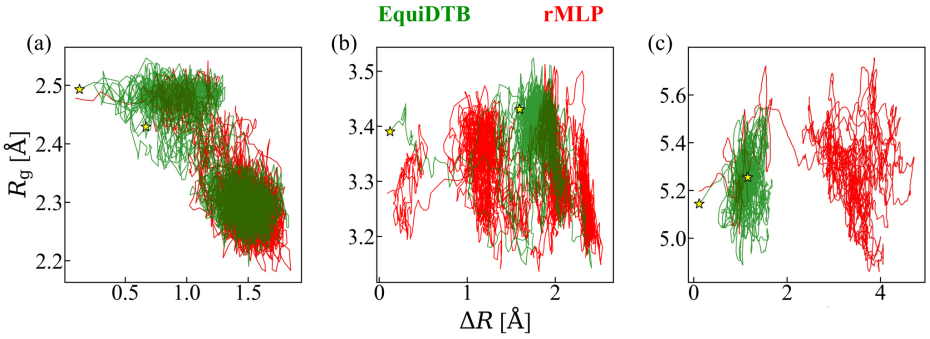

Figure 5.7: Figure from [63]. The structural evolution is evaluated via the two-dimensional space presented as defined by the RMSD w.r.t. the optimised geometry, ΔR, and the radius of gyration, R_{g}. The results are shown for (a) alanine dipeptide, (b) zaprinast, and (c) ligand 2Q5K, obtained using the EquiDTB model and the rMLP potential.

Next, the structural evolution of these molecules during the MD trajectory is evaluated. To this end, the radius of gyration R_{g} of a molecular structure was compared to its RMSD ΔR every 10 fs (see Figure 5.7). ΔR values were computed using the initial optimised structure as a reference for each approach. For alanine dipeptide, the EquiDTB and rMLP models drive the molecule to visit two well-defined structural states: one where the atomic struc-

Model	E_{int}		F_{at}				
	MAE	MARE	H-bond	$\pi-\pi$	London	Mixed	Total
DFTB3	1.04	38.0	5.06	6.32	3.08	3.93	4.52
GFN2-xTB	**0.86**	36.0	5.46	6.22	2.37	6.25	5.20
NN_{rep}	3.13	118.4	2.00	1.88	1.14	1.21	1.57
EquiDTB	0.97	**33.9**	0.70	**0.59**	**0.26**	**0.46**	**0.52**
rMLP	1.19	48.0	**0.65**	0.75	0.30	0.55	0.57

Table 5.3: Table from [63]. Performance of Δ_{TB} potentials in predicting the interaction energies E_{int} and atomic forces F_{at} for equilibrium and non-equilibrium small molecular dimers from S66x8 dataset. We show the MAE and mean absolute relative error (MARE) for the previously developed NN_{rep}[289] model and the best-performing EquiDTB model. For comparison, the error values for widely used TB methods (DFTB3 and GFN2-xTB) and the reference ML potential (rMLP) are also presented. All calculations include a many-body dispersion treatment, except for xTB, which considers D4 correction. The error values for energies and forces are given in kcal/mol and kcal/mol·Å, respectively.

ture is more extended and another where it is compressed. On the other hand, for more complex molecules, the difference between the conformational states gets more pronounced. In particular, for the branched ligand 2Q5k, the MD simulations for both models start from very similar conformations, with an initial $\Delta R = 1.0$ Å between the optimised structures. Nevertheless, their structural evolution differs, reaching a $\Delta R \approx 3.4$ Å between the corresponding structures at 100 ps. Moreover, the conformations for the EquiDTB model exhibit smaller fluctuations in both R_{g} and ΔR. This exposes an area of discrepancy, which needs to be further investigated by an alternatively, and possibly higher-level, method.

5.3 Results: molecular dimers

Finally, let us consider the application of the EquiDTB framework to small non-covalent systems in the s66x8 dataset [54].

5.3.1 Benchmarking S66x8 molecular dimers

The S66x8 dataset contains QM energetic and structural data of 66 small organic molecular dimers at eight different dimer separation distances $q \times d_{eq}$ (with d_{eq} as the equilibrium dimer distance and $q =$ values of 0.90, 0.95, 1.00, 1.05, 1.10, 1.25, 1.50, and 2.00), generating a total of 528 conformations. Since the target level of theory here is PBE0, E_{int} and atomic force F_{at} values for the S66x8 molecular dimers have been recalculated using PBE0+MBD, which serves as reference data. E_{int} is here calculated using the supramolecular approach,

$$E_{\text{int}} = E_{\text{dim}} - (E_{\text{mono1}} + E_{\text{mono2}}) \,, \tag{5.2}$$

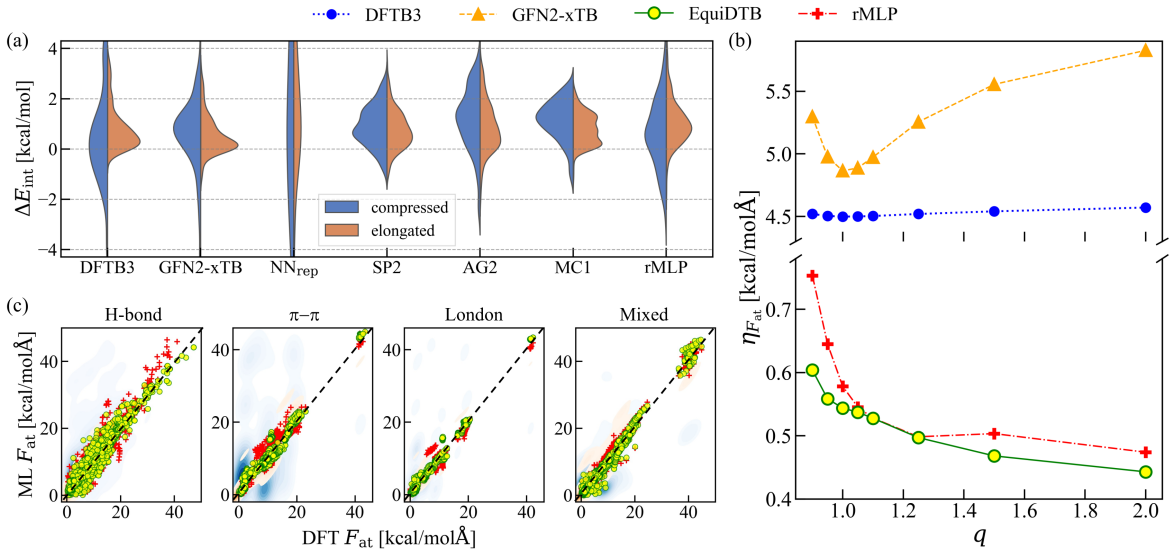

Figure 5.8: Figure from [63]. Benchmarking Δ_{TB} potentials for calculating interaction energies (E_{int}) and atomic forces ($\mathbf{F}_{\text{at}}$) in small molecular dimers from the S66x8 dataset. (a) Distributions of the error in computing E_{int} values for compressed ($q \leq 1.0$) and elongated ($q > 1.0$) molecular dimers. We show the distributions for widely used TB methods (DFTB3 and GFN2-xTB), NN_{rep} model, selected equivariant Δ_{TB} potentials (SP2, AG2, MC1), and the reference ML potential (rMLP). (b) Variation of $\eta_{F_{\text{at}}}$ (see Eq. 5.3) as a function of the relative distance between monomers (q). (c) Correlation plots between DFT and ML F_{at} values for each dimer group. Graphs (b) and (c) primarily show the results for the best-performing EquiDTB model (*i.e.*, MC1 model) and the rMLP potential. For comparison, we also include the corresponding values obtained with DFTB3 and GFN2-xTB. Reference values for E_{int} and $\mathbf{F}_{\text{at}}$ were calculated using PBE0+MBD. All calculations include a many-body dispersion treatment, except for xTB, which considers D4 correction.

where E_{dim} and E_{mono1} (E_{mono2}) are the total energies of the dimer configuration and the monomer 1 (monomer 2), respectively. To better understand the results for molecular dimers, error values were computed for the entire dataset and the four dimer groups, categorized based on their dominant non-covalent interaction, *i.e.*, H-bond (184 confs.), $\pi - \pi$ (80 confs.), London (104 confs.), and Mixed (160 confs.).

5.3.2 Properties of non-covalent systems

Next, the generalisability of the Δ_{TB} potentials is examined by calculating the interaction energy E_{int} of equilibrium and non-equilibrium small molecular dimers from the S66x8 dataset [54], which includes C, H, N, and O atoms. Figure 5.8(a) shows the error in the prediction of E_{int} for the different ML approaches investigated in this work. For comparison, the results obtained with widely used tight-binding (TB) methods (DFTB3 and GFN2-xTB) are also presented. The analysis is split similarly to that in Chapter 3 into compressed

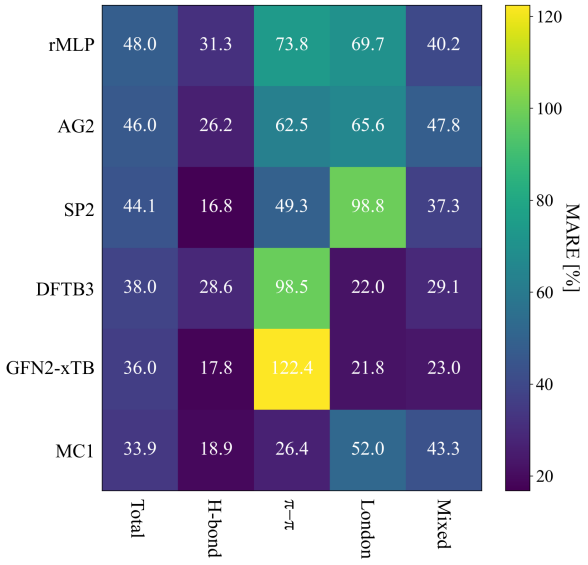

Figure 5.9: Figure from [63]. Benchmarking Δ_{TB} potentials for calculating interaction energies (E_{int}) in small molecular dimers from the S66x8 dataset. Heatmap plot of the mean absolute relative errors (MARE) for each studied model, split according to the predominant non-covalent interaction in the molecular dimer. Reference values for E_{int} were calculated using PBE0+MBD. All calculations include MBD treatment, except for xTB, which considers D4 correction.

($q \leq 1$) and elongated ($q > 1$) dimers to analyse their performance in capturing a diverse range of non-covalent interactions. Overall, the equivariant Δ_{TB} potentials perform better than the $\mathrm{NN}_{\mathrm{rep}}$ model for both compressed and elongated dimers. In particular, `MC1` model produces narrower error distributions than those obtained by standard TB methods, `SP2` and `AG2` models, and the reference ML potential (rMLP). Accordingly, from now on, the `MC1` model will be referred to as the EquiDTB model, as it has shown more consistent performance on single molecules and molecular dimers among the equivariant Δ_{TB} potential. When analysing the mean error values across the entire dataset, listed in Table 5.3, one can observe a slightly superior performance of GFN2-xTB (MAE = 0.86 kcal/mol) compared to EquiDTB model (MAE = 0.97 kcal/mol). A consequence of the more accurate description of the global structural and energetic features of London and mixed dimers when using GFN2-xTB, see Figure 5.9. These results also show that EquiDTB model performs better than rMLP potential in predicting non-covalent properties, demonstrating that combining the electronic DFTB Hamiltonian with an equivariant Δ_{TB} potential offers better generalisability than a ML potential trained on absolute total energies and atomic forces of both single molecules and molecular dimers.

To understand the applicability of the developed Δ_3 potentials in (bio)molecular simulations, the accuracy in computing atomic forces F_{at} of molecular dimers is analysed at dif-

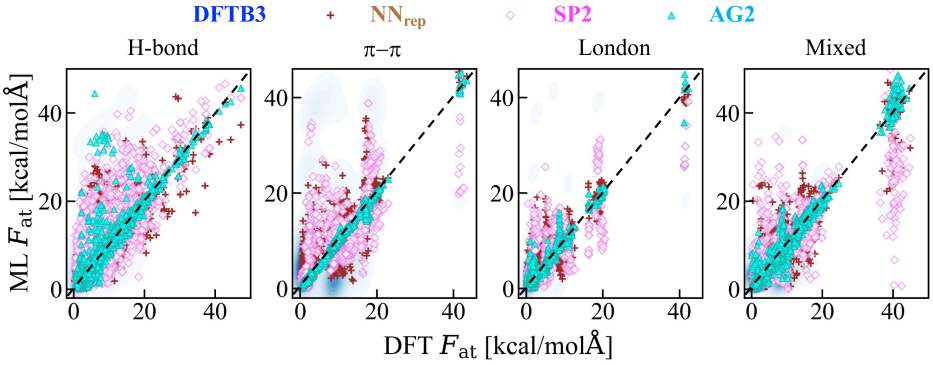

Figure 5.10: Figure from [63]. Benchmarking Δ_{TB} potentials for calculating atomic forces ($\mathbf{F}_{\mathrm{at}}$) in small molecular dimers from the S66x8 dataset. Correlation plots between DFT and ML F_{at} values for each dimer group computed by using the $\mathrm{NN_{rep}}$, SP2 and AG2 models. For comparison, we also include the corresponding values obtained with DFTB3. Reference values for $\mathbf{F}_{\mathrm{at}}$ were calculated using PBE0+MBD. All calculations include a many-body dispersion treatment.

ferent relative distances between monomers, q, by defining the parameter η as follows

$$\eta_{F_{\mathrm{at}}} = \frac{1}{66}\sum_{i=1}^{66}\left(\frac{1}{N_i}\sum_{j=1}^{N_i}|F_{\mathrm{j}}^{X} - F_{\mathrm{j}}^{DFT}|\right) , \tag{5.3}$$

where N_i is the number of atoms in a given dimer i. F_{j}^{DFT} and F_{j}^{X} are the atomic forces computed using the DFT reference method and the X=DFTB3, GFN2-xTB, EquiDTB, and rMLP models. Interestingly, the atomic forces obtained using DFTB3 and GFN2-xTB deviate significantly from the reference values obtained at the PBE0+MBD level, regardless of the q value (see Figure 5.8(b)). On the contrary, EquiDTB and rMLP show close agreement with the reference data, with MAE values for force prediction across the entire dataset of 0.52 and 0.57 kcal/mol·Å, respectively. The error in the forces for these models also decreases as a function of q, highlighting their higher accuracy in investigating the properties of single molecules compared to non-covalent systems. By performing the analysis per dimer group (see Figure 5.8(c) and Table 5.3), EquiDTB outperforms all other models for all dimer groups with the exception of the H-bond dimers, where rMLP exhibits slightly better performance with an MAE of 0.65 kcal/mol·Å. Although the `AG2` model exhibited the lowest errors in predicting F_{at} values for both training datasets, the selected EquiDTB model outperforms it when computing F_{at} values for the S66x8 molecular dimers (see Figure 5.10 and Table 5.4).

Among the different cases in which EquiDTB has improved on the DFTB accuracy, molecular dimers involving ethyne C_2H_2 (which contains C atoms with $F_{at} > 35$ kcal/mol·Å in $\pi-\pi$, London, and mixed dimers) are the most representative. This is because standard TB methods cannot properly describe their triple-bond electronic configuration (see ethyne-based systems in Figure 5.11). This analysis confirms the robustness of the EquiDTB models

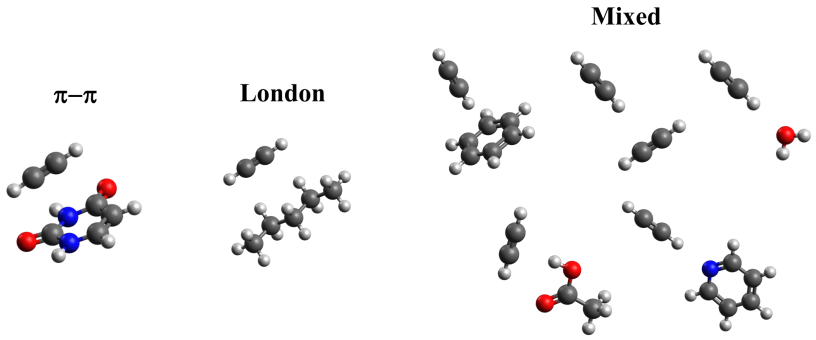

Figure 5.11: Figure from [63]. Ethyne-based molecular dimers, for which the standard tight-binding methods (DFTB3 and GFN2-xTB) failed to accurately compute atomic forces, $\mathbf{F}_{\mathrm{at}}$. Reference values for $\mathbf{F}_{\mathrm{at}}$ were calculated using PBE0+MBD.

Model	E_{int}			F_{at}
	Compressed	Elongated	Total	Total
DFTB3	1.39	0.83	1.04	4.52
GFN2-xTB	1.23	0.64	0.86	5.20
$\mathrm{NN_{rep}}$	3.45	2.94	3.13	1.57
SP2	0.99	0.86	0.91	1.68
AG2	1.45	1.03	1.19	0.55
EquiDTB	1.23	0.81	0.97	0.52
rMLP	1.25	1.16	1.19	0.57

Table 5.4: Figure from [63]. Performance of Δ_{TB} potentials in predicting the interaction energies E_{int} and atomic forces F_{at} for equilibrium and non-equilibrium small molecular dimers in S66x8 dataset. We show the mean absolute errors (MAE) for the $\mathrm{NN_{rep}}$, SP2, AG2, and EquiDTB (*i.e.*, MC1) models. For comparison, the error values for widely used TB methods (DFTB3 and GFN2-xTB) and the reference ML potential (rMLP) are also presented. All calculations consider MBD treatment, except for xTB, which considers D4 correction. The error values for energies and forces are given in kcal/mol and kcal/mol·Å, respectively.

in computing local features that influence the overall structure of non-covalent systems unseen by the model.

5.4 Conclusions

The introduced hybrid semi-empirical-ML framework EquiDTB parametrises many-body Δ_{TB} potentials using physics-inspired equivariant NNs, with the aim of improving the performance of the DFTB method by replacing the standard pairwise repulsive potential.

To this end, a systematic study was conducted integrating modern equivariant NNs with

QM datasets of small molecules, *i.e.*, QM7-X, and molecular dimers, *i.e.*, DES15K, to identify the most reliable and generalisable Δ_{TB} potential for achieving hybrid DFT-PBE0 level accuracy.

The results demonstrated that equivariant Δ_{TB} potentials improve the description of QM interactions with respect to the original DFTB Hamiltonian, and that the new models outperform previous ones based on similar procedures that use the deep tensor NN SchNet [289].

Although all models performed well on the training datasets, some Δ_{TB} potentials exhibited limited scalability and transferability when applied to non-trained or more flexible molecules.

Through a series of analyses on NN choices and both datasets, the best-performing EquiDTB model was identified as the one based on the MACE architecture and trained on small molecules with up to 7 heavy atoms, enabling accurate investigation of properties in larger molecules and molecular dimers.

The capabilities of the final EquiDTB model have been successfully validated through different and rigorous computational tasks involving molecules beyond those in QM7-X and DES15K datasets.

Indeed, the developed ML-corrected DFTB approach can now effectively explore the potential energy surface of more flexible molecules (*e.g.*, tyrosine, zaprinast, ligand 2Q5k) and determine atomic forces and interaction energies for small molecular dimers (*e.g.*, S66x8 dataset) as discussed here; as well as predict the energetic ranking for large drug-like molecules (*e.g.*, AQM molecules) and calculate the vibrational modes of α-amino acids, which is beyond the scope of this thesis but can be found in the pre-print associated with this work [63].

Notably, the EquiDTB model outperforms the widely used tight-binding methods (DFTB3 and GFN2-xTB) and the reference ML potential (rMLP) across the evaluated tasks, especially in atomic forces prediction.

Our findings thus indicate that with an optimal combination of an equivariant NN and QM datasets, the EquiDTB approach can advance the DFTB method to achieve DFT-PBE0 level accuracy with high computational efficiency on diverse simulations.

Although the computational cost increases by approximately a factor of two compared to the standard DFTB method, which is still considerably lower than a DFT calculation. The integration of more efficient large-scale ENN models–such as SO3krates [275, 305], which also include the JAX library [306]–can help mitigate this overhead and facilitate the development of the next generation of NN-enhanced tight-binding methods. New developments in multiple research directions continuously provide avenues for improvement and further research, such as a comparison against the newly released version of xTB, g-xTB tight-binding electronic models [307]. Moreover, the same strategy can be extended to other semi-empirical methods, in which the total energy and atomic forces are decomposed into multiple components (*e.g.*, Neglect of Differential-Diatomic Overlap (NDDO) [308, 309],

Modified Neglect of Diatomic Overlap (MNDO) [310, 311]). Hence, this work provides valuable insights and establishes a concrete framework for integrating and advancing semi-empirical and ML methods toward the development of generalisable and reliable data-driven electronic structure approaches for (bio)molecular simulations.

Chapter 6

Protein-protein interactions as applications outlook

This work was conducted in collaboration with Sergio Suárez Dou and with technical support from Grégory Cordeiro Fonseca. My contribution involved preparation of the protein systems, performing computational experiments, subsequent analysis and visualisation.

Equivariant NNs are key to reliable and stable predictions by ML models for scaling up to viable and accurate simulations of larger biochemical systems (*e.g.*, proteins), as discussed in Chapter 5. In Chapter 3, meanwhile, the limitations to the accuracy of current popular choices for property prediction in interacting large organic systems, like semi-empirical methods or classical FFs, are also highlighted, as well as the difficulty in creating large and reliable datasets of benchmark quality for modelling interacting biomolecular systems.

Thus, building on the goal of predicting the interaction energy of larger interacting molecular systems of biological significance, in the current chapter, the focus is on a particular biochemically relevant application involving protein-protein interactions. In particular, preliminary results on the description of protein-protein interactions via a recently developed MLFF are presented and these results are compared to those obtained via an optimised classical FF. These, similarly to protein-ligand bindings for most drug molecules, consist of non-covalent interactions propagating at a large distance. However, as computationally reliable benchmarks at DFT or higher levels are unobtainable for such systems, experimental data is the only opportunity for verification. Therefore, for biomolecular systems involving protein-protein interaction, experimental observables that correlate with the binding energy are pivotal as means to test of the results produces by models at the cutting edge of research. While such experimental data is expensive and therefore sparse, some proteins of particular significance are well-studied and well-documented, such as the crucially important in recent years SARS-CoV-2 virus [312] spike protein. Thus, it can serve as an appropriate system for testing newly developed MLFFs.

6.1 SARS-CoV-2 virus binding to the host

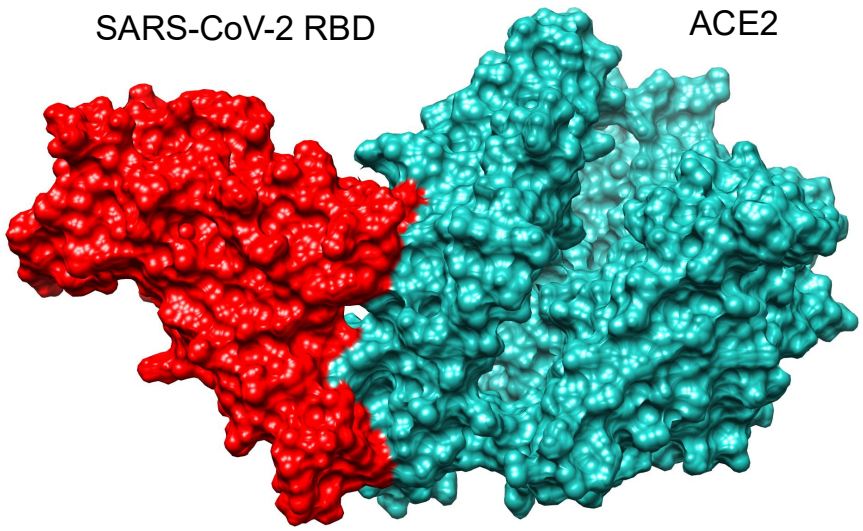

Figure 6.1: SARS-CoV-2 RBD interacting with human ACE2, molecular system from crystal structure 6LZG [313] - the image was produced with the Chimera software [314].

The spike proteins of viruses are responsible for attaching to the host cells and triggering entry, and are of particular interest given their key role in human host viruses of global health concern that employ spike proteins such as HIV-1, Coronaviruses, Influenza, Ebola, Measles, etc [315].

The SARS-CoV-2 (Severe Acute Respiratory Syndrome CoronaVirus 2) virus [313] responsible for the 2020 global pandemic attaches via a spike (S) trimer glycoprotein *i.e.*, consisting of three identical parts called protomers. This is a heavily glycosylated surface protein, with a large count of $\sim$22 specialised polysaccharide groups called glycans attached to each protomer and thus helping the virus evade detection by the immune system. Each protomer has two subunits, of which S1 includes the N-terminal domain (NTD) and receptor-binding domain (RBD). The RBD binds via non-covalent interactions to a human membrane protein called Angiotensin-Converting Enzyme 2 (ACE2). ACE2 is a Zinc metalloprotease, where the Zn^{2+} is a catalytic cofactor needed for the normal enzymatic function of ACE2. Thus, the Zn ion appears in some experimental data for RBD interacting with ACE2, such as the case for the 6LZG crystal structure involving the wild type (WT) of SARS-CoV-2 [312], which is depicted here on Fig. 6.1.

Multiple SARS-CoV-2 variants of concern have emerged since the initial documented WT [312], some of which include mutations linked adversely to public health, *e.g.*, increased transmitability or evasion of natural or vaccine-induced immune response [316–319]. The variants of interest to this chapter, for which pre-computed binding energy data is available [320] at the level of classical FF (obtained from the CHARMM36 FF [140] with the GROMACS

software [321]) are: Alpha, Beta, Delta, Gamma, Omicron, Deltacron, Omni, and P2, with subvariants explored for some of them.

The patterns in the relative binding energies to ACE2 estimated from classical FF MDs for the different variants were compared in literature to clusters in the genomic phylogenetic distances from Nextstrain [322] (based on experimental genomic sequencing data), and a good correlation was found between the evolution of the wide-spread virus variants of interest discussed and the predicted binding energy [320]. Additionally, dissociation curves data obtained at such a classical MMFF level were previously used in literature for comparison of the gas phase dissociation curves of the SARS-CoV-2 RBD with ACE2 employing a MLFF model, also showing promising results between experiment and theory, although at a notably high computational cost [23].Thus, the long-term aim of this project, which is still in its initial stages, is to investigate for which models a similar or better match can be obtained with a state-of-the-art MLFF model. Consequently, the first step is to compare those results to the ones obtained from traditional MMFF methods for binding measures between RBD and ACE2, such as the interaction energy predictions.

A representation of the interacting protein-protein system of the SARS-CoV-2 RBD with the human receptor ACE2 is shown in Fig. 6.1. While this image depicts the WT RBD, it is also representative for the different variants, for which the residue mutations do not result in structurally visible changes at this level of coarse-grained depiction.

6.2 Methodology

Two approaches are utilised here for the prediction of the interaction energy between RBD and ACE2 - a classical FF and a ML model. Based on results of the TEA Challenge 2023 and subsequent improvements in the addition of physics-inspired modules and long-range effects, the ML model SO3LR [168] is chosen based on one of the best performing architectures SO3krates [34], as observed in the results reported in Chapter 4. For the choice of the classical MMFF, a specialised model for protein simulations is picked, namely FF AMBER ff14SB [323].

In the initialisation process of the MLFF, the training data coverage of the relevant chemical compound space is key. Pre-trained and yet unpublished SO3LR [168] models generated in-house were employed here, trained on the QCML dataset [324] generated at PBE0+MBD(-NL) [58–60, 114] level of data. Two types of pre-trained SO3LR models are tested here, and two fine-tuning strategies were applied to each pre-trained MLFF model. The pre-trained models involved architectures using 2 and 3 message-passing layers, respectively. As the training of the 3 message passing layer models requires additional time to reach convergence compared to the 2 message passing layers models, only the latter are presented here. Fine-tuning was performed on all model parameters, as fine-tuning only the head layers resulted in insufficient validation accuracy for the same initial models. In all the cases, the loss function of the fine-tuning involved forces, Hirshfeld ratios, dipole moments, and

energies in a ratio of 10:1:1:0.05. The long-range interaction cut-off applied for all models was 12Å. For the models where the fine-tuning is applied to all hyperparameters, two were obtained per each MLFF to test the variance in the predictions from the same architecture (denoted here as models A and B), which is possible to occur due to the non-deterministic operations in the JAX software [306]. Meanwhile, for the models with 3 message passing layers, different seeds were used during the training of the fine-tuning stage.

The initial models were trained on single precision (float 32), while the fine-tuning was carried out in single and double precision (float 32 and 64), find details in the Results section below. The single-point calculations were carried out in float 64. The 2-layer models A and B, as well as the baseline pre-trained SO3LR model from the GitHub repository [168] were then used to calculate the interaction energy E_{int} with the supramolecular approach (as in Chapter 3.2) between the RBD and ACE2 protein fragments.

The choice of fine-tuning data was based on current extensive ablation studies carried out in-house, and benchmarked on a variety of organic systems, including the presented here QUID dimers dataset [57]. The fine-tuning dataset included at PBE0+MBD level data including SARS-CoV-2 fragments from the WT and mutated types (MT) of the RBD (including with Zn ion cofactor in the ACE2) [23], bottom-up fragments from the GEMS dataset [23] including molecules and peptides, dipeptides from the SPICE dataset [197], as well as at PBE0+MBD-NL [114] level water clusters of 1 to 100 water molecules with Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^- ions (obtained initially from the TIP3P model [325, 326]) from the yet unpublished QCell dataset generated in-house.

A preliminary investigation of the interaction energy for the RBD and ACE2 interactions is presented for one of the aforementioned different SARS-CoV-2 variants obtained in the study [320], namely Alpha. The protein-protein structure was extracted from the solvent (water and ions) for the performed tests. While the interaction interface between the RBD and ACE2 is a complex geometric object due to the curvature and the different distances between the individual residues, a vector broadly capturing the predominant non-covalent bonds direction was capture using the Chimera software [314] to ensure smoothness in the dissociation of the non-covalently interacting proteins. 8 structures were thus generated by moving one of the proteins along the chosen interaction vector by -1, -0.5, $+0.5$, $+1$, $+2$, $+3$, $+4$, $+5$, $+6$, $+7$, $+8$ Å. These structures were then directly calculated with single point calculations with the respective model.

6.3 Preliminary results

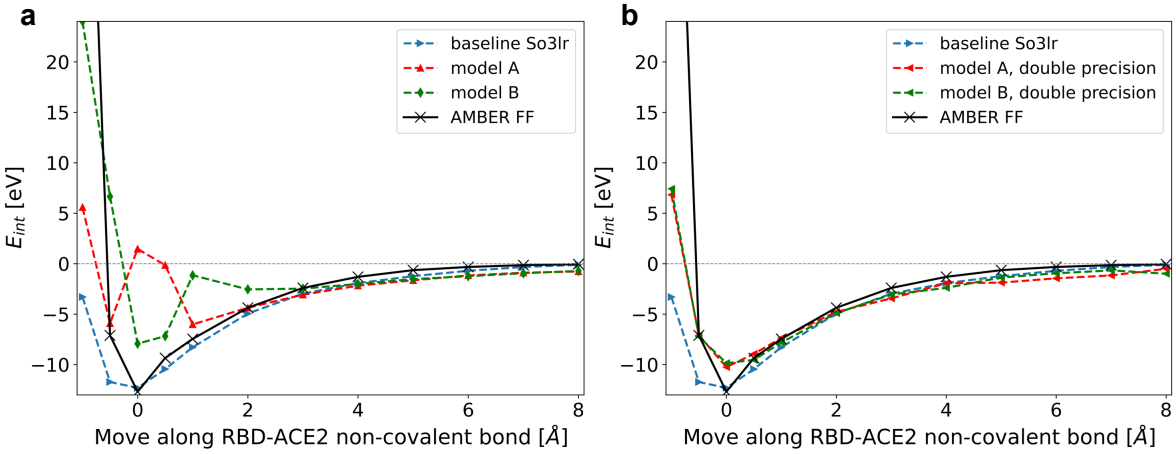

Figure 6.2: Predictions of interaction energies E_{int} along a dissociation of the non-covalent interactions of SARS-CoV-2 RBD with human ACE2. In both panels a and b, the prediction from AMBER classical FF and the baseline So3lr model is shown. In panel a, model A and B denote fine-tuned models with 2 message passing layers trained and fine-tuned on the same setting on single precision. In panel b, the same models are shown fine-tuned using double precision.

In this preliminary investigation, the interaction energy calculated by the ML model is obtained with a single point calculation and compared to the widely used classical FF AMBER ff14SB [323]. Double precision proved to be necessary in the single point calculation step to obtain consistent results from each ML model, with the dissociation curve results are shown in Fig. 6.2.

These results were further reinforced by the experiment with the same settings (including same random seeds) for models A and B. The comparison between the single and double precision indicates that the stability of the results may be influenced by the single precision fine-tuning procedure. It is also notable that the structures were not re-optimised at the ML level and therefore the ML results were not predicted for structures starting from a fully equilibrium point for them. This is relevant as the different models among the classical FF and ML methods may have different equilibrium points for the initial system. However, all models in Fig. 6.2 predict the same minimum and recover curves with the same shape, barring small deviations, mostly in the compressed regime where the RBD and ACE2 are brought together.

6.4 Outlook

From the planned experiments and presented preliminary results, the current work outlook paves the road towards efficient protein-protein interactions predictions with modern MLFFs.

The preliminary results here raise a number of meaningful questions pending for exploration of interacting biomolecular systems. The current work highlights the sensitivity of the So3lr model to single vs double precision operational regimes, not only for obtaining interaction energy predictions from single point calculations, but also within the fine-tuning itself. As the optimal fine-tuning approach is yet to be established, any future framework will necessarily have to consider the appropriate switching point to double precision as well as its scope. Hence, the questions raised here are particularly pertinent given the pivotal role of fine-tuning as an approach in improving the accuracy of predictions and the lower application cost to obtain them. Specifically, the effect on the results of using single vs double precision, or possibly a combination of the two, as well as the training and fine-tuning data, need further exploration. Moreover, while the qualitative agreement of all models and the classical FF model on the protein-protein dissociation curve is a good starting point, assessing the comparative accuracy of the predicted results remains an open question.

To that end, further work to recover the dissociation curves between the RBD spike protein region and the human receptor ACE2 for the different SARS-CoV-2 variants, as well as the clustering of their predicted interaction energies against the experimental data for phylogenetic tree building, would be an interesting area of study at the forefront of MLFF research for biomolecular targets.

Chapter 7
Summary and Outlook

This thesis explores and assesses the building blocks needed for reliable molecular property prediction in the drug discovery pipeline. In particular, we have focused on determining the ingredients necessary to construct accurate Machine Learning techniques that create or enhance models enabling molecular dynamics simulations of proteins including important quantum effects that arise from the nature of the interactions involved. Specifically, the molecular targets of interest are protein-ligand and protein-protein systems due to their pivotal role in *in silico* screenings. Furthermore, biochemical processes, *e.g.*, ligand docking to a protein pocket, involve molecules out of equilibrium.

With this final goal in mind, here three particular studies are presented, each addressing a specific question. In the work presented in Chapter 3, the newly developed high fidelity QUantum Interacting Dimer (QUID) framework for modelling ligand-protein pocket motifs is presented. QUID is comprised of large flexible chemically diverse molecular dimers in and out of equilibrium of up to 64 atoms and structures representing different pocket types. A single dimer involves multiple types of non-covalent interactions, governed predominantly by dispersion and electrostatics. As these effects are frequently captured inadequately by molecular mechanics (MM), improvements or alternative methods are needed. To that end, very high fidelity data is needed as a benchmark and this is why the interaction energy of the QUID dimers was obtained at the "gold standard" level of accuracy, namely Coupled Cluster and Quantum Monte Carlo methods, and verified further by contrasting the results. The approximately 0.5 kcal/mol disagreements are small, unlike those found previously for larger interacting systems [234] and correspond to predominantly electrostatic interactions, which match recent findings for the S66 dataset [235]. Among all the explored methods, the Density Functional Theory (DFT) ones (*e.g.*, PBE0+MBD, ωB97X-V, and PBE0+D4) achieve excellent agreement in interaction energy prediction with the more costly LNO-CCSD(T) reference. Additionally, certain limitations in the widely investigated semi-empirical and MM methods used were identified for the complex non-covalent motifs in QUID, which raises questions about those methods' reliability in binding affinity simulations, *e.g.*, in protein pockets with ligands. These results highlight the relevance of determining the appropriate level of theory to accurately characterise protein-ligand sys-

tems, particularly in the development and design of extensive and chemically complex QM datasets utilised in physical method benchmarking and ML-based investigations.

With the aim of exploring reference data's reliability, especially in the context of benchmark data for FF applicable to protein systems, further molecular properties were calculated for the QUID dimers at high level DFT. This allows for the electronic characterisation of chemical environments within ligand-pocket motifs—a common limitation of current benchmarks. The additional electronic properties can be used for verifying ML's characterisation of structure-property relations, as well as for filtering drug candidates. Among the calculated properties, the atomic forces were considered in depth, and in particular their vdW components due to their strong impact on the binding simulation and pose in ligand-pocket systems and their dynamics. Discrepancies were found in vdW components obtained from MBD, D4, and XDM methods, demonstrating the need for further research and reliable benchmarks in this direction.

Beyond a reliable reference with QM accuracy, a robust ML model is also needed for the simulations of interacting biomolecular systems. In the second study reported in this thesis (Chapter 4), the TEA Challenge 2023's Challenges I and II provided a comprehensive evaluation of a representative sample of modern MLFF models' performance for large flexible organic molecules like alanine tetrapeptide and N-acetylphenylalanyl-pentaalanyl-lysine with high level DFT reference. A selection of models, including kernel and neural network (NN) architectures, MACE, SO3krates, sGDML, SOAP/GAP, and FCHL19*, captured the current state of the field in 2023-24. Areas for improvement in the architectures were identified, some of which have already been updated, such as the inclusion of long-range effects. Challenges I and II aimed to asses the limits of MLFFs under various conditions and system complexities for large flexible organic molecules with tests on the prediction of potential energy surfaces and forces. The results indicate significant advancements in the accuracy of MLFFs, particularly with the investigated equivariant NN architectures showing marked improvements in MAE and RMSE for energy and force predictions compared to the kernel-based models. However, the high maximum force errors remain a critical challenge, highlighting the necessity for further robust predictions, particularly across different atom types in a given system, for which the heterogeneous prediction accuracy was found across all the investigated MLFFs.

A stability assessment revealed that the MACE and SO3krates models generally exhibited better stability in MD simulations, completing most trajectories under varying conditions, unlike the kernel-based models, which proved only reliable in well-sampled near-equilibrium regions. A comprehensive comparative analysis of MD simulations conducted under identical conditions found excellent agreement between the MD results of the MACE and SO3krates MLFFs trained on comprehensive datasets for both Challenge I and II, as well as sGDML for Challenge I. Discrepancies were discovered primarily in the transition regions between (meta)stable states or large atomic fluctuations, likely due to the coverage of the chemical compound space region of interest by the training data rather than the ML architecture itself, while the SOAP/GAP and FCHL19*, kernel-based methods were gener-

ally insufficiently stable. Furthermore, throughout common technical pitfalls proved to be another key hurdle and frequently had to be addressed with the help of the MLFF's developers - a list of guidelines for development and use of MLFFs is consequently provided in Chapter 4. Despite the success of MLFFs trained on comprehensive datasets, the MDs in Challenges I and II revealed noticeable artefacts when trained on incomplete datasets for both kernel-based models and NNs. This highlights the importance of reliable, high-quality, and comprehensive training data as a major bottleneck for MLFFs, linking these findings to the need for highly reliable benchmark data outlined in the first project (Chapter 3) as a pre-requisite for the MLFFs' integration into drug development pipelines.

The research direction of the third project presented in this thesis follows logically the findings of the first two projects, namely the need for improved accuracy in semi-empirical, as well as the predictive power of physics-inspired equivariant NN architectures. The hybrid semi-empirical - ML framework EquiDTB was introduced with the results obtained discussed in Chapter 5. The crux of the EquiDTB approach is the parameterisation of many-body Δ_{TB} potentials using the MACE, Allegro, and SpookyNet NNs, with the goal of improving the accuracy of the DFTB method by replacing its pairwise repulsive potential. A systematic study was performed on the reliability and generalisability of the thus-obtained EquiDTB models, where the optimal choice for architecture was explored in combination with the training choice from the QM datasets of small molecules QM7-X and molecular dimers DES15K. The results demonstrated that equivariant Δ_{TB} potentials improve the capture of QM interactions that are not adequately described by DFTB, outperforming the corresponding deep tensor NN Δ_{TB} potentials, especially for atomic force prediction. All models performed well on the training data and some Δ_{TB} potentials demonstrated some scalability and transferability limited to unseen or more flexible molecules. The best-performing model within the EquiDTB framework was found via a series of challenging tasks to be that of the MACE architecture when trained on small molecules, which proved capable of accurate investigation of larger molecules and molecular dimers across multiple properties.

The challenges presented in this thesis involved molecules beyond the training datasets, including potential energy surface exploration for flexible molecules like tyrosine, zaprinast, and ligand 2Q5k; and determining atomic forces and interaction energies for the small molecular dimers in the S66x8 dataset. Notably, the EquiDTB model outperforms the explored widely used semi-empirical methods and the reference ML potential across all tasks. Our findings thus indicate that with an optimal combination of an equivariant NN and QM datasets, the EquiDTB approach can advance the DFTB method to achieve DFT-PBE0 level accuracy with high computational efficiency on diverse simulations.

In conclusion, the complimentary findings of the experiments presented in this thesis build a coherent body of results demonstrating that the future of molecular property prediction and simulation of biomolecular systems rests on physics-informed ML or ML-enhanced models trained on highly reliable and validated as representative training and benchmark data.

However, improvements are still required until computational models can reach their full potential in accelerating the drug development pipeline. Datasets like QUID can serve as a way to ensure the building blocks of an *in silico* pipeline for a more accurate and reliable prediction of the binding affinity by both traditional physical models, as well as ones incorporating ML. Moreover, additional research in this area is required to create more comprehensive benchmarks, which also include charged structures, robust atomic forces, solvation effects, and larger more flexible pocket models to better represent the complex protein pocket-ligand systems. These are especially relevant as the scalability and transferability of ML models is improving, and the target systems can now include proteins of interest such as the Sars-Cov-2 spike protein. While in its initial phase, the results presented in Chapter 6 outline the approach towards a major application, for which data is largely scarce and computationally expensive to obtain. Furthermore, the force predictions of state-of-the-art atom types still need improvements for uniformly reliable homogenous accuracy based on atom types. The semi-empirical methods alternative involving QM effects still require improvement in accuracy and scalability. The framework presented here for ML-enhancement of semi-empirical methods, EquiDTB, is applicable to all those systems, which necessitate component decomposition of total energy and atomic forces.

Overall, the findings in this thesis provide a stable platform for improvements in the aforementioned directions, which form an actionable roadmap toward robust generalisable models for biomolecular simulations.

Appendix A

Ab initio: Coupled Cluster and Quantum Monte Carlo

This appendix offers a short introduction of the applied methods in this thesis: LNO-CCSD(T) (Linear Natural Orbitals (LNO) - Coupled Cluster (CC) with singles, doubles, and perturbative triplets) and Fixed Node Diffusion Monte Carlo (FN-DMC). This is relevant for the *ab initio* calculations in Chapter 3.4, which were performed by Peter Nagy and Balázs D. Lőrincz for Coupled Cluster and Jorge Charry for Quantum Monte Carlo.

A.1 Quantum Monte Carlo

This section is based on general review works [14, 196, 327, 328] and is focused on non-covalent interactions [195].

Monte Carlo (MC) methods are stochastic family of integration techniques that uses repeated random sampling for numerical solutions of deterministic problems, whose complexity is prohibitive for analytical ones. In particular, Quantum Monte Carlo (QMC) [329] is a class of algorithms that apply MC sampling for the solution of the many-body Schrödinger equation for molecular and solid state systems. Among the most common QMC approaches, Diffusion Monte Carlo (DMC) is specifically designed for accurate prediction of ground state properties in a system, and the Fixed Node approximation of DMC (FN-DMC) is the conventional approach used to solve the sign problem for fermionic systems such as electrons in molecules.

Under the premise of DMC, the Schrodinger Eq. 2.1 is transformed into a diffusion equation in imaginary time $\tau = it$ with and energy shift E_R

$$-\frac{\partial \Phi(\mathbf{X}, \tau)}{\partial \tau} = (\hat{H} - E_R)\Phi(\mathbf{X}, \tau), \tag{A.1}$$

so that a general solution is written as the linear combination of decaying components

$$\Phi(\mathbf{X}, \tau) = \sum_j c_j e^{-\tau(E_j - E_R)} \Psi_j(\mathbf{X}), \tag{A.2}$$

where $\mathbf{X}$ is a configuration of the systems degrees of freedom, in this case the electronic coordinates, and Ψ_j and E_j are respectively the set of eigenfunctions and eigenvalues of the time-independent Scrhödinger equation, end E_R is a energy shift used to preserve the norm (usually $E_R \approx E_0$). For a sufficiently long time evolution the only state that survives in the projection will be the one with the lowest eigenvalue, *i.e.* the ground state.

In order to construct the evolution of the wave function in time, the Green's function approach is applied, and discretized through the second order Trotter-Suzuki approximation. To avoid the systematic error of that comes with this last approximation, that introduces a time-step bias due to the choice of finite time step, one usually extrapolates the values of the physical observables computed with DMC to the $\tau \to 0$ limit. Due to computational limitations, for large systems, and especially when computing energy differences, it has been shown to be sufficient to compute only a single estimation with a relative small time-steps of 0.01 Ha - this is also the approach taken in the calculations performed for Chapter 3.4.

As anticipated above, the FN-DMC method employs a trial wavefunction that defines the nodal surface of the many-body wavefunction that remains fixed during the diffusion process, to solve the sign problem.

In general, the exact nodes of many-electron systems are unknown, and approximated wave functions have to be employed [330].

A successful and practical trial wavefunction under the FN approximation is found to be the Slater-Jastrow (SJ) wavefunction $\Psi_T(\mathbf{x}) = F(\mathbf{x})J(\mathbf{x})$, where $F(\mathbf{x})$ is an anti-symmetric function that describes the spatial and spin symmetries of a fermionic wavefunction and $J(\mathbf{x}$ is the Jastrow factor, a symmetric function dependent of the interelectronic distances, which is responsible of introducing electron correlation, and describing the electron-electron and electron-nucleus cusp conditions (the behavior of a wave function at small interparticle distances) [331, 332]. The Jastrow can only modify the amplitudes of the wavefunction, and not the fixed nodes described in the Slater determinant.

The results presented in this thesis are obtained through the FN-DMC implementation of QMeCha. The Slater determinant is obtained from DFT calculations using the molecular orbitals obtained from PBE0 calculations using the ccECP effective pseudopotentials, to substitute the core electrons of the heavier atoms (Chapter 3.4).

A.2 Coupled Cluster

The Coupled Cluster 'golden benchmark' method [333, 334] uses a different paradigm to the methods presented in the Theoretical background. Namely, it employs an exponential

wavefunction ansatz:

$$|\Psi_{\mathrm{CC}}\rangle = e^{\hat{T}}|\Phi_0\rangle, \tag{A.3}$$

where Φ_0 is a reference single Slater determinant, $\hat{T}$ is called a 'cluster operator', which is expressed as a series sum of operators, where all the encoded correlation effects reside. The exponential formalism in coupled cluster theory ensures the method is size-extensive, meaning the energy of a system composed of non-interacting fragments is correctly calculated as the sum of the energies of the individual fragments $e^{\hat{T}_{AB}} = e^{\hat{T}_A} + e^{\hat{T}_B} = e^{\hat{T}_A}e^{\hat{T}_B}$. The operators of increasing rank are given as follows:

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \dots, \tag{A.4}$$

which correspond to singles excitation operator $\hat{T}_1$ (*i.e.*, the operator acting on a ground state to result in the promotion of a single electron up a single energy level), doubles excitation operator $\hat{T}_2$, etc. In the second quantisation form, their mathematical form is given as follows

$$\hat{T}_1 = \sum_{i}^{\mathrm{occ}} \sum_{a}^{\mathrm{virt}} t_i^a \, a_a^\dagger a_i \quad \text{and} \quad \hat{T}_2 = \frac{1}{4} \sum_{i,j}^{\mathrm{occ}} \sum_{a,b}^{\mathrm{virt}} t_{ij}^{ab} \, a_a^\dagger a_b^\dagger a_j a_i, \tag{A.5}$$

where $i, j, k, \dots$ runs over the occupied orbitals in the reference determinant; $a, b, c, \dots$ runs on the virtual (unoccupied) orbitals. The $a_a^\dagger$ and a_i are creation and annihilation operators, respectively.

The exponential operator $e^{\hat{T}}$ may be expanded as a Taylor series

$$e^{\hat{T}} = 1 + \hat{T} + \frac{1}{2!}\hat{T}^2 + \dots = 1 + \hat{T}_1 + \hat{T}_2 + \frac{1}{2}\hat{T}_1^2 + \frac{1}{2}\hat{T}_1\hat{T}_2 + \frac{1}{2}\hat{T}_2\hat{T}_1 + \frac{1}{2}\hat{T}_2^2 + \dots . \tag{A.6}$$

Truncated at second order, a $\hat{T}$ operator can be uniquely defined by the expansion amplitude coefficients t_i^a and t_{ij}^{ab}, which are obtained by solving the CC equations of the energy and amplitudes, a system of polynomial equations. Thus, with the CC ansatz, the Schrödinger equation becomes

$$e^{-\hat{T}} \hat{H} \, e^{\hat{T}} |\Phi_0\rangle = E_{\mathrm{CC}} |\Phi_0\rangle, \tag{A.7}$$

where E_{CC} is the coupled cluster energy. The equation can be re-written with the similarity-transformed Hamiltonian $\bar{H}$ defined as

$$\bar{H} = e^{-\hat{T}} \hat{H} \, e^{\hat{T}}. \tag{A.8}$$

For the CCSD method, the considered projections are limited to single and double excited Slater determinants, *i.e.*, $|\Phi_i^a\rangle = a_a^\dagger a_i \, |\Phi_0\rangle$ and $|\Phi_{ij}^{ab}\rangle = a_a^\dagger a_b^\dagger a_j a_i \, |\Phi_0\rangle$. Given the orthonormalisation conditions, the similarity-transformed Hamiltonian applied to these orbitals produces the singles and doubles amplitude equations

$$0 = \langle \Phi_i^a | \bar{H} | \Phi_0 \rangle \quad \text{and} \quad 0 = \langle \Phi_{ij}^{ab} | \bar{H} | \Phi_0 \rangle. \tag{A.9}$$

These equations are solved iteratively until achieving convergence to find the amplitude coefficients and therefore the CC energy, which also drives the cost of this method. To maintain viability, instead of full tiplets CCSDT, CCSD(T) [335] is the established approach for organic molecules - with a perturbative triplets approximation addition due to the prohibitive scaling (O(N^8) compared to O(N^7) for CCSD(T)). The perturbative triplets approach involves computing non-iteratively an energy correction derived from Møller-Plesset perturbation theory, starting from the converged CCSD solution as the zeroth-order reference wave function.

Appendix B

Extended quantum-mechanical benchmark accuracy to biological ligand-pocket interactions: Computational details

---

B.0.1 Coupled Cluster and Quantum Monte Carlo results

Modern efforts on CC methods focus on achieving the "gold standard" for larger systems that were previously inaccessible. Among them, the local natural orbital (LNO) and domain-based local pair natural orbital (DLPNO) approaches take advantage of the local nature of electron correlation to reduce computational cost without significant loss of accuracy, enabling chemically accurate results for molecules containing hundreds of atoms. The results presented in this work were obtained using the LNO approximation of CCSD(T), LNO-CCSD(T) [240, 336]. For LNO-CCSD(T), starting again from Φ_0, localised (*i.e.*, spatially confined around atoms/bonds) MP2 molecular orbitals (MOs) are built using a localisation algorithm. The orbital space occupied by the system is redefined by creating natural orbitals from a diagonalised density matrix of pairs of orbitals. Using also a reworked perturbative triplets calculation scheme and an auxiliary basis set for speeding up the evaluation of two-electron integrals, the efficiency of the LNO-CCSD(T) allows for the energy prediction of larger organic molecules, including the systems of interest in Chapter 3 of ∼65 atoms. The interaction energy of the molecular dimers is calculated with LNO-CCSD(T) towards the Complete Basis Set (CBS) via basis set and local approximation free (LAF) extrapolation employing the `Normal` and `Tight` and the T–vT use the `Tight` and `very Tight` LNO settings as follows

$$E^{\mathrm{CBS}}_{\mathrm{CCSD(T)}} \approx \bar{E}^{\mathrm{CBS(T,Q)}}_{\mathrm{T-vT\ \ LNO-CCSD(T)}} \approx E^{\mathrm{CBS(D,T)}}_{\mathrm{T-vT\ \ LNO-CCSD(T)}} - E^{\mathrm{CBS(D,T)}}_{\mathrm{Tight\ \ LNO-CCSD(T)}} + E^{\mathrm{CBS(T,Q)}}_{\mathrm{Tight\ \ LNO-CCSD(T)}}, \tag{B.1}$$

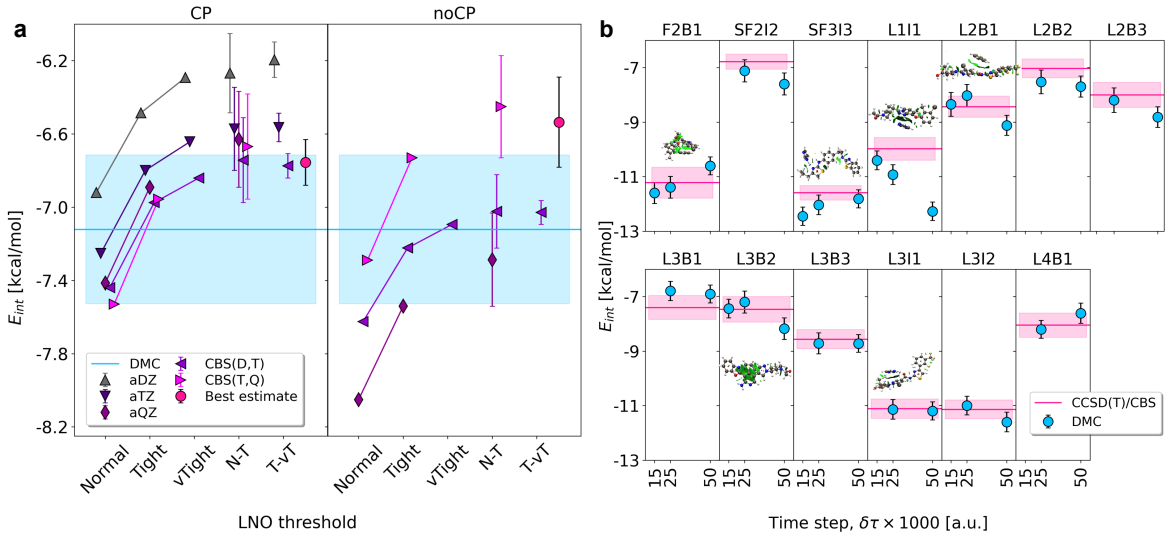

Figure B.1: Mutual agreement for interaction energies from benchmark *ab initio* methods LNO-CCSD(T) and FN-DMC. Figure from [57]. **a** LNO-CCSD(T) interaction energy convergence analysis with respect to the LNO thresholds (x axis) and aug-cc-pV(X+d)Z (aXZ) basis set choices with (CP) and without (noCP) counterpoise corrections for the SF2I2 dimer at equilibrium distance, including the best estimate interaction energy corresponding to Eq. (B.1); the horizontal line indicates the FN-DMC interaction energy ($0.025\ \delta\tau$) with its statistical error in a box. **a** FN-DMC time-step convergence plots for a selection of 13 dimers. The reference LNO-CCSD(T) values are added for comparison, represented as horizontal lines with their uncertainty estimates shown as boxes.

where CBS(D,T) and CBS(T,Q) denote the CBS extrapolation obtained from using the Double-Triple (D,T) and Triple-Quadruple (T,Q) Dunning's basis set aug-cc-pV(X+d)Z (X=D,T,Q) [337]. Moreover, an uncertainty estimate can be assigned to the LNO approximation as the difference between the usually monotonically converging steps, that is, $\pm 0.5|E^{\mathrm{XZ}}_{\mathrm{veryTight}} - E^{\mathrm{XZ}}_{\mathrm{Tight})}|$ for T-vT.

After an analysis on the R_{int} for 9 QUID dimers with the above analysis, a more affordable composite energy expression can be recommended

$$\bar{E}^{\mathrm{CBS(D,T)}}_{\mathrm{N-T\ \ LNO-CCSD(T)}} \approx E^{\mathrm{DZ}}_{\mathrm{N-T\ \ LNO-CCSD(T)}} - E^{\mathrm{DZ}}_{\mathrm{Normal\ \ LNO-CCSD(T)}} + E^{\mathrm{CBS(D,T)}}_{\mathrm{Normal\ \ LNO-CCSD(T)}}, \tag{B.2}$$

which was used to compute reference E_{int} values for all QUID dimers.

The convergence of the methods to obtain robust results is provided in the context of basis set choice and counterpoise (CP) correction in the Figures B.1a and B.2.

Fixed-Node Diffusion Monte Carlo (FN-DMC) calculations [194–196] were performed for QUID dimers corroborating the LNO-CCSD(T) E_{int} with another accurate QM method. The

FN-DMC wave function ansatz was built using a single Slater determinant in addition to a Jastrow factor with one-body, two-body, and 3/4-body terms to account for cusps conditions at the nuclei, fermionic pair correlations, and product of pair correlations at the field of the nuclei, respectively. See references [331, 332] for further details about the electronic wave function ansatz. The molecular orbitals of the Slater determinant were taken from a previous DFT calculation employing ORCA code [338] (version 5.0.4) with the Local-density approximation (LDA) exchange-correlation functional, a cc-pVTZ basis set for all atoms, and the ccECP pseudopotential [339].

All variational parameters of the Jastrow factor were variationally optimized at the Variational Monte Carlo (VMC) level with the stochastic reconfiguration optimization method [340], while molecular orbital coefficients, basis set contraction coefficients, and exponents were kept fixed from the initial DFT calculation. The optimized VMC wave functions were taken as guiding functions in the FN-DMC calculation, in which we also employed the ccECP pseudopotentials to approximate core electrons for each atom, integrated with the determinant localization approximation (DLA) [341]. Consequently, the FN-DMC calculations were computed at two time-steps of 0.050 and 0.025 (a.u.), using 12800 walkers divided into 300 blocks, each 100 steps long, for a total of 4×10^8 sampled configurations. FN-DMC statistical error bars represent one-σ standard error of the mean estimated using binning technique (reblocking) to avoid autocorrelation. For some systems, it was required to run an additional third calculation with a time-step of 0.015 a.u. to get statistical agreement within 1-σ in the observed binding energies. In Fig. B.1 we displayed the time-step convergence against the CCSD(T) reference values. Both VMC wave function optimization and FN-DMC calculations were performed with the QMeCha code [331, 332, 342] (version Dec22, 2024, commit b296fc0).

B.0.2 Basis sets

In this work the used basis sets are def2-QZVPPD, aug-cc-pVTZ, 'tight' in the `FHI-aims` software [217], a highly accurate numerical atom-centered orbitals including polarisation and diffuse functions and devised to reach 'chemical accuracy' *i.e.*, 1 kcal/mol [78].

The highly accurate def2-QZVPPD is from the def2 family [343] *e.g.*, and uses Quadruple Zeta *i.e.*, 4 basis functions per valence orbital, with Valence and two sets of Polarisation functions, and Diffuse functions for electron density far from nucleus. The aug-cc-pVTZ is one of the Dunning's basis sets, where aug stands for augmented with diffuse functions for each atom, cc- correlation consistent ensuring convergence, pV - polarised Valence, TZ - triple zeta). Counterpoise corrections were applied to PBE0+MBD, PBE-QIDH+D3, and CCSD(T) single-point calculations. The basis set superposition error was negligibly small (under 1.5%) for DFT and ca. 4% on the average for CCSD(T) when extrapolated to the complete basis set (CBS) limit (see results in Fig. B.3).

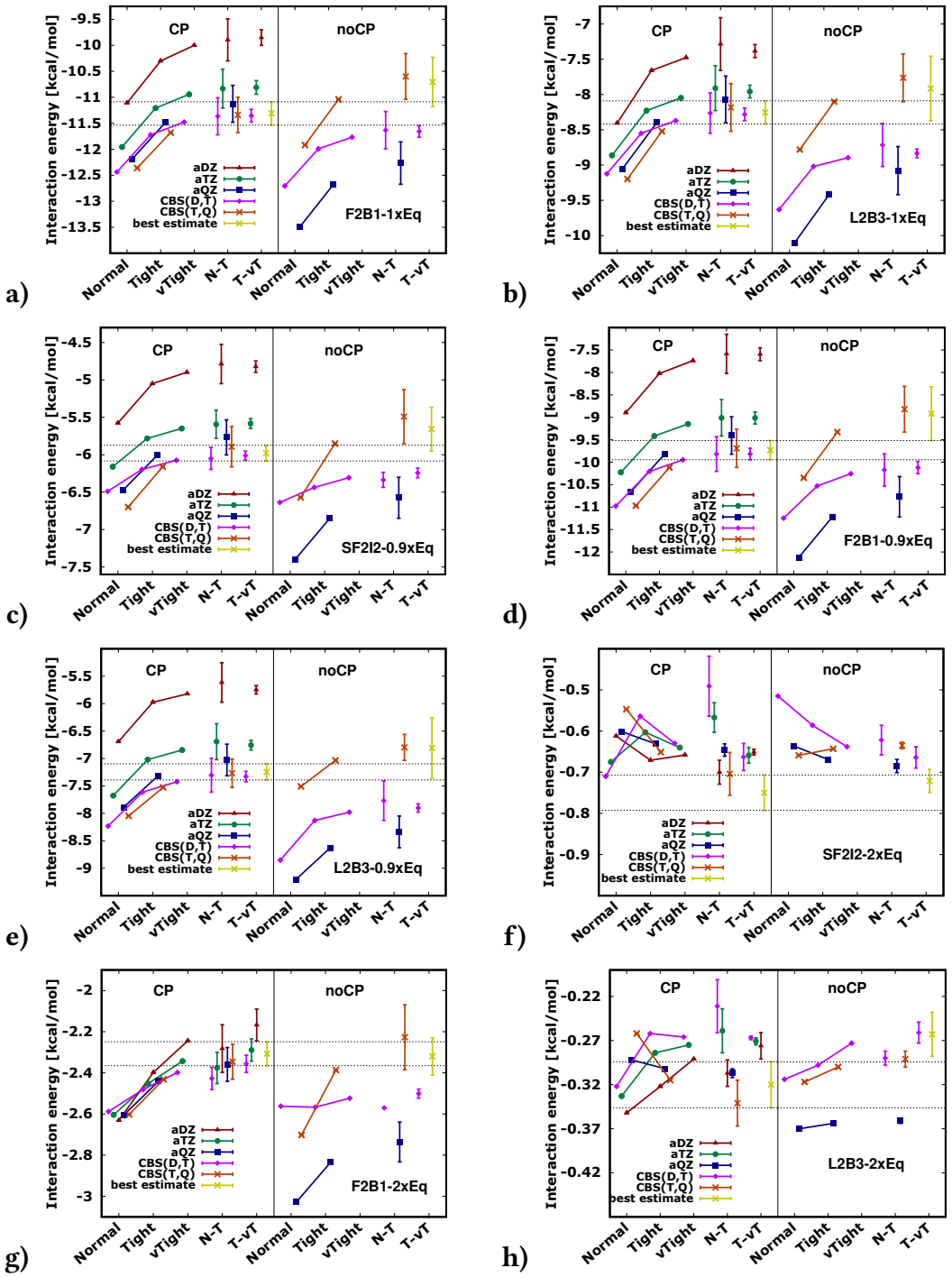

Figure B.2: Figure from [57]. LNO-CCSD(T) interaction energy convergence analysis with respect to the LNO thresholds (x axis) and aug-cc-pV(X+d)Z (aXZ) basis set choices with (CP, left) and without (noCP, right) counterpoise corrections, including the best estimate interaction energy [Eq. B.1]. In separate panels are presented the dimers F2B1 in panel a), L2B3 in panel b), and SF2I2 at equilibrium distance is given in Fig.7 of the main text for 1× the equilibrium geometry intermonomer distance; SF2I2 in panel c), F2B1 in panel d), and L2B3 in panel e) for 0.9× the equilibrium geometry intermonomer distance; and SF2I2 in panel f), F2B1 in panel g), and L2B3 in panel h) for 2.0× the equilibrium geometry intermonomer distance.

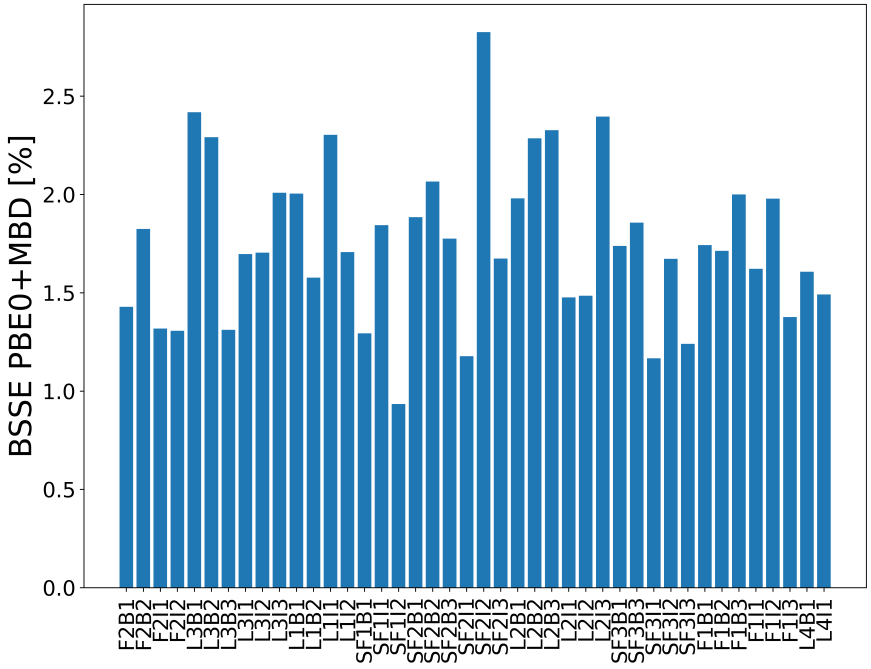

Figure B.3: Figure from [57]. BSSE, Basis set superposition error, (also known as counterpoise correction) for the interaction energies at PBE0+MBD for all 42 equilibrium QUID structures, shown in % by dimer.

B.0.3 Property calculation

The interaction energies E_{int} of QUID dimers were calculated using the supramolecular approach,

$$E_{\text{int}} = E_{dimer} - (E_{L_{monomer}} + E_{S_{monomer}}). \tag{B.3}$$

To investigate the level of agreement among QM methods for calculating E_{int} of QUID dimers, a selection of well-performing hybrid, double hybrid, range-separated hybrids DFT functionals is considered here, including M06-2X [98], ωB97X+D3 [100], ωB97M-V [101] ωB97X-V [218], PBE+MBD [219], PBE-QIDH+D3 [107], B3LYP+D3 [220, 221], CAM-B3LYP+XDM [103] and BH&HLYP+XDM [93]. Additionally, the PBE0 functional was combined with multiple two-body or many-body corrections: MBD [59, 60] (range-separated self-consistent screening (MBD@rsSCS) approach), MBD-NL [114], XDM [109], TS-vdW [111], D4 [113, 118],

ωB97X+D3 [100], ωB97M-V [101]. These calculations were performed using either the `FHI-aims` (version 221103) software [231] with "tight" settings, the `Psi4` software [222, 223] (version 1.9.1) with the quadruple-zeta def2-QZVPPD basis set or the `QChem` software [224] (version 6.1) with the quadruple-zeta def2-QZVPPD basis set except in the CAM-B3LYP and BH&HLYP cases using the XDM dispersion, where the aug-cc-pVTZ basis set was used in accordance with previous studies [246] and the available parametrisation from literature [110]. The PBE-QIDH+D3 implementation made use of the counterpoise correction as needed for the double hybrid functional method using MP2 correlation in the calcula-

tion [107]. All DFT calculations on Gaussian basis set employed the resolution of identity RI technique to accelerate the calculation of electron repulsion integrals. Notably, for the XDM method, a specific parametrisation for the PBE0 functional was applied and then computed using `FHI-aims` software on "tight" settings (a_1=0.4710 and a_2=2.3857) [344]. SAPT energy decomposition calculations were carried out at the sSAPT0/jaDZ level [225] employing the `Psi4` software [222] (version 1.9.1). At the semiempirical level, E_{int} was calculated via single-point calculations using DFTB3+MBD [129] with DFTB+ software [226] and GFN2-xTB with the xTB software [227].

Regarding MM methods, the results for AMBER [141] were obtained using Openbabel [228] for molecular format conversion. The parametrisation with AmberTools and GAFF2 [141] required manual assignment and adjustment of bonds for more complex cases, such as ring interactions, as well as modification of the self-consistent loop limits for the F2I1 dimer. The CHARMM-CGenFF [142] calculations were conducted using OpenMM [229] following a CGenFF2 [230] parametrisation. For these calculations, manual inclusion of the dihedral angles for the flexible side chains, such as the 'C-C-N' type, was necessary. An example is the L4B1 dimer, which was assumed to exhibit relatively low flexibility due to the nature of its bonds and chemical environment.

Additionally, the optimised structures of equilibrium and non-equilibrium QUID dimers were also utilised for more accurate QM single-point calculations using PBE0+MBD level of theory to compute other physicochemical properties (as detailed in Table B.1). For these calculations, the `FHI-aims` code [231] was used together with "tight" settings for basis functions and integration grids. Energies were converged to 10^{-6} eV and the accuracy of the forces was set to 10^{-4} eV/Å. The convergence criteria used during self-consistent field (SCF) optimisations were 10^{-3} eV for the sum of eigenvalues and 10^{-6} electrons/Å^3 for the charge density. The MBD energies and MBD atomic forces were here computed using the range-separated self-consistent screening (rsSCS) approach [60], while the atomic C_6 coefficients, isotropic atomic polarisabilities, molecular C_6 coefficients and molecular polarisabilities (both isotropic and tensor) were obtained *via* the SCS approach [59]. Here, also computed were the van der Waals forces using D4 and XDM methods. Hirshfeld ratios correspond to the Hirshfeld volumes divided by the free atom volumes. In the TS dispersion energy the vdW radii were also obtained using the SCS approach *via* $R_{\text{vdW}} = \left(\alpha^{\text{SCS}}/\alpha^{\text{TS}}\right)^{1/3} R^{\text{TS}}_{\text{vdW}}$, where α^{TS} and $R^{\text{TS}}_{\text{vdW}}$ are the atomic polarisability and vdW radius computed according to the TS scheme, respectively. Atomisation energies were obtained by subtracting the atomic PBE0 energies from the PBE0 total energy of each molecular conformation.

#	Symbol	Property	Unit	Type	Level	HDF5 keys
1	Z	Atomic numbers	-	S	-	'atNUM'
2	R	Atomic positions (coordinates)	Å	S	TB	'atXYZ'
3	E_{at}	Atomization energy	eV	M,G	P0	'eAT'
4	E_{PBE0}	PBE0 energy	eV	M,G	P0	'ePBE0'
5	E_{MBD}	MBD energy	eV	M,G	P0M	'eMBD'
6	E_{TS}	TS dispersion energy	eV	M,G	P0	'eTS'
7	E_{nn}	Nuclear-nuclear repulsion energy	eV	M,G	-	'eNN'
8	E_{kin}	Kinetic energy	eV	M,G	P0	'eKIN'
9	E_{ne}	Nuclear-electron attraction	eV	M,G	P0	'eNE'
10	E_{coul}	Classical coulomb energy (el-el)	eV	M,G	P0	'eEE'
11	E_{xc}	Exchange-correlation energy	eV	M,G	P0	'eXC'
12	E_{x}	Exchange energy	eV	M,G	P0	'eX'
13	E_{c}	Correlation energy	eV	M,G	P0	'eC'
14	E_{xx}	Exact exchange energy	eV	M,G	P0	'eXX'
15	E_{KS}	Sum of Kohn-Sham eigenvalues	eV	M,G	P0	'eKSE'
16	ϵ	Kohn-Sham eigenvalues*	eV	M,G	P0	'KSE'
17	E_{HOMO}	HOMO energy	eV	M,G	P0	'eH'
18	E_{LUMO}	LUMO energy	eV	M,G	P0	'eL'
19	E_{gap}	HOMO-LUMO gap	eV	M,G	P0	'HLgap'
20	D_{s}	Scalar dipole moment	$e \cdot$ Å	M,G	P0	'DIP'
21	D	Dipole moment	$e \cdot$ Å	M,G	P0	'vDIP'
22	Q_{tot}	Total quadrupole moment	$e \cdot$ Å^2	M,G	P0	'vTQ'
23	Q_{ion}	Ionic quadrupole moment	$e \cdot$ Å^2	M,G	P0	'vIQ'
24	Q_{elec}	Electronic quadrupole moment	$e \cdot$ Å^2	M,G	P0	'vEQ'
25	C_6	Molecular C_6 coefficient	$E_h \cdot a_0^3$	M,R	P0M	'mC6'
26	α_{s}	Molecular polarizability (isotropic)	a_0^3	M,R	P0M	'mPOL'
27	α	Molecular polarizability tensor	a_0^3	M,R	P0M	'mTPOL'
28	F_{tot}	Total PBE0+MBD atomic forces	eV/Å	A,G	P0M	'totFOR'
29	F_{PBE0}	PBE0 atomic forces	eV/Å	A,G	P0	'pbe0FOR'
30	F_{MBD}	MBD atomic forces	eV/Å	A,G	P0M	'FvdWMBD'
31	F_{D4}	D4 atomic forces	eV/Å	A,G	P0D4	'FvdWD4'
32	F_{XDM}	XDM atomic forces	eV/Å	A,G	P0XDM	'FvdWXDM'
33	V_{H}	Hirshfeld volumes	a_0^3	A,G	P0	'hVOL'
34	V_{ratio}	Hirshfeld ratios	-	A,G	P0	'hRAT'
35	q_{H}	Hirshfeld charges	e	A,G	P0	'hCHG'
36	$D_{\mathrm{H,s}}$	Scalar Hirshfeld dipole moments	$e \cdot a_0$	A,G	P0	'hDIP'
37	D_{H}	Hirshfeld dipole moments	$e \cdot a_0$	A,G	P0	'hVDIP'
38	$\widetilde{C_6}$	Atomic C_6 coefficients	$E_h \cdot a_0^6$	A,R	P0M	'atC6'
39	$\widetilde{\alpha}_{\mathrm{s}}$	Atomic polarisabilities (isotropic)	a_0^3	A,R	P0M	'atPOL'
40	R_{vdW}	vdW radii	a_0	A,R	P0M	'vdwR'

*The number of Kohn-Sham eigenvalues varies for each molecule.

Table B.1: Table from [57]. List of physicochemical properties found in the HDF5 QUID file under the 'properties' key of the types: structural (S), molecular (M), atom-in-a-molecule (A), ground-state (G), and response (R). The levels of theory are indicated as: DFTB3+MBD (TB), PBE0 (P0) +MBD (P0M) /+D4 (P0D4), /+XDM (P0XDM). P0D4 and P0XDM are only used to indicate the vdW force components at the corresponding level. E_h and a_0 are Hartree and Bohr radius, respectively.

Appendix C

Clustering procedure for Ramachandran plots

Parts of this chapter have been published in this or similar form in:

I. Poltavsky, **M. Puleva**, A. Charkin-Gorbulin, G. Fonseca, I. Batatia, N. J. Browning, S. Chmiela, M. Cui, J. T. Frank, S. Heinen, B. Huang, S. Käser, A. Kabylda, D. Khan, C. Müller, A. J. A. Price, K. Riedmiller, K. Töpfer, T. W. Ko, M. Meuwly, M. Rupp, G. Csányi, O. A. von Lilienfeld, J. T. Margraf,u K.-R. Müller, and A. Tkatchenko "Crash testing machine learning force fields for molecules, materials, and interfaces: molecular dynamics in the TEA challenge 2023" *Chemical Science* **16**, 3738-3754, 2025

and have been produced in collaboration with the above authors.

The clustering procedure used in Chapter 5 consists of the following steps:

- All dihedral angles are computed for each fifth configuration in all 12 MD trajectories, dataset **X**.
- Using **X**, a 2D histogram is produced, in which a pair of sequential dihedral angles along the peptide's backbone is taken with one-degree resolution.
- To focus subsequent analysis on the regions with the highest statistical contributions, eliminating transition states and outliers, we identify a threshold population for a bin to be considered statistically relevant. We locate a set of the least populated bins of the 2D histogram, comprising 5% of the histogram's overall population. A quarter of the population of the most populated bin from this set is chosen as a threshold value **t** in the following procedural steps.
- As a further focusing measure, the regions with low statistics are removed from the current analysis by setting the population to zero of all histogram bins visited less than **t** times within the entire dynamics.

  Afterward, within each bin, a normally distributed auxiliary points per every **t** visits

of the bin is generated (with a limit of 4 per bin at most). This creates a relatively small (10k - 15k) auxiliary dataset for dihedral angles, **Y**, preserving the distribution of the original MD trajectory but limiting the maximum density to a viable 4 per degree squared.

- The dataset **Y** is clustered using the DBSCAN algorithm. The DBSCAN neighbourhood radius, ϵ, is equal to 2; the threshold for a core point status is 25 samples in a neighbourhood (i.e. half of the maximum number of auxiliary points within a circle of radius ϵ). The metric is employed such that it accounts for the periodic boundary conditions for both dihedral angles during the clustering. The chosen setup is necessitated by the large size of the original dataset **X**, 2.4 million data points, too many for any density-based clustering algorithm.
- Upon obtaining the DBSCAN clustering result, each bin in the original dihedral histogram is labeled correspondingly. The label is chosen based on majority contribution - for instance, if a bin contains 1 point labeled Cluster 1, and 3 points labeled Cluster 2, it will be considered a part of Cluster 2.
- Once the classification of the data points in the original dataset **X** is completed based on their bin labels, the population of each cluster is computed. The directional transitions between all pairs of clusters is computed and at this step, the transition regions are removed from **X**.
- Iteratively, the pairs of clusters with more than 100 transitions in total in both directions are merged into single clusters. Subsequently, the cluster labels, populations, and information about transitions are updated.
- According to the proposed procedure, over 95% of the original dataset is typically classified as non-outliers. Information about the number of transitions between clusters in all directions is also retained, as well as the numbers of metastable and transition states in dynamics, thus giving a representative picture of the dynamics.

Appendix D

Symmetry-Adapted perturbation Theory

Symmetry-Adapted Perturbation Theory (SAPT) decomposes intermolecular interactions (or intramolecular, between separate molecular fragments) as a sum of individual but interlinked contributions from more intuitive physical [345, 346]. The SAPT interaction energy generally includes the following components: Electrostatics, Exchange, Induction (Polarisation), and Dispersion. Let us define them for simplicity for the case of an interacting dimer consisting of two monomers. The Electrostatics contribution arises from the Coulomb interactions between the charge densities of the monomers, which includes at long-range charge, dipole, quadrupole and higher multipole interactions, and at short range charge penetration due to electron density overlap between the monomers. The Induction contribution (also called polarisation) arises from the polarisation of the monomers due to the permanent electric field of the other, which end up primarily resulting in attractive interaction contributions at long range between a permanent multipole moment on one monomer and an induced multipole moment on the other. The Dispersion contribution arises from the electron correlation effects of the monomers and results in a weakly attractive force binding the monomers together. Lastly, the Exchange contribution arises from the Pauli exclusion principle, which prevents two or more identical electrons from occupying the same orbital, and results in short-range repulsion between the monomers, which can provide quenching of the Induction and Dispersion contributions at those ranges (dropping off exponentially with distance). To obtain those contributions, SAPT is built on a perturbative approach of solving the Schrödinger equation. While many versions of SAPT exist, SAPT(DFT), SAPT(FCI), in this work, the flavour of SAPT used is SAPT0 [347], the simplest many-body SAPT approximation, and the following explanation addresses this particular version.

SAPT builds on a perturbative treatment of the interaction potential between two non-interacting Hamiltonians, starting from a symmetrised version of the standard time- independent non-degenerate Rayleigh–Schrödinger perturbation theory. The symmetrised Rayleigh– Schrödinger (SRS) is needed to enforce the antisymmetric condition for the dimer

wavefunction due to the Pauli principle - this is achieved by introducing a intermolecular antisymmetriser operator $\hat{\mathcal{A}}$ that handles the sign change in the electron exchange between monomers A and B, such that the dimer wavefunction Φ would satisfy $\hat{\mathcal{A}}\,|\Phi\rangle = |\Phi\rangle$. If Φ_A and Φ_B are the antisymmetric monomer wavefunction (for SAPT0 those are the Hartree-Fock (HF) ground state wavefunctions for the pure monomers), then their product Ψ_0 is the 0th order dimer wavefunction that describes the non-interactive dimer. Introducing a multiplicative coupling parameter ζ as a measure of the degree of intermolecular interaction $\hat{V}$, the SAPT Hamiltonian $\hat{H}(\zeta)$ of the system can be written as

$$\hat{H}(\zeta) = \hat{H}_A + \hat{H}_B + \zeta\,\hat{V}, \tag{D.1}$$

with corresponding interaction energy eigenvalue and wavefunction $\hat{H}(\zeta)\,\Psi(\zeta) = E(\zeta)\,\Psi(\zeta)$. Notably, for SAPT0 the H_A and H_B monomer Hamiltonians are given by the Fock operators as the intramolecular electron correlation is neglected and therefore so are the Møller-Plesset (MP2) terms from many-body SAPT. Hence, the name SAPT0 indicates 0 order in MP2, and the index appearing in the interaction energy terms, vide infra. The interaction energy of the system E_{int} can be calculated perturbatively as

$$E_{\text{int}}(\zeta) = \frac{\langle \Psi_0 \,|\, \zeta\hat{V}\,\hat{\mathcal{A}}_{AB} \,|\, \Psi(\zeta)\rangle}{\langle \Psi_0 \,|\, \hat{\mathcal{A}}_{AB} \,|\, \Psi(\zeta)\rangle}. \tag{D.2}$$

Plugging in the above equation a perturbative expansion of the wavefunction w.r.t. the coefficient ζ,

$$|\Psi(\zeta)\rangle \;=\; |\Psi_0\rangle \;+\; \zeta\,|\Psi^{(1)}\rangle \;+\; \zeta^2\,|\Psi^{(2)}\rangle \;+\; \cdots, \tag{D.3}$$

the interaction energy is obtained as a series

$$E_{\text{int}}(\zeta) \;=\; \zeta\,E^{(1)} \;+\; \zeta^2\,E^{(2)} \;+\; \cdots, \tag{D.4}$$

whose elements first-order perturbation in V, $E^{(1)}$, second-order perturbation in V, $E^{(2)}$, etc., which resolve into sums of physically meaningful terms corresponding to interaction effects. $E^{(1)}$ can be decomposed into a sum of the electrostatics between unperturbed charge densities $E^{(10)}_{\text{elst}}$ and the exchange contribution from observing the Pauli principle $E^{(10)}_{\text{exch}}$, where the 0th index denotes the 0 intramolecular contribution from MP2 for SAPT0. $E^{(2)}$, meanwhile, decomposes into second order perturbation effects with induction, exchange, and dispersion contributions as follows:

$$\begin{aligned} E^{(1)} &= E^{(10)}_{\text{elst}} + E^{(10)}_{\text{exch}} \\ E^{(2)} &= E^{(20)}_{\text{ind,resp}} + E^{(20)}_{\text{exch-ind,resp}} + E^{(20)}_{\text{disp}} + E^{(20)}_{\text{exch-disp}}, \end{aligned} \tag{D.5}$$

where the notation 'resp' stands for 'response', *i.e.*, self-consistent relaxation of one monomer's HF virtual orbitals via virtual excitations, in response to the electrostatic potential of the other monomer. Here, $E^{(20)}_{\text{disp}}$ and $E^{(20)}_{\text{exch-disp}}$ are the only terms that includes true electron correlation, providing a simple approximation to van der Waals dispersion and their respective exchange quenching. The missing higher order induction and exchange-induction

terms are added in SAPT0 as obtained from the discrepancy between the HF-level computed interaction energy via the supermolecular approach (see 3.2) and the SAPT0 terms in Eq. D.5 - the missing part is accounted for in the additional $\delta E^{(2)}_{\mathrm{HF}}$. This term and setting the interaction parameter at unit value, $\zeta = 1$ produce the SAPT0 interaction energy decomposed as

$$E^{\mathrm{SAPT0}}_{\mathrm{int}} = E^{(10)}_{\mathrm{elst}} + E^{(10)}_{\mathrm{exch}} + E^{(20)}_{\mathrm{ind,resp}} + E^{(20)}_{\mathrm{exch\text{-}ind,resp}} + E^{(20)}_{\mathrm{disp}} + E^{(20)}_{\mathrm{exch\text{-}disp}} + \delta E^{(2)}_{\mathrm{HF}}. \quad \text{(D.6)}$$

As the simplest many-body approximation to provide a reasonable total interaction energy, the SAPT0 method's primary strength lies in its ability to decompose this energy into physically meaningful components. This decomposition provides crucial insight into the nature of intermolecular forces, making SAPT0 a key tool for classifying weakly interacting complexes as, for example, hydrogen-bonded, dispersion-dominated, or of mixed character.

Appendix E

Kernel ridge regression results

This work was conducted in collaboration with Leonardo Medrano Sandonas, Artem Kokorin, and Igor Poltavsky. My contribution involved data curation, performing computational experiments, subsequent analysis and visualisation.

Exploration of the vast chemical compound space has been widely assisted by ML approaches, both neural networks (NN) and kernel ridge regression (KRR). While the former performs better given large datasets, KRR is a more data-efficient method, and the trained models can achieve high accuracy in the small data regime[348]. This is crucial when the complexity of the study allows only limited sampling, *e.g.*, experimental data or prediction of high-fidelity electronic quantum-mechanical (QM) properties of diverse covalent and non-covalent systems. A benchmark study of modern kernel methods, as well as state-of-the-art NNs, can be found in Chapter 5.

Here, the influence of different KRR components is examined to improve our understanding of how they affect the performance of the predictive models. An extensive and comprehensive analysis of KRR components for the prediction of electronic QM properties has been performed for equilibrium and non-equilibrium conformations of single small organic molecules from the QM7-X dataset [202], on MD data of organic molecules from the MD17 dataset[349] and on the small organic molecular dimers from the NENCI-2021 dataset[209]. In particular, the focus is on the prediction of the energetics of the systems - the total energy of the MD17 molecules, the interaction energies of the NENCI-2021 dimers, and the atomisation energy as well as the HOMO-LUMO gap of the QM7-X molecules, for which also the related physiochemical properties dipole moment and molecular polarisability are investigated.

To that end, two- and three-body geometric representations (Bag-of-Bond, SLATM) in combination with Gaussian/Laplacian kernels and Minkowski metrics are employed to probe the impact of a distance metric and optimising the role of outliers in molecular data. A simple 'Generic' kernel was also tested as a middle ground between the Laplacian and the

Gaussian kernels and found more stable in performance with the different Minkowski metrics across all systems. The results for the molecular dimers are also checked against a Δ-learning model between DFTB and the reference CCSD(T) data. We found the Δ-ML model to improve the performance in terms of mean absolute error but also to increase the presence of outliers. This kernel study is conducted by training and testing on molecules from a single dataset as befitting a method renowned for its power in the sparse data regime, in contrast with our study of the DFTB method augmented with NN Δ_{TB} many-body potentials, presented in Chapter 6, where the interaction energy of molecular dimers was studied for models trained predominantly or exclusively on single molecules.

Overall, an in-depth analysis of KRR component was performed, which informed the creation of the 'Generic' kernel offering a more stable performance with a variety of Minkowski metrics than the Laplacian and Gaussian kernels. Given the role of the metrics in regulating the presence of outliers this makes it a great choice for a variety of systems with outliers without an increase in the computational cost.

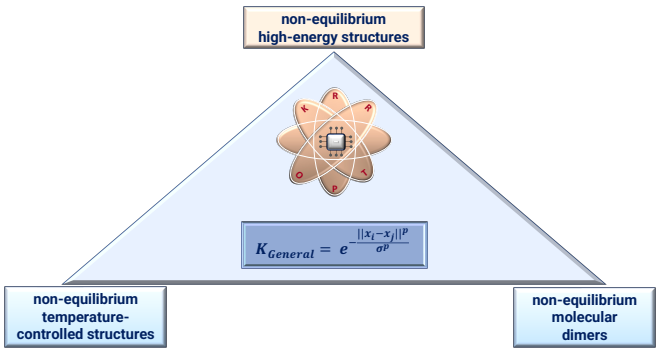

Figure E.1: The Kernel ridge regression algorithm `KRR-OPT` used with our suggested General kernel for the prediction of a range of organic chemical systems.

E.0.1 Computational experiments

Our focus is to analyse the effect of KRR components on the property prediction of molecules contained in three well-known benchmark datasets, *i.e.*, QM7-X, NENCI-2021, and MD-17 (see Fig. E.2). We have split the study cases depending on the diversity and complexity of the molecular systems.

QM7-X case. QM7-X dataset [202] includes $\approx$ 4.2 M (equilibrium and non-equilibrium) molecular structures containing up to seven heavy (non-H) atoms, namely C, N, O, S, Cl, as well as a large set of 42 physicochemical properties (per molecular structure) at benchmark level values with PBE0+MBD level of theory. For each equilibrium molecule, 100 non-equilibrium molecular geometries were generated by distorting the molecule along linear combination of vibrational modes. In the present work, we have studied the equilibrium and the most distorted structure per molecular conformer, *i.e.*, we have two QM7-X subsets

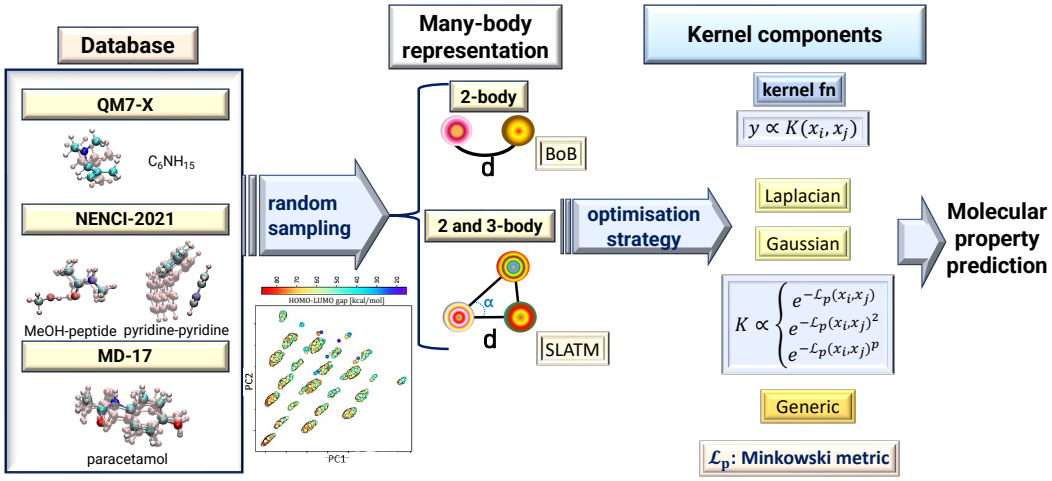

Figure E.2: Workflow of molecular property prediction: A dataset of molecules is produces from a database through random sampling. N number of samples is created. For each sample, the following steps are performed: 1) the information about the molecules is represented by a molecular descriptor, 2) the dataset is split into training, test and validation set with the ratios of their sizes chosen as appropriate by our *KRR-OPT* algorithm, 3) KRR is implemented for different kernel functions Laplacian, Gaussian, and Generic and Minkowski metrics with the hyperparameters optimised simultaneously via L-BFSG-B algorithm. The extensive or intensive property considered is predicted for the test set, whose MAE is calculated. MAE results of the samples are compared and the best one is taken as output.

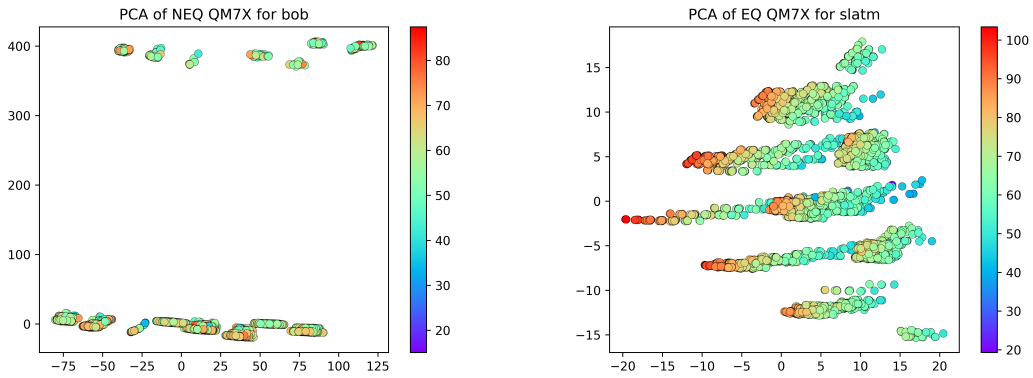

Figure E.3: a)2-component PCA for equilibrium subset of QM7-X CHON molecules for the SLATM representation, and b)2-component PCA for non-equilibrium subset of QM7-X CHON molecules for the Bag-of-Bond (BoB) representation.

and each of them contains circa 42k structures. An example of an equilibrium and most distorted molecular structure is depicted in Fig.E.2. We have also considered the extensive

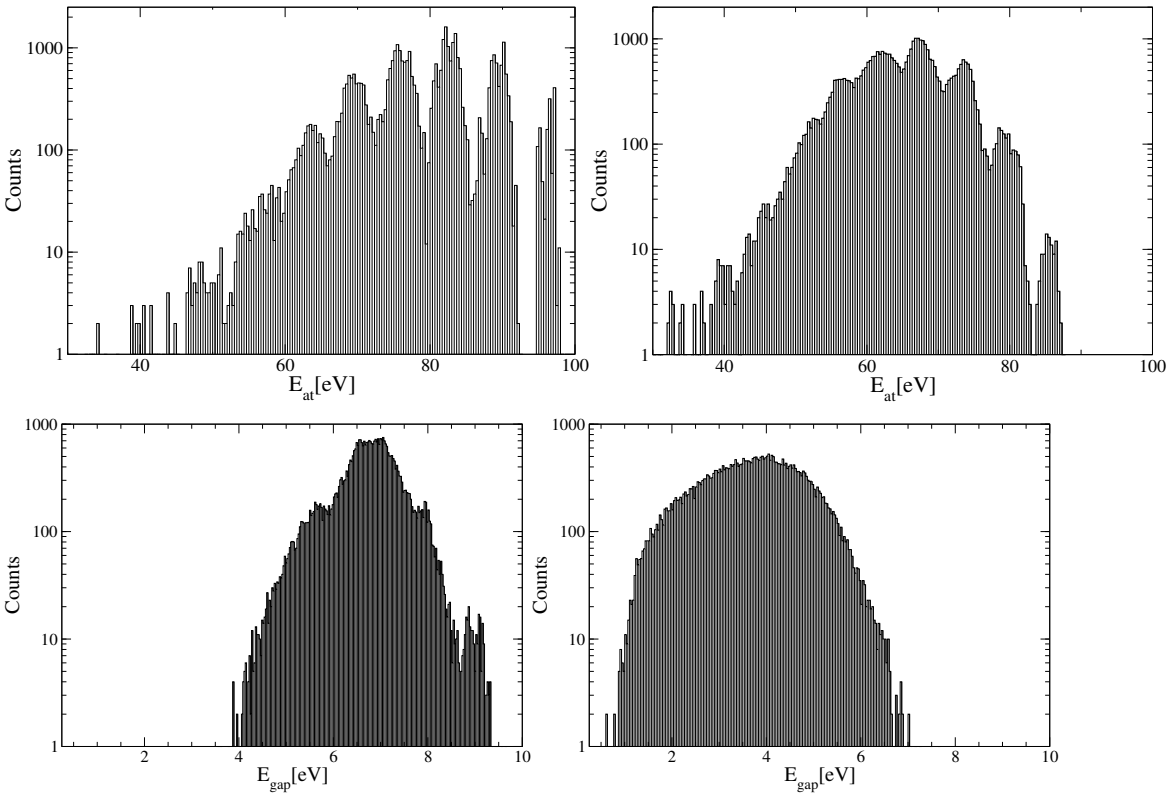

Figure E.4: a)Atomisation energy profile for the equilibrium QM7-X subset, b)Atomisation energy profile for the non-equilibrium QM7-X subset, c)HOMO-LUMO energy gap profile for the equilibrium QM7-X subset, d) HOMO-LUMO energy gap profile for the non-equilibrium QM7-X subset

property atomisation energy (E_{AT}) and the intensive property HOMO-LUMO (Highest Occupied Molecular Orbital - Lowest Unoccupied Molecular Orbital) gap (E_{gap}) as target QM properties. The energy distribution for both molecular subsets is shown in Fig. E.4.

NENCI-2021 case. The NENCI dataset[209] contains information on interaction energies of small molecular dimers of up to 34 atoms at different levels of theory, see an example of a dimer in Fig. E.2. For this study case, we have considered two subsets of NENCI: the first one consists of 6,748 of the 7,763 molecules with chemical composition C, N, O, H, S, P, Cl, and F atoms (i.e. without the Li, Na, and Br-containing compounds). The second dataset, known as 'filtered' subset, contains 4.8k molecules with the same chemical composition but without the dimers with monomer distance 0.7x, and 0.8x equilibrium to ensure the physical viability of the training set and no overlap of atoms).. Our target property will be the interaction energy E_{int} (intensive property) computed at the golden standard level of CCSD(T)/CBS. To gain more insights into the prediction of this property, E_{int} is also computed using Density Functional Tight Binding 3rd order [129, 350] supplemented by Many Body dispersion [121, 122] (DFTB3+MBD). Indeed, we will use this data to apply the

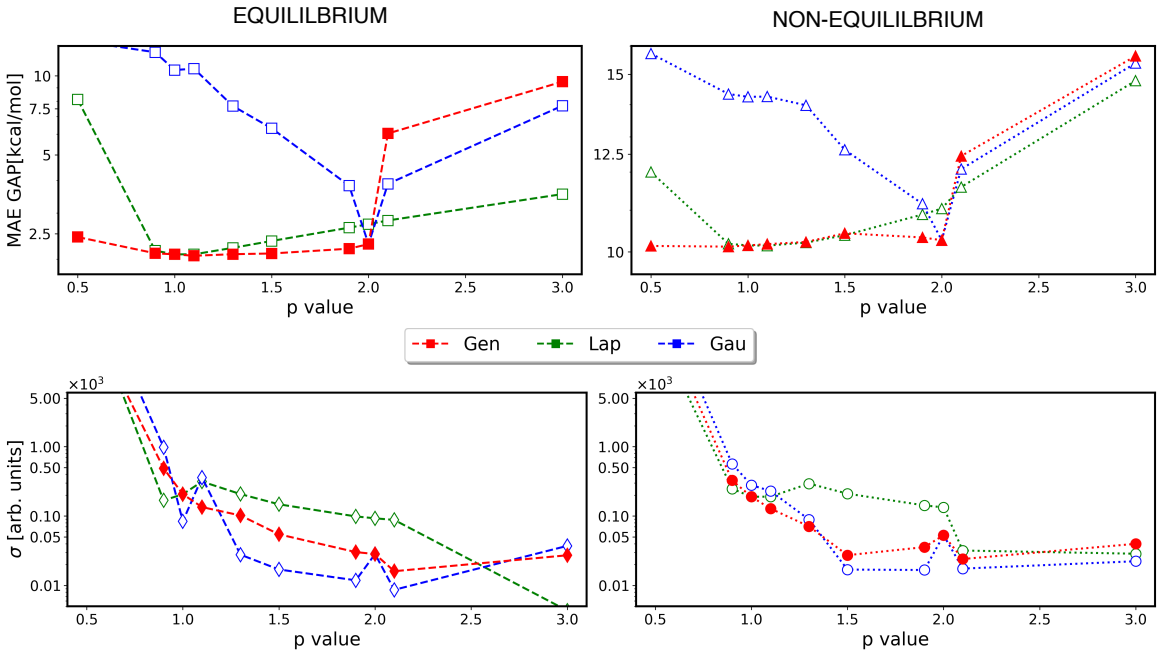

Figure E.5: An investigation with the **KRR-OPT** algorithm into the role of the different Minkowski metrics depending on their p values for the Laplacian, Gaussian, and Generic kernel for the prediction of the HOMO-LUMO gap of for 10k training set molecules from the equilibrium (left) and non-equilibrium (right) subsets of the QM7-X database using a SLATM representation.

delta learning (ΔML) method for the prediction of E_{int}.

MD17 case. The MD17 dataset [349] consists of molecular dynamics (MD) trajectories for small organic molecules of pharmacological relevance computed at PBE + vdW-TS level of theory (van der Waals - Tkatchenko/Sheffler)[351, 352] with the `FHI-Aims` [353, 354] software on *light* settings. The trajectories for each molecule include between 150k to 1M conformations, generated with a timestep of $0.5fs$ at 300K. MD17 considers 10 molecules containing from 3 up to 14 heavy atoms (including C, H, O, and N) and diverse potential energy surface (PES) characteristics. To understand the effect of KRR components in the energy prediction of non-equilibrium conformations obtained from MD runs, we will focus on the MD trajectories of malonaldehyde (MDA) and paracetamol with 5 and 11 heavy atoms, respectively.

$$K_{Generic} = e^{-\frac{||x_i - y_i||^p}{\sigma^p}} \quad . \tag{E.1}$$

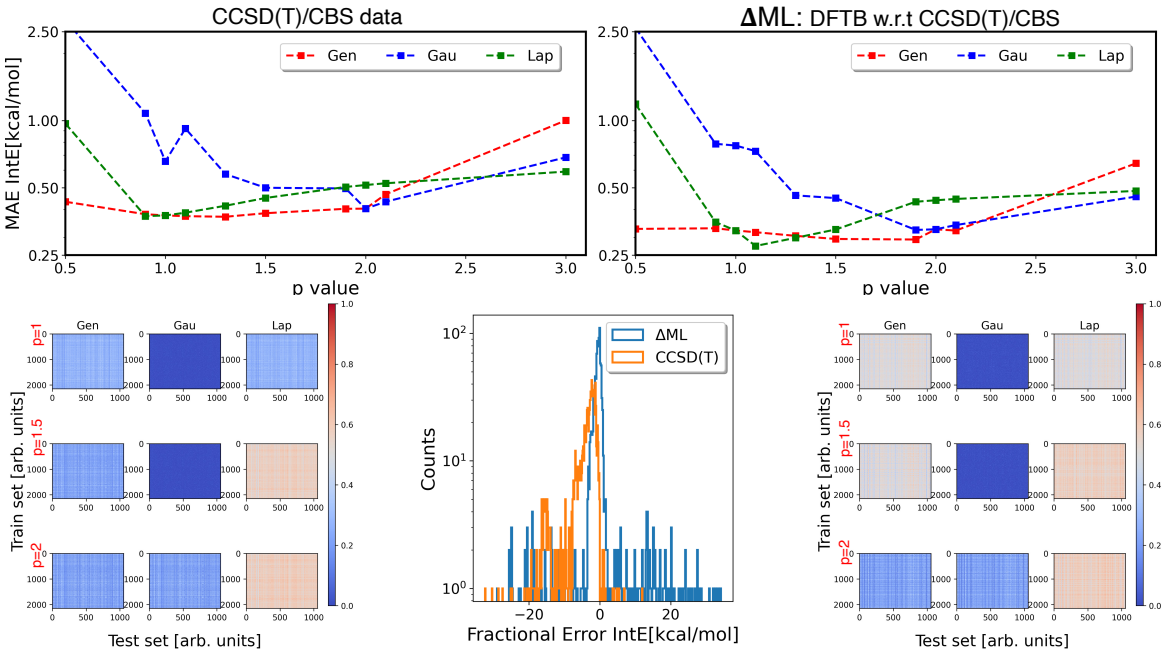

Figure E.6: Investigation with the KRR-OPT algorithm of the prediction of the interaction energy of NENCI-2021 molecular dimers containing C, H, N, O, Cl, F, and S for intermonomer distances in the range 0.9 to 1.1 x the equilibrium distance. a): MAE values for different p-value Minksowski metric experiments with Laplacian (Lap), Gaussian (Gau), and General (Gen) kernel for left) the CCSD(T)/CBS reference data, and b) ΔML data for DFTB3+MBD w.r.t. CCSD(T)/CBS levels of theory. c) 2-component PCA for the whole CCSD(T)/CBS reference data, d) histogram with the distributions for the Fractional Errors of the Laplacian kernel with L_1 norm for the ΔML and CCSD(T)/CBS datasets, and e) a heatmap for the kernel matrices for Lap, Gau, and Gen kernels for p=1, 1.5, and 2.0.

E.0.2 Optimal training set determination

The study cases discussed above were carried out using the developed `KRR-OPT` toolbox (see Fig. E.2). The overall performance and mechanism of the `KRR-OPT` algorithm can be exemplified by considering a representative case, *e.g.*, the prediction of the HOMO-LUMO gap E_{gap} for the subset of non-equilibrium structures from QM7-X, a more challenging case sure to the high energy conformers and the targeted intensive property. In doing so, we have examined three different training set sizes containing 0.5k, 2k, and 10k molecular structures. The results for the BoB descriptor are shown in Fig. 5.3, with the SLATM descriptor results behaving similarly. For the smaller training sets, the `KRR-OPT` algorithm utilises more randomly distributed sets (known as "samples" in the code) to determine the optimal training and validation set with minimal errors. The number of samples decreases with the increase of the training set size due to the required computational expenses. The rationale behind that implementation can be seen in the dataset profile visualised with a simple 2-component Principal Component Analysis (PCA) in Fig. E.2 and Fig.E.3. The presence of

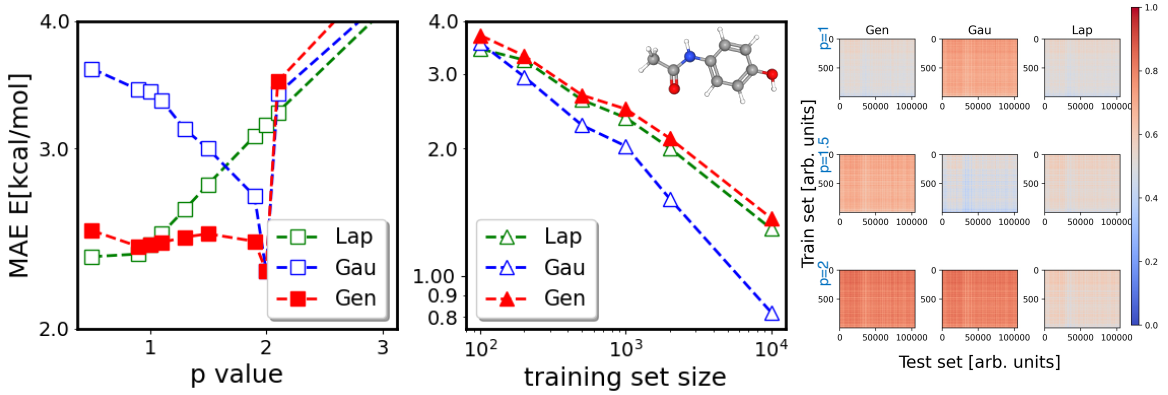

Figure E.7: Minkowski metric test w.r.t. p for paracetamol for Gaussian, Laplacian, and Generic kernel with the *KRR-OPT* method, a learning curve for the same kernels for paracetamol for their optimal Minkowski metrics respectively p=2.0, p=1.0, and p=2.0, and a kernel elements' heatmap for three Generic kernels with p=1.0, p=1.5, and p=2.0 metrics.

numerous distinct clusters is a striking feature, with the same differentiation present for both the BoB and SLATM descriptors applied to a non-equilibrium sample. This differentiation is retained, albeit less pronouncedly, also for the equilibrium dataset. The presence of separate groupings of molecules indicates that the choice of random sample for the training and validation sets is of critical importance for the performance of the algorithm. A representative sample for training and validation would include molecules from all the different clusters formed by considering the major PCA components. With the objective of efficient solutions for KRR, and thus without introducing extra pre-processing analysis and overheads due to training set engineering, we propose the `KRR-OPT` algorithm.

The capture of a representative training set can be explained with the aid of the plot of Mean Absolute Error (MAE) vs the Mean Square Error (MSE) values for the predictions trained on the different samples, see Fig. 5.8(a).

The figure showcases a representative example in the case of non-equilibrium QM7-X subset for the prediction of the HOMO-LUMO gap with a BoB descriptor. Here we see that the MAE and MSE errors differ more for the smaller datasets and vary less with increasing test size, reaching a negligible discrepancies for the four 10k dataset samples. The big variation for the small set can be seen more precisely in the distributions of the test set shown for the best and worst performing samples among the 0.5k training set case (see Fig. 5.8(b)). The large variation indicates that for the worse random sample, a non-representative selection is obtained, *e.g.*, excluding molecules belonging to different clusters in the PCA plot (see Fig. E.2). Hence, had we used a standard method as the 5-fold cross-validation, the model could still not scale to capturing the whole dataset due to the variation in the data. This is particularly true for the scarce data regime or for smaller training sets, *e.g.*, the 0.5k training set for the prediction of the intensive and thus harder-to-capture molecular properties as

can be seen in Fig. 5.8 depicting the results for the HOM0-LUMO gap prediction for the most distorted non-equilibrium subset of QM7-X molecules.

Training set	Validation set %	Samples (BoB)	Samples (SLATM)
500	4	22	11
1000	2	16	8
2000	1	12	6
4000	1	8	4
10000	0.5	4	2
20000	0.25	2	2

Table E.1: Specifications of the learning curve calculations for the QM7-X (non-) equilibrium molecules. The validation set size is given as a % of the training set size.

In our algorithm, representative sampling is ensured through multiple samplings of the dataset (see Table E.1 of for details). Moreover, when using bigger training sets with more expensive descriptors, e.g. SLATM, it is more efficient to perform KRR calculations for two 10k training samples (see Table E.1) with `KRR-OPT`'s approach instead of 5-fold cross-validation, *i.e.*, for five 10k training samples. Further, the application target of `KRR-OPT` toolbox is, in particular, for training sets that are structurally or chemically diverse, and still small compared to the requirements for training neural network models.

E.0.3 QM7-X case: Intensive and extensive properties

The first case study we discuss is single (equilibrium and non-equilibrium) molecules from the QM7-X database. Two subsets from the 101 available in the database are chosen: the equilibrium subset and the most distorted non-equilibrium subset, each 42k structures. We consider the impact of different KRR components on the prediction of an extensive and intensive property for each dataset, namely atomisation energy (E_{AT}) and HOMO-LUMO gap (E_{gap}). To obtain a full picture of the impact and entanglement of the separate KRR components, models with different combinations of KRR elements are examined, with the results presented here being for the SLATM descriptor (checks with the BoB descriptor have also been carried out). An investigation of the kernel function choice is carried out jointly with variations of the Minkowski metric, as defined by the different p values (Eq.(3)). The experiments are performed for the 10k training set size with 2 random samples for SLATM, and 4 random samples for BoB representation. We settle on a training set of 10k molecules for the QM7-X case as a representative capture among the 42k dataset molecules, as shown by the fact that the MAE and MSE values for the different samples differ little (see Fig. 5.8(a)). Hence, the 10k training set size ensures we are working with a representative training set, *e.g.*, containing molecules from all the distinct clusters identified by the PCA. In the following, we only discuss the results for E_{gap} prediction using the SLATM descriptor, as it captures the molecular features of the dataset better (see Fig. 5.3).

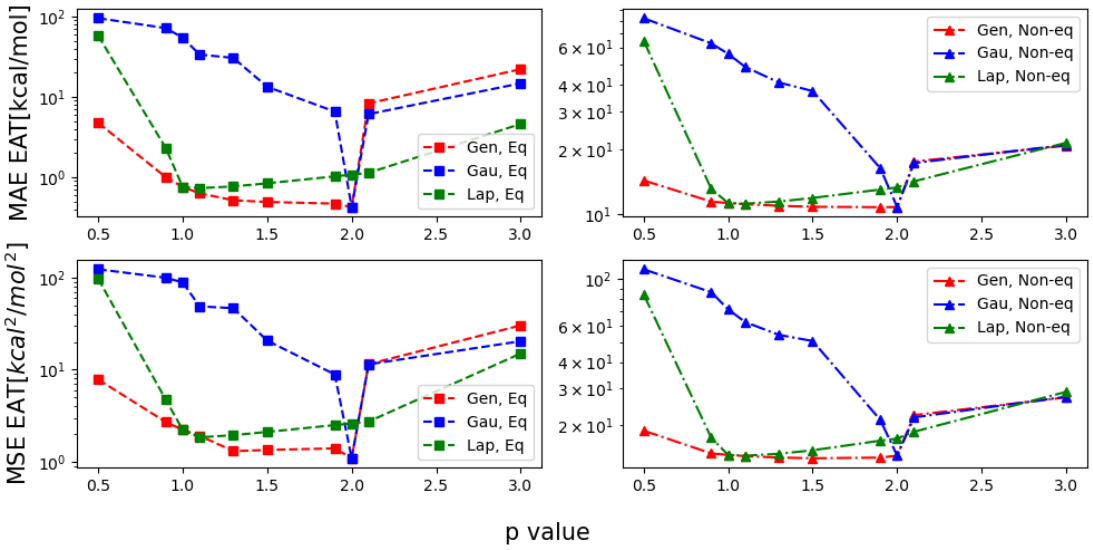

Figure E.8: Atomisation energy prediction for QM7-X equilibrium and non-equilibrium datasets in terms of MAE and MSE errors with different p values for the Minkowski metrics for the General, Gaussian, and Laplacian kernels with SLATM descriptor.

The results for E_{AT} predictions for the same set are available in Fig. E.8.

Across all cases, the Laplacian kernel outperforms the Gaussian one for the Minkowski p-value around the minimum for the kernel function, *i.e.*, $p = 1$ and $p = 2$, respectively. The Generic kernel (Eq. 2.93) exhibits a stable and reliable prediction for the range of $p = 1$ to $p = 2$, with a marked drop in accuracy beyond Minkowski metric $p = 2$. Notably, the tuning of the σ hyperparameter for that range for the Generic kernel is smoother and more consistent than that of the Laplacian and Gaussian kernel. As expected, both the σ and MAE values for Generic kernel with $p = 1$ match those of the Laplacian kernel and with $p = 2$ match the Gaussian kernel ones, validating the robustness of our implementation. The chosen region of p values to be examined is from 0.5 to 3 due to the sharp increase in errors beyond those values for the Laplacian and Gaussian kernels. This signifies that below $p = 0.5$ the molecular descriptors in the kernel space are overlaid too closely to be distinguishable from one another, while beyond $p = 3$, they are spread too far apart, diminishing their similarities beyond the threshold of effective capture for the models' learning processes and subsequently their predictive capabilities. This phenomenon occurs already for the Gaussian and Generic kernels beyond $p = 2$, where we can observe that the sigma hyperparameter does not adjust to the change of p value to optimise the model's learning resulting in a sharp jump in error between $p = 2$ and $p = 2.1$. The more pronounced changes in the Gaussian kernel's results for different metrics are due to the enhanced differences between the descriptors lifted to the power of two in the kernel (Eq. 2.93). Given the mathematical expression of the Gaussian kernel, even a small variation in the p-value of the Minkowski matrix is amplified more than in the Laplacian case. Most importantly, for both equilibrium and non-equilibrium datasets the use of the Generic kernel shows a

steady performance throughout the $p = 0.5$ to $p = 2$ range with smooth changes between the metrics' results. Therefore, the expected benefit is confirmed by the ability of the optimisation parameter to fit directly to the similarity measure between the molecular descriptors due to the matching powers of sigma and the kernel. However, even with an optimal choice of p-value or kernel function, the prediction of either an intensive or extensive property for the highly distorted non-equilibrium molecular conformations has a large MAE of 10 kcal/mol. As such, the benefit of using a Generic kernel is in the stability of the performance, but for more challenging cases like distorted structures, a different approach to training a KRR-OPT model is still needed (*e.g.*, more advanced molecular descriptors).

E.0.4 NENCI case: Intermolecular interactions

The second case study we consider is the prediction of intermolecular interactions, specifically the interaction energy (E_{int}) of equilibrium and non-equilibrium molecular dimers from the NENCI-2021 dataset. Among the different levels of theory for E_{int} contained in NENCI-2021, the high-fidelity golden-standard benchmark values of Coupled-Cluster Singles, Doubles, and Triplets in the Complete Basis Set regime (CCSD(T)/CBS[12]) were chosen as target data. Due to our focus on small organic systems, we choose to analyse a subset of the dataset, comprised of molecules only with chemical composition consisting of C, N, O, H, Cl, F, and S atoms, *i.e.*, removing the ones with non-organic elements like Li and Na, as well as the few with the large organic atom Br. The chosen subset consists of 6.7k out of the original 7.8k molecular dimers. In it, there are equilibrium and non-equilibrium configurations, such that for every equilibrium dimer with distance between the monomers at $q = 1.0$(factor x the equilibrium distance) there are non-equilibrium configurations with 7 non-equilibrium distances. They span the inter-monomer distance range of q: 0.7 to 1.1, namely 0.7, 0.8, 0.9, 0.95, 1, 1.05, and 1.1. Furthermore, for each interatomic distance, there are molecular dimers at equilibrium bond angle and two non-equilibrium bond angles: $\pm 30^{\circ}$ w.r.t equilibrium bond angle.

The results for the predictions of E_{int} for the 6.7k molecular dimers for Laplacian, Gaussian, and Generic kernels with different Minkowski metrics (p in range 0.5 to 3.0) are carried out and presented in Fig. E.9. Notably, the error of the predictions is lowest for the best performing case of Generic kernel at p=2.0 (matching as expected the form of the Gaussian kernel there). However, it remains $\sim$1kcal/mol, which given the weaker strength of the non-covalent interactions calls for a further improvement. Therefore, as a next step, we also explore the effect of filtering out strongly non-equilibrium molecular dimers on the physicochemical property prediction. Specifically, we filter out the molecular dimers with distances between the monomers of 0.7, 0.8x the equilibrium distance. This test of was chosen due to investigate a subset of NENCI comprised of 4,820 molecular dimers, which only includes non-equilibrium configurations in a bond range of 0.9-1.1x the equilibrium distance, thus resulting in a dataset, which is symmetric in the distortion from equilibrium, and hence hypothetically less complex to be learned. Furthermore, that ensures that there will not be highly energetic structures resulting from the reduction of the inter-monomers

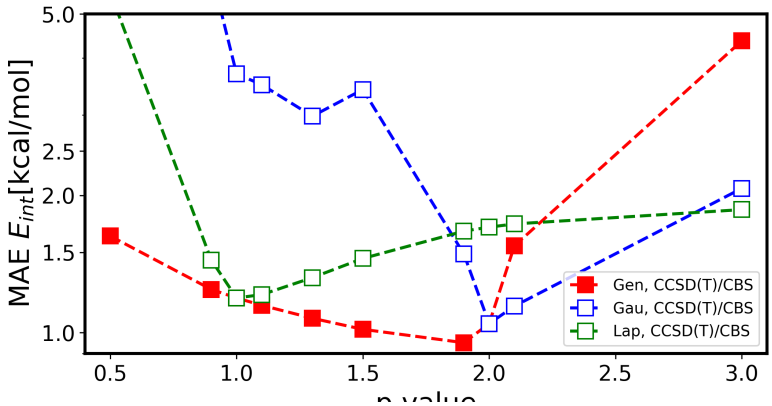

Figure E.9: MAE plot for Laplacian, Gaussian, and Generic kernels for the 6.7k NENCI-2021 subset for CCSD(T)/CBS level of reference data.

distance combined with distorted angles of $\pm 30^{\circ}$ w.r.t equilibrium bond angle. Property prediction of the interaction energies for this subset of equilibrium and non-equilibrium molecular dimers is shown on Fig.E.6a) for Laplacian, Gaussian, and Generic kernels with a investigation of the Minkowski metrics as usual for p in range 0.5 to 3.0. The MAE value for the E_{int} prediction is 0.35kcal/mol in the best performing case of Generic kernel with $p = 1.3$ metric. This is a marked improvement on the prediction outcome compared to the 6.7k NENCI subset and its optimal MAE of 1kcal/mol. The two tests were carried out with the same conditions for KRR-OPT software, with notably the fraction of the investigated dataset is used for training (46%). The challenge to learning the NENCI molecular dimers, especially in their more distorted configurations is in part due to the diversity of the dataset. As shown via a simple 2-component PCA analysis for the SLATM representation on Fig.E.6, NENCI-2021 presents a diffuse diverse dataset, unlike the clusters present in the QM7-X subsets (Fig.5.8). Notably, the Laplacian kernel with L_1 metric is performing better than the Gaussian kernel with L_2 metric, again consistent with a diverse training set.

Given the diverse dataset of challenging non-covalently interacting molecular dimers, we have tested the use of Δ-learning in combination with the KRR-OPT approach as another potential avenue for improvement of the learning process. As a typical choice, we aim to learn the difference between a high-fidelity level of theory (CCSD(T)/CBS) and a less expensive one, namely DFTB3+MBD (Density Functional Tight Binding 3rd order + Many Body Dispersion correction). We carried out the single point calculations using `DFTB+` to compute the E_{int} value at the level of DFTB3+MBD as a suitable choice for organic systems. The results using a SLATM descriptor for the filtered dataset of 4,820 molecules are presented on Fig.E.6c) for Laplacian, Gaussian, and Generic kernel for the filtered dataset for CCSD(T)/CBS and for ΔML for DFTB3+MBD w.r.t. CCSD(T)/CBS. Comparing the two performances, a similarity is directly notable - the Generic kernel exhibits a steady performance for both around optimal values for all Minkowski metrics with p values in range 0.9 to 2.0, reminiscent of the stability of the Generic kernel for the QM7-X (non-)equilibrium

datasets on Fig.E.8a) and b). For both the CCSD(T)/CBS and the ΔML property prediction, the Laplacian kernel outperforms the Gaussian, and notably for the $p = 1.1$ metric also the Generic kernel, where it reaches minimum for the MAE value. Interestingly, for CCSD(T)/CBS, $p = 1.0$ is optimal for Laplacian/Generic kernel, indicating that the similarity measure between the molecules decreases when considering their difference with DFTB3+MBD level.

The strong dips at $p = 1$ for the Laplacian kernels and $p = 2$ for the Gaussian kernel are consistent with a homogeneous dataset, which covers well the subset of CSS, corroborated by our PCA check in Fig.E.6c). However, the values of the MAE measure do not always tell the whole story [355]. Hence, we also investigate the fractional errors for both models, as shown on the histogram on Fig.E.6d). The optimal functional combination is taken for the CCSD(T)/CBS and the ΔML results, with the Laplacian kernel optimising for Minkowski metric p=1.1 and the Generic kernel at p=1.9. Notable on the histogram is that , the filter of those systems reduces the bigger fractional errors but also results in a skew to larger negative error values, so the peak is no longer centered at zero. The mean absolute relative error (MARE) for CCSD(T)/CBS and ΔML are respectively 0.08 and 0.07 with MAE values for the same datasets 0.37kcal/mol and ΔML 0.29kcal/mol, respectively. This indicates that the the fractional gains in using the ΔML approach for DFT3+MBD vs CCSD(T)/CBS data for the NENCI dimers is smaller (13%) than the MAE values indicate (22%), as well as having the ΔML approach more prone to higher error outliers. Hence, the CCSD(T)/CBS and the DFTB3+MBD levels of theory likely do not match well in the degree of errors for their predictions of the E_{int} property for small organic non-covalalent dimers and this results in a complication of learning the difference between the two levels corresponding to a higher mean absolute relative error.

A good measure for that degree of similarity between molecules can also be a heatmap of the kernel function representation for the optimal hyperparameters for each specific Minkowski metric and kernel combination. It also can be used to depict captured degree of complexity and diversity of the dataset as indicated by the homogeneity or lack thereof in the predictions. As the heatmap is a depiction of the values in the kernel function in arbitrary units, with each pixel represents the kernel function calculated for a pair of molecules in the dataset as proportional to their similarity measure, namely the employed Minkowksi metric. Such heatmaps depicting the the kernel elements for the Generic, Gaussian, and Laplacian kernels at p=1.0, p=1.5, and p=2.0 metrics are available for the CCSD(T)/CBS and theΔML of DFTB3+MBD w.r.t. CCSD(T)/CBS on Fig. E.6c) and Fig. E.6e), respectively. The kernel heatmaps for the selected Minkowski metrics for the other applications are also shown in Fig. E.10.

Notably, the Laplacian and Gaussian kernels we see pixels in the blue and orange spectra indicating similar and dissimilar molecules to a higher degree than the QM7-X heatmaps (see Fig.E.10). Further, the jump between optimal and less optimal kernel space is visible in a way reminiscent of a phase transition, with the optimal regime in one predominant colour changing with the metric p value. The uniformity of the blue colour for Gaussian kernel

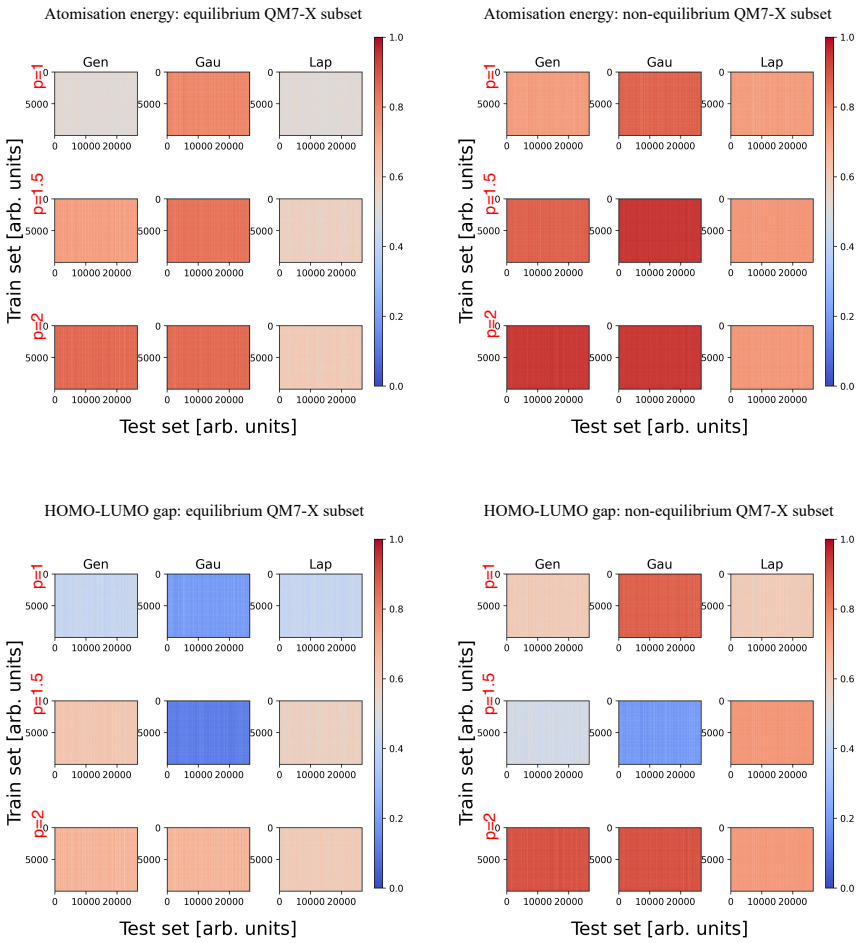

Figure E.10: A plot of the kernel matrices for atomisation energy and HOMO-LUMO gap prediction for the equilibrium and non-equilibrium dataset each.

with $p = 1.0$ and $p = 1.5$ metrics indicates the poor match of those kernel components as they fail to capture the inhomogeneity of the dataset for both the direct learning and the ΔML approach.

E.0.5 MD17 case: Prediction of energy in molecular dynamics

Thus far we have considered KRR in the context of property prediction in subsets spanning equilibrium and non-equilibrium conformational space for single molecules (QM7-X) and for molecular dimers (NENCI-2021). To complete the picture of the KRR components' impact on property prediction for small organic molecules, we investigate as a third and final case molecular dynamics datasets. In particular, we focus on the paracetamol ($C_8H_9NO_2$) set from the MD17 database [349]. The paracetamol dataset is comprised of $106,490$ configurations computed at *ab initio* PBE0+MBD level. In our exploration of the effect of partial

metrics and kernels on the prediction, we consider a training set of 1k molecules, as per the standard cited in literature [32, 179, 356–358]. Given the sizes of the MD17 datasets, more

Training set	Validation set %	Samples (SLATM)
100	6	22
200	4	18
500	4	16
1000	2	14
2000	1	10
10000	0.5	2

Table E.2: Specifications of the learning curve calculations for paracetamol structures from molecular trajectories. The validation set size is given as a % of the training set size.

training set samples are tested with the KRR-OPT model (see Table E.2) with 14 samples for the 1k training set. However, unlike the random sampling of the diverse datasets in the first and second case, the training sets here were chosen to sample regular points of the MD, with training set engineering present in the original reference paper[349]. The achieved MAE for the total energy of paracetamol E_{tot} is well within strict chemical accuracy <1kcal/mol and our aim in this work is not to outperform other available models but rather to explore the role of the different components in the KRR prediction while staying around chemical accuracy range of 1-2kcal/mol. We aim to create an exhaustive study informing the choice of kernel and metric in existing and future KRR models.

Our findings as exhibited on Fig.E.6 indicate that unlike in the predictions for individual molecules or molecular dimers, the Gaussian kernel greatly and consistently for both test molecules outperforms the Laplacian one. The Gaussian kernel with Euclidean metric is the only combination to reach or near the 1-2kcal/mol chemical accuracy threshold. Interestingly, the Gaussian kernel as expected performs better for a single molecule[359] with a steeper learning curve than the Laplacian and Generic kernel (see Fig.E.6 (b)). Most notable, however, is the sharp dip to optimal value at p=2.0 present not only in the Gaussian but also for the Generic kernel. While the Generic kernel is steady as before for p=0.9 to p=1.9 metrics, the optimal metric at p=2.0 also visible in the colour change of the kernel elements' heatmap on Fig.E.6. The Generic kernel specifically exhibits a steady transition from smaller to larger values in the p=1.0, p=1.5, and p=2.0 cases. Interestingly, there is a regular pattern noticeable here as vertical and horizontal stripes of more orange or bluer colours, i.e. higher and lower values. In all the plots, a fainter cross-shaped pattern is visible in the middle with lower values than its surroundings. All of this is consistent with a dataset composed of MD of a single molecule without any bond creation of breakage. The regularity is due to the closer of further geometries, which generally correlate with closer of further apart in time in the MD trajectory.

E.1 Summary

We conducted an exploration of the effect of the KRR components on the performance of different molecular properties for molecules systems in and out of equilibrium. Our investigation included the study of the Laplacian, Gaussian, and General kernel. The General kernels is introduced in this work as a balanced alternative to the standard choice Laplacian and Gaussian kernels. The General kernel offers a way to capitalise on the different Minkowski metric distances, for which the non-integer p values allow for an attunement of the kernel to the studied dataset. In particular, the role of the hard-to-capture outliers can be emphasised or de-emphasised by changing the value of p. We have noted the strongly correlation between the database profile and the optimal choice of kernel components, including the choice of algorithm. We focus on avenues for improvement in the KRR components choice, which would not rely on significant computational overheads and especially in the low-data regime (e.g. out of equilibrium).

We use the KRR-OPT algorithm as a way of selecting a training set without bespoke dataset engineering, while taking advantage of a simultaneous and elaborate hyperparameter optimisation process (see Fig.E.7). Our new introduction, the Generic kernel, for which the power in the kernel matches that of the Minkoswki metric, consistently showed a steadier performance due to the ability of the model to learn directly on the descriptors without further non-linearity from the kernel form. A diverse choice of organic systems were investigated in this work for covalent and non-covalent molecules and molecular dimers. The included property prediction for the equilibrium and non-equilibrium single molecules in the chemically diverse QM7-X database, tested separately. Further, a chemically diverse database was also treated in the context of non-covalent interactions with the NENCI-2021 molecular dimers database (filtered for Li, Na, and Br-containing ones). Finally, a single system dataset was chosen from MD17 for the molecular dynamics of paracetamol. The optimal minima were around the typical L_1 or L_2 norms for the metric and the kernel choice for the 2- and 3-body SLATM descriptor. It proved sufficient to reach chemical accuracy for the equilibrium subset of QM7-X for an extensive and intensive property when using only 1/4 of the dataset for training. For the non-covalent molecular dimers, 1/3 of the data was used for training for the results to reach 1 kcal/mol MAE. For the paracetamol MD dataset, as expected, the Gaussian kernel (Generic kernel at p=2.0) outperforms Laplacian, and is confirmed to be the optimal selection with L_2 norm for molecular dynamics data. An interesting exploration into the molecular dimers also considered Δ-learning for DFTB3+MBD and CCSD(T)/CBS showed that that avenue for improving ML models is not trivial given different physical models. In the property prediction of interaction energy, in absolute terms the improvement appears visible, however when we consider the small values inherent to ΔML, and consider the fractional error, the learning is better directly from the original CCSD(T)/CBS data. The Generic kernel m

Thus, a dataset with well-mapped chemical compound space for non-equilibrium conformations is key to their learning and prediction, which can inform prospective work in the

field. Capturing non-equilibrium confirmations in covalent or non-covalent systems is an outstanding challenge. The optimal choice of kernel and metric setup would depend on the application at hand, with the introduced General kernel offers an optimisation avenue for finding optimal spaces between Laplacian and Gaussian kernel's typical options, especially relevant in the presence of complex datasets in the low-data regime.

E.1.1 KRR-OPT toolbox

The KRR-OPT software was developed by Artem Kokorin and can be found at the GitHub repository https://github.com/arkochem/krr-opt.git.

The developed KRR-OPT toolbox that can be used to train ML models for property prediction using the KRR). Various features, including kernel functions, molecular descriptors, and metrics, have been implemented to capture the unknown structure-to-property relationships in complex molecular systems. KRR-OPT toolbox also considers a quasi-Newton algorithm such as the limited-memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) for hyperparameter optimization.

Acknowledgments

I would like to thank Prof. Dr. Alexandre Tkatchenko, my main PhD supervisor under my ""AQMA" PhD project funded by the "Young Academics" program of the Institute of Advanced Studies, and head of the Theoretical Chemical Physics (TCP) research group that was my academic home in the last five years. I would also like to thank Prof. Dr. Alexander Skupin as co-supervisor on the "AQMA" project and my CET committee (Comité d'encadrement de thèse) member. I would like to extend my gratitude to all esteemed members of the PhD defence jury for the evaluation of this thesis and contributions to the PhD defence, which includes also Prof. Dr. Reinhard Maurer, a long-standing CET committee member, Doc. Dr. Jan Řezáč, and Prof. Dr. Etienne Fodor.

I would also like to thank the key collaborators to the research projects conducted during this PhD. Specifically, Dr. Leonardo Medrano Sandonas for the projects in Chapters 3, and Appendix D, and leader of the project in Chapter 5, and also as contributing with supervision during the first half of this PhD. Further, I extend my gratitude to Dr. Igor Poltavsky as leader of the Chapter 4 project; Dr Jorge Charry as collaborator to the Project in Chapter 3; Sergio Suárez Dou for his support to the project in Chapter 3 and collaboration in the project in Chapter 6; Dr. Grégory Cordeiro Fonseca as collaborator in the project in Chapter 4, and his support to the projects in Chapter 3 and Chapter 6; and Prof. Dr. Peter R. Nagy for his collaboration in the project in Chapter 3.

I would like to express a heartfelt thanks to Dr. Matteo Barborini, Dr. Jorge Charry, Dr. Matteo Gori, Dr. Stefan Chmiela, Dr. Grégory Cordeiro Fonseca, and Dr. Josh Berryman for their support in the writing of this thesis.

I would like to also express my thanks to Dr. Thais Arns, Balázs D. Lőrincz, Dr. Anton Charkin-Gorbulin, Artem Kokorin, Dr. Szabolcs Góger, Dr. Matej Ditte, Dr. Alessio Fallani, Dr. Kyunghoon Han, all other collaborators and the TCP group as a whole for the scientific exchanges.

Large Language Models, *e.g.*, Microsoft Copilot, were used in the early drafting stage for brainstorming on how to introduce some theoretical concepts; for proof-reading and rephrasing individual paragraphs to improve clarity - all suggestions were verified and modified by the author.

I would like to gratefully acknowledge the financial support of the Institute of Advanced Studies for funding this PhD project "AQMA: Approaching Quantum Mechanical Accuracy for Drug-Protein Binding with Machine Learning" under the Young Academics grant and I would like to thank the project coordinator Dr. Sylvie Fromentin for her support. I would like to also acknowledge the early access to the yet-unpublished models and data used in Chapter 6 generated in-house. Generative AI was used for polishing the text with changes reviewed and corrected by the author. I would like to express appreciation for the support of the teams of the High Performance Computing (HPC) facilities of the University of Luxembourg and of LuxProvide of MeluXina, the Luxembourg national supercomputer, for their expert support.

Bibliography

[1] A. M. Edwards & D. R. Owen. "Protein–ligand data at scale to support machine learning". *Nat. Rev. Chem.* 1–12 (2025).

[2] C. G. Rousseaux, W. M. Bracken & S. Guionaud. "Chapter 1 - overview of drug development". In *Haschek and Rousseaux' s Handbook of Toxicologic Pathology*, 3–48 (Academic Press, 2023), fourth edn.

[3] N. Higashi-Kuwata *et al.* "Identification of SARS-CoV-2 Mpro inhibitors containing P1'4-fluorobenzothiazole moiety highly active against SARS-CoV-2". *Nat. Commun.* **14**, 1076 (2023).

[4] Y. Hu *et al.* "Naturally occurring mutations of SARS-COV-2 main protease confer drug resistance to nirmatrelvir". *ACS Cent. Sci.* **9**, 1658 (2023).

[5] N. A. Church & J. L. McKillip. "Antibiotic resistance crisis: challenges and imperatives". *Biologia* **76**, 1535 (2021).

[6] K. Uppalapati, E. Dandamudi, S. N. Ice, G. Chandra, K. Bischof, C. L. Lorson & K. Singh. "A comprehensive guide to enhancing antibiotic discovery using machine learning derived bio-computation". *arXiv* 2411.06009 (2024).

[7] R. G. Govindaraj & M. Brylinski. "Comparative assessment of strategies to identify similar ligand-binding pockets in proteins". *BMC Bioinform.* **19**, 1 (2018).

[8] J. Černỳ & P. Hobza. "Non-covalent interactions in biomacromolecules". *Phys. Chem. Chem. Phys.* **9**, 5291 (2007).

[9] P. Hobza & K. Müller-Dethlefs. *Non-covalent interactions: theory and experiment.* 2 (Royal Society of Chemistry, 2010).

[10] Y. S. Raouf. "Covalent inhibitors: To infinity and beyond". *J. Med. Chem.* **67**, 10513 (2024).

[11] R. Haunschild, A. Barth & W. Marx. "Evolution of DFT studies in view of a scientometric perspective". *J. Cheminformatics* **8**, 1 (2016).

[12] G. D. Purvis III & R. J. Bartlett. "A full coupled-cluster singles and doubles model: The inclusion of disconnected triples". *J. Chem. Phys.* **76**, 1910 (1982).

[13] J. B. Anderson. "A random-walk simulation of the Schrödinger equation: H+ 3". *J. Chem. Phys.* **63**, 1499 (1975).

[14] B. M. Austin, D. Y. Zubarev & W. A. Lester Jr. "Quantum Monte Carlo and related approaches". *Chem. Rev.* **112**, 263 (2012).

[15] N. Mardirossian & M. Head-Gordon. "Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals". *Mol. Phys.* **115**, 2315 (2017).

[16] M. Stöhr, T. Van Voorhis & A. Tkatchenko. "Theory and practice of modeling van der Waals interactions in electronic-structure calculations". *Chem. Soc. Rev.* **48**, 4118 (2019).

[17] D. J. Livingstone. "The characterization of chemical structures using molecular properties. A survey". *J. Chem. Inf. Comput. Sci.* **40**, 195 (2000).

[18] J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller & A. Tkatchenko. "Combining machine learning and computational chemistry for predictive insights into chemical systems". *Chem. Rev.* **121**, 9816 (2021).

[19] O. Wieder, S. Kohlbacher, M. Kuenemann, A. Garon, P. Ducrot, T. Seidel & T. Langer. "A compact review of molecular property prediction with graph neural networks". *Drug Discov. Today: Technol.* **37**, 1 (2020).

[20] A. Alakhdar, B. Poczos & N. Washburn. "Diffusion models in de novo drug design". *J. Chem. Inf. Model.* **64**, 7238 (2024).

[21] O. Zhang, H. Lin, H. Zhang, H. Zhao, Y. Huang, C.-Y. Hsieh, P. Pan & T. Hou. "Deep lead optimization: leveraging generative AI for structural modification". *J. Am. Chem. Soc.* **146**, 31357 (2024).

[22] D. C. Elton, Z. Boukouvalas, M. D. Fuge & P. W. Chung. "Deep learning for molecular design—a review of the state of the art". *Mol. Syst. Des. & Eng.* **4**, 828 (2019).

[23] O. T. Unke *et al.* "Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments". *Sci. Adv.* **10**, eadn4397 (2024).

[24] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner & G. Csányi. "MACE: Higher order equivariant message passing neural networks for fast and accurate force fields". *Adv. Neural Inf. Process. Syst.* **35**, 11423 (2022).

[25] R. Galvelis, A. Varela-Rial, S. Doerr, R. Fino, P. Eastman, T. E. Markland, J. D. Chodera & G. De Fabritiis. "NNP/MM: Accelerating molecular dynamics simulations with machine learning potentials and molecular mechanics". *J. Chem. Inf. Model.* **63**, 5701 (2023).

[26] K. Takaba *et al.* "Machine-learned molecular mechanics force fields from large-scale quantum chemical data". *Chem. Sci.* **15**, 12861 (2024).

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser & I. Polosukhin. "Attention is all you need". *Adv. Neural Inf. Process. Syst.* **30**, 5998 (2017).

[28] V. G. Satorras, E. Hoogeboom & M. Welling. "E (n) equivariant graph neural networks". In *International Conference on Machine Learning*, 9323–9332 (PMLR, 2021).

[29] J. Jumper *et al.* "Highly accurate protein structure prediction with AlphaFold". *Nature* **596**, 583 (2021).

[30] J. Pereira, A. J. Simpkin, M. D. Hartmann, D. J. Rigden, R. M. Keegan & A. N. Lupas. "High-accuracy protein structure prediction in CASP14". *Proteins: Struct. Funct. Bioinform.* **89**, 1687 (2021).

[31] O. T. Unke, S. Chmiela, M. Gastegger, K. T. Schütt, H. E. Sauceda & K.-R. Müller. "SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects". *Nat. Commun.* **12**, 7273 (2021).

[32] O. T. Unke & M. Meuwly. "PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges". *J. Chem. Theory Comput.* **15**, 3678 (2019).

[33] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner & G. Csányi. "MACE: Higher order equivariant message passing neural networks for fast and accurate force fields". *Adv. Neura.l Inf. Process. Syst.* **35**, 11423 (2022).

[34] J. T. Frank, O. T. Unke & K.-R. Müller. "So3krates–self-attention for higher-order geometric interactions on arbitrary length-scales". *arXiv* 2205.14276 (2022).

[35] A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth & B. Kozinsky. "Learning local equivariant representations for large-scale atomistic dynamics". *Nat. Commun.* **14**, 579 (2023).

[36] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt & B. Kozinsky. "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials". *Nat. Commun.* **13**, 2453 (2022).

[37] T. Plé, L. Lagardère & J.-P. Piquemal. "Force-field-enhanced neural network interactions: from local equivariant embedding to atom-in-molecule properties and long-range effects". *Chem. Sci.* **14**, 12554 (2023).

[38] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt & B. Kozinsky. "E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials". *Nat. Commun.* **13**, 2453 (2022).

[39] K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko & K.-R. Müller. "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions". *Adv. Neural Inf. Process. Syst.* **30** (2017).

[40] J. S. Smith, O. Isayev & A. E. Roitberg. "ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost". *Chem. Sci.* **8**, 3192 (2017).

[41] R. P. Pelaez, G. Simeon, R. Galvelis, A. Mirarchi, P. Eastman, S. Doerr, P. Tholke, T. E. Markland & G. De Fabritiis. "Torchmd-net 2.0: Fast neural network potentials for molecular simulations". *J. Chem. Theory Comput.* **20**, 4076 (2024).

[42] J. Gasteiger, S. Giri, J. T. Margraf & S. Günnemann. "Fast and uncertainty-aware directional message passing for non-equilibrium molecules". *arXiv* 2011.14115 (2020).

[43] G. Wang, C. Wang, X. Zhang, Z. Li, J. Zhou & Z. Sun. "Machine learning interatomic potential: Bridge the gap between small-scale models and realistic device-scale simulations". *Iscience* **27** (2024).

[44] M. Kulichenko *et al.* "Data generation for machine learning interatomic potentials and beyond". *Chem. Rev.* **124**, 13681 (2024).

[45] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko & K.-R. Müller. "Machine learning force fields". *Chem. Rev.* **121**, 10142 (2021).

[46] Y. Wang *et al.* "On the design space between molecular mechanics and machine learning force fields". *Appl. Phys. Rev.* **12** (2025).

[47] M. Amezcua, J. Setiadi & D. L. Mobley. "The SAMPL9 host-guest blind challenge: an overview of binding free energy predictive accuracy". *Phys. Chem. Chem. Phys.* **26**, 9207 (2024).

[48] S. Ackloo *et al.* "CACHE (Critical assessment of computational hit-finding experiments): A public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding". *Nat. Rev. Chem.* **6**, 287 (2022).

[49] H. Lm *et al.* "The seventh blind test of crystal structure prediction: Structure ranking methods". *Struct. Sci.* **80** (2024).

[50] H. Fooladi, T. N. L. Vu & J. Kirchmair. "Evaluating machine learning models for molecular property prediction: Performance and robustness on out-of-distribution data". *ChemRxiv* 10.26434/chemrxiv–2025–g1vjf–v2 (2025).

[51] V. Cărare, F. L. Thiemann, J. Morrow, D. J. Wales, E. O. Pyzer-Knapp & L. Dicks. "Global properties of the energy landscape: a testing and training arena for machine learned potentials". *arXiv* 2508.16425 (2025).

[52] A. Karton & M. T. De Oliveira. "Good practices in database generation for benchmarking density functional theory". *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **15**, e1737 (2025).

[53] J. Rezác, K. E. Riley & P. Hobza. "S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures". *J. Chem. Theory Comput.* **7**, 2427 (2011).

[54] J. Rezác, K. E. Riley & P. Hobza. "Extensions of the S66 data set: More accurate interaction energies and angular-displaced nonequilibrium geometries". *J. Chem. Theory Comput.* **7**, 3466 (2011).

[55] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi & S. Grimme. "A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions". *Phys. Chem. Chem. Phys.* **19**, 32184 (2017).

[56] L. R. M. Sandonas *et al.* "Dataset for quantum-mechanical exploration of conformers and solvent effects in large drug-like molecules". *Sci. Data* **11**, 742 (2024).

[57] M. Puleva, L. Medrano Sandonas, B. Lőrincz, J. Charry, D. M. Rogers, P. R. Nagy & A. Tkatchenko. "Extending quantum-mechanical benchmark accuracy to biological ligand-pocket interactions". *ChemRxiv* 10.26434/chemrxiv−2025−f6615 (2025).

[58] C. Adamo & V. Barone. "Toward reliable density functional methods without adjustable parameters: The PBE0 model". *J. Chem. Phys.* **110**, 6158 (1999).

[59] A. Tkatchenko, R. A. DiStasio Jr, R. Car & M. Scheffler. "Accurate and efficient method for many-body van der Waals interactions". *Phys. Rev. Lett.* **108**, 236402 (2012).

[60] A. Ambrosetti, A. M. Reilly, R. A. DiStasio & A. Tkatchenko. "Long-range correlation energy calculated from coupled atomic response functions". *J. Chem. Phys.* **140**, 18A508 (2014).

[61] I. Poltavsky *et al.* "Crash testing machine learning force fields for molecules, materials, and interfaces: model analysis in the tea challenge 2023". *Chem. Sci.* **16**, 3720 (2025).

[62] I. Poltavsky *et al.* "Crash testing machine learning force fields for molecules, materials, and interfaces: molecular dynamics in the TEA challenge 2023". *Chem. Sci.* **16**, 3738 (2025).

[63] L. R. M. Sandonas, M. Puleva, R. P. Payano, M. Stöhr, G. Cuniberti & A. Tkatchenko. "Advancing Density Functional Tight-Binding method for large organic molecules through equivariant neural networks". *ChemRxiv* 10.26434/chemrxiv–2025–z3mhh (2025).

[64] M. Gaus, Q. Cui & M. Elstner. "DFTB3: extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB)". *J. Chem. Theory Comput.* **7**, 931 (2011).

[65] M. Gaus, A. Goez & M. Elstner. "Parametrization and benchmark of DFTB3 for organic molecules". *J. Chem. Theory Comput.* **9**, 338 (2013).

[66] Z. Jin *et al.* "Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors". *Nature* **582**, 289 (2020).

[67] A. Szabo & N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory* (Dover Publications, 1989).

[68] F. Jensen. *Introduction to Computational Chemistry* (Wiley, 2017), 3rd edn.

[69] E. Schrödinger. "Quantisierung als eigenwertproblem (erste mitteilung)". *Ann. Der Phys.* **384**, 361 (1926).

[70] E. Schrödinger. "Quantisierung als eigenwertproblem (zweite mitteilung)". *Ann. Der Phys.* **384**, 489 (1926).

[71] E. Schrödinger. "Quantisierung als eigenwertproblem (dritte mitteilung)". *Ann. Der Phys.* **385**, 437 (1926).

[72] E. Schrödinger. "Quantisierung als eigenwertproblem (vierte mitteilung)". *Ann. Der Phys.* **386**, 109 (1926).

[73] L. D. Landau & E. M. Lifshitz. *Quantum mechanics: non-relativistic theory*, vol. 3 (Elsevier, 2013).

[74] M. Born & J. R. Oppenheimer. "Zur quantentheorie der molekeln". *Ann. Der Phys.* **389**, 457 (1927).

[75] V. Fock. "„Selfconsistent field" mit austausch für Natrium". *Zeitschrift FÜR Phys.* **62**, 795 (1930).

[76] J. C. Slater. "The theory of complex spectra". *Phys. Rev.* **34**, 1293 (1929).

[77] S. Giarrusso & A. Pribram-Jones. "Comparing correlation components and approximations in Hartree–Fock and Kohn–Sham theories via an analytical test case study". *J. Chem. Phys.* **157** (2022).

[78] J. G. Hill. "Gaussian basis sets for molecular applications". *Int. J. Quantum Chem.* **113**, 21 (2013).

[79] J. Olsen. "An introduction and overview of basis sets for molecular and solid-state calculations". In *Basis Sets in Computational Chemistry*, 1–16 (Springer, 2021).

[80] T. H. Dunning Jr. "Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen". *J. Chem. Phys.* **90**, 1007 (1989).

[81] R. A. Kendall, T. H. Dunning Jr & R. J. Harrison. "Electron affinities of the first-row atoms revisited. systematic basis sets and wave functions". *J. Chem. Phys.* **96**, 6796 (1992).

[82] D. E. Woon & T. H. Dunning Jr. "Gaussian basis sets for use in correlated molecular calculations. IV. Calculation of static electrical response properties". *J. Chem. Phys.* **100**, 2975 (1994).

[83] I. Y. Zhang, X. Ren, P. Rinke, V. Blum & M. Scheffler. "Numeric atom-centered-orbital basis sets with valence-correlation consistency from H to Ar". *New J. Phys.* **15**, 123033 (2013).

[84] P. Hohenberg & W. Kohn. "Density functional theory (DFT)". *Phys. Rev.* **136**, B864 (1964).

[85] W. Kohn & L. J. Sham. "Self-consistent equations including exchange and correlation effects". *Phys. Rev.* **140**, A1133 (1965).

[86] J. P. Perdew & K. Schmidt. "Jacob's ladder of density functional approximations for the exchange-correlation energy". In *AIP Conference Proceedings*, vol. 577, 1–20 (American Institute of Physics, 2001).

[87] A. J. Cohen, P. Mori-Sánchez & W. Yang. "Challenges for density functional theory". *Chem. Rev.* **112**, 289 (2012).

[88] J. P. Perdew, K. Burke & M. Ernzerhof. "Generalized gradient approximation made simple". *Phys. Rev. Lett.* **77**, 3865 (1996).

[89] A. D. Becke. "Density-functional exchange-energy approximation with correct asymptotic behavior". *Phys. Rev. A* **38**, 3098 (1988).

[90] C. Lee, W. Yang & R. G. Parr. "Development of the colle-salvetti correlation-energy formula into a functional of the electron density". *Phys. Rev. B* **37**, 785 (1988).

[91] E. H. Lieb & S. Oxford. "Improved lower bound on the indirect coulomb energy". *Int. J. Quantum Chem.* **19**, 427 (1981).

[92] M. Lewin, E. H. Lieb & R. Seiringer. "Improved Lieb–Oxford bound on the indirect and exchange energies". *Lett. Math. Phys.* **112**, 92 (2022).

[93] A. D. Becke. "A new mixing of Hartree-Fock and local density-functional theories". *J. Chem. Phys.* **98**, 1372 (1993).

[94] J. A. Pople, M. Head-Gordon, D. J. Fox, K. Raghavachari & L. A. Curtiss. "Gaussian-1 theory: A general procedure for prediction of molecular energies". *J. Chem. Phys.* **90**, 5622 (1989).

[95] L. A. Curtiss, K. Raghavachari, G. W. Trucks & J. A. Pople. "Gaussian-2 theory for molecular energies of first-and second-row compounds". *J. Chem. Phys.* **94**, 7221 (1991).

[96] S. H. Vosko, L. Wilk & M. Nusair. "Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis". *Can. J. Phys.* **58**, 1200 (1980).

[97] A. D. Becke. "Density-functional exchange-energy approximation with correct asymptotic behavior". *Phys. Rev. A* **38**, 3098 (1988).

[98] Y. Zhao & D. G. Truhlar. "The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals". *Theor. Chem. Acc.* **120**, 215 (2008).

[99] J. P. Perdew, M. Ernzerhof & K. Burke. "Rationale for mixing exact exchange with density functional approximations". *J. Chem. Phys.* **105**, 9982 (1996).

[100] Y.-S. Lin, G.-D. Li, S.-P. Mao & J.-D. Chai. "Long-range corrected hybrid density functionals with improved dispersion corrections". *J. Chem. Theory Comput.* **9**, 263 (2013).

[101] N. Mardirossian & M. Head-Gordon. "ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation". *J. Chem. Phys.* **144**, 214110 (2016).

[102] O. A. Vydrov & T. Van Voorhis. "Nonlocal van der Waals density functional: The simpler the better". *J. Chem. Phys.* **133**, 244103 (2010).

[103] T. Yanai, D. P. Tew & N. C. Handy. "A new hybrid exchange–correlation functional using the coulomb-attenuating method (CAM-B3LYP)". *Chem. Phys. Lett.* **393**, 51 (2004).

[104] X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter & M. Scheffler. "Resolution-of-identity approach to Hartree–Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions". *New J. Phys.* **14**, 053020 (2012).

[105] M. Head-Gordon, J. A. Pople & M. J. Frisch. "MP2 energy evaluation by direct methods". *Chem. Phys. Lett.* **153**, 503 (1988).

[106] K. E. Riley, J. A. Platts, J. Rezac, P. Hobza & J. G. Hill. "Assessment of the performance of MP2 and MP2 variants for the treatment of noncovalent interactions". *J. Phys. Chem. A* **116**, 4159 (2012).

[107] É. Brémond, J. C. Sancho-García, A. J. Pérez-Jiménez & C. Adamo. "Excitation energies of polycyclic aromatic hydrocarbons by double-hybrid functionals: Assessing the PBE0-DH and PBE-QIDH models and their range-separated versions". *J. Chem. Phys.* **158**, 044105 (2023).

[108] A. Stone. *The theory of intermolecular forces* (Oxford University Press, 2013).

[109] A. D. Becke & E. R. Johnson. "A density-functional model of the dispersion interaction". *J. Chem. Phys.* **123**, 154101 (2005).

[110] A. Otero-De-La-Roza & E. R. Johnson. "Non-covalent interactions and thermochemistry using XDM-corrected hybrid and range-separated hybrid density functionals". *J. Chem. Phys.* **138**, 204109 (2013).

[111] A. Tkatchenko & M. Scheffler. "Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data". *Phys. Rev. Lett.* **102**, 073005 (2009).

[112] S. Grimme, J. Antony, S. Ehrlich & H. Krieg. "A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-PU". *J. Chem. Phys.* **132** (2010).

[113] E. Caldeweyher, C. Bannwarth & S. Grimme. "Extension of the D3 dispersion coefficient model". *J. Chem. Phys.* **147**, 034112 (2017).

[114] J. Hermann & A. Tkatchenko. "Density functional model for van der Waals interactions: Unifying many-body atomic approaches with nonlocal functionals". *Phys. Rev. Lett.* **124**, 146401 (2020).

[115] H. B. Casimir & D. Polder. "The influence of retardation on the London-van der Waals forces". *Phys. Rev.* **73**, 360 (1948).

[116] E. R. Johnson & A. D. Becke. "A post-hartree-fock model of intermolecular interactions: Inclusion of higher-order corrections". *J. Chem. Phys.* **124** (2006).

[117] S. Grimme, S. Ehrlich & L. Goerigk. "Effect of the damping function in dispersion corrected density functional theory". *J. Comput. Chem.* **32**, 1456 (2011).

[118] E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth & S. Grimme. "A generally applicable atomic-charge dependent London dispersion correction". *J. Chem. Phys.* **150**, 154122 (2019).

[119] S. Grimme, S. Ehrlich & L. Goerigk. "Effect of the damping function in dispersion corrected density functional theory". *J. Comput. Chem.* **32**, 1456 (2011).

[120] B. Axilrod & E. Teller. "Interaction of the van der waals type between three atoms". *J. Chem. Phys.* **11**, 299 (1943).

[121] A. Tkatchenko, R. A. DiStasio Jr, R. Car & M. Scheffler. "Accurate and efficient method for many-body van der Waals interactions". *Phys. Rev. Lett.* **108**, 236402 (2012).

[122] A. Ambrosetti, A. M. Reilly, R. A. DiStasio & A. Tkatchenko. "Long-range correlation energy calculated from coupled atomic response functions". *J. Chem. Phys.* **140**, 18A508 (2014).

[123] M. Mortazavi, J. G. Brandenburg, R. J. Maurer & A. Tkatchenko. "Structure and stability of molecular crystals with many-body dispersion-inclusive density functional tight binding". *J. Phys. Chem. Lett.* **9**, 399 (2018).

[124] A. Tkatchenko & M. Scheffler. "Accurate molecular van der waals interactions from ground-state electron density format and free-atom reference data". *Phys. Rev. Lett.* **102**, 073005 (2009).

[125] S. I. Allec, Y. Sun, J. Sun, C.-e. A. Chang & B. M. Wong. "Heterogeneous CPU+ GPU-enabled simulations for DFTB molecular dynamics of large chemical and biological systems". *J. Chem. Theory Comput.* **15**, 2807 (2019).

[126] C. Bannwarth, S. Ehlert & S. Grimme. "GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions". *J. Chem. Theory Comput.* **15**, 1652 (2019).

[127] G. Seifert, D. Porezag & T. Frauenheim. "Calculations of molecules, clusters, and solids with a simplified LCAO-DFT-LDA scheme". *Int. J. Quantum Chem.* **58**, 185 (1996).

[128] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai & G. Seifert. "Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties". *Phys. Rev. B* **58**, 7260 (1998).

[129] M. Gaus, Q. Cui & M. Elstner. "DFTB3: Extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB)". *J. Chem. Theory Comput.* **7**, 931 (2011).

[130] P. Koskinen & V. Mäkinen. "Density-functional tight-binding for beginners". *Comput. Mater. Sci.* **47**, 237 (2009).

[131] R. S. Mulliken. "Electronic population analysis on LCAO–MO molecular wave functions. I". *J. Chem. Phys.* **23**, 1833 (1955).

[132] M. Gaus, A. Goez & M. Elstner. "Parametrization and benchmark of DFTB3 for organic molecules". *J. Chem. Theory Comput.* **9**, 338 (2013).

[133] M. Gaus, X. Lu, M. Elstner & Q. Cui. "Parameterization of DFTB3/3OB for sulfur and phosphorus for chemical and biological applications". *J. Chem. Theory Comput.* **10**, 1518 (2014).

[134] G. Seifert. "Tight-binding density functional theory: an approximate Kohn-Sham DFT scheme". *J. Phys. Chem. A* **111**, 5609–5613 (2007).

[135] A. S. Krishnapriyan, A. S. Krishnapriyan, P. Yang, A. M. N. Niklasson & M. J. Cawkwell. "Numerical optimization of density functional tight binding models: Application to molecules containing Carbon, Hydrogen, Nitrogen, and Oxygen.". *J. Chem. Theory Comput.* **13**, 6191–6200 (2017).

[136] N. Goldman, L. E. Fried, R. Lindsey, C. H. Pham & R. Dettori. "Enhancing the accuracy of density functional tight binding models through ChIMES many-body interaction potentials.". *J. Chem. Phys.* **158**, 144112 (2023).

[137] P. Ren & J. W. Ponder. "Polarizable atomic multipole water model for molecular mechanics simulation". *J. Phys. Chem. B* **107**, 5933 (2003).

[138] Y. Shi, Z. Xia, J. Zhang, R. Best, C. Wu, J. W. Ponder & P. Ren. "Polarizable atomic multipole-based AMOEBA force field for proteins". *J. Chem. Theory Comput.* **9**, 4046 (2013).

[139] X. Lu, J. Chen & J. Huang. "The continuous evolution of biomolecular force fields". *Structure* **33**, 1138 (2025).

[140] J. Huang & A. D. MacKerell Jr. "CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data". *J. Comput. Chem.* **34**, 2135 (2013).

[141] D. Case *et al.* "AMBER 2016" (2016). Publisher: University of California, San Francisco.

[142] K. Vanommeslaeghe *et al.* "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields". *J. Comput. Chem.* **31**, 671 (2010).

[143] M. A. González. "Force fields and molecular dynamics simulations". *Écol. thém. Soc. Fr. Neutron* **12**, 169 (2011).

[144] A. Jakalian, D. B. Jack & C. I. Bayly. "Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation". *J. Comput. Chem.* **23**, 1623 (2002).

[145] X. He, V. H. Man, W. Yang, T.-S. Lee & J. Wang. "A fast and high-quality charge model for the next generation general AMBER force field". *J. Chem. Phys.* **153** (2020).

[146] G. Henkelman, B. P. Uberuaga & H. Jónsson. "A climbing image nudged elastic band method for finding saddle points and minimum energy paths". *J. Chem. Phys.* **113**, 9901 (2000).

[147] D. H. Mathews & D. A. Case. "Nudged elastic band calculation of minimal energy paths for the conformational change of a GG non-canonical pair". *J. Mol. Biol.* **357**, 1683 (2006).

[148] A. W. Ruttinger, D. Sharma & P. Clancy. "Protocol for directing nudged elastic band calculations to the minimum energy pathway: Nurturing errant calculations back to convergence". *J. Chem. Theory Comput.* **18**, 2993 (2022).

[149] T. Schneider & E. Stoll. "Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions". *Phys. Rev. B* **17**, 1302 (1978).

[150] G. A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*, vol. 60 of *Texts Appl. Math.* (Springer, 2014).

[151] A. Hjorth Larsen *et al.* "The atomic simulation environment – a Python library for working with atoms". *J. Phys. Condens. Matter* **29**, 273002 (2017).

[152] S. Chmiela, H. E. Sauceda, I. Poltavsky & A. Müller, K.-R.and Tkatchenko. "sGDML: Constructing accurate and data efficient molecular force fields using machine learning". *Comput. Phys. Commun.* **240**, 38 (2019).

[153] I. Batatia, D. P. Kovács, G. N. Simm, C. Ortner & G. Csányi. "Mace: Higher order equivariant message passing neural networks for fast and accurate force fields". *arXiv* 2206.07697 (2022).

[154] E. Noether. "Invarianten beliebiger differentialausdrücke". *Nachr. Ges. Wiss. Göttingen, Math.-Phys. Kl.* **1918**, 37 (1918).

[155] F. Fuchs, D. Worrall, V. Fischer & M. Welling. "Se (3)-transformers: 3d roto-translation equivariant attention networks". *Adv. Neural Inf. Process. Syst.* **33**, 1970 (2020).

[156] F. Noé, A. Tkatchenko, K.-R. Müller & C. Clementi. "Machine learning for molecular simulation". *Annual review of physical chemistry* **71**, 361 (2020).

[157] M. Rupp, A. Tkatchenko, K.-R. Müller & O. A. von Lilienfeld. "Fast and accurate modeling of molecular atomization energies with machine learning". *Phys. Rev. Lett.* **108** (2012).

[158] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller & A. Tkatchenko. "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space". *J. Phys. Chem. Lett.* **6** (2015).

[159] B. Huang & O. A. Von Lilienfeld. "Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity". *J. Chem. Phys.* **145** (2016).

[160] W. Pronobis, A. Tkatchenko & K.-R. Müller. "Many-Body Descriptors for Predicting Molecular Properties with Machine Learning: Analysis of Pairwise and Three-Body Interactions in Molecules". *J. Chem. Theory Comput.* **14**, 2991 (2018).

[161] B. Huang & O. A. von Lilienfeld. "Quantum machine learning using atom-in-molecule-based fragments selected on the fly". *Nat. Chem.* **12**, 945 (2020).

[162] A. P. Bartók, R. Kondor & G. Csányi. "On representing chemical environments". *Phys. Rev. B* **87**, 184115 (2013).

[163] A. S. Christensen, L. A. Bratholm, F. A. Faber & O. A. von Lilienfeld. "FCHL revisited: Faster and more accurate quantum machine learning". *J. Chem. Phys.* **152** (2020).

[164] T. B. Blank, S. D. Brown, A. W. Calhoun & D. J. Doren. "Neural network models of potential energy surfaces". *J. Chem. Phys.* **103**, 4129 (1995).

[165] J. Behler & M. Parrinello. "Generalized neural-network representation of high-dimensional potential-energy surfaces". *Phys. Rev. Lett.* **98**, 146401 (2007).

[166] M. Geiger *et al.* "Euclidean neural networks: e3nn" (2022). Avialable at https://doi.org/10.5281/zenodo.6459381.

[167] F. L. Hirshfeld. "Bonded-atom fragments for describing molecular charge densities". *Theor. Chim. Acta* **44**, 129 (1977).

[168] A. Kabylda, J. T. Frank, S. S. Dou, A. Khabibrakhmanov, L. M. Sandonas, O. T. Unke, S. Chmiela, K.-R. Müller & A. Tkatchenko. "Molecular simulations with a pretrained neural network and universal pairwise force fields". *ChemRxiv* 10.26434/chemrxiv–2024–bdfr0–v3 (2025).

[169] Z. Wang *et al.* "Advanced Graph and Sequence Neural Networks for Molecular Property Prediction and Drug Discovery.". *Bioinformatics* btac112 (2022).

[170] M. Welling. "A first encounter with machine learning". *Irvine CA.: Univ. Calif.* **12** (2011).

[171] T. Hofmann, B. Schölkopf & A. J. Smola. "Kernel methods in machine learning". *Ann. Stat.* **36**, 1171 (2008).

[172] O. A. von Lilienfeld & K. Burke. "Retrospective on a decade of machine learning for chemical discovery". *Nat. Commun.* **11**, 4895 (2020).

[173] O. C., A. O. von Lilienfeld & B. Baumeier. "Wasserstein metric for improved quantum machine learning with adjacency matrix representations". *Mach. Learn. Sci. Technol.* **1** (2020).

[174] B. Matérn. "Spatial variation, volume 49 of meddelanden från statens skogsforskningsinstitut". *Stock. Statens Skogsforskningsinstitut* (1960).

[175] P. Whittle. "Stochastic-processes in several dimensions". *Bull. Int. Stat. Inst.* **40**, 974 (1963).

[176] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti & G. Csányi. "Gaussian process regression for materials and molecules". *Chem. Rev.* **121**, 10073 (2021).

[177] A. Subasi. "Chapter 3 - machine learning techniques". In A. Subasi (ed.) *Practical Machine Learning for Data Analysis Using Python*, 91–202 (Academic Press, 2020).

[178] J. F. Ziegler & J. P. Biersack. "The stopping and range of ions in matter". In *Treatise on heavy-ion science*, 93–129 (Springer, 1985).

[179] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller & A. Tkatchenko. "Quantum-chemical insights from deep tensor neural networks". *Nat. Commun.* **8**, 13890 (2017).

[180] A. Einstein, B. Podolsky & N. Rosen. "Can quantum-mechanical description of physical reality be considered complete?". *Phys. Rev.* **47**, 777 (1935).

[181] D. Lingė *et al.* "PLBD: protein–ligand binding database of thermodynamic and kinetic intrinsic parameters". *Database* **2023**, baad040 (2023).

[182] U. Ryde & P. Söderhjelm. "Ligand-binding affinity estimates supported by quantum-mechanical methods". *Chem. Rev.* **116**, 5520 (2016).

[183] D. Mucs & R. A. Bryce. "The application of quantum mechanics in structure-based drug design". *Expert. Opin. Drug Discov.* **8**, 263 (2013).

[184] F. Sohraby & A. Nunes-Alves. "Advances in computational methods for ligand binding kinetics". *Trends Biochem. Sci.* **48**, 437 (2023).

[185] I. Jarmoskaite, I. AlSadhan, P. P. Vaidyanathan & D. Herschlag. "How to measure and evaluate binding affinities". *Elife* **9**, e57264 (2020).

[186] U. Ryde & P. Soderhjelm. "Ligand-binding affinity estimates supported by quantum-mechanical methods". *Chem. Rev.* **116**, 5520 (2016).

[187] H. J. Davis & R. J. Phipps. "Harnessing non-covalent interactions to exert control over regioselectivity and site-selectivity in catalytic reactions". *Chem. Sci.* **8**, 864 (2017).

[188] G. A. Ross, C. Lu, G. Scarabelli, S. K. Albanese, E. Houang, R. Abel, E. D. Harder & L. Wang. "The maximal and current accuracy of rigorous protein-ligand binding free energy calculations". *Commun. Chem.* **6**, 222 (2023).

[189] R. Abel, L. Wang, D. L. Mobley & R. A. Friesner. "A critical review of validation, blind testing, and real-world use of alchemical protein-ligand binding free energy calculations". *Curr. Top. Med. Chem.* **17**, 2577 (2017).

[190] J. D. Chodera, D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson & V. S. Pande. "Alchemical free energy methods for drug discovery: progress and challenges". *Curr. Opin. Struct. Biol.* **21**, 150 (2011).

[191] P. Dauber-Osguthorpe & A. T. Hagler. "Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there?". *J. Comput. Mol. Des.* **33**, 133 (2019).

[192] M. Stöhr & A. Tkatchenko. "Quantum mechanics of proteins in explicit water: The role of plasmon-like solute-solvent interactions". *Sci. Adv.* **5**, eaax0024 (2019).

[193] R. J. Bartlett & M. Musiał. "Coupled-cluster theory in quantum chemistry". *Rev. Mod. Phys.* **79**, 291 (2007).

[194] W. M. C. Foulkes, L. Mitas, R. J. N. Rajagopal & G. Rajagopal. "Quantum Monte Carlo simulations of solids". *Rev. Mod. Phys.* **73**, 33 (2001).

[195] M. Dubecký, L. Mitas & P. Jurečka. "Noncovalent Interactions by Quantum Monte Carlo". *Chem. Rev.* **116**, 5188 (2016).

[196] F. Becca & S. Sorella. *Quantum Monte Carlo approaches for correlated systems* (Cambridge University Press, 2017).

[197] P. Eastman *et al.* "Spice, a dataset of drug-like molecules and peptides for training machine learning potentials". *Sci. Data* **10**, 11 (2023).

[198] S. A. Spronk, Z. L. Glick, D. P. Metcalf, C. D. Sherrill & D. L. Cheney. "A quantum chemical interaction energy dataset for accurately modeling protein-ligand interactions". *Sci. Data* **10**, 619 (2023).

[199] D. S. Levine *et al.* "The open molecules 2025 (OMol25) dataset, evaluations, and models". *arXiv* 2505.08762 (2025).

[200] S. Ackloo *et al.* "A target class ligandability evaluation of WD40 repeat-containing proteins". *J. Med. Chem.* **68**, 1092 (2024).

[201] E. C. Meng, T. D. Goddard, E. F. Pettersen, G. S. Couch, Z. J. Pearson, J. H. Morris & T. E. Ferrin. "UCSF ChimeraX: Tools for structure building and analysis". *Protein Sci.* **32**, e4792 (2023).

[202] J. Hoja, L. Medrano Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio Jr & A. Tkatchenko. "QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules". *Sci. Data* **8**, 43 (2021).

[203] J. S. Smith, O. Isayev & A. E. Roitberg. "ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules". *Sci. Data* **4**, 170193 (2017).

[204] C. Isert, K. Atz, J. Jiménez-Luna & G. Schneider. "QMugs, quantum mechanical properties of drug-like molecules". *Sci. Data* **9**, 273 (2022).

[205] S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko & K.-R. Müller. "Accurate global machine learning force fields for molecules with hundreds of atoms". *Sci. Adv.* **9**, eadf0873 (2023).

[206] P. Jurečka, J. Šponer, J. Černỳ & P. Hobza. "Benchmark database of accurate (MP2 and CCSD (T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs". *Phys. Chem. Chem. Phys.* **8**, 1985 (2006).

[207] P. Hobza. "Calculations on noncovalent interactions and databases of benchmark interaction energies". *Acc. Chem. Res.* **45**, 663 (2012).

[208] R. Sedlak, T. Janowski, M. Pitoňák, J. Řezáč, P. Pulay & P. Hobza. "Accuracy of quantum chemical methods for large noncovalent complexes". *J. Chem. Theory Comput.* **9**, 3364 (2013).

[209] Z. M. Sparrow, B. G. Ernst, P. T. Joo, K. U. Lao & R. A. DiStasio. "NENCI-2021. I. A large benchmark database of non-equilibrium non-covalent interactions emphasizing close intermolecular contacts". *J. Chem. Phys.* **155**, 184303 (2021).

[210] A. G. Donchev *et al.* "Quantum chemical benchmark databases of gold-standard dimer interaction energies". *Sci. Data* **8**, 55 (2021).

[211] C. Villot & K. U. Lao. "Ab initio dispersion potentials based on physics-based functional forms with machine learning". *J. Chem. Phys.* **160**, 184103 (2024).

[212] R. Wang, X. Fang, Y. Lu & S. Wang. "The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures". *J. Med. Chem.* **47**, 2977 (2004).

[213] A. Siwach & P. K. Verma. "Synthesis and therapeutic potential of imidazole containing compounds". *BMC Chem.* **15**, 12 (2021).

[214] R. Ferreira de Freitas & M. Schapira. "A systematic analysis of atomic protein-ligand interactions in the PDB". *Med. Chem. Commun.* **8**, 1970 (2017).

[215] M. Gao & J. Skolnick. "A comprehensive survey of small-molecule binding pockets in proteins". *PLoS Comput. Biol.* **9**, e1003302 (2013).

[216] D. K. Johnson & J. Karanicolas. "Druggable protein interaction sites are more predisposed to surface pocket formation than the rest of the protein surface". *PLoS Comput. Biol.* **9**, e1002951 (2013).

[217] V. Blum, F. H. Ralf Gehrke, P. Havu, V. Havu, X. Ren, K. Reuter & M. Scheffler. "Ab initio molecular simulations with numeric atom-centered orbitals". *Comput. Phys. Comm.* **180**, 2175 (2009).

[218] N. Mardirossian & M. Head-Gordon. "ωB97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy". *Phys. Chem. Chem. Phys.* **16**, 9904 (2014).

[219] J. P. Perdew, K. Burke & M. Ernzerhof. "Generalized gradient approximation made simple". *Phys. Rev. Lett.* **77**, 3865 (1996).

[220] A. D. Becke. "Density-functional thermochemistry. III. The role of exact exchange". *J. Chem. Phys.* **98**, 5648 (1993).

[221] C. Lee, W. Yang & R. G. Parr. "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density". *Phys. Rev. B* **37**, 785 (1988).

[222] J. M. Turney *et al.* "Psi4: An open-source ab initio electronic structure program". *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2**, 556 (2012).

[223] S. Lehtola, C. Steigemann, M. J. Oliveira & M. A. Marques. "Recent developments in libxc—A comprehensive library of functionals for density functional theory". *SoftwareX* **7**, 1 (2018).

[224] Y. Shao *et al.* "Advances in molecular quantum chemistry contained in the Q-Chem 4 program package". *Mol. Phys.* **113**, 184 (2015).

[225] T. M. Parker, L. A. Burns, R. M. Parrish, A. G. Ryno & C. D. Sherrill. "Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies". *J. Chem. Phys.* **140**, 094106 (2014).

[226] B. Hourahine *et al.* "DFTB+, a software package for efficient approximate density functional theory based atomistic simulations". *J. Chem. Phys.* **152**, 124101 (2020).

[227] C. Bannwarth, S. Ehlert & S. Grimme. "GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions". *J. Chem. Theory Comput.* **15**, 1652 (2019).

[228] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch & G. R. Hutchison. "Open Babel: An open chemical toolbox". *J. Cheminformatics* **3**, 1 (2011).

[229] P. Eastman *et al.* "OpenMM 8: molecular dynamics simulation with machine learning potentials". *J. Phys. Chem. B* **128**, 109 (2023).

[230] K. Vanommeslaeghe, E. P. Raman & A. D. MacKerell Jr. "Automation of the CHARMM general force field (CGenFF) II: assignment of bonded parameters and partial atomic charges". *J. Chem. Inf. Model.* **52**, 3155 (2012).

[231] V. Blum, M. Rossi, S. Kokott & M. Scheffler. "The FHI-aims Code: All-electron, ab initio materials simulations towards the exascale". *arXiv* 2208.12335 (2022).

[232] J. Contreras-García, E. Johnson, S. Keinan, R. Chaudret, J. Piquemal, D. Beratan & W. Y. NCIPLOT. "NCIPLOT: A program for plotting noncovalent interaction regions". *J. Chem. Theory Comput.* **7**, 625 (2011).

[233] K. Kumar, S. M. Woo, T. Siu, W. A. Cortopassi, F. Duarte & R. S. Paton. "Cation–π interactions in protein–ligand binding: theory and data-mining reveal different roles for lysine and arginine". *Chem. Sci.* **9**, 2655 (2018).

[234] Y. S. Al-Hamdani, P. R. Nagy, A. Zen, D. Barton, M. Kállay, J. G. Brandenburg & A. Tkatchenko. "Interactions between large molecules pose a puzzle for reference quantum mechanical methods". *Nat. Commun.* **12**, 3927 (2021).

[235] B. X. Shi, F. Della Pia, Y. S. Al-Hamdani, A. Michaelides, D. Alfè & A. Zen. "Systematic discrepancies between reference methods for noncovalent interactions within the S66 dataset". *J. Chem. Phys.* **162** (2025).

[236] V. Fishman, M. Lesiuk, J. M. L. Martin & A. D. Boese. "A new angle on benchmarking noncovalent interactions". *arXiv* 2410.12603 (2024).

[237] P. R. Nagy & M. Kállay. "Optimization of the linear-scaling local natural orbital CCSD(T) method: Redundancy-free triples correction using Laplace transform". *J. Chem. Phys.* **146**, 214106 (2017).

[238] P. R. Nagy, G. Samu & M. Kállay. "Optimization of the linear-scaling local natural orbital CCSD(T) method: Improved algorithm and benchmark applications". *J. Chem. Theory Comput.* **14**, 4193 (2018).

[239] P. R. Nagy & M. Kállay. "Approaching the basis set limit of CCSD(T) energies for large molecules with local natural orbital coupled-cluster methods". *J. Chem. Theory Comput.* **15**, 5275 (2019).

[240] P. R. Nagy. "State-of-the-art local correlation methods enable affordable gold standard quantum chemistry for up to hundreds of atoms". *Chem. Sci.* **15**, 14556 (2024).

[241] Y. S. Al-Hamdani & A. Tkatchenko. "Understanding non-covalent interactions in larger molecular complexes from first principles". *J. Chem. Phys.* **150**, 010901 (2019).

[242] T. Schäfer, A. Irmler, A. Gallo & A. Grüneis. "Understanding discrepancies of wavefunction theories for large molecules". *arXiv* 2407.01442 (2024).

[243] V. Fishman, M. Lesiuk, J. M. L. Martin & A. Daniel Boese. "Another angle on benchmarking noncovalent interactions". *J. Chem. Theory Comput.* **21**, 2311 (2025).

[244] E. Semidalas, A. D. Boese & J. M. Martin. "Post-CCSD(T) corrections in the S66 noncovalent interactions benchmark". *Chem. Phys. Lett.* **863**, 141874 (2025).

[245] J. E. Alfonso-Ramos, C. Adamo, É. Brémond & T. Stuyver. "Improving the reliability of, and confidence in, dft functional benchmarking through active learning". *J. Chem. Theory Comput.* **21**, 1752 (2025).

[246] M. Gray & J. M. Herbert. "Density functional theory for van der Waals complexes: Size matters". *Ann. Rep. Comp. Chem.* **20**, 1 (2024).

[247] R. Sedlak, T. Janowski, M. Pitonak, J. Rezac, P. Pulay & P. Hobza. "Accuracy of quantum chemical methods for large noncovalent complexes". *J. Chem. Theory Comput.* **9**, 3364 (2013).

[248] J.-C. Sancho-Garcia, É. Brémond, M. Savarese, A. Pérez-Jiménez & C. Adamo. "Partnering dispersion corrections with modern parameter-free double-hybrid density functionals". *Phys. Chem. Chem. Phys.* **19**, 13481 (2017).

[249] K. Tang & J. P. Toennies. "An improved simple model for the van der Waals potential based on universal damping functions for the dispersion coefficients". *J. Chem. Phys.* **80**, 3726 (1984).

[250] J.-D. Chai & M. Head-Gordon. "Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections". *Phys. Chem. Chem. Phys.* **10**, 6615 (2008).

[251] A. Khabibrakhmanov, D. V. Fedorov & A. Tkatchenko. "Universal pairwise interatomic van der Waals potentials based on quantum drude oscillators". *J. Chem. Theory Comput.* **19**, 7895 (2023).

[252] K. Lejaeghere *et al.* "Reproducibility in density functional theory calculations of solids". *Science* **351**, aad3000 (2016).

[253] G. Fonseca, I. Poltavsky & A. Tkatchenko. "Force field analysis software and tools (FFAST): Assessing machine learning force fields under the microscope". *J. Chem. Theory Comput.* **19**, 8706 (2023).

[254] L. Medrano Sandondas, J. Hoja, B. G. Ernst, A. Vásquez-Mayagoitia, R. A. DiStasio & A. Tkatchenko. ""Freedom of design" in chemical compound space: towards rational in silico design of molecules with targeted quantum-mechanical properties". *Chem. Sci.* **14**, 10702 (2023).

[255] D. Hait & M. Head-Gordon. "When is a bond broken? The polarizability perspective.". *Angew. Chemie* **62**, e202312078 (2023).

[256] D. Mester *et al.* "An overview of developments in the MRCC program system". *J. Chem. Phys. A* **129**, 2086 (2025).

[257] E. Slootman, I. Poltavsky, R. Shinde, J. Cocomello, S. Moroni, A. Tkatchenko & C. Filippi. "Accurate quantum Monte Carlo forces for machine-learned force fields: Ethanol as a benchmark". *J. Chem. Theory Comput.* **20**, 6020 (2024).

[258] M. Hilfiker, L. Medrano Sandonas, M. Klähn, O. Engkvist & A. Tkatchenko. "Leveraging quantum mechanical properties to predict solvent effects on large drug-like molecules". In *AI in Drug Discovery*, 47–57 (Springer Nature Switzerland, Cham, 2025).

[259] S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko & K.-R. Müller. "Accurate global machine learning force fields for molecules with hundreds of atoms". *Sci. Adv.* **9**, eadf0873 (2023).

[260] P. Pracht *et al.* "CREST—A program for the exploration of low-energy molecular chemical space". *J. Chem. Phys.* **160**, 114110 (2024).

[261] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter & M. Scheffler. "Ab initio molecular simulations with numeric atom-centered orbitals". *Comput. Phys. Commun.* **180**, 2175 (2009).

[262] R. Jinnouchi, J. Lahnsteiner, F. Karsai, G. Kresse & M. Bokdam. "Phase transitions of hybrid perovskites simulated by machine-learning force fields trained on the fly with Bayesian inference". *Phys. Rev. Lett.* **122**, 225701 (2019).

[263] S. Stocker, H. Jung, G. Csányi, C. F. Goldsmith, K. Reuter & J. T. Margraf. "Estimating free energy barriers for heterogeneous catalytic reactions with machine learning potentials and umbrella integration". *J. Chem. Theory Comput.* **19**, 6796 (2023).

[264] H. Jung, L. Sauerland, S. Stocker, K. Reuter & J. T. Margraf. "Machine-learning driven global optimization of surface adsorbate geometries". *Npj Comput. Mater.* **9**, 114 (2023).

[265] X. Fu, Z. Wu, W. Wang, T. Xie, S. Keten, R. Gomez-Bombarelli & T. Jaakkola. "Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations". *arXiv* 10.1063/5.0147023 (2022).

[266] S. Stocker, J. Gasteiger, F. Becker, S. Günnemann & J. T. Margraf. "How robust are modern graph neural network potentials in long and hot molecular dynamics simulations?". *Mach. Learn. Sci. Technol.* **3**, 045010 (2022).

[267] J. Gawlikowski *et al.* "A survey of uncertainty in deep neural networks". *Artif. Intell. Rev.* **56**, 1513 (2023).

[268] A. Cheng *et al.* "How to do impactful research in artificial intelligence for chemistry and materials science". *arXiv E-Prints* arXiv–2409 (2024).

[269] A. Sultan, J. Sieg, M. Mathea & A. Volkamer. "Transformers for molecular property prediction: Lessons learned from the past five years". *J. Chem. Inf. Model.* **64**, 6259 (2024).

[270] G. Ramachandran, C. Ramakrishnan & V. Sasisekharan. "Stereochemistry of polypeptide chain configurations". *J. Mol. Biol.* **7**, 95 (1963).

[271] C. M. Venkatachalam. "Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units". *Biopolymers* **6**, 1425 (1968).

[272] G. T. Ramachandran & V. Sasisekharan. "Conformation of polypeptides and proteins". *Adv. Protein Chem.* **23**, 283 (1968).

[273] M. Ester, H.-P. Kriegel, J. Sander & X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise". In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, 226–231 (AAAI Press, 1996).

[274] E. Schubert, J. Sander, M. Ester, H. P. Kriegel & X. Xu. "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN". *ACM Trans. Database Syst.* **42** (2017).

[275] J. T. Frank, O. T. Unke, K.-R. Müller & S. Chmiela. "A Euclidean transformer for fast and stable machine learned force fields". *Nat. Commun.* **15**, 6539 (2024).

[276] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen & X. Wang. "A survey of deep active learning". *ACM Comput. Surv.* **54**, 1 (2021).

[277] L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay & C. W. Coley. "Uncertainty quantification using active learning for molecular property prediction". *J. Chem. Inf. Model.* **60**, 3770 (2020).

[278] I. Batatia, D. P. Kovacs, G. N. C. Simm, C. Ortner & G. Csanyi. "MACE" (2025). Available at https://github.com/ACEsuit/mace/.

[279] A. Samtsevych, Y. Song, T. van der Heide, B. Aradi, B. Hourahine, R. Maurer, K. Reuter, C. Scheurer & C. Panosetti. "DSKO: dancing through DFTB parameterization". *ChemRxiv* 10.26434/chemrxiv−2025−4cmx9 (2025).

[280] P. Zheng, R. Zubatyuk, W. Wu, O. Isayev & P. O. Dral. "Artificial intelligence-enhanced quantum chemical method with broad applicability". *Nat. Commun.* **12**, 7022 (2021).

[281] E. H. E. Farrar & M. N. Grayson. "Machine learning and semi-empirical calculations: a synergistic approach to rapid, accurate, and mechanism-based reaction barrier prediction". *Chem. Sci.* **13**, 7594 (2022).

[282] M. Nováček & J. Řezáč. "PM6-ML: The synergy of semiempirical quantum chemistry and machine learning transformed into a practical computational method". *J. Chem. Theory Comput.* **21**, 678 (2025).

[283] Y. Chen, W. Yan, Z. Wang, J. Wu & X. Xu. "Constructing accurate and efficient general-purpose atomistic machine learning model with transferable accuracy for quantum chemistry". *J. Chem. Theory Comput.* **20**, 9500 (2024).

[284] W. Sun, G. Fan, T. van der Heide, A. McSloy, T. Frauenheim & B. Aradi. "Machine learning enhanced DFTB method for periodic systems: Learning from electronic density of states". *J. Chem. Theory Comput.* **19**, 3877 (2023).

[285] J. Zhu, V. Q. Vuong, B. G. Sumpter & S. Irle. "Artificial neural network correction for density-functional tight-binding molecular dynamics simulations". *MRS Commun.* **9**, 867 (2019).

[286] J. J. Kranz, M. Kubillus, R. Ramakrishnan, O. A. von Lilienfeld & M. Elstner. "Generalized density-functional tight-binding repulsive potentials from unsupervised machine learning". *J. Chem. Theory Comput.* **14**, 2341 (2018).

[287] A. McSloy, G. Fan, W. Sun, C. Hölzer, M. Friede, S. Ehlert, N.-E. Schütte, S. Grimme, T. Frauenheim & B. Aradi. "TBMaLT, a flexible toolkit for combining tight-binding and machine learning". *J. Chem. Phys.* **158**, 034801 (2023).

[288] C. Panosetti, A. Engelmann, L. Nemec, K. Reuter & J. T. Margraf. "Learning to use the force: Fitting repulsive potentials in density-functional tight-binding with gaussian process regression". *J. Chem. Theory Comput.* **16**, 2181 (2020).

[289] M. Stöhr, L. Medrano Sandonas & A. Tkatchenko. "Accurate many-body repulsive potentials for density-functional tight binding from deep tensor neural networks". *J. Phys. Chem. Lett.* **11**, 6835 (2020).

[290] A. Hofstetter, L. Böselt & S. Riniker. "Graph-convolutional neural networks for (QM)ML/MM molecular dynamics simulations.". *Phys. Chem. Chem. Phys.* **24**, 22497 (2022).

[291] D. J. Burrill, C. Liu, M. G. Taylor, M. J. Cawkwell, D. Perez, E. R. Batista, N. Lubbers & P. Yang. "MLTB: Enhancing transferability and extensibility of density functional tight-binding theory with many-body interaction corrections". *J. Chem. Theory Comput.* **21**, 1089 (2025).

[292] A. S. Christensen *et al.* "OrbNet Denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy". *J. Chem. Phys.* **155**, 204103 (2021).

[293] G. Zhou, N. Lubbers, K. Barros, S. Tretiak & B. Nebgen. "Deep learning of dynamically responsive chemical hamiltonians with semiempirical quantum mechanics". *Proc. Natl. Acad. Sci.* **119**, e2120333119 (2022).

[294] Z. Qiao, A. S. Christensen, M. Welborn, F. R. Manby, A. Anandkumar & T. F. Miller. "Informing geometric deep learning with electronic interactions to accelerate quantum chemistry". *Proc. Natl. Acad. Sci.* **119**, e2205221119 (2022).

[295] I. Batatia *et al.* "A foundation model for atomistic materials chemistry". *arXiv* 2401.00096 (2024).

[296] D. P. Kovács, J. H. Moore, N. J. Browning, I. Batatia, J. T. Horton, V. Kapil, I.-B. Magdău, D. J. Cole & G. Csányi. "MACE-OFF23: Transferable machine learning force fields for organic molecules". *J. Am. Chem. Soc.* **147**, 17598 (2025).

[297] B. Huang, G. F. von Rudorff & O. A. von Lilienfeld. "The central role of density functional theory in the AI age". *Science* **381**, 170 (2023).

[298] A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth & B. Kozinsky. "Learning local equivariant representations for large-scale atomistic dynamics". *Nat. Commun.* **14**, 579 (2023).

[299] B. Hourahine *et al.* "DFTB+, a software package for efficient approximate density functional theory based atomistic simulations". *J. Chem. Phys.* **152**, 124101 (2020).

[300] S. R. Bahn & K. W. Jacobsen. "An object-oriented scripting interface to a legacy electronic structure code". *Comput. Sci. & Eng.* **4**, 56 (2002).

[301] E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth & S. Grimme. "A generally applicable atomic-charge dependent London dispersion correction". *J. Chem. Phys.* **150** (2019).

[302] A. G. Donchev *et al.* "Quantum chemical benchmark databases of gold-standard dimer interaction energies". *Sci. Data* **8**, 55 (2021).

[303] C. Adamo & V. Barone. "Toward reliable density functional methods without adjustable parameters: The PBE0 model". *J. Chem. Phys.* **110**, 6158 (1999).

[304] V. Havu, V. Blum, P. Havu & M. Scheffler. "Efficient O (N) integration for all-electron electronic structure calculation using numeric basis functions". *J. Comput. Phys.* **228**, 8367 (2009).

[305] J. T. Frank, O. T. Unke & K.-R. Müller. "So3krates: Equivariant attention for interactions on arbitrary length-scales in molecular systems". *Adv. Neural Inf. Process. Syst.* **35**, 29400 (2022).

[306] J. Bradbury *et al.* "JAX: composable transformations of Python+NumPy programs" (2018). Available at http://github.com/jax-ml/jax.

[307] T. Froitzheim, M. Müller, A. Hansen & S. Grimme. "g-xTB: A general-purpose extended tight-binding electronic structure method for the elements H to Lr (Z= 1–103)". *ChemRxiv* 10.26434/chemrxiv–2025–bjxvt (2025).

[308] J. A. Pople, D. P. Santry & G. A. Segal. "Approximate self-consistent molecular orbital theory. I. Invariant procedures". *J. Chem. Phys.* **43**, S129 (1965).

[309] T. Husch & M. Reiher. "Comprehensive analysis of the neglect of diatomic differential overlap approximation". *J. Chem. Theory Comput.* **14**, 5169 (2018).

[310] M. J. S. Dewar & W. Thiel. "Ground states of molecules. 38. The MNDO method. approximations and parameters". *J. Am. Chem. Soc.* **99**, 4899 (1977).

[311] M. J. S. Dewar & W. Thiel. "A semiempirical model for the two-center repulsion integrals in the NDDO approximation". *Theor. Chim. Acta* **46**, 89 (1977).

[312] Q. Wang *et al.* "Structural and functional basis of SARS-CoV-2 entry by using human ACE2". *Cell* **181**, 894 (2020).

[313] D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, C.-L. Hsieh, O. Abiona, B. S. Graham & J. S. McLellan. "Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation". *Science* **367**, 1260 (2020).

[314] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng & T. E. Ferrin. "UCSF Chimera—a visualization system for exploratory research and analysis". *J. Comput. Chem.* **25**, 1605 (2004).

[315] J. M. White, A. E. Ward, L. Odongo & L. K. Tamm. "Viral membrane fusion: a dance between proteins and lipids". *Annu. Rev. Virol.* **10**, 139 (2023).

[316] B. Chen, M. Farzan & H. Choe. "SARS-CoV-2 spike protein: structure, viral entry and variants". *Nat. Rev. Microbiol.* 1–14 (2025).

[317] M. Bushman, R. Kahn, B. P. Taylor, M. Lipsitch & W. P. Hanage. "Population impact of SARS-CoV-2 variants with enhanced transmissibility and/or partial immune escape". *Cell* **184**, 6229 (2021).

[318] D. Planas *et al.* "Reduced sensitivity of SARS-CoV-2 variant delta to antibody neutralization". *Nature* **596**, 276 (2021).

[319] F. Grabowski, G. Preibisch, S. Giziński, M. Kochańczyk & T. Lipniacki. "SARS-CoV-2 variant of concern 202012/01 has about twofold replicative advantage and acquires concerning mutations". *Viruses* **13**, 392 (2021).

[320] T. Arns, A. Fouquier d'Hérouël, P. May, A. Tkatchenko & A. Skupin. "Mechanism-based classification of SARS-CoV-2 variants by molecular dynamics resembles phylogenetic tree". *bioRxiv* 10.1101/2023.11.28.568639 (2023).

[321] S. Páll, A. Zhmurov, P. Bauer, M. Abraham, M. Lundborg, A. Gray, B. Hess & E. Lindahl. "Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS". *J. Chem. Phys.* **153** (2020).

[322] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford & R. A. Neher. "Nextstrain: real-time tracking of pathogen evolution". *Bioinformatics* **34**, 4121 (2018).

[323] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser & C. Simmerling. "ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB". *J. Chem. Theory Comput.* **11**, 3696 (2015).

[324] S. Ganscha, O. T. Unke, D. Ahlin, H. Maennel, S. Kashubin & K.-R. Müller. "The QCML dataset, quantum chemistry reference data from 33.5 M DFT and 14.7 B semi-empirical calculations". *Sci. Data* **12**, 406 (2025).

[325] A. D. MacKerell Jr *et al.* "All-atom empirical potential for molecular modeling and dynamics studies of proteins". *J. Phys. Chem. B* **102**, 3586 (1998).

[326] D. J. Price & C. L. Brooks III. "A modified TIP3P water potential for simulation with Ewald summation". *J. Chem. Phys.* **121**, 10096 (2004).

[327] M. Towler. "The quantum Monte Carlo method". *Phys. Status Solidi B* **243**, 2573 (2006).

[328] W. M. Foulkes, L. Mitas, R. Needs & G. Rajagopal. "Quantum Monte Carlo simulations of solids". *Rev. Mod. Phys.* **73**, 33 (2001).

[329] N. Metropolis & S. Ulam. "The Monte Carlo method". *J. Am. Stat. Assoc.* **44**, 335 (1949).

[330] J. C. Grossman. "Benchmark quantum Monte Carlo calculations". *J. Chem. Phys.* **117**, 1434 (2002).

[331] J. A. Charry Martinez, M. Barborini & A. Tkatchenko. "Correlated wave functions for electron–positron interactions in atoms and molecules". *J. Chem. Theory Comput.* **18**, 2267 (2022).

[332] M. Ditte, M. Barborini, L. Medrano Sandonas & A. Tkatchenko. "Molecules in environments: Toward systematic quantum embedding of electrons and Drude oscillators". *Phys. Rev. Lett.* **131**, 228001 (2023).

[333] J. Čížek. "On the correlation problem in atomic and molecular systems. calculation of wavefunction components in Ursell-type expansion using quantum-field theoretical methods". *J. Chem. Phys.* **45**, 4256 (1966).

[334] F. M. Faulstich & M. Oster. "Coupled cluster theory: Toward an algebraic geometry formulation". *SIAM J. Appl. Algebr. Geom.* **8**, 138 (2024).

[335] K. Raghavachari, G. W. Trucks, J. A. Pople & M. Head-Gordon. "A fifth-order perturbation comparison of electron correlation theories". *Chem. Phys. Lett.* **157**, 479 (1989).

[336] P. R. Nagy, G. Samu & M. Kállay. "Optimization of the linear-scaling local natural orbital CCSD (T) method: Improved algorithm and benchmark applications". *J. Chem. Theory Comput.* **14**, 4193 (2018).

[337] T. H. Dunning Jr., K. A. Peterson & A. K. Wilson. "Gaussian basis sets for use in correlated molecular calculations. X. The atoms aluminum through argon revisited". *J. Chem. Phys.* **114**, 9244 (2001).

[338] F. Neese, F. Wennmohs, U. Becker & C. Riplinger. "The ORCA quantum chemistry program package". *J. Chem. Phys.* **152**, 224108 (2020).

[339] G. Wang, A. Annaberdiyev, C. A. Melton, M. C. Bennett, L. Shulenburger & L. Mitas. "A new generation of effective core potentials from correlated calculations: 4s and 4p main group elements and first row additions". *J. Chem. Phys.* **151**, 144110 (2019).

[340] S. Sorella. "Wave function optimization in the variational Monte Carlo method". *Phys. Rev. B* **71**, 241103 (2005).

[341] A. Zen, J. G. Brandenburg, A. Michaelides & D. Alfè. "A new scheme for fixed node diffusion quantum Monte Carlo with pseudopotentials: Improving reproducibility and reducing the trial-wave-function bias". *J. Chem. Phys.* **151**, 134105 (2019).

[342] M. Barborini. "Quantum Mecha (**QMeC**ha) package https://github.com/qmecha (private repository)" (2024). Available at https://github.com/QMeCha.

[343] F. Weigend & R. Ahlrichs. "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy". *Phys. Chem. Chem. Phys.* **7**, 3297 (2005).

[344] C. J. Nickerson, K. R. Bryenton, A. J. Price & E. R. Johnson. "Comparison of density-functional theory dispersion corrections for the DES15K database". *J. Phys. Chem. A* **127**, 8712 (2023).

[345] B. Jeziorski, R. Moszynski & K. Szalewicz. "Perturbation theory approach to intermolecular potential energy surfaces of van der Waals complexes". *Chem. Rev.* **94**, 1887 (1994).

[346] K. Patkowski. "Recent developments in symmetry-adapted perturbation theory". *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **10**, e1452 (2020).

[347] T. M. Parker, L. A. Burns, R. M. Parrish, A. G. Ryno & C. D. Sherrill. "Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies". *J. Chem. Phys.* **140** (2014).

[348] M. Pinheiro, F. Ge, N. Ferré, P. O. Dral & M. Barbatti. "Choosing the right molecular machine learning potential". *Chem. Sci.* **12**, 14396 (2021).

[349] S. Chmiela, A. Tkatchenko, H. Sauceda, I. Poltavsky, K. Schütt & K. Müller. "Machine learning of accurate energy-conserving molecular force fields. science advances". *Sci. Adv.* **3** (2017).

[350] Y. Yang, H. Yu, D. York, Q. Cui & M. Elstner. "Extension of the self-consistent-charge density-functional tight-binding method: third-order expansion of the density functional theory total energy and introduction of a modified effective coulomb interaction". *J. Phys. Chem. A* **111**, 10861 (2007).

[351] A. Tkatchenko & M. Scheffler. "Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data". *Phys. Rev. Lett.* **102**, 073005 (2009).

[352] J. P. Perdew, K. Burke & M. Ernzerhof. "Generalized gradient approximation made simple". *Phys. Rev. Lett.* **77**, 3865 (1996).

[353] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter & M. Scheffler. "Ab initio molecular simulations with numeric atom-centered orbitals". *Comput. Phys. Commun.* **180**, 2175 (2009).

[354] X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter & M. Scheffler. "Resolution-of-identity approach to hartree–fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions". *New J. Phys.* **14**, 053020 (2012).

[355] P. Pernot, B. Huang & A. Savin. "Impact of non-normal error distributions on the benchmarking and ranking of quantum machine learning models". *Mach. Learn. Sci. Technol.* **1**, 035011 (2020).

[356] K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko & K.-R. Müller. "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions". *Adv. Neural Inf. Process. Syst.* **30** (2017).

[357] N. Lubbers, J. S. Smith & K. Barros. "Hierarchical modeling of molecular energies using a deep neural network". *J. Chem. Phys.* **148**, 241715 (2018).

[358] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt & K.-R. Müller. "Machine learning of accurate energy-conserving molecular force fields". *Sci. Adv.* **3**, e1603015 (2017).

[359] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Muller & A. Tkatchenko. "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space". *J. Phys. Chem. Lett.* **6**, 2326 (2015).