PhD-FSTM-2026-002
The Faculty of Science, Technology and Medicine

DISSERTATION

Defense held on 16/01/2026 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN PHYSIQUE

by

Adil KABYLDA

Born on 15th February 1997 in Pavlodar (Kazakhstan)

GENERAL-PURPOSE MACHINE LEARNING FORCE FIELDS FOR (BIO)MOLECULAR SIMULATIONS

Dissertation defense committee

Dr. Alexandre Tkatchenko, Supervisor
Professor, University of Luxembourg

Dr. Etienne Fodor, Chairman
Professor, University of Luxembourg

Dr. Massimiliano Esposito
Professor, University of Luxembourg

Dr. Johannes T. Margraf
Professor, University of Bayreuth

Dr. Siewert J. Marrink
Professor, University of Groningen

Abstract

The accurate and efficient simulation of large (bio)molecular systems with quantum mechanical fidelity represents a grand challenge in computational science. *Ab initio* quantum chemistry methods provide accuracy but remain prohibitively expensive at realistic scales, whereas classical force fields achieve efficiency but sacrifice accuracy. Machine learning force fields (MLFFs) promise to close this gap, yet their predictive power is often limited by locality assumptions that miss the long-range effects governing the structure, dynamics, and function of complex (bio)molecular systems. This thesis develops a framework for general-purpose MLFFs that preserves quantum mechanical fidelity while scaling to large systems by combining quantum-mechanical data, efficient atomic representations, and models explicitly designed to capture long-range interactions.

To advance model development beyond small molecules, we introduce two quantum mechanical datasets that span the chemical space of cellular components: MD22 and QCell. MD22 offers a benchmark featuring molecular dynamics trajectories for six biomolecular units and two supramolecular complexes. It represents a significant increase in system size (up to 370 atoms) and conformational flexibility, and is specifically designed to probe nonlocal correlations. To support the training of broadly applicable, general-purpose models, QCell takes this a step further by significantly expanding coverage across all major classes of biomolecules, with ∼500k diverse fragments of carbohydrates, nucleic acids, lipids, as well as noncovalent dimers and ion-water clusters.

We then make collective effects tractable in global MLFFs that couple all atomic degrees of freedom by developing an efficient interatomic descriptor. The resulting algorithm, reduced descriptor gradient-domain machine learning (rGDML), automatically constructs the minimal set of interatomic features required to capture long-range fluctuations, converting the quadratic growth of global descriptors into linear scaling. rGDML improves accuracy over both local and baseline global models, and its efficiency and stability are demonstrated through a 50 ns molecular dynamics simulation of a tetrapeptide. Its enhanced interpretability enables systematic analysis across MD22 molecules, revealing that nonlocal features (atoms separated by up to 15 Å in the studied systems) are essential to retain overall accuracy for peptides, DNA base pairs, fatty acids, and supramolecular complexes.

Building on these insights, we introduce SO3LR, a pretrained general-purpose MLFF that couples a fast SO(3)-equivariant neural network for semi-local interactions with universal, physically grounded pairwise potentials for short-range repulsion, long-range electrostatics, and dispersion. SO3LR is trained on a diverse set of four million neutral and charged molecular complexes computed at the PBE0+MBD level of quantum mechanics, ensuring broad coverage of covalent and noncovalent interactions. The model scales to 200k atoms on a single GPU and achieves reasonable to high accuracy across the chemical space of organic (bio)molecules. We validate this performance with polyalanine simulations from 300 to 800 K, accurate structural and spectroscopic observables across both high and low vibrational frequencies for a solvated protein, and consistent local and global structural properties for a glycoprotein and a lipid bilayer.

This thesis establishes a complete route from data to long-range-aware, general-purpose MLFFs that bring quantum accuracy to the biomolecular scale. The synthesis of machine learning and physics marks the beginning of realistic modeling of biological processes with quantum-level fidelity, with important implications for understanding health and disease.

Preface

Acknowledgements

I was very fortunate to be advised by Alexandre Tkatchenko. He gave me the freedom to pursue the research directions I found most compelling and offered guidance and feedback that shaped my thinking. I especially value our conversations about science and life beyond the office, shared on walks, hikes, runs, and skis in many parts of the world. I left each discussion feeling more enthusiastic and inspired. I am deeply grateful.

I thank the members of the Theoretical Chemical Physics Group for creating a supportive and stimulating environment. In my first year, Igor Poltavsky and Valentin Vassilev-Galindo were generous with their time and helped me find my footing in a new field. I enjoyed our spirited scientific debates. I will miss the quiet camaraderie with Almaz Khabibrakhmanov, organizing events with Carolin Müller, solving debugging conundrums with Sergio Suárez-Dou, and playing chess with Nils Davoine. I also thank Ariadni Boziki, Matthieu Sarkis, Dmitry Fedorov, and Gregory Fonseca for the humor and kindness that made daily life in the group lighter. Beyond the university, I am grateful to the Kazakh community in Luxembourg for the warmth they brought to my life here.

Collaboration with Klaus-Robert Müller and his group in Berlin shaped this thesis significantly. I am grateful for his encouragement and for the long-term vision he shared, which broadened my own. I spent three months as a research fellow in his group, an experience that was both enjoyable and scientifically enriching. Much of what I know about machine learning comes from my interactions with J. Thorben Frank, Stefan Chmiela, Oliver T. Unke, and Malte Esders, and I am grateful for many technically illuminating discussions.

Throughout my PhD, I benefited greatly from workshops and conferences that brought together engaged colleagues. My time at the Institute for Pure and Applied Mathematics (IPAM) was particularly formative, and the IPAM reunion conference was a highlight. I also look back fondly on the Lindau, Telluride, and CECAM meetings, where the most valuable moments often came in conversations that continued long after the talks ended. I thank Yolande Edjogo for her help with administrative matters over the years. I gratefully acknowledge the Luxembourg National Research Fund for funding my doctoral fellowship.

I would also like to acknowledge the mentors who shaped my path on the way to my PhD. I thank my high school teacher, Almas Ordabayev, for fostering my early interest in chemistry. I am especially indebted to my master's advisor, Anastasia V. Bochenkova, for the many evenings she spent explaining quantum chemistry, discussing my work, and sharpening my talks.

Finally, I thank my family. I am grateful to my parents, who have encouraged my curiosity for as long as I can remember and believed in me throughout. Most of all, I thank my wife, Lisa, whose love and support have been a constant source of strength and mean more to me than I can express.

Note on publications

This thesis is based on the following publications:

1. 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).

2. **A. Kabylda**, S. Suárez-Dou, N. Davoine, F. N. Brünig, A. Tkatchenko. QCell: Comprehensive Quantum-Mechanical Dataset Spanning Diverse Biomolecular Fragments. *arXiv* 2510.09939 (2025).

3. **A. Kabylda**, V. Vassilev-Galindo, S. Chmiela, I. Poltavsky, A. Tkatchenko. Efficient Interatomic Descriptors for Accurate Machine Learning Force Fields of Extended Molecules. *Nat. Commun.* **14**, 3562 (2023).

4. **A. Kabylda**, J. T. Frank, S. Suárez-Dou, A. Khabibrakhmanov, L. Medrano Sandonas, O. T. Unke, S. Chmiela, K.-R. Müller, A. Tkatchenko. Molecular Simulations with a Pretrained Neural Network and Universal Pairwise Force Fields. *J. Am. Chem. Soc.* **147**, 37, 33723 (2025). Cover article.

Other publications, which are not covered in this thesis, are listed below:

5. **A. Kabylda**, B. Mortazavi, X. Zhuang, A. Tkatchenko. Mechanical Properties of Nanoporous Graphenes: Transferability of Graph Machine-Learned Force Fields Compared to Local and Reactive Potentials. *Adv. Funct. Mater.* **35**, 2417891 (2025).

6. M. Esders, T. Schnake, J. Lederer, **A. Kabylda**, G. Montavon, A. Tkatchenko, K.-R. Müller. Analyzing Atomic Interactions in Molecules as Learned by Neural Networks. *J. Chem. Theory Comput.* **21**, 2, 714 (2025).

7. 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, K.-R. Müller, A. Tkatchenko. Crash Testing Machine Learning Force Fields for Molecules, Materials, and Interfaces: Model Analysis in the TEA Challenge 2023. *Chem. Sci.* **16**, 3720 (2025).

8. 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, K.-R. Müller, A. Tkatchenko. Crash Testing Machine Learning Force Fields for Molecules, Materials, and Interfaces: Molecular Dynamics in the TEA Challenge 2023. *Chem. Sci.* **16**, 3738 (2025).

9. S. G. Dale, N. Kazeev, A. J. A. Price, V. Posligua, S. Roche, O. A. von Lilienfeld, K. S. Novoselov, X. Bresson, G. Mengaldo, X. Chen, T. J. O'Kane, E. R. Lines, M. J. Allen, A. E. Debus, C. Miller, J. Zhou, H. H. Dodge, D. Rousseau, A. Ustyuzhanin, Z. Yan, M. Lanza, F. Sciarrino, R. Yoshida, Z. Leong, T. L. Tan, Q. Li, **A. Kabylda**, I. Poltavsky, A. Tkatchenko, S. A. Tawfik, P. D. Kamath, T. J. Inizan, K. A. Persson, B. Y. Li, V. Karan, C. Duan, H. Jia, Q. Zhao, H. Hayashi, A. Seko, I. Tanaka, O. M. Yaghi, T. Gould, B. Chan, S. Vuckovic, T. Li, M. Lin, Z. Tang, Y. Li, Y. Xu, A. Joshi, X. Wang, L. W. T. Ng, S. V. Kalinin, M. Ahmadi, J. Zhang, S. Zhang, A. Lapkin, M. Xiao, Z. Wu, K. Hippalgaonkar, L. Wong, L. Bastonero, N. Marzari, D.

L. E. Cordoba, A. Tomut, A. Q. Andrade, J.-H. Garcia. AI4X Roadmap: Artificial Intelligence for the advancement of scientific pursuit and its future directions, *arXiv* 2511.20976 (2025).

Articles related to prior research:

10. D. P. Zarezin, **A. M. Kabylda**, V. I. Vinogradova, P. V. Dorovatovskii, V. N. Khrustalev, V. G. Nenajdenko. Efficient Synthesis of Tetrazole Derivatives of Cytisine Using the Azido-Ugi Reaction. *Tetrahedron* **74**, 32, 4315 (2018).

11. D. P. Zarezin, O. I. Shmatova, **A. M. Kabylda**, V. G. Nenajdenko. Efficient Synthesis of the Peptide Fragment of the Natural Depsipeptides Jaspamide and Chondramide. *Eur. J. Org. Chem.* **34**, 4716 (2018).

12. K. F. Chin, X. Ye, Y. Li, R. Lee, **A. M. Kabylda**, D. Leow, X. Zhang, E. C. X. Ang, C.-H. Tan. Bisguanidinium-Catalyzed Epoxidation of Allylic and Homoallylic Amines under Phase Transfer Conditions. *ACS Catal.* **10**, 4, 2684 (2020).

13. J. Langeland, N. W. Persen, E. Gruber, H. V. Kiefer, **A. M. Kabylda**, A. V. Bochenkova, L. H. Andersen. Controlling Light-Induced Proton Transfer from the GFP Chromophore. *ChemPhysChem* **22**, 9, 833 (2021). Cover article.

14. D. A. Gorbachev, E. F. Petrusevich, **A. M. Kabylda**, E. G. Maksimov, K. A. Lukyanov, A. M. Bogdanov, M. S. Baranov, A. V. Bochenkova, A. S. Mishin. A General Mechanism of Green-to-Red Photoconversions of GFP. *Front. Mol. Biosci.* **7**, 176 (2020).

15. E. Gruber, **A. M. Kabylda**, M. B. Nielsen, A. P. Rasmussen, R. Teiwes, P. A. Kusochek, A. V. Bochenkova, L. H. Andersen. Light Driven Ultrafast Bioinspired Molecular Motors: Steering and Accelerating Photoisomerization Dynamics of Retinal. *J. Am. Chem. Soc.* **144**, 1, 69 (2022).

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The structure and dynamics of biological macromolecules are governed by the forces acting between their constituent atoms. These forces drive the system across a multidimensional potential-energy surface (PES), giving rise to emergent phenomena such as protein folding, ligand binding, membrane organization, and viral capsid assembly. The ability to model this energy landscape with high fidelity is therefore not merely a computational challenge but a central route to understanding disease mechanisms, designing novel therapeutics, and engineering biomaterials with controlled function [1]. For decades, however, progress has been constrained by an apparent trade-off between simulation accuracy and computational efficiency.

At one end of the spectrum lie methods of *ab initio* quantum mechanics (QM). By solving, in approximation, the Schrödinger equation for a given arrangement of nuclei, QM approaches provide a predictive and physically grounded description of the PES. Consistent with this description, the forces are calculated as the gradient of the energy, directly reflecting the underlying electronic structure. This accuracy comes at a steep computational cost, with formal scaling that is typically cubic or worse with system size. As a result, direct *ab initio* molecular dynamics remains intractable for the system sizes and time scales required to capture many biologically relevant events, which often involve tens of thousands of atoms and microsecond to millisecond timescales.

At the opposite end is classical molecular mechanics, which employs empirical force fields. These models replace the explicit quantum-mechanical calculation with a fixed analytical energy function composed of bonded and non-bonded terms. Their simplicity underpins their outstanding efficiency, with effectively linear scaling in system size and the ability to simulate entire organelles or viruses [2, 3]. This efficiency is obtained at the expense of physical fidelity. Fixed functional forms and parameterizations limit the description of essential quantum-mechanical effects such as electronic polarization, charge transfer, and the many-body interactions. Rather than constituting a controlled approximation to the true Born–Oppenheimer PES, classical force fields define an alternative model with intrinsically limited accuracy and transferability to new chemical environments [4].

In recent years, machine learning force fields (MLFFs) have emerged as a promising way to mitigate this long-standing trade-off [5]. By learning the high-dimensional PES directly from QM reference data, MLFFs can, in principle, approach quantum accuracy at a cost comparable to classical force fields. Realising this promise in a robust and transferable way, however, requires addressing three interconnected challenges:

1. **Data.** The performance and reliability of any MLFF are critically dependent on the quality, diversity, and physical relevance of its underlying QM training data. At the outset of this work, significant gaps existed in the available datasets, particularly for large, flexible systems where nonlocal interactions are important and for entire classes of biomolecules such as lipids, carbohydrates, and nucleic acids.
2. **Representation.** To enable learning, molecular geometries must be mapped to numerical descriptors. These representations have either been computationally scalable but physically incomplete (local, short-range) or physically more complete but poorly scaling with system size (global, fully coupled).
3. **Scaling and long-range physics.** An MLFF should respect basic physical symmetries and capture interactions across all relevant length scales, from covalent bonding to long-range electrostatics and dispersion, while remaining efficient enough for large-scale biomolecular simulations. This requires architectural choices and model decompositions that avoid uncontrolled extrapolation when moving from small training systems to realistic biological assemblies.

This thesis presents a series of contributions addressing these challenges. **Chapter 2** provides the theoretical background on molecular simulations and machine learning force fields. **Chapter 3** introduces the MD22 dataset, a benchmark designed to probe the role of nonlocal interactions in larger and more flexible molecules, and demonstrates a new algorithm for scaling the global sGDML model to such systems. **Chapter 4** describes the QCell dataset, a chemically diverse QM dataset spanning the building blocks of all major biomolecular classes. **Chapter 5** addresses the representation problem by introducing the rGDML method, which achieves effectively linear scaling for a compressed global descriptor while retaining essential nonlocal information. **Chapter 6** presents the SO3LR model, a pretrained MLFF that combines an SO(3)-equivariant neural network with universal pairwise force fields for long-range interactions, delivering accuracy and transferability for large-scale biomolecular simulations. Finally, **Chapter 7** summarizes the main results and outlines directions for future research.

Taken together, these contributions help to narrow the traditional trade-off between accuracy and efficiency in (bio)molecular simulation and provide a coherent framework for constructing predictive, QM-informed force fields that can accelerate discovery in chemistry, biology, and medicine.

Chapter 2

Theoretical Background

In this chapter, we introduce the theoretical foundations of molecular simulations and machine learning force fields (MLFFs) that underpin the developments presented in this thesis. We begin with the quantum-mechanical description of molecules and its standard approximations, proceed through classical molecular mechanics, and then discuss modern ML-based models, with emphasis on those used and extended in Chapters 3–6.

2.1 Foundations of molecular simulations

2.1.1 Quantum description and the Born–Oppenheimer surface

At the fundamental level, a molecular system of electrons and nuclei is described by the non-relativistic, time-independent Schrödinger equation

$$\hat{H}\,\Psi(\mathbf{r},\mathbf{R}) = E\,\Psi(\mathbf{r},\mathbf{R}), \tag{2.1}$$

where $\Psi(\mathbf{r},\mathbf{R})$ is the total many-body wavefunction, $\mathbf{r}$ and $\mathbf{R}$ denote electronic and nuclear coordinates, and $\hat{H}$ is the full molecular Hamiltonian containing kinetic-energy operators and Coulomb interactions [6].

Practical electronic structure calculations almost universally rely on the Born–Oppenheimer approximation, which exploits the large mass ratio between nuclei and electrons to decouple their motion. The molecular wavefunction is factorized as

$$\Psi(\mathbf{r},\mathbf{R}) \approx \psi_0(\mathbf{r};\mathbf{R})\,\chi(\mathbf{R}), \tag{2.2}$$

with $\psi_0(\mathbf{r};\mathbf{R})$ representing the electronic ground state for clamped nuclei at positions $\mathbf{R}$, and $\chi(\mathbf{R})$ the nuclear wavefunction. Assuming the electrons adapt instantaneously to the nuclear motion allows one to neglect non-adiabatic couplings, yielding the electronic Schrödinger equation

$$\hat{H}_e(\mathbf{r};\mathbf{R})\,\psi_0(\mathbf{r};\mathbf{R}) = E_0(\mathbf{R})\,\psi_0(\mathbf{r};\mathbf{R}). \tag{2.3}$$

Here, $E_0(\mathbf{R})$ defines the ground-state Born–Oppenheimer potential-energy surface (PES) governing the nuclear motion. In principle, Eq. (2.3) provides the exact PES, but a direct solution is computationally intractable for most systems of chemical interest because the underlying Hilbert space grows exponentially with system size. Consequently, all electronic structure methods and machine learning force fields discussed in this thesis aim to approximate this surface efficiently.

2.1.2 Classical dynamics and statistical ensembles

Given a potential-energy surface $E(\mathbf{R})$, classical molecular dynamics (MD) integrates Newton's equations of motion

$$m\frac{d^2\mathbf{R}}{dt^2} = \mathbf{F}(\mathbf{R}) = -\nabla_{\mathbf{R}} E(\mathbf{R}), \tag{2.4}$$

to generate trajectories of the nuclear coordinates. In the absence of coupling to an external reservoir, these equations define Hamiltonian dynamics that conserve the total energy and sample the microcanonical (NVE) ensemble, characterized by the phase-space distribution

$$\rho_{\mathrm{NVE}}(\Gamma) \propto \delta\big(E - H_n(\Gamma)\big), \tag{2.5}$$

where $\Gamma = (\mathbf{R}, \mathbf{P})$ denotes a microstate in phase space and H_n is the classical nuclear Hamiltonian [7].

In many applications, one wishes to simulate at fixed temperature or pressure. This is achieved by augmenting Eq. (2.4) with additional degrees of freedom or stochastic terms that model coupling to a heat or pressure bath. Extended-system methods of the Nosé–Hoover type and their generalizations provide deterministic equations of motion that generate canonical (NVT) or isothermal–isobaric (NPT) ensembles, while Langevin dynamics introduces stochastic and frictional forces that drive the system towards a prescribed temperature [7–10]. In all cases, the equations of motion are constructed such that the target equilibrium ensemble is preserved as a stationary distribution in phase space.

The connection to macroscopic or experimentally accessible observables is made through statistical averages. Given an observable $A(\Gamma)$, its equilibrium expectation value in a chosen ensemble with distribution $\rho(\Gamma)$ is

$$\langle A \rangle = \int d\Gamma\, \rho(\Gamma)\, A(\Gamma), \tag{2.6}$$

which, under ergodic dynamics, is estimated by a time average along an MD trajectory. For example, structural properties such as the radial distribution function $g(r)$ are computed as histograms of interatomic distances averaged over snapshots. Dynamical quantities are accessed via time-correlation functions. For instance, the self-diffusion coefficient D can be determined from the long-time limit of the mean squared displacement (Einstein relation)

$$D = \lim_{t\to\infty} \frac{1}{6t} \langle |\mathbf{R}_i(t) - \mathbf{R}_i(0)|^2 \rangle, \tag{2.7}$$

while vibrational spectra are obtained from the Fourier transform of the velocity autocorrelation function $\langle \mathbf{v}(t) \cdot \mathbf{v}(0) \rangle$.

In the remainder of this thesis, we will repeatedly use this framework: forces derived from a given potential-energy surface are propagated in time by MD, and ensemble or time averages of suitable microscopic observables are compared to quantum-mechanical reference data and, where available, to experimental measurements. The reliability of these macroscopic predictions is therefore directly tied to the accuracy of the underlying potential $E(\mathbf{R})$, motivating the development of high-fidelity machine learning force fields.

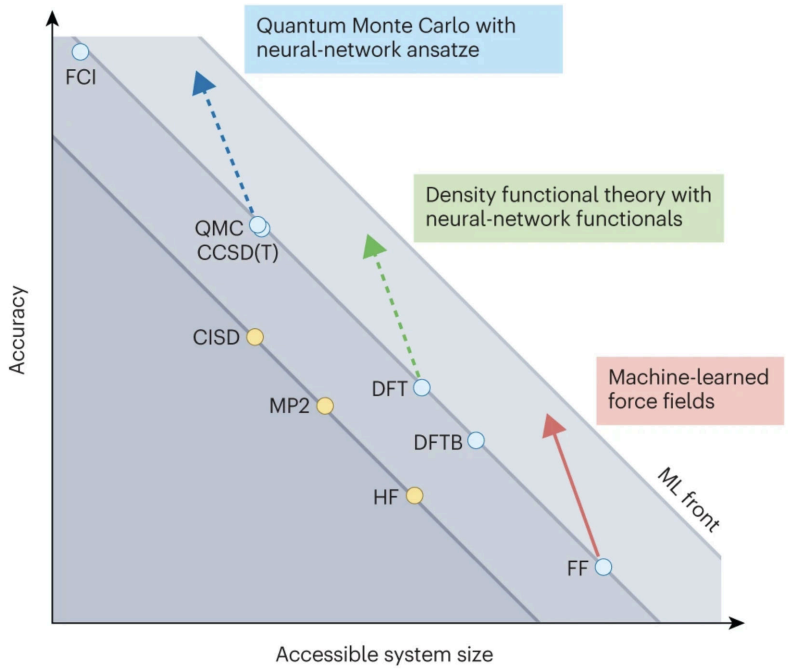

Figure 2.1: Hierarchy of atomistic simulation methods. Reproduced with permission from Ref. 11. © Nature Publishing Group.

2.1.3 Electronic structure methods

The potential-energy surface of a molecular system is obtained from solutions to the electronic Schrödinger equation, Eq. (2.3). For all but the simplest systems, however, exact solutions are out of reach and practical calculations rely on systematic approximations. Electronic structure methods are commonly grouped into two broad families: wavefunction-based approaches and density-functional theory (DFT). Wavefunction methods construct explicit approximations to the many-electron wavefunction, ranging from mean-field Hartree–Fock to correlated post-Hartree–Fock expansions, while DFT reformulates the problem in terms of the ground-state electron density as the fundamental variable. Together, these approaches span a wide range of computational cost and accuracy, as summarized in Fig. 2.1.

Wavefunction-based methods.

Wavefunction methods can, in principle, be systematically improved towards the exact solution by including electron correlation, at the price of rapidly increasing computational cost. The Hartree–Fock (HF) approximation serves as the standard mean-field starting point. While HF treats exchange interactions exactly, it neglects the instantaneous repulsion between electrons; its formal cost scales as $O(N^4)$ with system size N. To recover these missing many-body effects, Post-HF strategies such as Møller–Plesset perturbation theory (MP2) [12] and Coupled-Cluster (CC) theory [13] are employed, forming a hierarchy that targets the Full Configuration Interaction (FCI) limit. The "gold standard" of this hierarchy, CCSD(T) [14] (coupled-cluster with single and double excitations augmented by perturbative triples) exhibits a formal scaling of $O(N^7)$. Consequently, canonical implementations are typically restricted to molecules containing at most a few dozen atoms.

These high-level methods can routinely reach so-called *chemical accuracy* (total energy errors of less than 1 kcal/mol) for small and medium-sized molecules and are often treated

as *de facto* reference standards. Over the past decade, a combination of explicitly correlated techniques, density fitting, and reduced orbital and auxiliary spaces (e.g. frozen natural orbitals and related schemes) has significantly extended the practical reach of CCSD(T), making it possible to obtain near basis-set-limit interaction energies for non-covalent complexes containing on the order of 50–75 atoms and a few thousand orbitals on modern high-performance computing architectures [15]. Nevertheless, even these advanced implementations remain restricted to relatively small fragments compared to typical biomolecular and condensed-phase systems and are therefore primarily used as benchmarks and training data for more scalable approaches.

Density-functional theory.

Density-functional theory (DFT) offers a more favorable accuracy–efficiency trade-off for larger systems by reformulating the electronic structure problem in terms of the ground-state electron density $\rho(\mathbf{r})$ rather than the many-electron wavefunction [6]. The Hohenberg–Kohn theorems guarantee that the ground-state energy can be written as a functional $E[\rho]$ of the density and that there is a one-to-one correspondence between the external potential $v_{\text{ext}}(\mathbf{r})$ and $\rho(\mathbf{r})$. This formal framework is realized through the Kohn–Sham (KS) construction, which maps the interacting many-electron problem onto an auxiliary system of non-interacting electrons moving in an effective local potential

$$\left[-\frac{1}{2}\nabla^2 + v_{\text{ext}}(\mathbf{r}) + v_{\text{H}}[\rho](\mathbf{r}) + v_{\text{xc}}[\rho](\mathbf{r})\right]\phi_i(\mathbf{r}) = \varepsilon_i\phi_i(\mathbf{r}), \tag{2.8}$$

where v_{ext} is the external potential due to the nuclei, v_{H} is the classical Hartree (Coulomb) potential, and v_{xc} is the exchange–correlation potential derived from the exchange–correlation energy functional $E_{\text{xc}}[\rho]$. The electron density is obtained self-consistently from the occupied KS orbitals as $\rho(\mathbf{r}) = \sum_i^{\text{occ}} |\phi_i(\mathbf{r})|^2$. The exact form of $E_{\text{xc}}[\rho]$ is unknown and must be approximated, and the choice of this functional is the dominant source of error in Kohn–Sham DFT and therefore a crucial modeling decision.

The key approximation in practical DFT calculations is thus the choice of $E_{\text{xc}}[\rho]$. A convenient organizing principle is "Jacob's ladder" [16], in which successive rungs introduce additional ingredients derived from the density and the Kohn–Sham orbitals:

1. *Local spin density approximation (LSD):* functionals that depend only on the local value of $\rho(\mathbf{r})$ and are constructed from the uniform electron gas.
2. *Generalized gradient approximations (GGAs):* functionals that depend on $\rho(\mathbf{r})$ and its gradient $\nabla\rho(\mathbf{r})$, improving the description of inhomogeneous systems.
3. *Meta-GGAs:* functionals that add further local ingredients such as the kinetic energy density, providing greater flexibility and enabling the enforcement of additional exact constraints.
4. *Hybrid functionals:* approaches that mix a fraction of exact Hartree–Fock exchange with a semi-local functional (GGA or meta-GGA), which often reduces self-interaction errors and improves the description of charge localization, band gaps, and reaction barriers.
5. *Double hybrids and RPA-type approaches:* functionals that supplement a hybrid description with an explicit perturbative correlation term or employ the random-phase approximation for correlation, offering higher accuracy at a cost approaching that of low-order wavefunction methods.

Independent of their place on Jacob's ladder, functionals also differ in their degree of empiricism: some are constructed primarily to satisfy known exact conditions with minimal parameter fitting (e.g. PBE [17], PBE0 [17, 18], SCAN [19]), while others are more heavily parameterized against experimental data or high-level calculations. For the reference calculations in this thesis, we utilize minimally empirical functionals, specifically PBE and PBE0. Further justification for this choice is detailed in Section 4.1.2.

Long-range interactions and many-body dispersion.

Many properties of molecular and condensed-phase systems are controlled by interactions acting well beyond covalent bond lengths, including electrostatics, induction, and van der Waals (vdW) dispersion [20, 21]. Although semi-local and hybrid density-functional approximations (DFAs) recover most of the total electronic energy, they miss a substantial fraction of the long-range correlation contribution and often fail to bind weakly interacting systems quantitatively. A classic example is that standard DFAs capture nearly all of the total energy of rare-gas dimers but only a small fraction of their interaction energy [20]. As a result, conventional DFT can misrepresent interactions in molecular crystals, layered materials, and biomolecular complexes [22].

Long-range correlation is therefore often added to DFT explicitly, for example through atom-pairwise London-type terms (e.g. DFT-D [23], TS [24], XDM [25, 26]). These electronic-structure-based schemes obtain dispersion coefficients and effective vdW radii from the underlying density or Kohn–Sham orbitals; this introduces some environment dependence and improves transferability over fixed-parameter force fields, but they remain fundamentally pairwise-additive and thus neglect collective screening and higher-order many-body effects [22, 27, 28].

The many-body dispersion (MBD) formalism provides an efficient route to include these collective long-range correlations within DFT [21, 29]. MBD represents each atom by an effective polarizable quantum oscillator and couples these oscillators through a dipole–dipole interaction tensor, so that the resulting collective normal modes describe correlated charge-density fluctuations across the entire system. Change in zero-point energy of oscillators due to dipolar interactions yields a many-body dispersion energy that naturally incorporates screening and non-additivity to infinite order, while retaining a computational cost that scales approximately as $O(N^3)$. In this thesis, such DFA+MBD schemes are employed as reference methods for training ML models.

2.1.4 Empirical potentials

Classical force fields, also called empirical potentials or molecular mechanics models, provide inexpensive, empirical representations of molecular potential-energy surfaces and are widely used when explicit quantum-mechanical treatments are too costly. In these models, one forgoes an explicit electronic description and instead represents atoms as point masses interacting through simple analytical functions [30–32]. The functional forms are parametrized to reproduce selected reference data, which may include experimental observables (e.g., equilibrium bond lengths, vibrational frequencies, heats of vaporization, densities, lattice constants) and/or quantum-chemical benchmarks (e.g., atomization energies, conformational energetics, *ab initio* forces, or potential-energy scans of dimers and torsions) [4, 33]. Modern biomolecular force fields typically rely on a carefully calibrated balance of both types of data.

The potential energy in a molecular-mechanics model is written as a sum of contributions associated with different classes of bonded and nonbonded interactions:

$$
\begin{aligned}
U(\mathbf{R}) = & \underbrace{\sum_{\text{bonds}} k_i^{\text{bond}} \left(r_i - r_{0,i}\right)^2}_{U_{\text{bond}}} + \underbrace{\sum_{\text{angles}} k_i^{\text{angle}} \left(\theta_i - \theta_{0,i}\right)^2}_{U_{\text{angle}}} \\
& + \underbrace{\sum_{\text{dihedrals}} k_i^{\text{dihe}} \left[1 + \cos\left(n_i \phi_i + \delta_i\right)\right]}_{U_{\text{dihedral}}} \\
& + \underbrace{\sum_i \sum_{j>i} \left[4\epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right) + \frac{q_i q_j}{\varepsilon\, r_{ij}} \right]}_{U_{\text{nonbond}}},
\end{aligned}
\tag{2.9}
$$

where r_i, θ_i, and ϕ_i denote bond lengths, angles, and dihedral angles, respectively, r_{ij} is the distance between atoms i and j, ϵ_{ij} and σ_{ij} are Lennard–Jones parameters, and q_i are fixed partial charges. The first three terms represent bonded interactions (covalent bond stretching, angle bending, and torsional rotations), while the last term accounts for pairwise nonbonded van der Waals and electrostatic interactions. Long-range Coulomb interactions are usually evaluated with Ewald or particle–mesh Ewald techniques (PME, SPME) to maintain accuracy while achieving favorable $O(Nlog(N))$ scaling with system size [34–36].

The simplicity of such fixed-topology, fixed-charge models underpins their efficiency and enables simulations of biomolecular assemblies containing hundreds of thousands to millions of atoms on microsecond and longer timescales. Classical force fields such as AMBER [32], CHARMM [30], and OPLS [31] have become the workhorses of biomolecular simulation and have been instrumental in establishing molecular dynamics as a "computational microscope" for proteins, nucleic acids, and membranes [1]. The first molecular dynamics simulation of a folded protein, bovine pancreatic trypsin inhibitor, already demonstrated that an empirical potential can yield a time-averaged structure close to the X-ray model and provide detailed information on the magnitude, spatial correlations, and decay of internal fluctuations that are inaccessible to static experimental structures [37]. Since then, increasingly refined parametrizations have made large-scale biomolecular simulations almost routine in structural biology and materials science [33, 38].

At the same time, it is important to recognise the approximations inherent in classical force fields, which ultimately limit their accuracy and transferability [4, 33]. The fixed functional form in Eq. (2.9) assumes that the PES can be decomposed into a small number of simple bonded and pairwise nonbonded terms, so that higher-order many-body couplings are either neglected altogether or absorbed into fixed, environment-independent parameters. Electronic degrees of freedom are not treated explicitly; their effects on bonding, lone pairs, polarization, and charge transfer are encoded in effective quantities ($r_{0,i}$, k_i, ϵ_{ij}, σ_{ij}, q_i) rather than emerging from an underlying electronic structure calculation. Standard biomolecular force fields also assume a fixed covalent topology and typically employ non-polarizable point charges, so they cannot describe bond breaking and formation and may struggle in strongly heterogeneous or highly polar environments.

A variety of extensions relax some of these assumptions. Polarizable force fields introduce induced dipoles or Drude oscillators to account for electronic response (for example AMOEBA and related models) [39, 40]; reactive force fields such as ReaxFF add bond-order-dependent

terms to enable bond rearrangements [41]; and coarse-grained models like MARTINI trade atomic detail for efficiency to access much larger length and time scales, up to cellular level [42, 43]. These approaches expand the reach of empirical potentials, but they retain a prescribed functional form and are usually parametrized for specific classes of systems, which limits their transferability across different chemistries and thermodynamic conditions.

These challenges motivate the development of alternative potential-energy models. In this thesis, machine learning force fields trained on quantum-mechanical data are employed to approximate the underlying Born–Oppenheimer surface while retaining much of the computational efficiency of classical models. Such machine learning force fields aim to capture short- and medium-range many-body effects, reduce reliance on fixed topologies and hand-crafted atom types, and offer a route toward more transferable and systematically improvable potentials.

2.2 Machine learning force fields

2.2.1 General workflow

Machine learning force fields (MLFFs) have emerged over the past two decades as a promising route to bridge the accuracy gap between classical force fields and quantum-mechanical methods [5]. The basic idea is to replace the explicit electronic structure calculation by a flexible statistical model that learns the mapping from nuclear configurations to energies and forces, using data from high-level quantum calculations as a reference. Once trained, the ML model serves as a surrogate for the underlying *ab initio* method, delivering energies and forces with near-quantum accuracy at a cost that approaches that of classical force fields.

Machine learning offers essentially unrestricted functional flexibility. Sufficiently expressive neural networks and kernel methods serve as universal approximators capable of representing arbitrarily complex relationships within training data. For atomistic modeling, this flexibility means that many-body effects and subtle environmental dependencies can be learned directly from quantum-mechanical reference data rather than being imposed through fixed analytical functional forms. This versatility is, however, a double-edged sword: while it allows MLFFs to capture patterns inaccessible to traditional empirical potentials, it also increases the risk of overfitting and learning spurious correlations (i.e., the Clever Hans effect[1]) that degrade transferability. Therefore, careful model design, regularization, and validation are essential to ensure the learned PES remains physically meaningful and robust.

The general workflow for constructing and using an MLFF is illustrated in Fig. 2.2 and consists of four main stages:

1. **Data generation.** A set of reference atomic configurations is generated and then evaluated using a chosen electronic structure method to obtain energies, forces, and other observables (e.g., stress tensors). The quality and diversity of this dataset are crucial, as they define the domain of validity of the final MLFF.

2. **Representation (descriptor).** Each configuration is converted into a numerical representation or descriptor to serve as input for the ML model. This descriptor is often designed to encode relevant physical symmetries, such as translational, rotational,

[1]The name comes from "Clever Hans," a horse in early-1900s Germany that appeared to solve arithmetic problems but was later shown to respond to subtle, unintentional cues from humans rather than genuinely understanding mathematics.

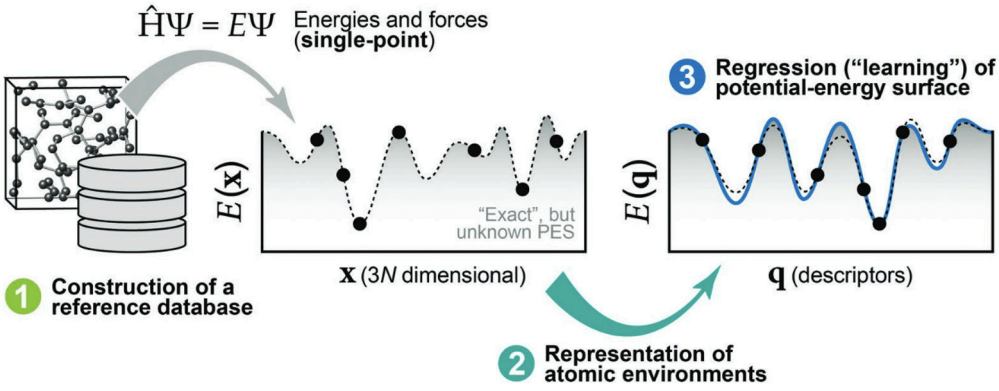

Figure 2.2: General MLFF workflow. Schematic illustration of data generation, representation, and learning steps in the construction of a machine learning force field. Reproduced with permission from Ref. 44. © Wiley.

and permutation invariance, to improve accuracy. Examples include local atomic environment descriptors or graph-based representations.

3. **Learning algorithm.** A supervised learning model (e.g., a neural network or kernel method) is trained on the reference data by minimizing a loss function. This step involves selecting hyperparameters and regularization to manage model complexity and prevent overfitting. Often, data generation and training are performed iteratively in an active-learning loop to optimize the reference calculations.

4. **Validation and deployment.** The final trained model is rigorously assessed using independent test data and, critically, by observing its behavior during MD simulations to confirm its accuracy and stability.

Over the past years, MLFFs have significantly evolved, transitioning from small, system-specific models to more generalized architectures applicable across diverse chemistries and materials. Early milestones include the Behler–Parrinello Neural Network (BPNN), which represents the total energy as a sum of individual atomic contributions, each modeled by a feedforward neural network operating on handcrafted local symmetry functions (descriptors) [45]. Another prominent early approach is the Gaussian Approximation Potential (GAP) framework, which combines kernel ridge regression with the Smooth Overlap of Atomic Positions (SOAP) descriptor, successfully applied to systems such as carbon [46, 47]. More recent deep-learning methods began replacing these fixed, handcrafted descriptors with learned representations obtained through message passing on atomistic graphs, exemplified by the SchNet model [48]. Building on this graph-based concept, highly advanced E(3)-equivariant architectures (such as PaiNN [49], NequIP [50], MACE [51], and SO3krates [52, 53]) further improved data efficiency and have since established the current state of the art in accuracy for many classes of systems.

The following subsections discuss the core ingredients common to these and other MLFFs: the construction of suitable training datasets, the design of atomic representations, and the choice of learning architectures.

2.2.2 Quantum data

Reference data form the foundation of MLFF development. These data are obtained either from *ab initio* molecular dynamics trajectories or from carefully curated collections of molecular geometries that span a targeted region of chemical compound space. Both well-designed benchmark sets and high-quality training datasets have been crucial for driving the field forward and for providing a common, standardized ground on which different MLFF models can be rigorously compared.

It is useful to distinguish two interconnected but conceptually distinct data regimes that have played a central role in the evolution of MLFFs. The first is the *single-system* (or configurational) regime, in which the goal is to reproduce the PES of a given molecule or small set of molecules with the highest possible precision using as little reference data as possible. Datasets like MD17 [54], which contains *ab initio* MD trajectories for prototypical small molecules (ethanol, aspirin, and others), and its later refinements have been instrumental in this context. They enabled systematic studies of how different models trade-off data efficiency, stability, and accuracy for well-defined systems and have served as standard benchmarks for many early architectures.

As MLFFs became more accurate, a second *chemical space* (or multi-system) regime gained prominence. Here, the emphasis shifts from optimizing performance on a single molecule to training on chemically diverse sets of structures so that one model can make reliable predictions for configurations and compounds that were not explicitly included in the training data. Representative examples include ANI-1 [55] and QM7-X [56], which provide tightly converged DFT reference data and a wide range of physicochemical properties for millions of small organic molecules. These datasets emphasize coverage of chemical and conformational space and are specifically designed to foster the development of general-purpose MLFFs with robust extrapolation behavior.

This thesis contributes to both directions. The MD22 benchmark dataset, introduced in Chapter 3, extends the single-system regime to substantially larger and more flexible molecules, probing the limits of MLFFs in terms of molecular size and complexity. The QCell dataset, developed in Chapter 4, targets the multi-system regime by sampling a broad range of biomolecular fragments. Together, these datasets provide complementary testbeds for assessing the accuracy and generalizability of the MLFFs developed in the subsequent chapters.

2.2.3 Atomic representations

A crucial component of any MLFF is the representation used to encode atomic structures for input to the model. The descriptor must map a set of atomic coordinates and chemical species to numerical features in a way that preserves the fundamental symmetries of the underlying physics and, ideally, reflects local chemical similarity and extensivity [57]. For a scalar property such as the total energy E, the representation and model should satisfy

$$E(\{\mathbf{R}_i\}) = E(\{\mathbf{R}_i + \mathbf{a}\}) = E(\{\mathbf{Q}\mathbf{R}_i\}) = E(\{\mathbf{R}_{\pi(i)}\}) \tag{2.10}$$

for any translation vector $\mathbf{a}$, rotation matrix $\mathbf{Q} \in \mathrm{SO}(3)$, and permutation π of identical atoms. Translational and rotational invariance follow from the fact that the internal energy depends solely on the relative positions of the nuclei, rather than their absolute coordinates or orientation. Permutation symmetry arises from the indistinguishability of identical particles. For vector and tensor properties (e.g., forces, multipole moments, stresses), the

corresponding observables must also transform consistently (equivariantly) under these operations.

Encoding these basic symmetries at the level of the representation and model is essential: it reduces the effective size of the configuration space that must be explored during training, helps enforce physical plausibility, and significantly improves accuracy.

Much of the progress in MLFFs stems from the development of increasingly sophisticated atomic representations [57]. These are broadly categorized as *global* or *local*, a classification that reflects the fundamental strategic choice between resolving the full topology of the system simultaneously or decomposing it into independent atomic neighborhoods. As detailed below, this distinction dictates the model's intrinsic trade-off between capturing long-range physical correlations and achieving the computational scalability required for large systems.

Global descriptors

Global descriptors are characterized by the absence of an intrinsic cutoff for the interaction range. A canonical example is the Coulomb matrix, which collects pairwise terms into a matrix representation of a molecule [58]. In its original form, a molecule with nuclear charges $\{Z_i\}$ at positions $\{\mathbf{R}_i\}$ is represented by a matrix M with elements:

$$M_{ij} = \begin{cases} \frac{1}{2} Z_i^{2.4}, & i = j, \\ \dfrac{Z_i Z_j}{|\mathbf{R}_i - \mathbf{R}_j|}, & i \neq j, \end{cases} \tag{2.11}$$

where off-diagonal elements correspond to internuclear Coulomb repulsion and diagonal elements encode a smooth fit to atomic energies as a function of nuclear charge [58]. Permutation invariance is typically enforced either by sorting the rows and columns of M according to their norm or by using its eigenvalue spectrum as the descriptor. Another example is the family of explicit many-body representations, such as the SLATM descriptor, which aggregates one-, two-, and three-body contributions into a fixed-size global vector [59]. When combined with kernel methods, these representations have achieved high accuracy across diverse molecular chemical spaces.

The main drawback of global descriptors is their unfavorable scaling with system size, which incurs high computational costs for both training and prediction. Extending such representations to systems containing hundreds or thousands of atoms is challenging. Furthermore, accommodating variable system sizes requires *ad hoc* architectural choices, such as zero-padding or pooling. These limitations are particularly severe in condensed-phase and biomolecular simulations, where one wishes to exploit linear scaling and reuse information across chemically similar environments.

Chapter 5 revisits global descriptors in the context of the rGDML model. In this approach, the original high-dimensional global representation is compressed to its most informative components, retaining essential nonlocal information while improving scaling. This dimensionality reduction allows the global descriptor to be extended toward larger molecules without sacrificing the ability to describe long-range correlations.

Local descriptors

Local, atom-centred descriptors represent each atom in terms of its neighborhood within a finite cutoff radius. This decomposition naturally enforces translational invariance and

enables the construction of size-extensive models with linear scaling. It also facilitates the transferability of the model to larger systems than those used in training, provided that the relevant local environments are adequately sampled [57].

Prominent examples of local descriptors include the Behler–Parrinello atom-centred symmetry functions (ACSFs), which encode the radial and angular distributions of neighboring atoms using a hand-crafted basis of functions [45, 60], and the Smooth Overlap of Atomic Positions (SOAP) descriptor, which represents the local neighbor density via an expansion in radial basis functions and spherical harmonics, followed by rotationally invariant contractions [47]. The FCHL (named after their developers) representation further refines these concepts, constructing local many-body features that are tailored for kernel methods and designed to achieve broad chemical transferability [59, 61].

More recently, the Atomic Cluster Expansion (ACE) has been developed as a complete, systematically improvable basis for local invariant features, providing a unifying language for many existing local representations [62]. In ACE, the environment of atom i is first encoded in atomic base functions $A^{(i)}_{nlm}$, obtained by projecting the neighbor density onto a radial–angular basis

$$A^{(i)}_{nlm} = \sum_{j \in \mathcal{N}(i)} R_{nl}(r_{ij})\, Y_{lm}(\hat{\mathbf{r}}_{ij}), \tag{2.12}$$

where $R_{nl}(r)$ are radial basis functions, Y_{lm} are spherical harmonics, and the sum runs over neighbors j within the cutoff radius. Symmetrized products of these moments are then combined into rotation- and permutation-invariant cluster basis functions $B^{(i)}_{\nu}$, which form a linear expansion for the atomic energy

$$E_i = \sum_{\alpha} c^{(1)}_{\alpha}\, B^{(i)}_{\alpha} + \sum_{\alpha\beta} c^{(2)}_{\alpha\beta}\, B^{(i)}_{\alpha\beta} + \sum_{\alpha\beta\gamma} c^{(3)}_{\alpha\beta\gamma}\, B^{(i)}_{\alpha\beta\gamma} + \dots, \tag{2.13}$$

where $B^{(i)}$ denotes the symmetrized cluster functions of increasing body order, and the coefficients $c^{(n)}$ are fitted to reference data. In this manner, ACE provides a hierarchy of invariant descriptors with explicit control over both body order and angular resolution. It recovers many existing local descriptors as special or truncated cases. An important practical feature is that the many-body expansion can be evaluated with a computational cost that is effectively linear in the number of neighbors, despite formally including high body orders, which makes ACE a competitive basis for developing highly accurate and transferable interatomic potentials.

While the locality assumption is the key to the efficiency of these methods, it also constitutes their primary physical limitation. By construction, interactions beyond the cutoff radius are ignored, which complicates the description of long-range phenomena that decay slowly with distance and are inherently collective [57]. Section 2.3 will discuss various approaches to recover these missing long-range contributions.

Equivariant graph-based representations

Recent developments in MLFFs have moved beyond purely invariant descriptors toward representations and models that are explicitly *equivariant* under rotations. Equivariant neural networks achieve this by systematically constructing internal features that transform as irreducible tensor representations of the SO(3) rotation group [49]. By utilizing specialized layers that consistently propagate these transformation properties, these models ensure that the output transforms consistently [63].

Formally, let $\mathbf{R} = \{\mathbf{R}_i\}$ denote a set of atomic positions and f a feature map (or network layer) that produces feature vectors $\mathbf{h}_i = f(\mathbf{R})$ for each atom. Equivariance with respect to a rotation $\mathbf{Q} \in \mathrm{SO}(3)$ means that rotating the input configuration by $\mathbf{Q}$ results in a transformed feature vector that is consistent with the rotation, expressed as

$$f\big(\{\mathbf{Q}\mathbf{R}_i\}\big) = \mathcal{D}(\mathbf{Q})\, f\big(\{\mathbf{R}_i\}\big), \tag{2.14}$$

where $\mathcal{D}(\mathbf{Q})$ is a (block-diagonal) matrix representation of $\mathbf{Q}$ acting on the feature space. For scalar (rotation-invariant) features, $\mathcal{D}(\mathbf{Q})$ is the identity matrix, $\mathbb{I}$, while for vector features, it is the rotation matrix itself, $\mathbf{Q}$. Higher-rank tensor features transform according to their corresponding irreducible representations $\mathcal{D}^{(l)}(\mathbf{Q})$. Equation (2.14) ensures that rotating the input geometry rotates the tensorial features, rather than arbitrarily altering their numerical values. This is particularly powerful when learning vector and tensor quantities such as forces, multipole moments, and stress, and has been demonstrated to significantly improve both data efficiency and stability compared with invariant architectures [49, 50, 53].

In this thesis, we explore both types of descriptors: sGDML is based on a global invariant descriptor, while SO3krates employs a local, SO(3)-equivariant graph representation. This links directly to the next subsections, which discuss kernel methods (global, system-specific, quadratically scaling) and message-passing neural networks (local, transferable, linearly scaling).

2.2.4 Gradient-domain machine learning

Gradient-Domain Machine Learning (GDML) is a kernel-based framework that trains directly on atomic forces to reconstruct a smooth, global potential-energy surface (PES) [54, 64–66]. It combines the non-parametric flexibility of kernel methods with a strict enforcement of energy conservation by construction. GDML employs an inverse-distance global descriptor (inspired by the Coulomb matrix) to encode the molecular geometry. The resulting model fits interatomic forces and energies with high fidelity to quantum-mechanical reference data, while guaranteeing that the learned force field, derived from a scalar potential, is conservative.

For small, drug-like molecules, GDML and its symmetrized extension (sGDML) have been shown to reproduce high-level PESs with excellent accuracy ($\sim$0.3 kcal/mol for energies and $\sim$1 kcal/mol/Å for forces) using $\sim 10^3$ training geometries [54]. This level of data efficiency is substantially better than that of energy-only models and has established GDML as an accurate and robust framework.

Kernel ridge regression

GDML is built upon Kernel Ridge Regression (KRR), a non-parametric method that performs linear regression in a high-dimensional reproducing kernel Hilbert space. The core component is a positive-definite *kernel function*, $k(\mathbf{x}, \mathbf{x}')$, which implicitly defines an inner product $\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$ in a feature space via the "kernel trick."

In standard scalar KRR, one seeks a function $\hat{f}$ that minimizes the regularized squared error over a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$. The representer theorem [67] states that the optimal solution is a linear combination of kernel evaluations centered at the training points

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i\, k(\mathbf{x}_i, \mathbf{x}). \tag{2.15}$$

The coefficients $\boldsymbol{\alpha}$ are obtained by solving the linear system $(\mathbf{K} + \lambda \mathbf{I})\,\boldsymbol{\alpha} = \mathbf{y}$, where $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel matrix and λ is a regularization parameter that governs the trade-off between fitting accuracy and model complexity.

Kernel ridge regression in the gradient domain

The key innovation of GDML is the extension of KRR to vector-valued force fields while strictly enforcing the condition that the force $\mathbf{F}$ is the negative gradient of a scalar potential V. To achieve this, the method defines a scalar covariance kernel $\kappa(\mathbf{x}, \mathbf{x}')$ representing the underlying PES, and derives a matrix-valued kernel for the forces via differentiation

$$\mathrm{Cov}\big(F_p(\mathbf{x}), F_q(\mathbf{x}')\big) = \frac{\partial^2}{\partial x_p \, \partial x'_q}\, \kappa(\mathbf{x}, \mathbf{x}'). \tag{2.16}$$

This operation generates a block-structured *Hessian kernel matrix*, $\mathbf{K}_{\mathrm{Hess}(\kappa)}$, which naturally encodes the physical constraints of a conservative field.

Given a dataset of M geometries and their associated force vectors $\{(\mathbf{x}_i, \mathbf{F}_i)\}_{i=1}^{M}$, the coefficients are determined by solving the regularized normal equations

$$(\mathbf{K}_{\mathrm{Hess}(\kappa)} + \lambda \mathbf{I})\,\boldsymbol{\alpha} = -\mathbf{F}, \tag{2.17}$$

where $\mathbf{F}$ is the concatenated vector of all training force components. Once the coefficients $\boldsymbol{\alpha}$ have been fitted, the predicted force at a new configuration $\mathbf{x}$ is given by

$$\hat{\mathbf{F}}(\mathbf{x}) = -\sum_{i=1}^{M} \mathbf{K}_{\mathrm{Hess}(\kappa)}(\mathbf{x}, \mathbf{x}_i)\,\boldsymbol{\alpha}_i. \tag{2.18}$$

The corresponding energy can be recovered (up to an additive constant) by integrating these forces or by evaluating the scalar kernel expansion using the same coefficients $\boldsymbol{\alpha}$ derived from the force training. By operating directly in the gradient domain, GDML exploits the fact that a single reference configuration provides $3N$ force constraints compared to only one energy value. This construction yields a substantial gain in data efficiency and guarantees that the learned model is strictly conservative (curl-free) by design.

Descriptors and kernel choice

To embed molecular geometries in a way that incorporates basic symmetries, GDML employs a global descriptor inspired by the Coulomb matrix representation [58] but omits nuclear charges, which are instead encoded implicitly via element-specific kernels or by using element labels in the descriptor. For a system of N atoms with Cartesian coordinates $\{\mathbf{r}_i\}$, the descriptor matrix $\mathbf{D}$ is defined as

$$D_{ij} = \begin{cases} \|\mathbf{r}_i - \mathbf{r}_j\|^{-1}, & i > j, \\ 0, & i \le j, \end{cases} \tag{2.19}$$

so that all pairwise inverse distances appear in the strictly lower triangular part of $\mathbf{D}$. Permutation invariance is enforced by symmetrization procedures that average over relevant permutations, as implemented in sGDML [64].

For the covariance kernel κ, GDML uses members of the Matérn family, which provides smooth, stationary kernels with tunable differentiability and spectral content. A Matérn

kernel with parameter $\nu = n + \frac{1}{2}$ and length scale σ can be written in the form

$$\kappa(d) = \exp\Big(- \sqrt{2\nu}\,\tfrac{d}{\sigma}\Big)\, P_n(d), \qquad d = \|\mathbf{x} - \mathbf{x}'\|, \tag{2.20}$$

where $P_n(d)$ is a polynomial of degree n

$$P_n(d) = \sum_{k=0}^{n} \frac{(n+k)!}{(2n)!} \binom{n}{k} \Big(\sqrt{2\nu}\,\tfrac{d}{\sigma}\Big)^{n-k}. \tag{2.21}$$

Specifically, GDML uses the $\nu = 5/2$ (i.e. $n = 2$) Matérn kernel, which is twice mean-square differentiable and therefore sufficiently smooth to support stable second derivatives with respect to all Cartesian coordinates. This ensures that the Hessian kernel blocks entering Eq. (2.17) are well defined and numerically well behaved.

The price for this expressive, globally coupled representation is computational cost. Because the descriptor couples all atom pairs, its dimensionality grows quadratically with system size ($O(N^2)$), and the Hessian kernel matrix in Eq. (2.17) scales quadratically with the number of atoms and quadratically with the number of training geometries ($O(M^2N^2)$). As a result, (s)GDML models are effectively limited to molecules with a few dozen atoms and training sets of at most a few tens of thousands of configurations. Chapter 3 extends (s)GDML to larger molecules in the MD22 benchmark, while Chapter 5 introduces the rGDML model, which reduces the dimensionality of the global descriptor to its most informative components and thereby improves scaling without sacrificing the ability to capture long-range correlations.

2.2.5 Message-passing neural networks

Graph neural networks, and in particular equivariant message-passing architectures, have become leading approaches for constructing MLFFs [48–51, 53, 68]. In these models, a molecular structure is encoded as a geometric graph $G = (V, \mathcal{E})$, where nodes V correspond to atoms and edges $\mathcal{E}$ connect atoms within a specified cutoff radius r_{cut}. Node features encode, for example, the chemical species and possibly charge and spin information, while edge features encode geometric information such as interatomic distances and, in more expressive variants, directional information (e.g., spherical harmonic expansions).

The network performs a sequence of message-passing iterations (Fig. 2.3). At each layer, every atom (node) aggregates information from its neighbors to update its internal state (feature vector). In this way, an atomic descriptor is not fixed a priori but is learned by the network. A typical (invariant) message-passing layer updates the atomic feature vectors $\mathbf{h}_i^{(\ell)}$ in two steps:

$$\mathbf{m}_{ij}^{(\ell)} = \phi_m\big(\mathbf{h}_i^{(\ell)}, \mathbf{h}_j^{(\ell)}, \mathbf{r}_{ij}\big) \tag{2.22}$$

$$\mathbf{h}_i^{(\ell+1)} = \phi_h\Big(\mathbf{h}_i^{(\ell)}, \sum_{j \in \mathcal{N}(i)} \mathbf{m}_{ij}^{(\ell)}\Big), \tag{2.23}$$

where ϕ_m and ϕ_h are learned functions (typically small neural networks), $\mathbf{r}_{ij}$ encodes the relative geometry, and $\mathcal{N}(i)$ denotes the neighborhood of atom i within the cutoff. Stacking multiple such layers allows information to propagate over increasing distances, so that the final atomic embeddings encode not only the immediate local but also a broader semi-local environment.

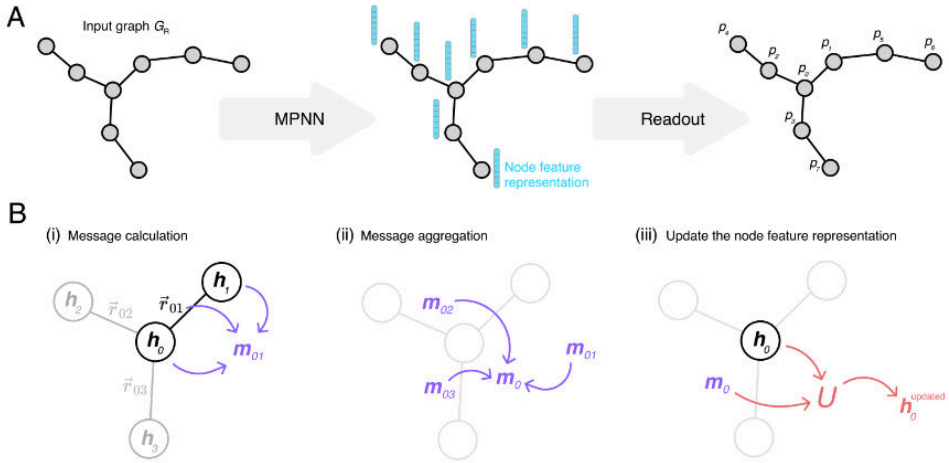

Figure 2.3: Message passing in a graph neural network. Schematic illustration of atoms as nodes in a graph and iterative exchange of information along edges during message-passing updates. Reproduced with permission from Ref. 69.

Equivariant message-passing architectures extend this idea by explicitly encoding how internal features transform under rotations. In these models, the atomic feature vector $\mathbf{h}_i^{(\ell)}$ is decomposed into scalar and tensor components that transform according to irreducible representations of SO(3) [70]. The update functions ϕ_m and ϕ_h are then constrained so that this transformation behavior is preserved across layers, ensuring that the learned representation is *equivariant*. Architectures such as PaiNN [49] and NequIP [50] implement this principle using rotation-covariant message passing and have demonstrated substantial gains in data efficiency and accuracy compared with purely invariant networks.

After several message-passing iterations, the final atomic embeddings $\mathbf{h}_i^{(L)}$ are used to predict atomic contributions to the energy

$$\varepsilon_i = \phi_{\text{out}}\left(\mathbf{h}_i^{(L)}\right), \tag{2.24}$$

where ϕ_{out} is a learned output function. The total energy E is then obtained as a sum of these contributions

$$E = \sum_i \varepsilon_i. \tag{2.25}$$

Conservative forces are subsequently computed by (automatic) differentiation of E with respect to the atomic positions. This construction naturally enforces translational and permutational invariance of the energy and, in equivariant architectures, rotational covariance of vector and tensor quantities such as forces and multipole moments.

In this thesis, we employ SO3krates as a representative equivariant message-passing model [52, 53]. In SO3krates, each atom carries both high-dimensional invariant feature channels and low-dimensional equivariant channels that transform under SO(3). Sparse, attention-based update layers exchange information between these channels and aggregate messages from neighbouring atoms defined by a local cutoff, resulting in an architecture that achieves a favorable balance between expressiveness, stability, and computational cost. Chapter 6 extends this idea in the SO3LR model by coupling the semi-local SO3krates energy with explicit physical models for long-range electrostatics and dispersion [71]. A more detailed description of the SO3krates implementation is provided in Appendix A4.

2.2.6 Model validity and generalization

General-purpose MLFFs must ultimately be judged by their predictive performance on systems and configurations not seen during training. To rigorously assess this, one should distinguish between two fundamental sources of error. The first is a systematic reference error: since an ML model cannot exceed the accuracy of its training data, the chosen electronic structure method (e.g., DFT functional, basis set) imposes a hard ceiling on predictive performance. The second is model approximation error, which arises from limited model capacity, finite or noisy training data, and imperfect optimization. In standard practice, the latter is quantified by in-distribution metrics (e.g., root-mean-square error on held-out test data), which measure how well the model interpolates within the densely sampled regions of the training distribution.

However, low test errors do not guarantee reliability in practical applications, which often require extrapolation to sparsely sampled or entirely new regions of configuration space. This generalization challenge manifests in two distinct forms. Configurational generalization refers to predicting new conformations of fixed molecular systems (e.g., in datasets like MD17 or MD22). Chemical-space generalization, by contrast, involves predicting properties across diverse compositions and topologies not present in the training set (e.g., in QM7-X or QCell). In high-dimensional spaces, the boundary between interpolation and extrapolation is subtle; models may perform well on random test splits yet fail when simulations visit under-represented rare-event regions or thermodynamic conditions different from those in the training set.

Consequently, it is essential to look beyond statistical test set metrics and evaluate deployment error: the discrepancy between reference and ML observables under actual simulation conditions. Deployment error captures the model's stability and physical fidelity in tasks such as long MD trajectories, free-energy calculations, and transport properties. Chapters 3–6 emphasize this distinction, designing datasets and models not only to minimize in-distribution errors but also to withstand stringent out-of-distribution assessments that reflect the realities of (bio)molecular simulations.

2.3 Long-range interactions in machine learning force fields

2.3.1 Relevance of long-range interactions

Long-range (LR) interactions, which prominently include electrostatic forces arising from charge distributions and dispersion forces driven by electronic fluctuations, extend far beyond the typical ∼5 Å cutoff radius employed in standard atomistic models. Although the energetic contribution of these terms is often small relative to covalent bonding, their collective influence pervades condensed and biomolecular matter, driving the collective phenomena that are qualitatively inaccessible to local models.

For example, van der Waals dispersion is a key contributor to the cohesion of layered and porous materials, the adsorption of molecules at surfaces, and the stability of molecular crystals [72–74]. In aqueous systems, many-body dispersion is essential for reproducing the experimental density and hydrogen-bond network of liquid water, correcting structural artifacts present in local DFT approximations [75, 76]. In the biological domain, collective fluctuations can shift protein stability by several kcal/mol and qualitatively alter hydration structure [77]. In medicinal chemistry, this sensitivity is vividly illustrated by the "Magic Methyl" effect, where the addition of a single methyl group ($-CH_3$) can increase binding

potency by orders of magnitude by optimizing van der Waals packing within a target protein pocket [78].

The significance of long-range electrostatics is equally profound. Their influence extends far beyond the mere stabilization of native structures via ion pairing and hydrogen bond networks. Indeed, these forces underpin the dynamics of binding through electrostatic steering, dictate the stability of transition states in enzymatic catalysis, and govern the complex liquid-liquid phase separation observed in intrinsically disordered proteins [79]. The sensitivity of soft matter to these potentials is clearly manifested in lipid membranes, where the artificial truncation of electrostatic terms is known to yield spurious headgroup ordering and unphysical phase behavior [80]. Furthermore, the molecular pathology of sickle cell anemia, the first disease understood at the molecular level, originates from a single charge mutation [81, 82]. Similarly, the double-helix model of DNA was correctly built only after placing the negatively charged phosphate groups on the surface [83]. These examples underscore that the importance of long-range interactions must never be underestimated when modeling biological phenomena.

2.3.2 Challenges of learning long-range interactions

Despite their importance, incorporating long-range interactions into MLFFs presents distinct challenges that stem from both statistical and practical considerations.

From a statistical perspective, learning long-range interactions requires isolating a subtle signal from a noisy background. The long-range dispersion and electrostatic energy is often orders of magnitude smaller than local covalent and repulsive interactions. Thus, their omission is not evident in static errors (RMSE) on test sets, in MD simulations of small systems, or even in short MD simulations of larger systems. Nevertheless, the gradient changes from these small energy terms can have meaningful effects on structure and dynamics. Learning such a smooth, slowly varying signal in the presence of dominant local fluctuations is intrinsically difficult, requiring models with high numerical resolution and correct asymptotic behaviour (e.g., $1/r$ for electrostatics).

Another challenge arises from the mismatch between the scale of available training data and the scale of intended deployment. Accurate quantum-mechanical (QM) reference calculations are limited to systems containing at most a few hundred atoms. For instance, a water sphere of radius 10 Å contains roughly 400 atoms, approaching the upper limit for conventional QM methods. Consequently, training data rarely sample mesoscopic, long-wavelength collective modes that govern large conformational changes. From a machine learning perspective, this results in an extrapolation problem. Most state-of-the-art MLFFs rely on local descriptors with finite cutoffs, effectively truncating physics at ∼5–15 Å. Without inductive biases that enforce correct physical asymptotics, models trained on small clusters may struggle to reliably extrapolate to the "global" physics of the macroscale.

2.3.3 Approaches to learning long-range interactions

Three main strategies have emerged to handle long-range interactions in MLFFs:

Explicit/Implicit Global models

Global kernel models (e.g., sGDML) or neural networks with a global representation [84] learn the total energy as a function of all atomic coordinates without hard cutoffs. The primary advantage is simplicity: no specific functional form for LR interactions needs to

be encoded. While this approach naturally captures cooperative motions, it suffers from poor scalability, often becoming intractable for large systems. Furthermore, these models require comprehensive training data covering all relevant long-range configurations, which is computationally prohibitive to generate with high-level QM.

Scaling semi-local Message-Passing Neural Networks (MPNNs)

Deep MPNNs theoretically extend the receptive field by stacking interaction layers. However, this adds significant computational overhead and faces severe gradient and information bottlenecks. In particular, deep networks suffer from *oversmoothing* [85], where node representations become indistinguishable, and *oversquashing* [86], where fixed-size vectors fail to compress information from exponentially growing neighborhoods (e.g., ∼50k atoms in a 50 Å water sphere). Moreover, stacking layers does not guarantee correct asymptotics [87] and fails to transmit interactions across a vacuum. Several approaches, such as adaptive and spectral message-passing methods, are under active development to mitigate these issues [88, 89].

Explicit incorporation of long-range physics (hybrid models)

This strategy augments local ML models with additional terms for electrostatics and dispersion [68, 71, 90–94]. By enforcing physically correct asymptotic behavior (e.g., Coulombic decay), the ML component is relieved of learning long-range trends from limited data. This greatly improves data efficiency and transferability, as the model no longer requires large fragments in the training set; the parameters for the LR terms, such as partial charges and dispersion coefficients, are themselves semi-local quantities. The primary limitation is that analytical terms are typically pairwise, potentially neglecting anisotropic or many-body long-range effects not captured by the local ML model.

Overall, long-range interactions remain a central challenge in MLFFs. Capturing them requires either global models, novel architectures that adaptively handle long-range information, or hybrid models with explicit physical terms. The rGDML model introduced in Chapter 5 tackles the quadratic scaling of descriptors in global models, while the SO3LR model introduced in Chapter 6 adopts a hybrid strategy, augmenting a semi-local equivariant graph neural network with explicit electrostatics and dispersion.

2.4 Summary

This chapter has outlined the theoretical background underlying the developments presented in the remainder of this thesis. We described how the Born–Oppenheimer approximation defines the target potential-energy surface (PES) that MLFFs seek to approximate, and how classical molecular dynamics employs this PES to generate trajectories from which thermodynamic and dynamical observables can be computed. We reviewed the main classes of electronic structure methods and classical force fields, highlighting their complementary strengths and weaknesses in terms of accuracy, computational cost, and transferability.

Subsequently, we introduced MLFFs, in which quantum-mechanical reference data are used to construct flexible statistical models of the PES. We emphasized the central roles of data quality and diversity, atomic representations, and learning architectures, alongside the need to encode basic physical symmetries directly into the model. Global kernel methods, such as (s)GDML, exemplify accurate, system-specific models based on global descriptors, whereas

local and message-passing neural networks provide scalable architectures that operate on atom-centred representations suitable for large-scale systems.

Finally, we discussed the importance of long-range interactions and the challenges associated with their modeling. We outlined strategies for incorporating long-range physics into MLFFs, ranging from fully global models and deep semi-local message-passing neural networks to hybrid approaches that augment short-range ML energies with explicit physical terms. These considerations motivate the specific models and datasets developed in this thesis. In the following chapters, we demonstrate how global kernel models (sGDML and rGDML) and graph neural network-based architectures (SO3krates, SO3LR) address these challenges in complementary ways, and how their respective advantages can be combined to achieve accurate, transferable MLFFs for chemically and biologically relevant systems.

Chapter 3

MD22 & sGDML: Global machine learning force fields for molecules with hundreds of atoms

Parts of this chapter have been published in this or similar form in Ref. 66:

- S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko, K.-R. Müller, *Science Advances* **9**, eadf0873 (2023).

  A.K. designed the MD22 dataset, performed the reference calculations, and contributed to the analysis and writing of the published work.

Modern machine learning force fields (MLFFs) bridge the accuracy gap between highly efficient, but exceedingly approximate classical force fields (FF) and prohibitively expensive high-level *ab initio* methods [5, 95, 96]. This optimism is based on the universal nature of ML models, which gives them virtually unrestricted descriptive power compared to the statically parametrized interactions in classical mechanistic FFs. Traditional ML approaches strive towards general assumptions about the problem at hand, such as continuity and differentiability, when constructing models. In principle, any physical property of interest captured in a dataset can be parametrized this way, including collective interactions that are too intricate to extract from the many-body wavefunction. As such, MLFFs can give unprecedented insights into quantum many-body mechanisms [5]. Albeit, the exceptional expressive power of global MLFFs goes along with a stark increase in parametric complexity [97–99] over classical FFs.

As a trade-off, many ML models reintroduce some of the classical mechanistic restrictions on the allowed interactions between atoms. It is unclear to which extent this departure from unbiased ML models compromises their advantages over classical FFs. In particular, localization assumptions are often made to allow a reduction of degrees of freedom across large structures. For example, message-passing neural networks only allow mean-field exchanges between local atomic neighborhoods, which leads to information loss over long distances [100]. This causes a truncation of long-range interactions, which in local models are assumed to have a rather small contribution to the overall dynamics of the system. Nonetheless, it has been shown that long-range effects can play a significant role [71, 77, 91, 101, 102], limiting the predictive power of local models in nanoscale and mesoscale systems [103, 104]. In fact, several recent MLFF models [68, 71, 91, 92, 105–108] introduce empirical correction terms for specific long-ranged effects (e.g. electrostatics),

yet long-range electron correlation effects remain poorly characterized. The number of available MLFF approaches augmented with physical interaction models indicates that we are observing an emerging field which has not yet settled on a universal solution.

In contrast, global models [54, 58, 61, 64, 109, 110] are able to include all interaction scales, but they face the challenge of having to couple at least a quadratic set of atom-atom interactions (Fig. 3.1). Such scaling behavior provides a hard computational constraint and has therefore slowed the development of global models in recent years. Current global models are thus restricted to system sizes of only a few dozen atoms, even though accurate *ab initio* reference data are available for much bigger systems. Here, we develop a combined closed-form and iterative approach to train global MLFF kernel models for large molecules. Our spectral analysis of these models demonstrates that the number of effective degrees of freedom in large molecules is substantially reduced compared to N^2 and can be captured using a low dimensional representation [111, 112]. Using this insight, our large-scale framework lowers the memory and computational time requirements of the model simultaneously. In a two step procedure, the effective degrees of freedom are solved in closed-form, before iteratively converging the remaining fluctuations to the exact solution for the full problem.

Our focus is on ML models based on Gaussian Processes (GPs), since they posses several unique properties such as linearity and loss-function convexity that can be exploited in pursuit of our goal. We demonstrate the effectiveness of our solution on the symmetric gradient domain machine learning (sGDML) FF [54, 64]. It allows us to reliably reconstruct sGDML FFs for significantly larger molecules and materials than previously possible [113]. Our new training scheme can handle systems that contain several hundreds of atoms, all of which are fully coupled within the model. We demonstrate that this parametric flexibility is indeed leveraged to let all atoms participate in generating the energy and force predictions. Our development allows us to study supramolecular complexes, nanostructures, as well as four major classes of biomolecular systems in stable nanosecond-long molecular dynamics (MD) simulations. All of these systems present phenomena with far-reaching characteristic correlation lengths. We offer these datasets as a benchmark (called MD22) that presents new challenges with respect to system size (42 to 370 atoms), flexibility and degree of nonlocality. As such, MD22 can be regarded as the next generation of the now well-established MD17 dataset [54].

3.1 Large-scale sGDML algorithm

The large-scale symmetric gradient domain machine learning (sGDML) algorithm reconstructs molecular force fields by embedding physical invariances and conservation laws into a Gaussian process (GP) framework, enabling strong generalization from limited reference data. It leverages the fact that many complex quantum-mechanical interactions can be expressed through linear constraints, which GPs naturally accommodate, and employs a kernel that models forces as gradients of a latent potential energy surface defined by a prior mean and covariance. Training is performed directly on forces, which can be obtained analytically from *ab initio* calculations via the Hellmann–Feynman theorem with minimal computational overhead, providing a more data-efficient approach than energy-only fitting. Despite being non-parametric, sGDML achieves high accuracy with significantly fewer parameters than deep neural networks, making it faster to evaluate in production-scale simulations.

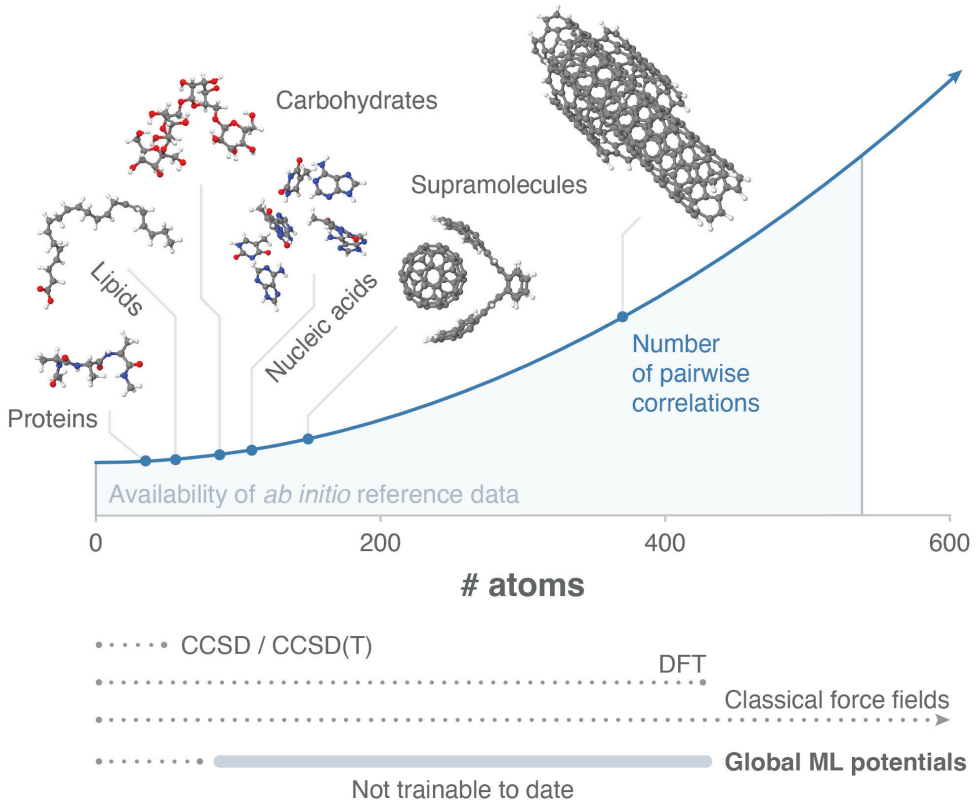

Figure 3.1: Current global MLFFs only scale to system sizes of a few dozen atoms, restricted by the computational challenge of having to couple a quadratic amount of atom-atom interactions. However, accurate *ab initio* reference data are available for much bigger systems (light blue area). This work scales global models with *ab initio* accuracy to hundreds of atoms, as is demonstrated on examples from newly developed MD22 dataset that cover units of four major classes of biomolecules and supramolecules.

For large datasets, direct GP inference via Cholesky decomposition becomes infeasible due to quadratic memory scaling, so sGDML adopts an iterative conjugate gradient (CG) solver that requires only matrix-vector products, avoiding full kernel storage. Convergence is improved by preconditioning the CG solver with a Nyström approximation of the kernel, constructed from a carefully chosen subset of inducing points selected using approximate statistical leverage scores. This preconditioner reduces the condition number of the system, enabling faster convergence without sacrificing the exactness of GP predictions. The resulting approach maintains the numerical stability required for noise-free *ab initio* datasets, with runtime and memory complexities scaling as $\mathcal{O}(mk^2)$ and $\mathcal{O}(mk)$, respectively, where m is the training size and k the number of inducing points. This combination of physically informed kernel design, force-based training, memory-efficient iterative solvers, and effective preconditioning allows large-scale sGDML to extend exact GP inference to systems and datasets that would otherwise be computationally prohibitive.

3.2 MD22 benchmark dataset

Although our solver enables significantly larger training datasets than before, we focus on the more practical example of scaling up system size. After all, the cost generating massive reference datasets overbears any speed up that a ML model could provide. At large scales, atomic interactions become potentially more complex, as they involve a broader spectrum

Table 3.1: Content and computational details of the MD22 datasets. The potential energy and atomic force labels were calculated at the PBE+MBD [17, 29] level of theory. The keywords *light* and *tight* denote different basis set options in FHI-aims. All trajectories were sampled at a resolution of 1 fs.

Dataset	# atoms	Size	Temp. [K]	Basis set
$AcAla_3NHMe$	42	85,109	500	PBE+MBD/tight
Docosahexaenoic acid	56	69,753	500	PBE+MBD/tight
Stachyose	87	27,272	500	PBE+MBD/tight
AT-AT	60	20,001	500	PBE+MBD/tight
AT-AT-CG-CG	118	10,153	500	PBE+MBD/tight
Buckyball catcher	148	6,102	400	PBE+MBD/light
Double-walled nanotube	370	5,032	400	PBE+MBD/light

of length-scales. It is this scenario, in which the combined scalabilty and data efficiency of a ML model is really needed.

To put the iterative sGDML solver to the ultimate test, we have generated a new set of MD trajectories (MD22) that cover systems of up to several hundred atoms. MD22 includes examples of four major classes of biomolecules and supramolecules, ranging from a small peptide with 42 atoms, all the way up to a double-walled nanotube with 370 atoms (see Tab. 3.1). The trajectories were sampled at temperatures between 400 K and 500 K at a resolution of 1 fs, with corresponding potential energy and atomic forces calculated at the PBE+MBD [17, 29] level of theory. Compared to the well-established MD17 benchmark [54], the standard deviations of the potential energies are significantly larger, varying between $\sim$ 8–77 kcal mol^{-1} (MD17: $\sim$ 2–6 kcal mol^{-1}). The standard deviations of the forces are however close, between $\sim$ 21–28 kcal mol^{-1} Å^{-1} (MD17: $\sim$ 20–30 kcal mol^{-1}).

We set the training dataset size for each of the systems, such that root mean squared test error for predicting atomic forces is around 1 kcal mol^{-1} A^{-1}. For some systems like the buckyball catcher (148 atoms) or the double-walled nanotube (370 atoms), this error is already achieved with small training set sizes of only a few hundred points. Other systems, e.g. DHA (docosahexaenoic acid) (56 atoms), stachyose (87 atoms) or the $AcAla_3NHMe$ peptide (42 atoms) require several thousands of training points for the same force prediction accuracy.

The corresponding energy mean absolute errors (MAEs) range between 0.39 kcal mol^{-1} ($AcAla_3NHMe$) and 4.01 kcal mol^{-1} (double-walled nanotube), which is in line with our previous results on MD17 [54] when normalized per atom. The (independent) random errors made for each atomic contributions to the overall energy prediction approximately propagate as the square root of sum of squares, which causes the energy error to scale with system size [114]. This scaling behavior is confirmed when comparing the energy MAE per atom, which is consistently around 0.01 kcal mol^{-1} for most datasets in our study. We observe, that the complexity of the learning task is neither correlated with the number of atoms, nor the simulation temperature of the reference trajectory. Rather, the difficulty to reconstruct a force field is determined by the complexity of the interactions within the system (see Tab. 3.2).

Table 3.2: sGDML prediction performance on large-scale datasets. All (test) RMSE errors are in kcal mol^{-1} (Å^{-1}) per molecule (energy) or component (forces). The training set sizes were chosen such that the root-mean-square error (RMSE) of the force prediction is around 1 kcal mol^{-1} Å^{-1}. $\|G\|$ denotes the cardinality of the leveraged permutation group for each respective dataset (see Refs. [64, 115] for details).

System	# atoms	$\|G\|$	# train.	% data	Energy	Force
Proteins - Tetrapeptide						
$AcAla_3NHMe$	42	18	6k	7%	0.50	1.21
Lipids - Fatty acid						
DHA	56	6	8k	12%	1.68	1.17
Carbohydrates - Tetrasaccharide						
Stachyose	87	1	8k	29%	4.52	1.07
Nucleic acids - DNA base pairs						
AT-AT	60	36	3k	15%	0.90	1.12
AT-AT-CG-CG	118	96	2k	20%	1.77	1.22
Supramolecules						
Buckyball catcher	148	48	600	10%	1.47	1.02
Double-walled nanotube	370	28	800	16%	4.99	0.97

3.3 Assessment of large-scale molecular force fields

3.3.1 Representation of nonlocal interactions

To investigate whether the trained sGDML models provide chemically meaningful predictions, we apply sGDML to a donor-bridge-acceptor type molecule consisting of two phenyl rings connected by an *E*-ethylene moiety forming a conjugated π-system (bridge). The phenyl rings are substituted in *para*-position with an electron-donating dimethylamine group (donor) and an electron-withdrawing nitro group (acceptor), respectively. When the phenyl rings are coplanar, electrons are delocalized over the whole molecule and can freely "flow" from donor to acceptor. However, when the two phenyl rings are rotated against each other, the conjugation of the π-orbitals is broken and the favorable interaction between donor and acceptor is lost, increasing the potential energy of the molecule. A chemically meaningful model should predict that this energy change is delocalized over the whole π-system (as opposed to explaining it by local changes in the vicinity of the center of the rotation).

To get a qualitative understanding of how these interactions are handled within the sGDML model, we investigate how individual atoms contribute towards the prediction. Being a linear combination of pairwise correlations between atoms, a partial evaluation of the model reveals the atomic contributions to each prediction (Fig. 3.2). We observe that all atoms in the system participate in generating the prediction with sGDML, which would not be possible with a model that partitions the energy into localized atomic contributions. Fig. 3.2 demonstrates that an sGDML model learns to delocalize changes in energy upon ring rotation across the whole molecule, which is in accordance with chemical intuition. Note that, starting from the global minimum structure of this system, rotating by π does not return to the starting position (despite the apparent symmetry of the molecule), because

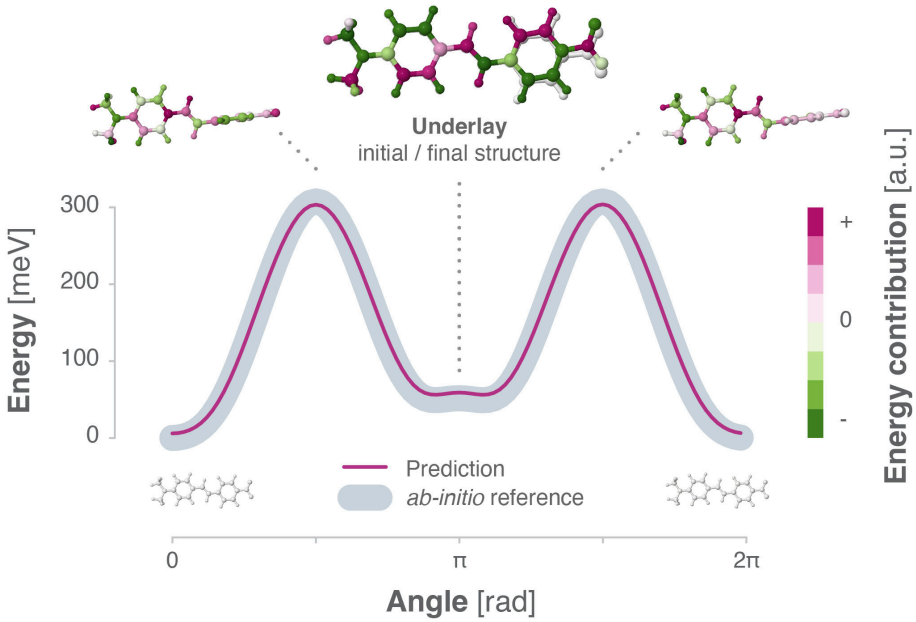

Figure 3.2: Energy contributions as predicted by the sGDML FF for a donor-bridge-acceptor type molecule (4-dimethylamino-4'-nitrostilbene). The energy profile for a full rotation around the single bond between the acceptor and the ethylene moiety is shown. When the conjugation of the π-system is broken upon rotating the phenyl rings by 90° against each other, the sGDML model predicts that the energy change is delocalized across the whole molecule.

of a slight asymmetry about the central C=C bond (see overlay of structures in Fig. 3.2). Thus, a full rotation is necessary to return to the starting point, explaining the somewhat counter-intuitive rotational energy profile.

3.3.2 Molecular dynamics

One of the biggest advantages of employing MLFFs is that they can enable accurate large-scale simulations. Here, we test our sGDML FFs by running nanosecond-long classical MD and path integral MD (PIMD) simulations for the double-walled carbon nanotube saturated with hydrogen atoms at its edges. All simulations were run at a constant temperature of 300 K with a Langevin thermostat and a time-step of 0.2 fs. The number of beads of the PIMD simulations was set to 16.

To confirm the reliability of any MLFF, it is first and foremost essential to assess its capability to yield stable (PI)MD simulations. In this regard, Fig. A1 shows the cumulative potential-energy ($V_{step} = \frac{1}{N_{step}} \sum_{k=1}^{N_{step}} V_k$; where N_{step} is the number of completed time steps, and V_k is the potential-energy at the k^{th} step) along both MD and PIMD simulations. After thermalization (roughly 500 ps for MDs and 100 ps for PIMDs), the simulations reach equilibrium.

Next, we compare the difference in geometry fluctuations between MD and PIMD simulations in Fig. 3.3, measured by root-mean-squared deviations (RMSDs) from the initial geometry. The RMSDs of the PIMD present a series of peaks that are higher than those observed in the classical MD simulation. The first of such peaks appears at ∼60 ps and then there is one every 100 ps (the highest one corresponding to a RMSD greater than 3.0 Å). The origin of these RMSD fluctuations is the relative angle of rotation (Φ) of the inner nanotube with respect to the outer one (Fig. 3.3). The outer and inner nanotubes have a 7- and 4-fold

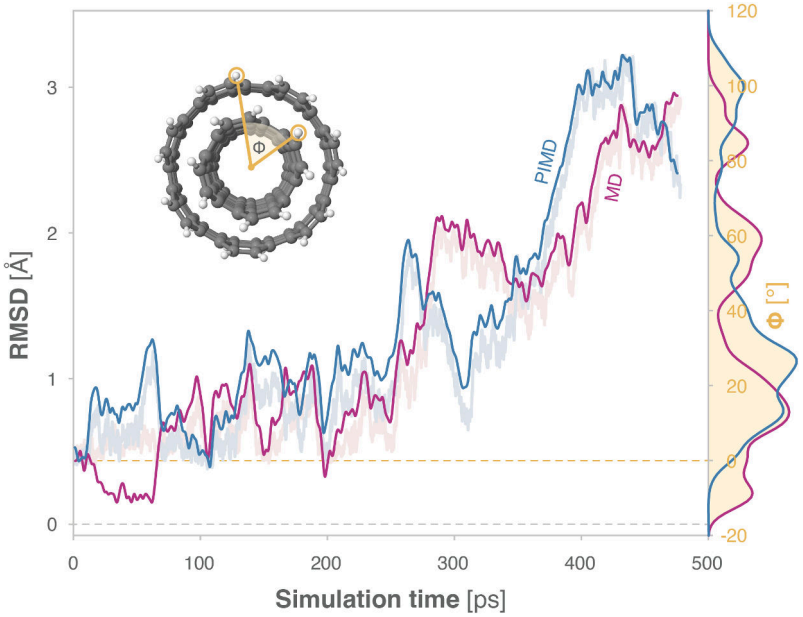

Figure 3.3: Instantaneous root-mean squared deviation (RMSD in Ångström; transparent lines in the background) and angle of the relative rotation (Φ in degrees) between both nanotubes as a function of simulation time (in ps). The RMSDs were computed with respect to the initial configurations used for the simulations.

axis of rotation (the axis parallel to the nanotubes), respectively, meaning that we have the same configuration every ∼13°. However, RMSDs are dependent on atom indices and uncoupled rotations of the nanotubes lead to a different arrangement of the atoms with respect to each other. This causes the increments observed in RMSDs along the simulations since no other degrees of freedom fluctuate as much as the angle Φ.

Indeed, we observe a strong correlation between the RMSD variations and the evolution of the angle Φ in a scale of 360° (a full rotation). The differences in the values of Φ between the MD and PIMD simulations suggest that a coupling of nuclear quantum effects (NQEs) and long-range interactions (resembling existing studies on the stability of different aspirin crystal polymorphs [116]) eases the rotation of one of the nanotubes with respect to the other. The distribution of values of Φ further confirms that NQEs smoothen the rotational profile of the nanotubes. While in the PIMD values of Φ from 40° to 80° are equally sampled, in the MD one can observe two pronounced peaks at around 60° and 80°. It is important to note that the rather large time-scale (100 ps) between each of these rotations indicates that this motion corresponds to low-frequency vibrational modes.

As a final demonstration of the capability of sGDML MLFF models for providing insights into large systems, we computed the molecular vibrational spectra of the buckyball catcher and the double-walled nanotube from both MD and PIMD simulations (Fig. A2). These spectra correspond to velocity auto-correlation functions. The spectra feature peaks corresponding to =C-H stretching (at around 3000 cm^{-1}) and bending (close to 1000 cm^{-1}) modes, as well as those that correlate to C=C vibrations at around 500 and 1500 cm^{-1} accounting for expansions and contractions of the buckyball, the "hands" of the catcher and the nanotubes (for instance, see Refs. [117, 118] for a discussion of the vibrational spectra of the buckyball).

Although MD and PIMD simulations provide similar spectra, the inclusion of NQEs yields more accurate frequencies. Namely, nuclear quantum delocalization leads to a shift of the =C-H stretching mode, which puts the peak closer to 3000 cm^{-1}. This value is in agreement to that of other aromatic and π-π interacting systems. For instance, with experimental and theoretical values of the fundamental C-H stretching modes of benzene and the benzene dimer (mainly the $\nu_{13}(\mathbf{B}_{1u})$ mode) [119]. Hence, PIMD simulations capture some of the anharmonic behavior of the systems and correct the overestimated value of ≈ 3100 cm^{-1} in the classical MD. Differently to $=C-H$ stretching, the parts of the spectra corresponding to "long-range" vibrations (i.e., low-frequency modes) are, in general, consistent among MD and PIMD. This agreement aligns with the fact that both the buckyball catcher and the double-walled nanotube are relatively symmetric systems and that the differences between configuration spaces sampled by MD and PIMD simulations are mostly of local nature. Therefore, even though NQEs promote a low-frequency mode, such as the relative rotation between the nanotubes, these modes are smoothed out in the low-frequency part of the vibrational spectrum.

3.4 Conclusion

Kernel-based FFs are known to be sample efficient, but believed to be limited in their ability to scale well with training set or system size. A key reason is the common use of direct solvers, which factorize the kernel matrix in order to solve the associated optimization problem to train the model. While this approach is numerically stable, it quickly incurs a prohibitive memory and runtime complexity. This dilemma can be evaded using iterative solvers, which essentially allow kernel-based models to be trained similar to neural networks. However, the straightforward application of iterative solvers is difficult when large kernel matrices are involved, due to their notoriously poor numerical conditioning.

In this work, we propose an iterative scheme that enables the robust application of sGDML to significantly larger systems (both, in terms of training set and system size) without introducing any approximations to the original model. This is achieved with a numerical preconditioning scheme which drastically reduces the conditioning number of the learning problem, enabling rapid convergence of a CG iteration. With this advance, we are now able to apply our kernel-based model to large-scale learning tasks that have previously only been accessible to neural networks, while carrying over the sample efficiency and accuracy of sGDML. We attribute the latter to the model's unique ability to represent global interactions on equal footing with local interactions, as a series of numerical experiments demonstrate. Now, molecular systems that exhibit phenomena with far-reaching characteristic correlation lengths can be studied in long-timescale MD simulations.

Our technical development allows future cross-fertilization between both MLFF development approaches: Now, kernel-based MLFFs can capitalize on the massive parallelism available on GPUs and the software infrastructure that enabled scalability of deep neural networks. On the other hand, modeling principles from kernel methods inspire the development of new architectures such as transformers using self-attention [52] and pave the way out of overly restrictive localization assumptions. The exclusive use of on-the-fly model evaluations by iterative solvers also represents a paradigm shift in the way kernel-based MLFFs are typically trained, which opens up new avenues for further developments. We have recently shown, how this makes them amenable to strong differential equation constraints via algorithmic differentiation techniques to simplify descriptor development and further improve data efficiency [120]. With the ability to reconstruct MLFFs for larger systems, the need for

better management of the growing set of molecular features arises. To this end, we have recently proposed a novel descriptor pruning scheme to contract trained models and make them easier to evaluate [121]. One could also envision a systematic construction of local and nonlocal fragments (by generalizing from non-interacting to *interacting* amons [122]) that would enhance the scalability and transferability of global MLFFs. Future research will furthermore explore our significantly better scaling behavior across a broad range of application fields in the physical sciences.

Breakthroughs in MLFF development are often driven by the creation of benchmark datasets that offer ever evolving challenges. Early quantum chemistry datasets such as QM7 [58], MD17 [54], the valence electron densities for small organic molecules [123], ISO17 [124], SN2 [68], or SchNOrb [125, 126], focused on defining useful inference problems and opportunities in quantum chemistry, whereas later benchmarks (e.g. QM7-X [56]) steered the field towards developing more robust transferable models. This chapter presents the MD22 benchmark dataset, which offers new challenges for atomistic models with regard to molecular size and flexiblity that could further advance research on novel MLFF architectures similarly to previous datasets of quantum-mechanical calculations. Despite these advances, MD22 focuses on a limited set of representative biomolecular and supramolecular systems. Many chemically and biologically important motifs remain outside its scope. Realistic modeling of the cellular environment requires datasets with broad coverage across all four major biomolecular classes. To address this need, in the next chapter we develop the QCell dataset, which builds upon the MD22 by extending the chemical and structural diversity of available high-quality QM data for biomolecular fragments and assemblies.

Chapter 4

QCell: Quantum-Mechanical Dataset Spanning Diverse Biomolecular Fragments

---

Parts of this chapter have been published in this or similar form in Ref. 127:

- A. Kabylda, S. Suárez-Dou, N. Davoine, F. N. Brünig, A. Tkatchenko, *arXiv* 2510.09939 (2025).

The accurate modeling of molecular interactions in (bio)chemical systems has long been a central challenge in computational chemistry and biophysics. Existing methods span a spectrum of approaches that introduce tradeoffs between efficiency and accuracy. At one end are quantum mechanical (QM) methods, ranging from highly accurate techniques such as coupled cluster and quantum Monte Carlo to density functional approximations, which vary from non-empirical to heavily parameterized variants trained on curated datasets. At the other end, empirical atomistic force fields achieve high efficiency through fixed functional forms and parameter sets. These approaches have been invaluable for simulating the structure, dynamics, and function of biomolecules, providing either high accuracy or access to biologically relevant timescales [38]. Recently, machine learning force fields (MLFFs) have emerged as a promising alternative, aiming to combine the accuracy of QM methods with the efficiency of classical force fields [5].

However, successful MLFF applications are critically dependent on the availability of diverse and high-quality QM datasets that faithfully represent the chemical space encountered in (bio)molecular systems [128, 129]. Substantial progress has been made in the development of MLFFs, fueled by datasets such as QM7-X [56], MD22 [66], Splinter [130], GEMS [102], SPICE [131, 132], AQM [133], QCML [134], AIMNet2 [135], and OMol25 [136], among many others. These datasets provide extensive coverage for small molecules and proteins, spanning broad elemental diversity, sizes, conformations, charge and protonation states.

The GEMS, QCML, and OMol25 datasets exemplify recent efforts to extend QM coverage across diverse chemical and biomolecular spaces. GEMS employs a hierarchical fragmentation strategy, combining small, transferable fragments of proteins in gas-phase and aqueous environments with larger, system-specific fragments extending up to 18 Å to capture long-range interactions. QCML systematically maps small-molecule chemical space by enumerating species with up to eight heavy atoms across a wide range of elements and

electronic states, providing chemically diverse bonding motifs. OMol25 offers chemically heterogeneous collection spanning small molecules, biomolecular fragments, metal complexes, and electrolytes; its biomolecular subset includes fragmented protein pockets, gas-phase DNA/RNA fragments, protein–protein and protein–ligand complexes. Despite this progress, significant gaps persist for three of the four major biomolecular classes, namely nucleic acids, lipids, and carbohydrates, which together constitute roughly 40% of cellular biomass (Fig. 4.1A).

Biomolecular chemical space possesses distinct characteristics compared to that of small organic molecules or materials. Instead of vast elemental and topological diversity, biomolecular complexity arises primarily from the conformational space accessible to a relatively limited set of recurring chemical building blocks [137]. For instance, proteins are composed primarily of about 20 canonical amino acids, and their intricate functions are dictated by backbone conformations and side-chain rotamer preferences. Similarly, nucleic acids utilize repeating sugar-phosphate backbones and four main nucleobases, with critical conformational variations in sugar pucker and backbone torsions determining their overall structure and interactions. Polysaccharides are formed via various glycosidic linkages between a few monosaccharide types, and their properties depend heavily on the conformations around these linkages. Lipids typically combine a finite set of head groups and fatty acid tails, whose composition and flexibility determine membrane behavior.

In this context, we introduce the QCell dataset, a collection of quantum mechanical data that covers the three major biomolecular classes beyond proteins: lipids, carbohydrates, and nucleic acids, along with relevant ion clusters, water molecules, and non-bonded dimers. The dataset includes 525k newly generated biomolecular fragments, ranging from 2 to 402 atoms, computed at the PBE0+MBD(-NL) level of theory (Fig. 4.1C). By focusing on fundamental building blocks, the QCell dataset provides an accurate quantum description of the semi-local chemical environments and interaction motifs that recur in larger, more complex biological assemblies.

The chemical element distribution in QCell focuses mainly on biologically relevant elements (H, C, N, O, P, and S) with additional coverage of important biological ions (Na^+, K^+, Cl^-, Mg^{2+}, Ca^{2+}). This composition provides deeper conformational sampling of the specific chemical environments most relevant to biomolecular systems and allows the QCell dataset to serve as a specialized complement to existing datasets like QCML [134], QM7-X [56], AQM [133], and GEMS [102]. When combined, they provide extended coverage of chemical space relevant to (bio)molecular simulation, comprising over 41 million data points and spanning 82 chemical elements. The consistent use of the PBE0+MBD(-NL) level of theory across these datasets facilitates their integration into unified training sets for MLFF development. By expanding the coverage to core biomolecular components, QCell enables the development of more comprehensive and transferable MLFFs capable of modeling diverse biological systems.

4.1 Methods

The QCell dataset was generated using a multi-step workflow (Fig. 4.1B): (1) curating a library of biomolecular building blocks and generating initial 3D structures; (2) performing extensive conformational sampling using molecular dynamics or dedicated conformer-generation tools; (3) selecting representative fragments from the resulting ensembles; (4) briefly optimizing the selected fragments with the semi-empirical DFTB+MBD method; and (5) running high-quality quantum-mechanical PBE0+MBD(-NL) calculations.

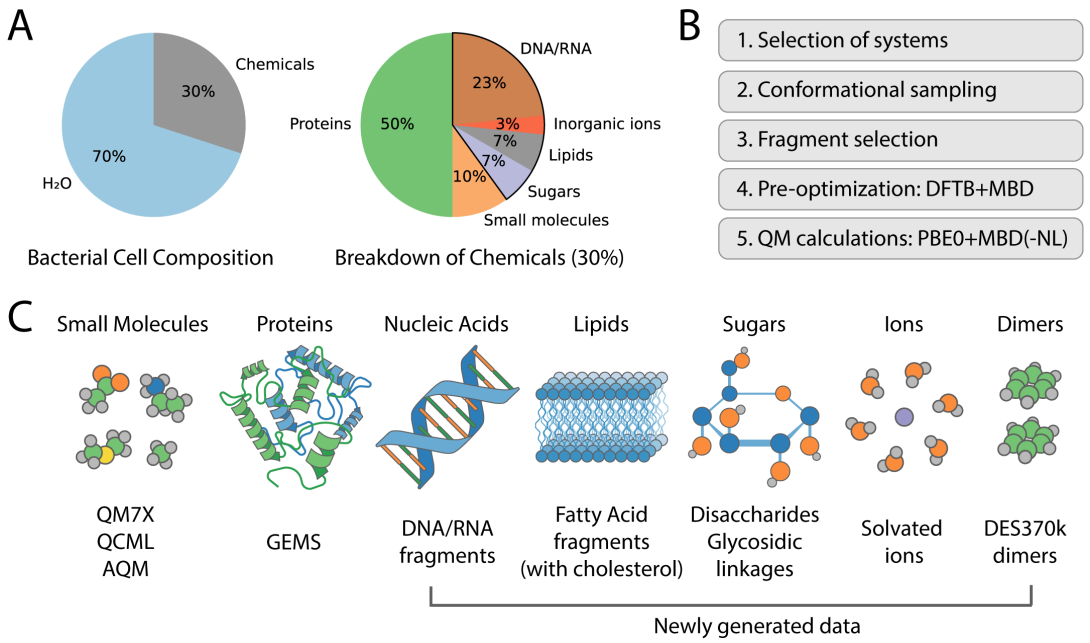

Figure 4.1: Overview. **A)** Composition of a bacterial cell by weight, with a breakdown of the chemical constituents [137]; about 40% of these compounds are not properly covered in existing datasets. **B)** Multi-step workflow used to construct QCell, beginning with the selection of building blocks, followed by conformational sampling and fragment selection, pre-optimization with DFTB+MBD, and finally hybrid PBE0+MBD(−NL) calculations. **C)** Coverage of molecular species at the PBE0+MBD(-NL) level of theory, including entries from existing databases and newly generated QCell data for nucleic acid fragments, lipids, sugars, solvated ions, and dimers.

4.1.1 Generation of representative fragments

The current subsection describes steps 1–4 for each molecular class, detailing the specific methods used for initial structure generation, conformational sampling, fragment selection, and pre-optimization.

Nucleic Acids. Solvated double-helical DNA heptamers in canonical A-, B-, and Z-DNA forms [138, 139] with Na^+ counterions were built using Nucleic Acid Builder [140] and simulated with the OL21 force field [141]. The central base-pair triplets covered all base-pair combinations. Each system was equilibrated for 1 ns in NVT, with the temperature ramped from 100 K to 300 K in 10 ps steps, then run for 10 ns in NPT at 300 K.

From the heptamer trajectories, snapshots saved every 100 ps were used to extract central double-stranded trimer fragments. These trimers were then simulated for 10 ps in NPT at 300 K with strong positional restraints on nucleotide atoms to relax the surroundings. In addition to trimers, solvated DNA base pair dimers were taken from Ref. 142, and a subset of smaller gas-phase RNA fragments were taken from OMol25/rna which were processed from BioLiP2 [136, 143].

Lipids. Initial structures of lipid membranes composed of POPC, POPE, POPG, and POPS phospholipids were generated with the CHARMM-GUI Membrane Builder [144]. These lipids provide a representative set of phospholipid head groups with a palmitoyl–oleoyl–glycerol fatty-acid backbone. To probe sterol–lipid interactions that significantly influence membrane packing and dynamics, we also prepared mixed membranes containing cholesterol: POPC and POPS bilayers were generated at a 3:1 phospholipid-to-cholesterol ratio.

The selected membranes were simulated with the Lipid21 force field [145]. Equilibration involved 20 ps of NVT at 100 K followed by 100 ps of NPT at 300 K using an anisotropic XY–Z barostat. During equilibration, heavy atoms were restrained with a harmonic potential of 5 kcal/mol/Å^2. A 500 ns production simulation was then performed.

The resulting trajectories were sampled randomly over 25000 frames, from which fatty acid monomers, dimers and trimers, 1000 per -mer and each phospholipid type, were selected for subsequent steps. Multimers were identified based on geometric proximity: molecules with geometric centers within 5 Å were considered dimers, and those within 6 Å of a dimer were classified as part of a trimer. For cholesterol-containing fragments, only those clusters including at least one cholesterol molecule were retained.

Carbohydrates. A library of 52 common monosaccharides, including both pentose and hexose structures in α and β anomeric configurations, was used to construct disaccharides in PyMOL [146]. Additionally, we sampled saccharide–peptide linkages, including N-glycosylation involving arginine residues and O-glycosylation involving threonine and serine. These glycosylated residues were capped with ACE and NME groups to mimic peptide termini. In total, 2959 disaccharide structures were generated, representing one unique combination of pentose/hexose and α/β configuration for each glycosidic linkage and 150 saccharide–peptide molecules.

Conformers were generated with the CREST [147–149] program, employing a 12 kcal/mol maximum energy threshold. The resulting ensembles were clustered by linkage dihedral angles, and cluster representatives were selected to ensure broad conformational coverage, retaining at most 100 conformers per amino acid linkage and 20 per disaccharide.

Ions and Water. Solvated ion systems were prepared by placing a single ion at the center of a water box. Bulk water and monovalent ions (Na^+, Cl^-, K^+) in water were simulated in LAMMPS [150] under NPT using the MBpol force field implemented in MBX [151, 152]. Temperature was maintained at 298 K with a Nosé–Hoover thermostat, pressure at 1 bar with a Nosé–Hoover barostat, and the time step was 0.5 fs. Divalent ions (Ca^{2+} and Mg^{2+}) were simulated for 50 ns under NPT with the AMBER force field [32, 153].

To capture solvation effects across different hydration levels, bulk water and water-ion clusters were cut to contain 1–100 water molecules. Trajectories were sampled every 5 ps for monovalent ions and bulk water, and every 10 ps for divalent ions.

General MD settings. All molecular mechanics simulations were carried out using OpenMM [154] under NPT conditions at 300 K, with a 2 fs time step. A Langevin thermostat was applied with a friction coefficient of 1 ps^{-1}, and pressure was maintained at 1 atm using a Monte Carlo barostat. For solvated biomolecules, the TIP3P water model [155] was used, with Na^+ ions serving as the counterions.

Summary and pre-optimization Overall, fragments ranged in size from 2 to 402 atoms, with larger fragments chosen to represent important biological motifs such as DNA base-pair stacking and lipid packing interactions. The selected fragments were pre-optimized using the DFTB+MBD method to avoid high-energy clashes [156]. In addition to that, motivated by the importance of dimers in the early stages of developing general-purpose machine-learned force field SO3LR [157], we also sourced DES370K dimers from Ref. 158 for further calculations.

Table 4.1: Composition of the QCell dataset shown alongside existing PBE0+MBD(-NL) datasets (separated by a double line). Structures are gas-phase unless denoted as solvated (solv.). Abbreviations: bp (base pair), FA (fatty acid), chol. (cholesterol), frag. (fragments).

Dataset	Type	Size	Atoms	Elements	Theory level	Basis set
Small molecules						
QCML	Small molecules	33.5m	2–36	79 elements	PBE0+MBD-NL	tight
QM7-X	Small organic molecules	4.2m	6–23	H, C, N, O, S, Cl	PBE0+MBD	tight
AQM	Drug-like molecules	60k	2–92	H, C, N, O, F, P, S, Cl	PBE0+MBD	tight
Proteins						
GEMS	Bottom-up frag. (solv.)	2.7m	2–120	H, C, N, O, S	PBE0+MBD	def2-TZVPP
SPICE	Dipeptides	34k	26–60	H, C, N, O, S	PBE0+MBD	tight
GEMS	Top-down frag. (solv.)	12k	162–321	H, C, N, O, S	PBE0+MBD	def2-TZVPP
Nucleic acids						
QCell	DNA duplex (2 bp, solv.)	5.3k	186–246	H, C, N, O, Na, P	PBE0+MBD-NL	intermediate
QCell	DNA duplex (3 bp, solv.)	9.5k	297–382	H, C, N, O, Na, P	PBE0+MBD-NL	intermediate
QCell	RNA frag.	20k	14–282	H, C, N, O, Na, Mg, S, P	PBE0+MBD-NL	intermediate
Lipids						
QCell	FA clusters (1–3)	12k	125–402	H, C, N, O, P	PBE0+MBD	intermediate
QCell	FA clusters (1–2) + chol.	4k	148–342	H, C, N, O, P	PBE0+MBD	intermediate
Carbohydrates						
QCell	Disaccharides	59k	35–75	H, C, N, O	PBE0+MBD	tight
QCell	Glycosidic linkages	15k	38–52	H, C, N, O	PBE0+MBD	tight
Ions/Water						
QCell	Solvated ions	25k	4–301	H, O, Na, Cl, K, Mg, Ca	PBE0+MBD-NL	tight
QCell	Water clusters	5k	6–303	H, O	PBE0+MBD-NL	tight
Non-covalent dimers						
QCell	DES370K dimers	371k	2–34	20 elements	PBE0+MBD-NL	tight
QCell New		526k	2–402	20 elements	PBE0+MBD(-NL)	–
Total		41m	2–402	82 elements	PBE0+MBD(-NL)	–

4.1.2 Quantum mechanical calculations

Within the landscape of electronic structure methods, density functional theory (DFT) offers one of the best tradeoffs between efficiency and accuracy and is widely used for generating large QM datasets. DFT provides a hierarchy of approaches that vary in accuracy and theoretical sophistication, as described by the "Jacob's ladder" [16]. Each rung of Jacob's ladder represents a higher level of refinement: local density approximation (LDA), generalized gradient approximation (GGA), meta-GGA, hybrid functionals, and advanced formulations such as the random phase approximation or double-hybrid functionals. Ascending the ladder generally improves accuracy but also increases computational cost.

Separate from this hierarchy, functionals also differ in their degree of empiricism. Minimally empirical functionals (such as PBE [17] or SCAN [19]) are constructed from physical constraints with little or no fitting to reference data, while highly empirical functionals are trained on large datasets of experimental or high-level quantum chemical results. Highly empirical methods can achieve impressive accuracy for systems similar to their training data; however, their transferability to new or unseen systems raises concerns. In particular, the reliability of highly empirical functionals in molecular dynamics simulations remains unclear.

Moreover, advanced empirical functionals are typically optimized against coupled cluster data, which is considered a gold standard for small- to medium-sized molecules. However, there is an ongoing debate on how well this method performs for larger systems, where electronic complexity increases. Recently, it has been shown that the most accurate quantum-mechanical methods, CCSD(T) and quantum Monte Carlo, agree for ligand–pocket motifs within 0.5 kcal/mol [159], but can struggle to provide consistent reference data for larger molecules or supramolecular complexes with extended π–π interactions. Specifically, the disagreement in the binding energy of a 132-atom buckyball ring complex between the two methods can be up to 12 kcal/mol [160].

For this reason, in the current dataset we employed the non-empirical hybrid PBE0 functional with a many-body treatment of dispersion interactions to accurately capture non-covalent interactions MBD(-NL) [17, 18, 29, 161]. Single-point hybrid DFT calculations were performed with the FHI-aims code [162, 163]. For systems involving ions, many-body dispersion interactions were described using the MBD-NL method due to its superior performance for charged systems, while neutral systems employed the MBD approach. Calculations used "tight" basis sets for small- and medium-sized systems and "intermediate" basis sets for subsets containing molecules with more than 350 atoms (see Tab. 4.1). Scalar-relativistic corrections were included via the atomic ZORA formalism for subsets containing ions. The self-consistent field convergence criteria were set to the following values (or tighter): 10^{-5} eV for the total energy, 10^{-3} eV for the eigenvalue sum, 10^{-5} electrons/Å^3 for the charge density, and 10^{-4} eV/Å for the forces. The iteration limit was set to 200, and unconverged calculations were discarded.

4.2 Data records

The resulting QCell dataset contains a total of 525,881 QM calculations for biomolecular fragments spanning diverse conformations (Tab. 4.1).

The QCell dataset is provided in five HDF5 archive files hosted on a Zenodo data repository and is organized according to the classes listed in Tab. 4.1 (lipids, carbohydrates, nucleic acids, ions/water, and dimers) [164]. Each molecule in the HDF5 files includes the 34–35 properties listed in Tab. 4.2. A README file is also provided, containing technical usage details and examples illustrating how to access the information stored in the archives (see the `h5_to_extxyz.py` file).

4.3 Technical validation

To ensure the reliability and consistency of structures in the QCell dataset, we validated the structural diversity across biomolecular classes. Our analysis focused on key geometric descriptors.

Nucleic acids. For DNA fragments, we analyzed intra-strand phosphate–phosphate (P–P) distances and backbone bending angles (Fig. 4.2A). These parameters directly reflect the global geometry of DNA helices: P–P distances measure the spacing of the sugar–phosphate backbone, while bending angles characterize backbone flexibility and helical form. The distributions reproduce the expected values for A-, B-, and Z-DNA (P–P peaks and bending angles spanning canonical ranges), confirming that the QCell captures the conformational diversity of real nucleic acids [165, 166].

Table 4.2: List of properties stored in the QCell dataset. The number of Kohn–Sham eigenvalues varies for each molecule. h_i ratios are present only in MBD data, whereas C_6 and a_0 ratios appear only in MBD-NL data. File-level information is contained in metadata (`metadata/free_atom_energy` and `metadata/fhi_aims_settings`).

#	Symbol	Property	HDF5 key	Unit	Shape
Structure					
1	Z	Atomic numbers	`atomic_numbers`	—	(N)
2	R	Atomic positions	`positions`	Å	(N, 3)
Energies					
3	E_{tot}	Total energy	`total_energy`	eV	()
4	E_{form}	Formation energy	`formation_energy`	eV	()
Total energy components:					
5	$\sum_i \varepsilon_i$	Sum of KS eigenvalues	`sum_of_eigenvalues`	eV	()
6	ΔE_{XC}	XC energy correction	`xc_energy_correction`	eV	()
7	ΔV_{XC}	XC potential correction	`xc_potential_correction`	eV	()
8	E_{FA}	Free-atom electrostatic energy	`free_atoms_elec`	eV	()
9	ΔE_{H}	Hartree energy correction	`hartree_correction`	eV	()
10	E_{vdW}	van der Waals dispersion energy	`vdw_energy`	eV	()
Derived energy quantities:					
11	E_{kin}	Kinetic energy	`kinetic_energy`	eV	()
12	E_{elst}	Electrostatic energy	`electrostatic_energy`	eV	()
Decomposition of the XC energy:					
13	E_{HF}	Hartree–Fock energy	`hf_energy`	eV	()
14	E_x	Exchange energy	`x_energy`	eV	()
15	E_c	Correlation energy	`c_energy`	eV	()
16	E_{XC}	Total XC energy	`total_xc_energy`	eV	()
Forces					
17	F_{tot}	Total forces	`total_forces`	eV/Å	(N, 3)
Total forces components:					
18	F_{HF}	Hellmann–Feynman forces	`hellmann_feynman_forces`	eV/Å	(N, 3)
19	F_{ion}	Ionic forces	`ionic_forces`	eV/Å	(N, 3)
20	F_{mult}	Multipole forces	`multipole_forces`	eV/Å	(N, 3)
21	F_{HFx}	HF exchange forces	`hf_exchange_forces`	eV/Å	(N, 3)
22	F_{Pulay}	Pulay+GGA forces	`pulay_gga_forces`	eV/Å	(N, 3)
23	F_{vdW}	van der Waals forces	`vdw_forces`	eV/Å	(N, 3)
Dipoles and Multipoles					
24	μ	Dipole vector	`dipole`	e × Å	(3)
25	Q_{tot}	Total quadrupole moment	`quadrupole`	e × Å^2	(3)
26	Q_{el}	Electronic quadrupole moment	`electronic_quadrupole`	e × Å^2	(3)
27	Q_{ion}	Ionic quadrupole moment	`ionic_quadrupole`	e × Å^2	(3)
Electronic structure					
28	E_{HOMO}	HOMO energy	`homo_energy`	eV	()
29	E_{LUMO}	LUMO energy	`lumo_energy`	eV	()
30	E_{gap}	HOMO–LUMO gap	`homo_lumo_gap`	eV	()
31	$\{\varepsilon_i\}$	Kohn–Sham eigenvalues	`ks_eigenvalues`	eV	(*)
Other / Atomic properties					
32	Q	Total charge	`charge`	e	()
33	h_i	Hirshfeld ratios	`hirshfeld_ratios`	—	(N)
34	C_6	Atomic C_6 ratios	`c6_ratios`	—	(N)
35	a_0	Atomic polarizability ratios	`a0_ratios`	—	(N)
36	—	Source tag	`source`	—	(text)

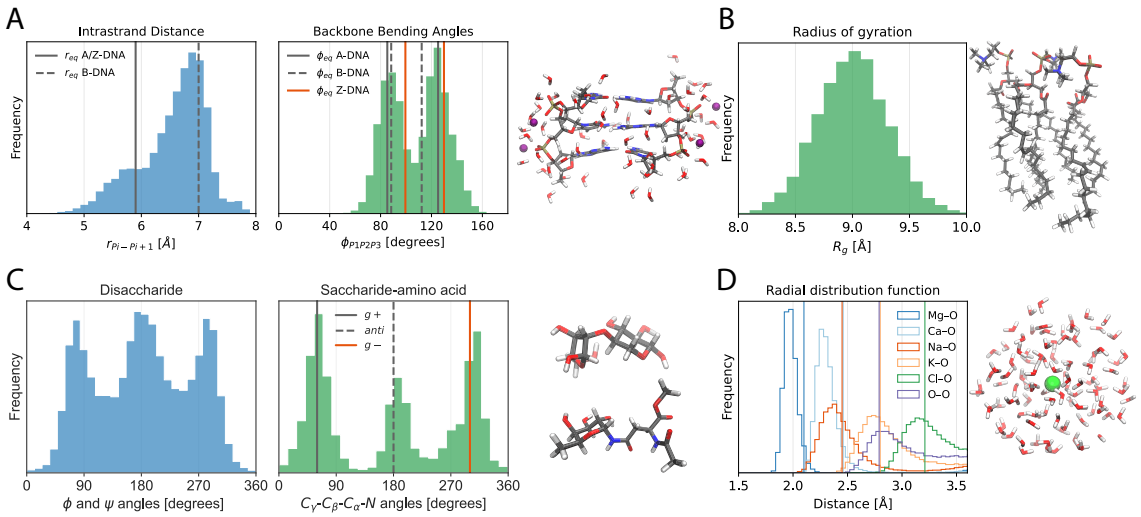

Figure 4.2: Structural distributions across (bio)molecular datasets, with representative structures. A) Distribution of intra-strand phosphate–phosphate distances (left) and backbone bending angles (middle) in DNA trimers, compared to reference values of A-, B-, and Z-DNA [165, 166]. **B)** Radius of gyration distribution of fatty acid fragments with more than 300 atoms. **C)** Distribution of O/N-glycosidic linkage dihedrals in carbohydrates. **D)** Pair distance distributions for ions and water [167–170].

Lipids. For lipids, we examined the radius of gyration of fatty acid fragments, which provides a measure of chain extension and packing (Fig. 4.2B).

Carbohydrates. Carbohydrate conformations are primarily governed by N/O-glycosidic torsional angles, which determine linkage geometry and flexibility. Thus, we verified that the structures span the full torsional space, reproducing all major rotameric states (Fig. 4.2C).

Ions/Water. To assess solvation and intermolecular structure, we analyzed radial distribution functions (RDFs, Fig. 4.2D). RDFs quantify the probability of finding neighboring atoms at a given distance and directly reflect solvation-shell organization. The monovalent ion–oxygen and O−O peaks in water match experimental hydration distances [167, 168, 170]. For the divalent ions Mg^{2+} and Ca^{2+}, the RDF peaks are slightly shifted relative to experimental values, reflecting that the initial structures were generated using empirical force-field simulations. Despite these deviations, the distributions capture the correct solvation-shell organization and remain within physically meaningful ranges.

Machine learning models. To evaluate the dataset in a realistic application, we trained a state-of-the-art machine learning force field on all subsets listed in Tab. 4.1 and measured its accuracy on held-out test configurations. As a representative model, we employed the SO3LR architecture (SO3krates with long-range terms) [53, 157], which is particularly well suited for biomolecular systems because it explicitly incorporates long-range electrostatic and dispersion interactions, as well as electronic degrees of freedom, and can therefore describe charged and open-shell structures. The model was trained to predict formation energies and atomic forces from atomic numbers, coordinates, and electronic state information (total charge and spin multiplicity). Fig. 4.3 summarizes the force mean absolute errors (MAEs) across different molecular classes and model sizes. The errors decrease systematically with increasing model capacity, reaching values below 1 kcal/mol/Å for most subsets. This highlights both the internal consistency of the QCell and the ability of modern MLFFs to generalize across chemically diverse systems.

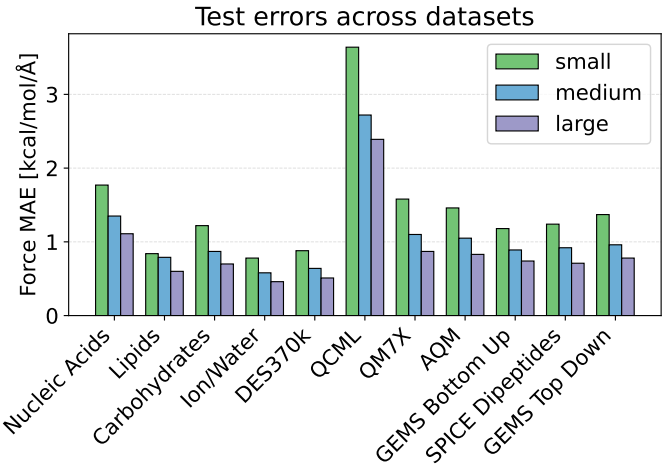

Figure 4.3: Test set errors for machine learning force fields. Force mean absolute errors [kcal/mol/Å] for SO3LR models of increasing size (small, medium, large) across all the training subsets, illustrating systematic error reduction with model capacity and consistent data quality across chemically diverse systems.

4.4 Conclusion

This chapter addresses the critical data bottleneck that has hindered the development of truly general-purpose machine learning force fields for biomolecular simulation. We introduced the QCell dataset, a comprehensive and diverse collection of over 525,000 new quantum-mechanical calculations spanning the fundamental building blocks of all four major biomolecular classes: proteins, lipids, carbohydrates, and nucleic acids, along with relevant ions and water clusters. By employing a consistent and high-quality PBE0+MBD(-NL) level of theory, QCell accurately describes the semi-local chemical environments and interaction motifs that recur throughout complex biological assemblies. It serves as a complement to existing datasets, contributing to a combined resource of over 40 million datapoints that covers a broad chemical space relevant to life sciences. Ultimately, QCell provides the essential foundation required to train the next generation of accurate and transferable MLFFs, paving the way for models capable of simulating complex, heterogeneous (sub)cellular environments with quantum fidelity.

While QCell addresses the challenge of chemical diversity through a systematically constructed quantum-mechanical dataset for biomolecular fragments, the training of global machine learning force fields for large molecules remains limited by the scaling and redundancy of fully coupled representations. The following chapter therefore introduces a reduced global descriptor that preserves essential nonlocal information while enabling efficient learning for extended molecular systems.

Chapter 5

rGDML: Efficient interatomic descriptors for accurate machine learning force fields of extended molecules

Parts of this chapter have been published in this or similar form in Ref. 121:

- A. Kabylda, V. Vassilev-Galindo, S. Chmiela, I. Poltavsky, A. Tkatchenko, *Nature Communications* **14**, 3562 (2023).

Reliable atomistic force fields are essential for the study of dynamics, thermodynamics, and kinetics of (bio)chemical systems. Machine learning force fields (MLFFs) are lately becoming a method of choice for constructing atomistic representations of energies and forces [5, 45, 50, 54, 64, 71, 95, 171–181]. Contrary to traditional computational chemistry methods, MLFFs use datasets of reference calculations to estimate functional forms which can recover intricate mappings between molecular configurations and their corresponding energies and/or forces. This strategy has allowed to construct MLFFs for a wide range of systems from small organic molecules to bulk condensed materials and interfaces with energy prediction errors below 1 kcal/mol with respect to the reference *ab initio* calculations [5, 54, 64, 106, 113, 171, 182–188]. Applications of MLFFs already include understanding the origins of electronic and structural transitions in materials [182], computing molecular spectra [106, 183–185], modeling chemical reactions [186], and modeling electronically excited states of molecules [187, 188]. Despite these great successes of MLFFs, many open challenges remain [5, 189, 190]. For instance, the applicability of MLFF models to larger molecules is limited, partly due to the rapid growth in the dimensionality of the descriptor (i.e. a representation used to characterize atomic configurations).

A descriptor used to encode the molecular configurations determines the capability of an MLFF to capture the different types of interactions in a molecule. Therefore, descriptors are designed to contain features that emphasize particular aspects of a system or to highlight similar chemical or physical patterns across different molecules or materials. Many different descriptors have been proposed to construct successful MLFFs for specific subsets of the vast chemical space [47, 57–59, 61, 62, 64, 173, 191–194]. However, there is no guarantee that a given descriptor is capable of accurately describing all relevant features throughout

high-dimensional potential-energy surfaces (PESs) that characterize flexible molecular systems [189]. The main challenge here is to balance the number of features required for a given ML model to describe simultaneously the interplay between short and long-range interactions. One possible approach to address this challenge is to increase the complexity of descriptors by adding explicit features to model specific interactions [92, 189]. However, such solution usually yields descriptors that are high-dimensional and inefficient for large systems. As an alternative solution, several approaches have been proposed to generate reduced descriptors, targeting specific properties of interest [195–199]. Such reduced descriptors have led to insights into complex materials, and this approach has also been applied to MLFFs, specifically to reduce ACSFs and SOAP representations [200–203].

The descriptors discussed above correspond to local MLFF models, where only a certain neighborhood of atoms is considered within a specified cutoff distance. Such locality approximation is usually employed in MLFFs to enhance their transferability and applicability for larger systems than the given training set. However, as a downside, accounting for long-range interactions requires additional effort. Therefore, some recent MLFF models [71, 91, 92, 106–108, 204] have integrated correction terms to account for certain long-range effects (e.g. electrostatics), but long-range electron correlation effects are still not well characterized. It is evident that the field of MLFF combined with physical interaction models is rapidly growing and developing, but a definitive solution to these challenges has not yet been found. In general, ML models should be able to correctly describe i) the nonadditivity of long-range interactions, ii) the strong dependence of such interactions on the environment of interacting objects, and iii) the nonlocal feedback effects that give rise to their multiscale nature. Addressing these features requires the development of flexible yet accurate and efficient MLFFs that do not rely on strictly predefined functional forms for interactions or impose characteristic length scales.

Alternatively, one can switch to so-called global descriptors, such as the Coulomb matrix, where all interatomic distances are considered. Unfortunately, such global descriptors scale quadratically with system size. In addition, reducing the descriptor dimensionality in global models is an unsolved challenge. For example, it is evident that most short-range features (e.g. covalent bonds, angles, and torsions) should be preserved when constructing accurate MLFFs. In fact, the number of local features scales linearly with the system size. In contrast, the number of nonlocal (long-range) features scales quadratically and a general coarse-graining procedure to systematically reduce nonlocal features does not exist yet.

To address these challenges, in this work we propose an automatic procedure for identifying the essential features in global descriptors that are most relevant for the description of large and flexible molecules. We apply the developed approach to identify efficient representations for various systems of interest, including a small molecule, a supramolecular complex, and units of all four major classes of biomolecules (i.e. proteins, carbohydrates, nucleic acids, and lipids): aspirin (21 atoms), buckyball catcher (148 atoms), alanine tetrapeptide ($AcAla_3NHMe$, 42 atoms), lactose disaccharide (45 atoms), adenine-thymine DNA base-pairs (AT-AT, 60 atoms), and palmitic fatty acid (50 atoms). Employing the reduced descriptor results in an improvement of prediction accuracy and a two- to four-fold increase in computational efficiency. Moreover, an analysis of the features that are selected by our reduction procedure suggests that these features follow certain patterns that are explained by both interaction strength and statistical information they provide about atomic fluctuations. In particular, while most short-ranged features are essential for the PES reconstruction, a linearly scaling number of selected nonlocal features are enough for an ML model to describe collective long-range interactions.

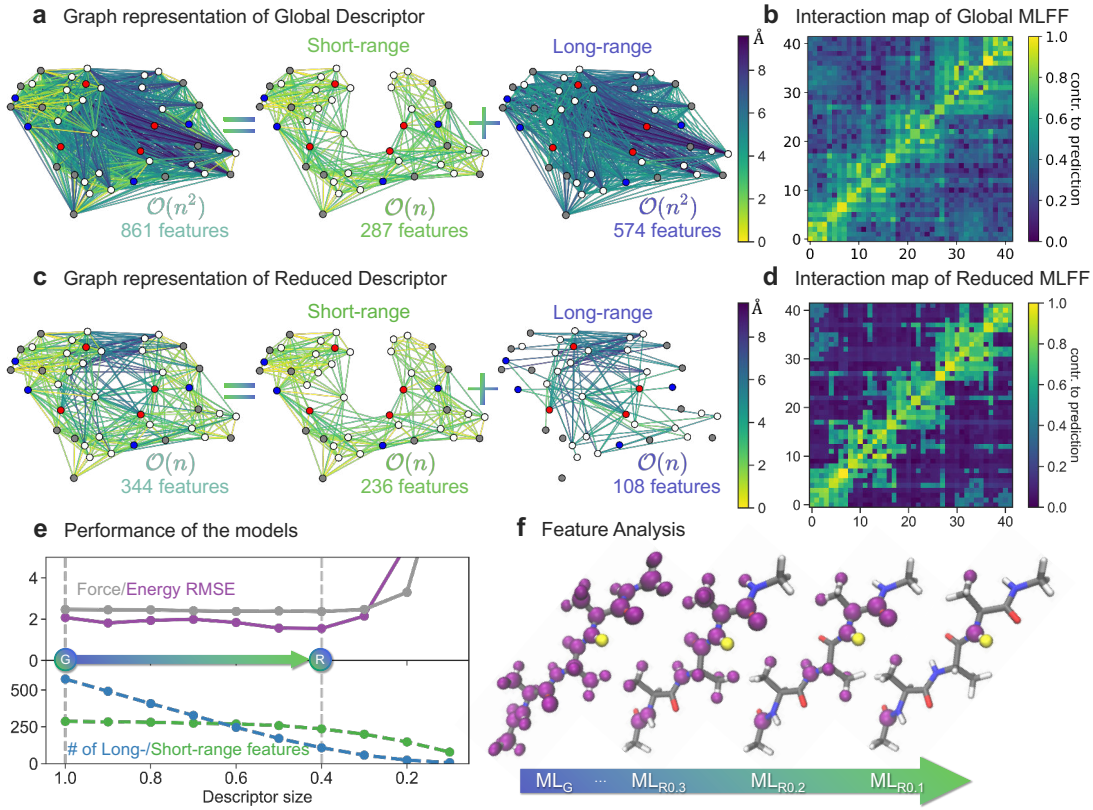

Figure 5.1: Overview of the descriptor reduction scheme. a, c Graph representation of global and reduced descriptors for $AcAla_3NHMe$, and its decomposition into short- and long-range features with corresponding scaling with the number of atoms, n. The color of nodes indicates atom type: H - white, C - grey, N - blue, O - red. The color of the edges indicates average distance between atoms. **b, d** Interaction map of global and reduced Machine Learning Force Fields (MLFFs). Each square in the heatmaps represents a given pair of atoms in the molecule (atom indices start from 0). The colorbar is in log scale normalized to the 0-1 range and goes from dark magenta (small) to yellow (large) for the average contributions to the force prediction. **e** Performance of the global and reduced models: energy (in kcal/mol) and force (in kcal/mol/Å) root mean square errors (RMSEs) as a function of the size of the descriptor (upper panel). The RMSE values were calculated on a test set of ∼80k points, distinct from the training (1k) and validation sets (1k). Decomposition of the descriptors by short- and long-range features (lower panel). Descriptor sizes in x-axis go from 1 to 0, where 1 corresponds to a default global descriptor and 0 to an empty descriptor. **f** Feature analysis in the global (denoted as ML_G) and reduced models (denoted as ML_{RX}, where X indicates the descriptor size). Hydrogen atom highlighted in yellow keeps interactions with atoms highlighted in purple. The arrow indicates transition of the employed descriptor from default global to substantially reduced ones.

5.1 Reduced GDML algorithm

The quadratic scaling of global descriptors with molecular size, especially their long-range part, becomes a considerable challenge with the increasing number of atoms. For molecules containing just a few dozen of atoms, such descriptors are, in fact, substantially over-defined. For example, the number of degrees of freedom (DOF) uniquely defining a configuration of a molecule with N atoms is $3N - 6$. At the same time, the Coulomb matrix and related global descriptors contain $N(N-1)/2$ DOFs. Thus, such descriptors will span a much larger space than what is effectively needed, making ML models harder to optimize and compromise their performance/accuracy. In the case of a complete interatomic inverse

distances descriptor (a simplified version of the Coulomb matrix [58]), the interatomic interactions can be visualized as a fully connected graph with atoms as nodes and descriptor features as edges. For example, Fig. 5.1a shows such a descriptor for the $AcAla_3NHMe$ molecule containing 861 features. Each edge of the graph represents a dimension in the descriptor space, where an ML model should be trained.

The large dimensionality of the descriptor significantly complicates the learning task. The interaction map (Fig. 5.1b) shows how the (s)GDML model interprets the interatomic interactions when the entire global interatomic inverse distance descriptor is employed (values are averaged over 1000 configurations, see Methods for further details). As a projection of complex many-body forces into atomic components, this partitioning is non-unique and is mainly determined by the chosen descriptor. In turn, the simpler the descriptor space, the more straightforward the task for the ML model. One can see that the interaction map shown in Fig. 5.1b is rather non-uniform and complex, meaning the (s)GDML model needs to be able to reproduce a complex mapping between the descriptor (861 dimensions) and force (126 dimensions) spaces.

5.1.1 Reduced descriptors

The automatized descriptor reduction procedure proposed in this work significantly simplifies the learning task and noticeably decreases the complexity of the interaction map. To reduce the size of the descriptor, we employ a definition of similarity between system states, which plays a pivotal role in kernel-based ML models. Namely, we assume that the least important descriptor features for the similarity measure can be omitted without losing generality in an MLFF model (see Methods for further details). The reduced descriptor space of the optimal ML model (344 features) is shown as a graph in Fig. 5.1c. Interestingly, the short-range part (236 features) of the graph is practically unaltered by the reduction procedure. In contrast, a small fraction of long-range features (108 out of 574 in the full descriptor) enables an accurate account of all relevant long-range forces while greatly simplifying the interaction map (Fig. 5.1d). The reduced descriptor still completely and uniquely represents the molecular configurations. For $AcAla_3NHMe$, we can remove up to 60% of the initial global descriptor while preserving the accuracy of the (s)GDML model (Fig. 5.1e). This is a remarkable result since many approaches for reducing the dimensionality of the learning task (e.g. low-rank approximations of the kernel matrix) typically lead to performance degradation because the model has to compensate for omitted features in some arbitrary reduced representation [205].

We also analysed the features that are kept to interpret the content of the reduced descriptor (Fig. 5.1f). One sees that the reduced descriptors are not a simple localization because features' importance is not necessarily correlated with the distance between atoms. The proposed selection scheme considers both the strength of the interactions between atoms and the information the features provide about the molecular structure. The latter means that the optimal nonlocal features depend on the training dataset and the respective sampled region of PES. As a possible future outlook, one could consider switching from selecting atom-centered nonlocal features from the initial global descriptor to projecting them into more efficient and general collective coordinates. This would provide us with effective interaction centers for large molecules, similar to those employed by the TIP4P [206] or Wannier centroid [204] models of water. In turn, this would enable the construction of automatized coarse-grained representations preserving the MLFFs accuracy, a long-desired tool for simulating complex and large systems.

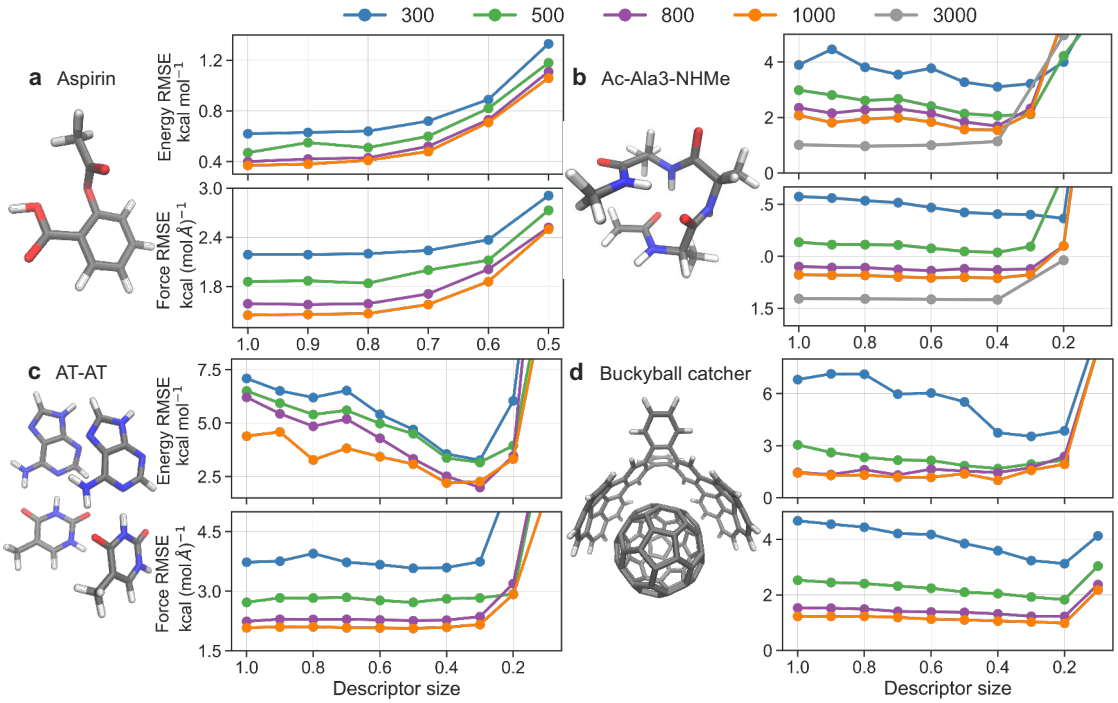

Figure 5.2: Accuracy of the models with reduced descriptors. Energy (in kcal/mol) and force (in kcal/mol/Å) root mean square errors (RMSEs) as a function of the size of the descriptor. RMSEs of Gradient Domain Machine Learning (GDML) models for aspirin (**a**), $AcAla_3NHMe$ (**b**), AT-AT (**c**), and the buckyball catcher (**d**) trained on 300, 500, 800, 1000, and 3000 configurations. Descriptor sizes in x-axis go from 1 to 0, where 1 corresponds to a default global descriptor and 0 to an empty descriptor.

5.2 Assessment and analysis of the rGDML model

The proposed descriptor reduction scheme is general and applicable to a wide range of systems. Fig. 5.2 shows GDML performance curves of energy and forces for aspirin, $AcAla_3NHMe$, AT-AT, and the buckyball catcher as a function of the size of the descriptor for different sizes of the training set. The aspirin molecule represents a rather small semi-rigid molecule, for which one can already build accurate and data-efficient MLFFs [50, 54, 61, 64, 71, 171, 207]. The other molecules represent large and flexible systems that constitute a challenge for existing ML models. For each of these systems, GDML models with 300, 500, 800, and 1000 training points were trained using descriptors of different sizes. For the $AcAla_3NHMe$ molecule, due to its size and flexibility, we have also constructed the model using 3000 training points.

For a small molecule such as aspirin (210 features in the original descriptor), the descriptor showing the lowest RMSEs is the default global descriptor. Nevertheless, removing up to 30% of the descriptor only slightly affects the predictions of the model. Whereas, for $AcAla_3NHMe$ (861 features), AT-AT (1770 features), and the buckyball catcher (10878 features) one can significantly reduce the size of the descriptor while obtaining even more reliable predictions regardless of the training set size (Fig. 5.2b-d). For instance, models trained on 1000 training samples with a descriptor size reduced by 60% provide energy and force RMSEs that are up to 2.2 kcal/mol and 0.2 kcal/mol/Å lower than those of the models employing default global descriptors. The different behavior in prediction accuracy with decreasing size of the descriptor between aspirin and other bigger molecules is mainly

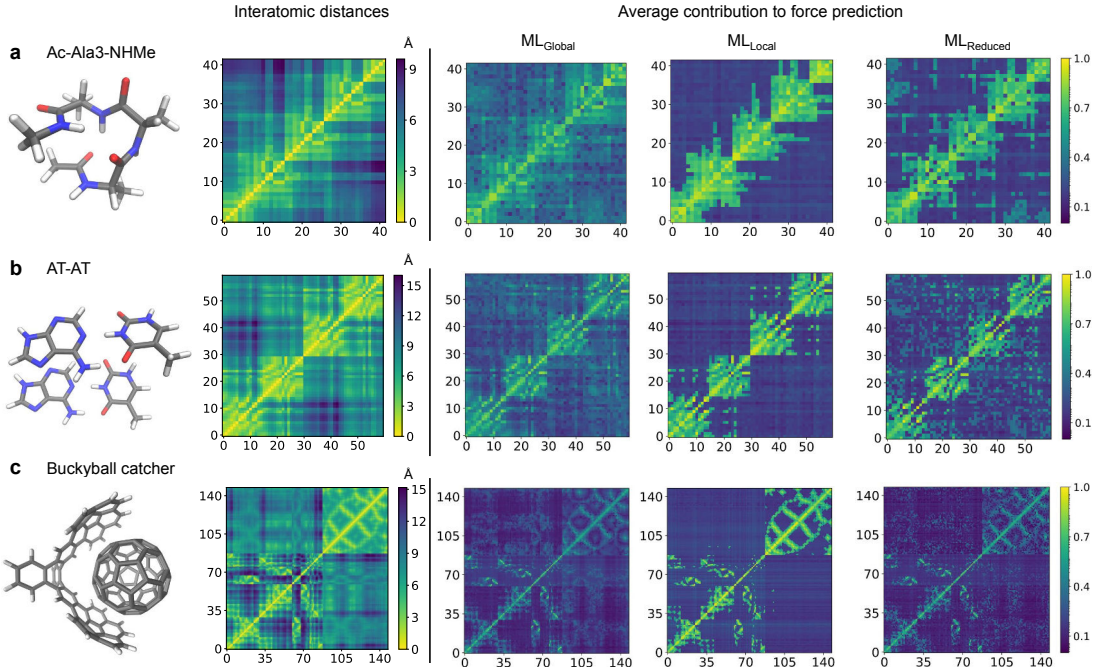

Figure 5.3: Complexity of interaction patterns. Heatmaps of average interatomic distances (in Å) and average contributions (in log scale normalized to the 0-1 range) of each atom to the force prediction of all atoms computed from 3000 configurations of $AcAla_3NHMe$ (**a**), AT-AT (**b**), and 1000 configurations of the buckyball catcher (**c**). Each square in the heatmaps represents a given pair of atoms in the molecule (atom indices start from 0). The scale goes from yellow (short distances) to dark magenta (long distances) for interatomic distances, and from dark magenta (small contributions) to yellow (large contributions) for the contributions to the force prediction.

caused by the differences in their size. Indeed, with increasing molecule size, the quadratic redundancy of the feature space offers greater reduction potential. Therefore, reducing the number of features contained in a global descriptor should be a routine task for building ML models of large molecules.

5.2.1 Improved description of interactions

The improved accuracy of models trained using reduced descriptors is a consequence of how well those models describe the interatomic interactions. Fig. 5.3 shows the interaction heatmaps and interatomic-distance heatmaps averaged over 1000 conformations for $AcAla_3NHMe$, AT-AT, and the buckyball catcher. For each of the molecules, we use the following GDML models trained with 1000 configurations: i) the ML_{Global} model, ii) a model trained using a $\frac{1}{r}$ descriptor mimicking a local descriptor by removing all features involving distances greater than 5.0 Å (the typical value for the cutoff radius in local descriptors) in at least one configuration in the dataset (ML_{Local}), and iii) a $ML_{Reduced}$ model. We remark that the prediction accuracy of our reduced models for large molecules is superior to state-of-the-art kernel-based local GAP/SOAP [175] ML model (Fig. A3).

For the ML_{Global} models (containing 861 features for $AcAla_3NHMe$, 1770 for the AT-AT, and 10878 for the buckyball catcher) the contributions are evenly distributed among different pairs of atoms regardless of the distance between them. This allows the model to effectively capture long-range interactions, but as a downside may degrade the ability to optimally resolve all short-range ones. Conversely, the ML_{Local} models (with a size equal to 33%,

17%, and 15% of the size of the default global descriptor for $AcAla_3NHMe$, AT-AT, and the buckyball catcher, respectively) only rely on the local environment of the molecule. This is confirmed by the contributions of the atoms to the force prediction of other atoms, which are directly related to the magnitude of the corresponding interatomic distances. Thus, the ML_{Local} models offer a more adequate description of short-range interactions but completely neglect those interactions arising from distances greater than the selected cutoff. One of the drastic consequences of such neglect is the instability of MD simulations performed using these local MLFFs. Finally, the $ML_{Reduced}$ models offer an improvement over both ML_{Global} and ML_{Local} models by achieving an adequate description of the local environment of the molecule and, at the same time, keeping the relevant information for describing nonlocal interactions. Therefore, using a reduced descriptor leads to ML models that provide a balanced, faithful description of all essential interactions in a given system.

We further compare the transferability of the global and reduced models by training them on compact structures and testing on extended structures of the tetrapeptide (and vice-versa). To do that, we split the tetrapeptide dataset based on the distance between the furthest atoms (ranges from ~ 8 to 14 Å) with a threshold of 12 Å (and 9.5 Å). The comparison of the force and energy RMSEs shows that the reduced models are more accurate than global models when dealing with unseen outlier extended or compact structures of the tetrapeptide (see Tab. A3). This suggests that overdetermined global ML model underperforms due to conflicting information from excessive features and that the model with a reduced descriptor indeed provides an improved description of interactions.

5.2.2 Efficiency and stability of reduced-descriptor models

The models obtained using reduced descriptors, together with the increment in accuracy provide up to a ten-fold increase in efficiency during training and four-fold during deployment (Tab. A4). Improvement in efficiency results from the fact that there are less noisy features in the reduced model, which leads to lower per-iteration costs. For training, such efficiency can only be obtained with a recently developed iterative solver [66], while the evaluation speedup is always present when using the GDML model.

We also checked the stability of molecular dynamics simulations employing reduced models. We found that optimally reduced models for $AcAla_3NHMe$ ($ML_{R0.6}$ with 3000 training points, 0.3 fs timestep) and the buckyball catcher ($ML_{R0.2}$ with 1000 training points, 0.5 fs timestep) are stable and the corresponding energy is conserved during the dynamics at 300 K for 3 ns.

In complex systems, long simulations can be unstable due to incomplete dataset even with the default global descriptor. For example, AT-AT shows degraded stability when encountering rare/new configurations that are not well sampled in the dataset (decomposes due to leaving the planar configuration or due to hydrogen transfer from T to A). Still, we find that simulations can remain stable for 3 ns with the default $ML_{R1.0}$ and the reduced $ML_{R0.5}$ models (1000 training points, 0.1 fs timestep). Further stability depends on the accuracy of the underlying original model. Thus, with increasing complexity of the PES, one should consider using active learning to detect "dark" states and adding them to the training process regardless of the employed descriptor.

Models with substantially reduced descriptors can only describe a smaller part of the PES and lead to artificial behaviour (e.g. steric clashes or fragmentation). Such artifacts happen with a higher probability in flexible molecules where atom pairs corresponding to removed features might come close, and their relative position cannot be neglected anymore. For

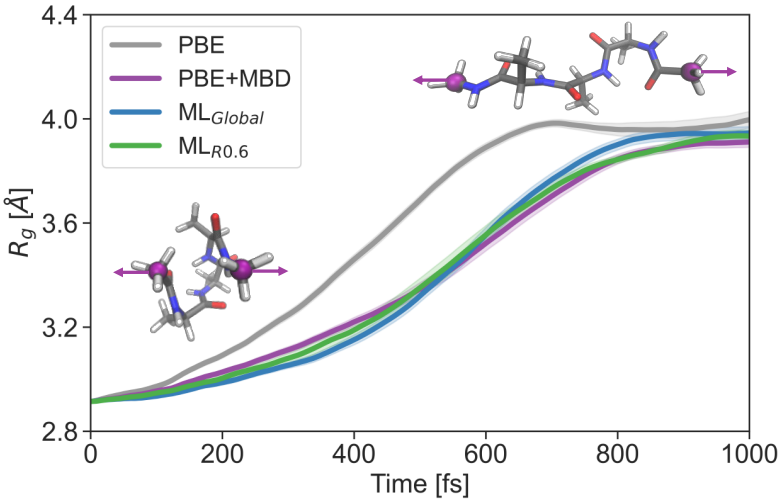

Figure 5.4: Steered dynamics between folded and extended states of the tetrapeptide. The tetrapeptide undergoes unfolding due to an external force acting parallel to the connecting line between two terminal carbon atoms. The gyration radius, averaged over 30 runs, is represented by the solid lines, with the shaded areas indicating the standard error.

example, we encounter steric clashes in AcAla$_3$NHMe when using ML$_{R0.4}$ trained on 1000 configurations at 0.5 fs timestep, even though test errors are lower than those of the global ML$_{R1.0}$ model. Therefore, smaller prediction errors do not always lead to a more reliable ML model when tested in an extended simulation of several nanoseconds (see also Ref. 208).

In order to further demonstrate the stability and the broad applicability of reduced GDML models, we study the evolution of the tetrapeptide molecule from a compact to an extended structure under a constant external force of 10 pN applied in opposite directions to the two terminal carbon atoms (Fig. 5.4). Statistics were collected using 30 simulations with different initial velocities following the Boltzmann distribution at 300 K (Fig. A5). We ran simulations using the global and reduced models trained with 5000 training points. In addition, we ran simulations at two levels of theory - PBE and PBE+MBD - and used the resulting data for validation (see Methods for further details). We measured the structural compactness using the gyration radius and compared the dynamical properties of the models. As expected, due to the absence of attractive dispersion interactions in the PBE simulations, the tetrapeptide unfolded faster than in the PBE+MBD ones (on average it took ∼550 and ∼750 fs, respectively, to reach $R_g = 3.8$ Å). Both the global and reduced models agreed well with the PBE+MBD results, indicating their accuracy and reliability. Also, this confidently shows that the reduced model preserves all the information needed to describe long-range interactions with *ab initio* accuracy.

After confirming the reliability of the reduced model, we further investigated the conformational space of the tetrapeptide to enhance our understanding of its behavior. To achieve this, we conducted multiple simulations in parallel, with an accumulated time of 50 ns (Fig. A6). This approach allowed us to obtain a converged folding and unfolding distribution, which was visualized using the ψ_2 angle in the probability distributions for the central residue. Our analysis reveals that the tetrapeptide populates the extended state with a probability of 13% on the 50 ns time scale (Fig. A7).

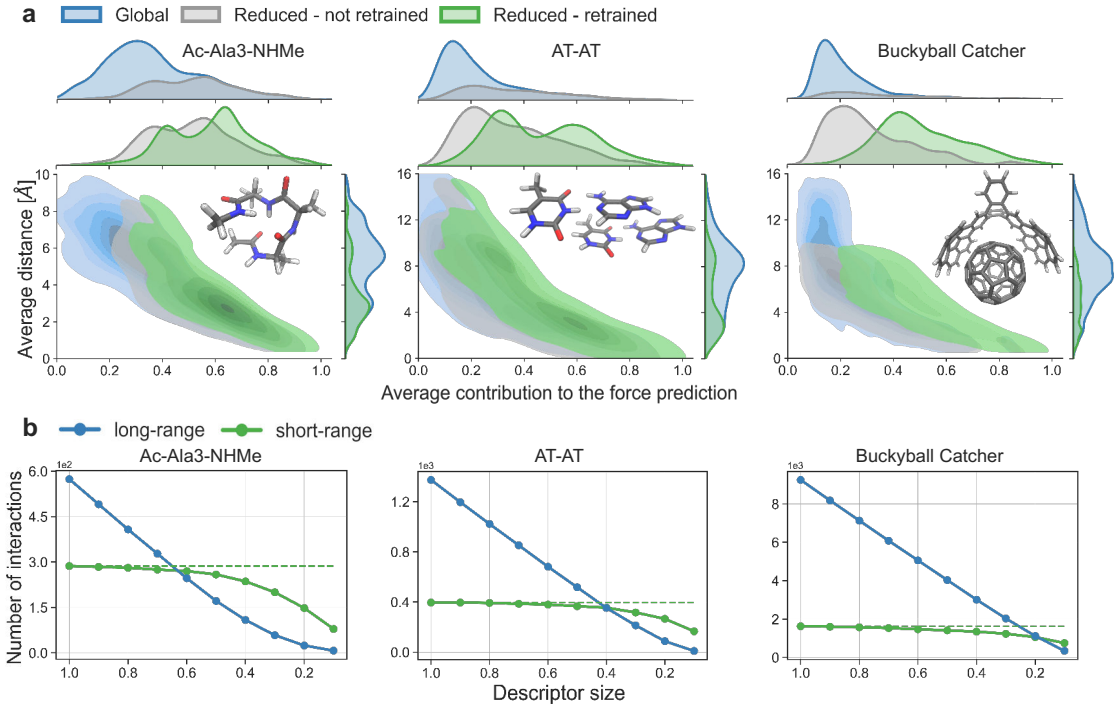

Figure 5.5: Analysis of relevant interatomic features. **a** Distributions of average distance of pairwise features and average contribution of features to the force prediction for $AcAla_3NHMe$, AT-AT and the buckyball catcher using bivariate kernel density estimate plots (Machine Learning models: global - green, reduced before retraining - gray, reduced after retraining - green). The marginal charts on the top and right show the distribution of the two variables using density plot. The average values were obtained from all configurations in the datasets. The x-axis is in log scale normalized to the 0-1 range. **b** Decomposition of the reduced descriptor by short- and long-features for $AcAla_3NHMe$, the AT-AT, and the buckyball catcher. Pairwise features with the average distance below 5 Å across all configurations in the dataset are counted as short-range (green line), long-range otherwise (blue line). Dashed green line represent number of short-range features in the global descriptor. Descriptor sizes in x-axis go from 1 to 0, where 1 corresponds to a default global descriptor and 0 to an empty descriptor.

5.2.3 Relevance of interatomic descriptor features

The importance of descriptor features is not always related to the magnitude of their contribution to the model predictions (Fig. 5.5a). As expected, the features contributing the most to the force predictions (above 0.5-0.6 a.u.) in the global model are all included in the reduced descriptor (see the top marginal plots). These strongly contributing features are primarily associated with short interatomic distances. In contrast, the selected features corresponding to medium- and long-range interatomic distances span almost all contribution ranges. For instance, some weakly contributing features that describe an average distance as large as 15 Å are included in the reduced descriptor of the AT-AT system. The distribution of the selected features is skewed towards the weak contributions upon increasing the molecular size (compare gray density distribution of three molecules in top marginal plots, Fig. 5.5a). Interestingly, the distribution of the contribution of the selected features is significantly shifted towards larger values after retraining the ML model (center marginal plots, Fig. 5.5a).

Further analysis reveals that contribution of particular features in the global model can range from linear to stochastic with respect to the interatomic distance (Fig. A8a). The

proportion of stochastic features increases with the size of the systems and the size of training set (Fig. A8b). In the reduced models after retraining most of the selected features have a high coefficient of determination, R^2 (Fig. A8c). Contribution of "linear" features to the force prediction decreases quadratically with distance (slope ≈ -2), suggesting the prominence of Coulombic contributions to the interatomic forces.

These findings are general and valid for all descriptor reduction degrees. Although we do not rely on any characteristic lengthscale, we show the effect of descriptor reduction approach on different types of interactions as conventionally defined when imposing lengthscales. Fig. 5.5b shows a decomposition of the reduced descriptor in short- and long-range features for different descriptor sizes. We consider the feature as short-ranged if the distance between two atoms across all configurations in the dataset is below 5 Å, and long-ranged otherwise. In all the cases, feature selection removes prevalently long-range features and 10-20% of the local features that always lie under 5 Å (compare dashed and solid green lines in Fig. 5.5b). Nevertheless, we emphasize that removing local features might worsen the stability in flexible systems and the best-practice solution is to keep them all in the reduced descriptor.

We construct our datasets using dynamics simulations at the PBE+MBD level of theory (see Methods for further details). However, long-range descriptor features are also kept in PBE calculations, as the PBE functional includes both long-range electrostatics and polarization, despite the semi-local nature of the exchange-correlation term. When we account for MBD, up to 22% of removed features can change depending on the degree of reduction (∼5% for the optimal $\mathrm{ML}_{R0.4-0.6}$ models). This is consistent with the fact that MBD contribution to the energy is relatively small compared to the PBE energy. Nevertheless, MBD contribution can greatly influence the dynamics of chemical systems, particularly in the case of large and flexible molecules. For example, MBD is essential for accurately evaluating the stability of aspirin polymorphs [116], standing molecules on surfaces [209], and interlayer sliding of 2D materials [210]. Therefore, it is essential to perform PBE+MBD calculations in order to generate a reliable dataset in the first place.

5.2.4 Analysis of patterns in relevant interatomic features

We analyze particular atoms and their selected chemical environment in the reduced descriptors to identify the trends in the features that an ML model considers essential. To make our results general for a wide range of (bio)molecules, we include in our discussion lactose and palmitic acid. Fig. 5.6 shows examples of such features for all of our test systems. One can see some general patterns. For example, shielded atoms (C, N, O) in backbone chains usually keep solely local features (Fig. 5.6a, e, d). Most interactions between the first-, second-, and third-nearest neighbors are intact. Such behavior is expected and reflects the importance of the local environment in describing interatomic interactions.

The outer atoms are responsible for accounting for the relevant nonlocal features in the molecules (Fig. 5.6c, d and Fig. 5.1f). The flexibility of the molecule defines the number of such features in the reduced descriptor. For instance, outer hydrogen atoms in semi-rigid molecules (e.g. lactose) only require local information in the descriptor. In contrast, flexible molecules (e.g. $\mathrm{AcAla_3NHMe}$ and palmitic acid) present a combination of short-range features to describe local bond fluctuations and a substantial number of nonlocal features for accurately characterizing essential conformational changes, such as the folding and unfolding of peptide chains (Fig. 5.6c, e, f).

There are more complex patterns like those observed in the AT-AT base pairs and the buckyball catcher (Fig. 5.6c, b). In the former, two hydrogen atoms in the imidazole ring

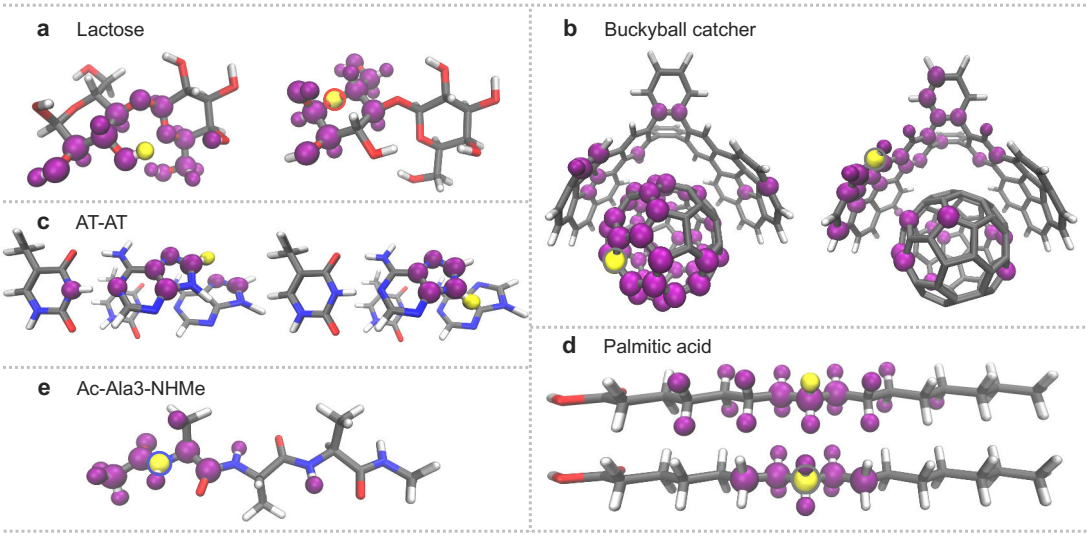

Figure 5.6: Examples of features in the reduced descriptors. The features were obtained from the reduced $\mathrm{ML}_{R0.3}$ descriptor for lactose (**a**), AT-AT (**c**), palmitic acid (**d**) and $\mathrm{AcAla_3NHMe}$ (**e**); the $\mathrm{ML}_{R0.2}$ descriptor for buckyball catcher (**b**). Root mean square errors as a function of the descriptor size for lactose and palmitic acid can be found in Fig. A4. Atoms highlighted in yellow keep in the reduced descriptor the features that correspond to interactions with purple atoms. Outline colors on reference atoms highlighted in yellow indicate their chemical symbols (hydrogen - no outline, carbon - grey, nitrogen - blue, oxygen - red).

of adenine retain contrasting sets of features. This is because some features contained in an optimal descriptor depend on the phenomena sampled in the datasets (e.g. MD trajectories at certain temperatures). In the buckyball catcher, the reduced descriptor reveals that the symmetry of the system is important. Only a few features from the catcher are needed to effectively describe the interaction with any atom in the buckyball (and vice-versa).

5.2.5 Linear scaling of descriptors with molecular size

As a result of the descriptor reduction procedure, we obtain reduced descriptors that scale linearly with the number of atoms (Fig. 5.7). This is achieved by revealing a minimal complete set of nonlocal features that describe long-range interactions. The number of such features is similar to the number of short-range ones (Fig. 5.5b). Therefore, reduced descriptors not only scale linearly with the system size but also the corresponding prefactor (~ 10) is a few orders of magnitude smaller to that of local descriptors (~ 1000). However, we must note that the Hessian of the GDML model is still of the same size ($3N \times 3N$, where N is the number of atoms), though many of the entries are omitted in the reduced model (as shown in Fig. 5.3). This noticeably reduces the computational cost of global ML models (up to a factor of four for studied systems) and paves the road to constructing efficient global MLFFs for systems consisting of hundreds of atoms.

Linear-scaling electronic structure methods, such as linear scaling density functional theory, are valuable tools for *ab initio* simulations of large systems. These methods assume that the electronic structure has a short-range nature and achieve linear-scaling by truncating elements beyond a given cutoff radius or below a given threshold [211]. In contrast, our approach does not impose any localization constraints - selected features span a wide range of distances and contributions. Descriptor reduction procedure allows us to find the right low-dimensional embedding of the high-dimensional PES. Furthermore, the linear-scaling

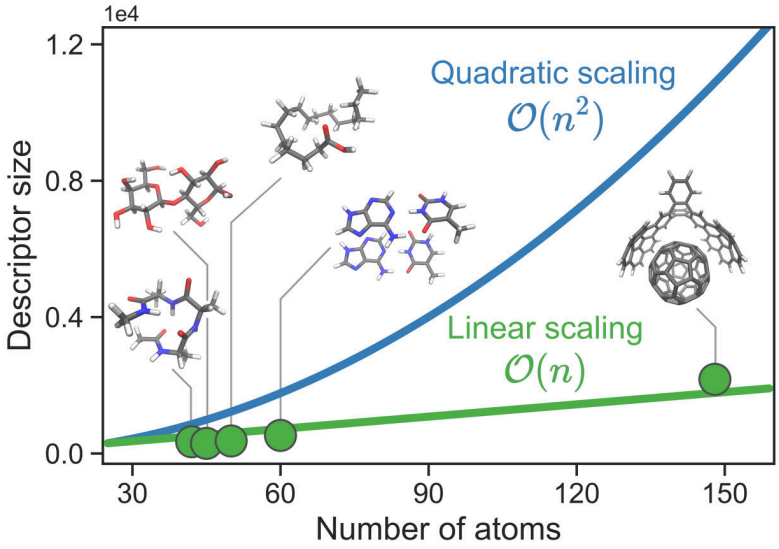

Figure 5.7: Scaling of the default and reduced global descriptors. Dots represent reduced descriptors for the molecules used in this study.

electronic structure methods are less accurate than the original $\mathcal{O}(N^3)$ approaches by design. The models trained with reduced descriptors provide predictions with equal or better accuracy than the original models since the deprecated features, as we have shown, constitute noise in the model.

5.3 Conclusion

Efficient modeling of large molecules requires descriptors of low dimensionality that include relevant features for a particular prediction task. Our results show that beyond increasing the efficiency, such descriptors improve the accuracy of ML models compared to those constructed with default global or local descriptors. This is the consequence of simplifying the interaction patterns which should be learned by ML models in the reduced descriptor spaces. The resulting MLFFs allow long-time molecular dynamic simulations demonstrating stable behavior in the regions of the PES represented in the training sets.

A detailed analysis of the nonlocal descriptor features relevant for accurate energy/force predictions shows non-trivial patterns. These patterns are related to the molecular structure and composition, balancing the strength of the interactions associated with the descriptor features and statistical information about atomic fluctuations these features provide. In particular, we show that the descriptor features related to interatomic distances as large as 15 Å can play an essential role in describing nonlocal interactions. Our examples cover units of all four major classes of biomolecules and supramolecules, making the conclusions general for a broad range of (bio)chemical systems.

The key outcome of the proposed descriptor reduction scheme is the linear scaling of the resulting global descriptors with the number of atoms. We found that global descriptors for large molecules are over-defined and equally accurate models can be constructed with just a handful of long-range features that describe collective long-range interactions. This behavior seems to be general for large molecular systems, provided that reliable reference data is available.

Overall, our work makes substantial advances in the broad domain of machine learning force fields. These advances include (i) demonstrating the potential for linear scaling in global MLFFs for large systems, (ii) analyzing the nonlocal interatomic features that contribute to accurate predictions, and (iii) demonstrating the accuracy, efficiency, and stability of reduced models in long time-scale molecular dynamics simulations. As such, this is a critical step for building accurate, fast, and easy-to-train MLFFs for systems with hundreds of atoms without sacrificing collective nonlocal interactions.

The rGDML framework demonstrates that global machine learning force fields can be made both accurate and computationally tractable by compressing nonlocal descriptors to their physically relevant components. Extending such approaches to chemically diverse, large-scale biomolecular simulations, however, ultimately requires models that combine transferable local learning with explicit, physically grounded long-range interactions, which motivates the hybrid architecture introduced in the next chapter.

Chapter 6

SO3LR: Molecular Simulations with a Pretrained Neural Network and Universal Pairwise Force Fields

Parts of this chapter have been published in this or similar form in Ref. 157:

- A. Kabylda, J. T. Frank, S. S. Dou, A. Khabibrakhmanov, L. M. Sandonas, O. T. Unke, S. Chmiela, K.-R. Müller, A. Tkatchenko, *Journal of the American Chemical Society* **147**, 37, 33723 (2025).

The desire to perform quantitative molecular dynamics simulations based solely on nuclear charges and electron numbers has been expressed by many researchers, including Schrödinger [212], Dirac [213], and Feynman [214]. Despite a century filled with groundbreaking advances, this vision has yet to be fully realized in the realm of molecular simulations. Existing approaches often make significant trade-offs concerning **E**fficiency, **A**ccuracy, **S**calability, or **T**ransferability (EAST) [215]. In this work, we argue that several methodological advances in the field of atomistic modeling have coalesced to bring us closer to achieving fully quantitative, quantum-accurate molecular simulations. While the journey toward this ultimate goal may be lengthy and complex, it is a pursuit that is undeniably worthwhile and requires a collaborative community-based effort.

A key challenge in molecular simulations is the construction of an atomistic force field model that satisfies the EAST requirements mentioned above [4, 5, 33, 190, 216–219]. Traditionally, force fields are obtained either from approximate but fast mechanistic expressions, or accurate but computationally prohibitive *ab initio* electronic structure calculations. Both approaches compromise either accuracy or efficiency, restricting the scope of problems that can be addressed. Recently, machine-learned force fields (MLFFs) have started to bridge this gap by exploiting statistical models with high flexibility [5, 95, 190, 219, 220]. Unlike classical force fields, MLFFs exhibit unprecedented transferability across chemical space; however, scalability with system size remains an issue.

Many challenges remain to be addressed to enable EAST-compliant and MLFF-driven general molecular simulations. Among these we mention the development of data and computationally efficient semilocal interatomic interaction models [45, 46, 50–52, 60, 61, 68, 71, 171], explicit treatment of (many-body) long-range interactions [94, 190, 221], building datasets with comprehensive coverage of chemical space [55, 56, 131, 133, 158, 222–225], and development of modern GPU-enabled molecular simulation frameworks [226–228].

Within this work, we take decisive steps towards solving the aforementioned challenges for organic (bio)molecules. Our solution combines recent advances from chemical and computational physics, machine learning (ML), and established techniques from the force field community. Semi-local interactions are described by the SO3krates ML model [53] using a many-body anharmonic treatment. The physical pairwise terms include short-range Ziegler-Biersack-Littmark repulsion [229], long-range electrostatic interactions, and a recently derived universal interatomic van der Waals (vdW) dispersion potential [230]. Complementarity between the different terms is achieved through careful parametrization on a curated and comprehensive dataset of 4M molecular structures computed with essentially non-empirical and widely applicable PBE0+MBD functional, leading to the SO3LR model (we suggest pronunciation "solar").

We demonstrate the applicability and stability of SO3LR in nanosecond-long simulations of small biomolecular units, polyalanine systems, bulk water, crambin protein, N-linked glycoprotein, and a lipid bilayer. SO3LR can be scaled to simulations involving up to ∼200k atoms with a latency of ∼3 μs/atom/step on a single H100 GPU, thus approaching sizes and timescales relevant for realistic biomolecules.

6.1 SO3LR components

Generally applicable molecular simulations can be directly related to an accurate description of interactions across systems and length-scales. To achieve these objectives, SO3LR decomposes the potential energy into four contributions (Fig. 6.1A):

$$E_{\text{Pot}} = \underbrace{E_{\text{ZBL}}}_{\text{short-range}} + \underbrace{E_{\text{SO3k}}}_{\text{semi-local}} + \underbrace{E_{\text{Elec}} + E_{\text{Disp}}}_{\text{long-range}}, \tag{6.1}$$

where E_{ZBL} is a short-ranged term inspired by Ziegler-Biersack-Littmark (ZBL) repulsion between nuclei (see Supporting Information for more details), E_{SO3k} is the semi-local many-body potential learned by the SO3krates model, and E_{Elec} and E_{Disp} are the long-ranged electrostatic and dispersion energies, respectively. All potential terms influence each other, and a careful optimization procedure based on a diverse dataset of ∼4 million points ensures a broad applicability. The proposed combination of model design, dataset curation and joint optimization, resolves the trade-offs in the EAST requirement and is described in the following paragraphs.

6.1.1 SO3krates

The cornerstone of our approach, which enables high computational efficiency and accuracy (**EA**ST), is the SO3krates model [52, 53] – an MLFF based on an equivariant graph neural network (a compact introduction into invariance and equivariance is given in the Supporting Information). Given atomic positions R, atomic numbers Z, total charge Q, and total spin S, it predicts atomic quantities

$$E_i, q_i, h_i = \text{SO3krates}(R, Z, Q, S), \tag{6.2}$$

where E_i are atomic energies, q_i are partial charges and h_i are Hirshfeld ratios (ratio of effective and free-atom volume, $V_{\text{eff}}/V_{\text{free}}$) [231]. The semi-local energy contribution is then

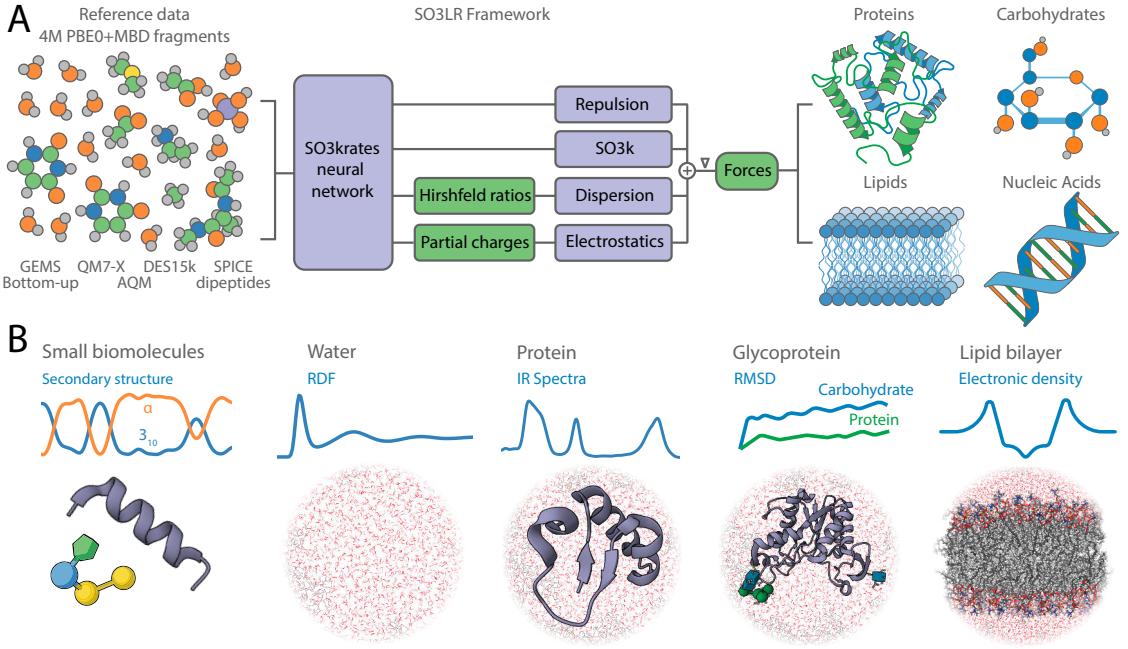

Figure 6.1: Overview of the SO3LR model and simulation results. (A) SO3LR combines the SO3krates neural network with physically inspired interactions, including ZBL repulsion, electrostatics, and a universal pairwise van der Waals potential for dispersion which interact directly with the neural network model. All building blocks are jointly trained on a carefully curated data set which covers a broad range of chemical space and interaction classes. SO3LR enables simulations of small biomolecular units of all four major types of biomolecules, and large-scale simulations of three types. (B) This includes large-scale simulations of liquid water, protein, glycoprotein, and a lipid bilayer.

calculated as the sum over the atomic energies

$$E_{\mathrm{SO3k}} = \sum_{i=1}^{N} E_i. \tag{6.3}$$

The predicted atomic energies contain information about atoms in the direct local neighborhood *and* beyond via mean field updates, which is why we refer to the energy prediction as semi-local. The mean-field nature of these updates cannot account for all types of interactions and is limited by an effective cutoff, which is *upper bounded* by the local cutoff times the number of update steps (the effective cutoff in SO3LR is 13.5 Å).

6.1.2 Long-range dispersion and electrostatics

To improve the description of long-range effects and extend the description beyond semi-local environments, we explicitly incorporate electrostatics and universal pairwise interatomic vdW potentials. Both partial charges and vdW parameters depend on the atomic environment and are predicted by the SO3krates neural network. As shown in Fig. A18, the distributions of Hirshfeld ratios and partial charges for $\mathrm{AcAla_{15}NHMe}$ exhibit substantial element- and environment-specific variability. For example, Hirshfeld ratios for hydrogen span a broad range from 0.55 to 0.8, partial charges of oxygen and nitrogen range from -0.4 to -0.3. The contributions of these long-range interactions to atomic forces are computed using automatic differentiation tools, hence the variation of charges and vdW parameters

with atomic displacements are fully accounted for. Both long-range terms follow the correct pairwise asymptotic decay. This is an important requirement for the scalability (EA**S**T) to length scales that exceed those covered by the training data.

Dispersion interactions are calculated using universal pairwise interatomic vdW potentials derived from quantum Drude oscillators (QDO) [230]:

$$E_{\text{Disp}} = -\sum_{i<j}\sum_{n=3}^{5} \frac{C_{2n}^{ij}}{r_{ij}^{2n} + R_{d,ij}^{2n}}, \tag{6.4}$$

where C_{2n}^{ij} are long-range interatomic dispersion coefficients, and $R_{d,ij}^{2n}$ are the vdW radii of the Becke-Johnson damping function [232]. The radii are defined based on atomic polarizabilities as [233]

$$R_{d,ij} = \gamma R_{\text{vdW}}^{ij} = 2\gamma \left(\frac{a_0^4 \alpha_{\text{fsc}}^{-4/3}}{4\pi\varepsilon_0} \frac{\alpha_i + \alpha_j}{2} \right)^{1/7}, \tag{6.5}$$

where a_0 is the Bohr radius, $\alpha_{\text{fsc}} = e^2/4\pi\varepsilon_0\hbar c$ is the fine-structure constant, with γ being a single tunable parameter in the dispersion module that controls the damping strength. Atomic polarizabilities α_i and dipole-dipole dispersion coefficients C_6^{ij} are obtained using the Tkatchenko-Scheffler method [24] with the ML-predicted Hirshfeld ratios h_i, whereas the scaling relations from the QDO model [230, 234] are applied to generate higher-order dispersion coefficients C_8^{ij} and C_{10}^{ij}.

Electrostatic interactions are modeled using a damped Coulomb potential

$$E_{\text{Elec}} = \sum_{i<j} q_i q_j \frac{\text{erf}(r_{ij}/\sigma)}{r_{ij}}, \tag{6.6}$$

where q_i are the ML-predicted partial charges, and σ is a hyperparameter that controls the damping strength. We remark that the semi-local SO3k module has the capacity to accurately describe multipolar interactions, hence we limit our model to leading-order electrostatics.

Coupling between long-range and semi-local energy contributions arises from the structure of the potential energy prediction (Eq. 6.1). The long-range modules have a non-zero energy contribution for all atomic pairs, including those within the local cutoff of the MLFF. As such, the functional forms of the long-range potentials alter the potential which is learned by the SO3krates model. The choice of damping hyperparameters γ and σ controls the fine balance between semi-local, electrostatic and dispersion interactions. In principle, SO3krates can learn to correct for arbitrary choices of σ and γ up to the local cutoff; however, damping particularly impacts dynamical behavior in MD simulations, although the overall model performance remains largely unaffected. This can be attributed to the fact that semi-local and long-range interactions are coupled nonlinearly through parameter and hyperparameter optimization in both modules. Hence, the damping hyperparameters were fine-tuned on the S66x8 benchmark dataset [235].

6.1.3 Optimization on diverse training data

All SO3LR modules are jointly optimized on a diverse dataset that spans a broad chemical space and various interaction classes. This enables transferability (EAS**T**) between all four

major types of biomolecules.

The comprehensive dataset has been a key factor in the development of our MLFF. It is a collection of extensive quantum mechanical data from both small and large molecules, as well as non-covalent systems with and without explicit solvation. To this end, we combined five datasets: 2.7M bottom-up GEMS fragments [102], 1M QM7-X molecules [56], 60k AQM gas-phase molecules [133], 33k SPICE dipeptides [131], and 15k DES molecular dimers [158] (Fig. A9 and Tab. A5 for more details). The first three datasets were originally computed at the PBE0+MBD level of theory [18, 29]. For consistency, we recomputed the remaining two datasets using the same reference method.

The PBE0+MBD method combines the non-empirical hybrid functional with an explicit treatment of many-body long-range dispersion interactions. This level of theory has been shown to yield excellent agreement with both high-level quantum chemistry methods and experimental data. Its accuracy has been demonstrated for a wide range of systems, including polypeptides [236, 237], supramolecular complexes [160], and molecular crystals [238, 239].

The datasets are complementary in terms of conformational space and chemical diversity, covering 8 elements predominantly present in biosystems (H, C, N, O, F, P, S, and Cl). Specifically, the QM7-X dataset encompasses the chemical space of small organic molecules, while the AQM dataset includes medium-sized drug-like molecules. DES molecular dimers were incorporated to improve the description of non-covalent interactions. SPICE dipeptide structures were added to enhance the accuracy for the protein-containing systems. Lastly, the GEMS bottom-up dataset contains gas-phase and explicitly micro-solvated protein fragments, as well as structures with gas-phase water clusters.

A natural and fundamental question concerns whether the 4M molecular conformations used to train SO3LR adequately span the chemical space relevant to (bio)molecular systems. Although a comprehensive assessment of this coverage remains an open challenge for future investigation, approximate estimates offer valuable insight. For example, chemically accurate simulations of medium-sized peptides, such as alanine tetrapeptide, have been demonstrated using fewer than 1,000 conformations when using SO3krates as the underlying MLFF [53]. Extrapolating this to the entire combinatorial space of tetrapeptides composed of the 20 natural amino acids yields a naïve estimate of approximately 160M conformations required for complete coverage. This is a significant overestimate, primarily because of the high redundancy of local chemical environments [122], a property that underpins the transferability of foundational MLFF models across families of molecules or materials.

Indeed, a meaningful measure of chemical diversity can be captured by the number of *orbits* – distinct equivalence classes of atoms that possess identical local environments across different molecular configurations [240]. The number of orbits depends on the effective distance cutoff used to build molecular subgraphs. For the alanine tetrapeptide, there are 10 orbits up to second neighbors, meaning that 100 conformations per orbit is a sufficient training size. We took several datasets with published geometries (the dataset used to train SO3LR, SPICE [131], GDB-13 [222]) and calculated the number of orbits by building graphs up to second neighbors. By doing so, we obtained a range of 10,000 – 50,000 orbits for molecular datasets containing 8-10 atomic species. This preliminary analysis demonstrates that between 1M and 5M molecular configurations should be enough to cover a broad chemical space of (bio)molecular systems. This analysis of course holds true only because SO3LR uses message passing and an explicit physical model for long-range interactions, meaning that only shorter-range orbits need to be accurately captured by the graph neural

network architecture.

Optimization of the model parameters is done by minimizing a combined loss

$$\begin{aligned}\mathcal{L} = &\frac{\lambda_F}{B}\sum_{b=1}^{B}\frac{1}{N_b}\sum_{i=1}^{N_b}\left|\left|\vec{F}_{i,\text{true}} - \vec{F}_{i,\text{pred}}\right|\right|_2^2 \\ &+ \frac{\lambda_\mu}{B}\sum_{b=1}^{B}\left|\left|\vec{\mu}_{b,\text{true}} - \vec{\mu}_{b,\text{pred}}\right|\right|_2^2 \\ &+ \frac{\lambda_h}{B}\sum_{b=1}^{B}\frac{1}{N_b}\sum_{i=1}^{N_b}\left(h_{i,\text{true}} - h_{i,\text{pred}}\right)^2,\end{aligned} \tag{6.7}$$

where $\vec{F}_i$ are atomic forces, $\vec{\mu}$ are molecular dipoles, and h_i are Hirshfeld ratios, with λ as trade-off parameters between the individual loss terms. The Hirshfeld ratios and partial charges are predicted by SO3krates (Eq. 6.2), and the forces are obtained as the gradient *w.r.t.* the atomic positions of the potential energy (Eq. 6.1). The partial charges are indirectly trained based on dipole moments, instead of direct fitting to reference partial charges. This approach reduces the model's sensitivity to the choice of charge-equilibration scheme and enhances transferability [241]. It should be noted that the model is trained on forces, rather than on energies and forces, which ensures accuracy of relative energy predictions only. Further training details can be found in the Supporting Information.

6.2 SO3LR evaluation

A force field that is truly EAST-compliant should be able to accurately simulate systems of varying nature and size. To demonstrate the capabilities and limitations of SO3LR, we first evaluate its performance in test and benchmark sets to assess its precision in predicting forces, binding energies, dipole moments, and Hirshfeld ratios. This is followed by an analysis of the dynamics of small biomolecular units from the MD22 benchmark dataset [66]. We then investigated the folding and stability of polyalanine systems *in vacuo*, which depend on a delicate interplay of various interactions. Before transitioning to simulations of larger biosystems, we performed a detailed analysis of water dynamics. Finally, we extend the evaluation to large-scale molecular dynamics simulations of more complex systems, including a protein, a glycoprotein, and a lipid bilayer, all in explicit water (Fig. 6.1B).

6.2.1 Test set and benchmark errors

We begin the evaluation of the model by analyzing its accuracy *w.r.t.* quantum mechanical reference data (Tab. 6.1). The test set comprises 10k randomly sampled structures from each of the QM7-X and GEMS bottom-up fragments (all other training sets were fully utilized during training). Furthermore, we recalculated 100 random structures from six MD22 reference molecules at the PBE0+MBD/tight level of theory and we evaluated the model using ∼300 $AcAla_{15}NHMe$ structures and ∼5600 top-down fragments of crambin that were used in the training of system-specific models in [102].

The model demonstrates good performance in predicting forces, dipole moments, and Hirshfeld ratios. A closer examination of Tab. 6.1 reveals two key observations. First, fragments from curved carbon-based systems, such as the buckyball catcher and double-walled nanotube, are absent from the training set, which is reflected in the increased errors. This suggests that further expansion of the data set would be necessary to achieve complete

Table 6.1: Root Mean Square Error of the model on various test sets. Force (eV/Å), dipole moment vector (e × Å), and Hirshfeld ratios. Dash indicates no data.

Dataset	**Size**	**# atoms**	**Force**	**Dipole**	**Hirsh. rat.**
QM7-X	10000	6–23	0.069	0.031	0.012
GEMS bottom-up	10000	2–120	0.086	0.048	–
$AcAla_3NHMe$	100	42	0.052	0.051	0.012
DHA	100	56	0.053	0.072	0.012
AT-AT	100	60	0.168	0.238	0.025
Stachyose	100	87	0.105	0.119	0.016
Buckyball Catcher	100	148	0.384	4.030	0.029
Nanotube	100	370	0.717	2.950	0.039
$AcAla_{15}NHMe$	312	162	0.055	–	–
Crambin top-down	5624	230–321	0.057	–	–
TorsionNet500	12000	13–37	0.088	0.061	0.019

transferability across the chemical space. However, the molecular dynamics simulations of the six MD22 molecules remain stable, as discussed in the following subsection. It is important to note that the commonly reported MD22 errors [50, 51, 53] correspond to system-specific models, which evaluate the performance in distribution and are therefore lower. Second, by comparing the errors of $AcAla_3NHMe$, $AcAla_{15}NHMe$, and crambin top-down fragments, which are identical, we conclude that the SO3LR model is scalable to large solvated protein fragments and that long-range modules effectively describe intermolecular interactions, despite being trained only on small fragments.

To assess the model's accuracy on conformational energetics, we evaluated torsional energy profiles using the TorsionNet500 benchmark [242], recomputed at the PBE0+MBD level of theory (Fig. A19). The model achieves a mean absolute error (MAE) of 1.03 kcal/mol, demonstrating accurate performance across diverse torsional motifs commonly encountered in biosimulations. It should be mentioned that the absence of certain functional groups, such as triazole and trifluoromethylthio moieties, in the training set significantly increases the average errors.

To evaluate the quality of electrostatic interactions, we benchmarked partial charge prediction using the QM7b and AlphaML datasets, which were computed at the LR-CCSD/d-aug-cc-pVDZ level of theory [243]. SO3LR accurately predicts dipole moments with mean absolute errors (MAE) of 0.13 D in magnitude and 5.1° in angle orientation (Fig. 6.2A). This performance is comparable to hybrid DFT at the B3LYP/d-aug-cc-pVDZ level of theory, which attains 0.09 D [243]. Our training set contains molecules from the QM7x dataset, which includes perturbed structures from QM7b. The AlphaML benchmark, on the other hand, contains a wider set of compounds, including DNA/RNA nucleobases, amino acids, carbohydrates, drugs, and hydrocarbons. Both B3LYP/d-aug-cc-pVDZ and PBE0+MBD/tight methods yield an MAE of 0.10 D on this dataset, while SO3LR achieves an MAE of 0.14 D (Fig. 6.2A), showcasing transferable and accurate prediction of dipole moments, which is crucial to calculate reliable electrostatic interactions.

Next, we evaluate noncovalent interaction energies on a comprehensive SAPT10k benchmark computed at the SAPT2+(3)(CCD)/aug-cc-pVTZ level of theory [244]. It consists of 70 subsets, featuring challenging binding motifs dominated by electrostatics and/or dispersion

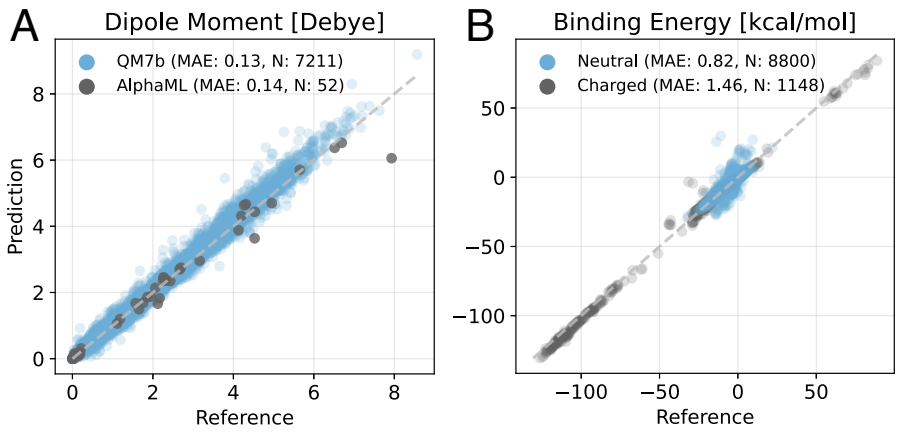

Figure 6.2: Evaluation of the SO3LR long-range modules' performance. (A) Evaluation of the model on dipole moment prediction for 7k QM7b molecules and AlphaML showcase database [243]. (B) Performance of the model evaluated on the unseen SAPT10k dataset [244], separated into neutral and charged subsets.

interactions and offering substantial diversity across chemical space. We exclude 34 out of 9982 complexes because they contain atom types beyond the 8 elements our model was trained on (hence, predictions on those structures are not meaningful). Overall, the model performs well, achieving sub-chemical accuracy with a MAE of 0.90 kcal/mol (Fig. 6.2B). Rare outliers with errors up to 40 kcal/mol include complexes with exotic molecules absent from the training set, such as ClF, $P(CNO)_3$, PH_2NO_2. Recalculation of these outliers at the PBE0+MBD level confirms the errors arise from missing training data rather than from the reference method (Fig. A11). This is a remarkable performance overall, particularly since part of the error comes from the difference between the CCD and PBE0+MBD reference levels.

6.2.2 Small biomolecular units

Molecular dynamics simulations are the ultimate test for evaluating force fields. We simulated six molecular systems from the MD22 benchmark, encompassing four major biomolecule types and two supramolecular complexes: the $AcAla_3NHMe$ tetrapeptide, stachyose tetrasaccharide, AT-AT DNA base pairs, docosahexaenoic fatty acid (DHA), the Buckyball Catcher, and the double-walled nanotube. The first two systems underwent 500 ps of simulation at 500 K to compare with the PBE+MBD references computed at 500 K, other systems were simulated at 300 K. The model demonstrated robust conformational exploration across all molecules. In particular, the free-energy surface exploration of tetraalanine and stachyose closely aligns with MD22 *ab initio* results, computed at the PBE+MBD level of theory, as shown in Ramachandran plots (Fig. 6.3). Note that in this figure we report only short molecular dynamics simulations to match the length of DFT simulations, and the comparison between PBE+MBD and SO3LR dynamics is only provided as a guide to the eye. Full 500 ps trajectories are shown in Fig. A12 [245]. The tetrapeptide explores all 'allowed' (ϕ/ψ) regions found in experimental protein structures [245]. The buckyball catcher and double-walled nanotube complexes remained stable (Fig. A13), despite larger errors on the test set. This highlights the stability of SO3LR, even when applied to systems far outside its training domain. Overall, these results suggest that our model can reliably explore conformational landscapes of small molecules even in the absence of the system-specific training data.

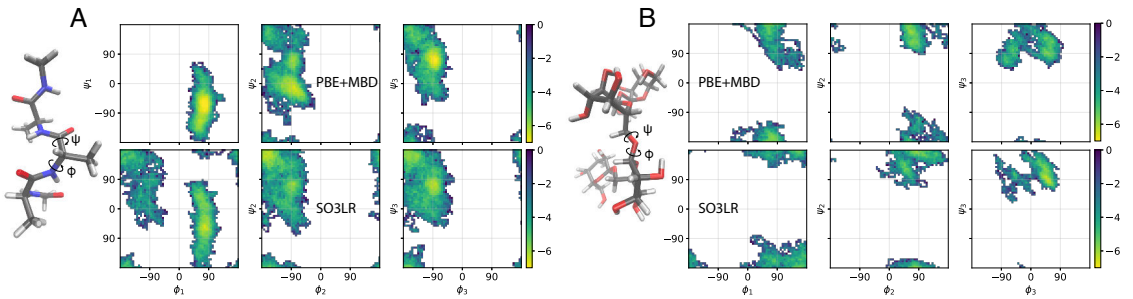

Figure 6.3: Simulations of small biomolecular fragments. Ramachandran plots (ϕ/ψ dihedrals) for (A) $AcAla_3NHMe$ and (B) stachyose from the MD22 dataset [66]. PBE+MBD and SO3LR simulations at 500 K with 85 ps for $AcAla_3NHMe$ and 27 ps dynamics for stachyose. SO3LR simulations of 500 ps are shown in Fig. A12. Trajectory is sampled every 1 fs. The Boltzmann-inverted scale is shown in kcal/mol. The comparison between PBE+MBD from MD22 and SO3LR (trained on PBE0+MBD) is only shown as a guide to the eye.

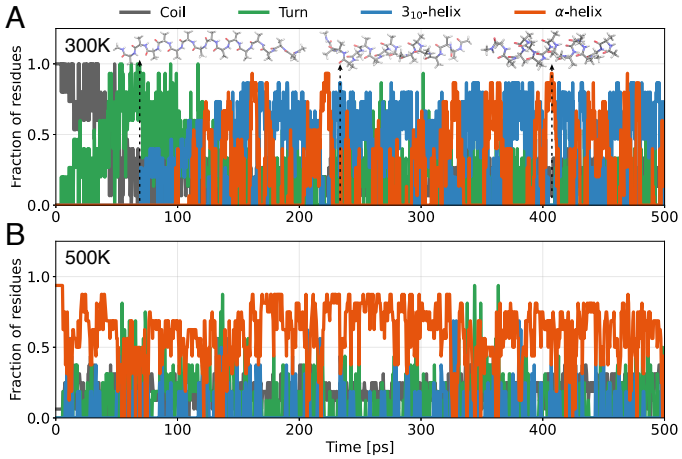

Figure 6.4: Simulations of polyalanines. (A) Secondary structural motifs observed along a typical folding trajectory of $AcAla_{15}NHMe$ at 300 K in gas phase. (B) Secondary structural motifs observed along a trajectory of $AcAla_{15}LysH^+$ at 500 K in gas phase, starting from the folded α-helix conformation.

6.2.3 Polyalanine systems

We further investigate polyalanines, focusing on the folding of extended $AcAla_{15}NHMe$ and the stability of the folded $AcAla_{15}LysH^+$ at elevated temperatures. These systems present significant challenges due to the delicate interplay of hydrogen bonding, polarization, and dispersion interactions. Previous attempts to simulate them without incorporating top-down fragments either failed to correctly fold $AcAla_{15}NHMe$ or overstabilized the α-helix, and in some cases, predicted diminished stability for $AcAla_{15}LysH^+$ [102, 246].

For each system, we performed four runs of 500 ps. The extended $AcAla_{15}NHMe$ structure folded in all cases (Fig. 6.4A and Fig. A14A). The timescales and folding mechanisms were similar to those observed in Ref. 102: initially, the peptide primarily consists of turns, then passes through a "wavy" intermediate, and finally folds into a helical form with dynamic transitions between α- and 3_{10}-helices. The latter is particularly noteworthy, as empirical force fields tend to overestimate the stability of α-helices [247, 248].

For the folded $AcAla_{15}LysH^+$, we observe that the α-helical motifs are preserved up to

500–600 K (Fig. 6.4B and Fig. A14B). These findings agree with experimental measurements, which observe scattering cross sections for $AcAla_{15}LysH^+$ consistent with an α-helical structure up to ≈725 K when subject to interactions with the helium buffer gas [249]. Direct comparison with gas-phase experiments would have to explicitly include the helium environment and quantum nuclear effects. Overall, the two polyalanine systems provide a good evaluation of scalability to medium-sized systems in dynamics, complementing the observed scalability in terms of test errors.

6.2.4 Liquid water

Liquid water plays a crucial role in biosystems, making it an essential subject for SO3LR's evaluation. We performed a simulation of a water box containing 4096 water molecules in the NPT ensemble. Observables were averaged over 300 ps following an initial 200 ps equilibration phase. Our analysis focused on three aspects: radial distribution function, density, and self-diffusion coefficient.

The oxygen-oxygen radial distribution function shows the expected shell structure (Fig. A15), which indicates, however, that the liquid phase is slightly overstructured. Increasing the temperature to 330 K allows for an approximate treatment of missing nuclear quantum effects and improves agreement with the experimental data [75, 169]. The water density varies between 1.04 and 0.97 g/cm^3 for long-range cutoffs of 10–20 Å (Fig. A16). We adopted a cutoff of 12 Å for all subsequent biosimulations in explicit water, balancing accuracy and computational efficiency. The calculated self-diffusion coefficient is 0.079 Å^2/ps at 300 K and 0.224 Å^2/ps at 330 K with the 12 Å long-range cutoff. For comparison, the experimental diffusion coefficient at room temperature is 0.23 Å^2/ps [250].

The SO3LR results agree well with explicit ab initio molecular dynamics using the PBE0+vdW functional [75]. This is notable given that the training set contains only gas-phase water clusters with at most 40 molecules (∼10k clusters or ∼0.26% of the combined dataset). It is known that *ab initio* MD simulations with the PBE0+vdW functional struggle to fully capture many experimental properties of water, mainly due to the tetrahedral H-bond arrangement that amplifies the slight overestimation of PBE0 for individual hydrogen bonds [75]. Consequently, the MLFF performance cannot and should not exceed the accuracy of the underlying ab initio calculations. The description of water could be improved by using higher-level ab initio data, such as coupled-cluster or quantum Monte Carlo methods, and by explicitly incorporating nuclear quantum effects in MD simulations. For biomolecules in water, hydrogen bonding is just one of many contributing interactions, and we have shown that accurate biomolecular dynamics can be carried out with MLFFs trained on PBE0+MBD data provided that the density of water is correctly reproduced [102].

6.2.5 Large biomolecules

Finally, we showcase the potential of SO3LR by simulating large biomolecules in explicit water. The selected systems encompass various classes of biomolecular components, each characterized by distinct structural and functional properties that can be validated against existing simulations or experimental data. The systems include the crambin protein, glycoprotein (PDB: 1K7C), and the POPC lipid bilayer.

For crambin (25k atoms including water), we compute the power spectrum from 125 ps of dynamics at a temporal resolution of 2.5 fs, after 1 ns equilibration period. The experimental water vibrations at 1640 cm^{-1} and 3200–3600 cm^{-1} are reproduced in SO3LR with better

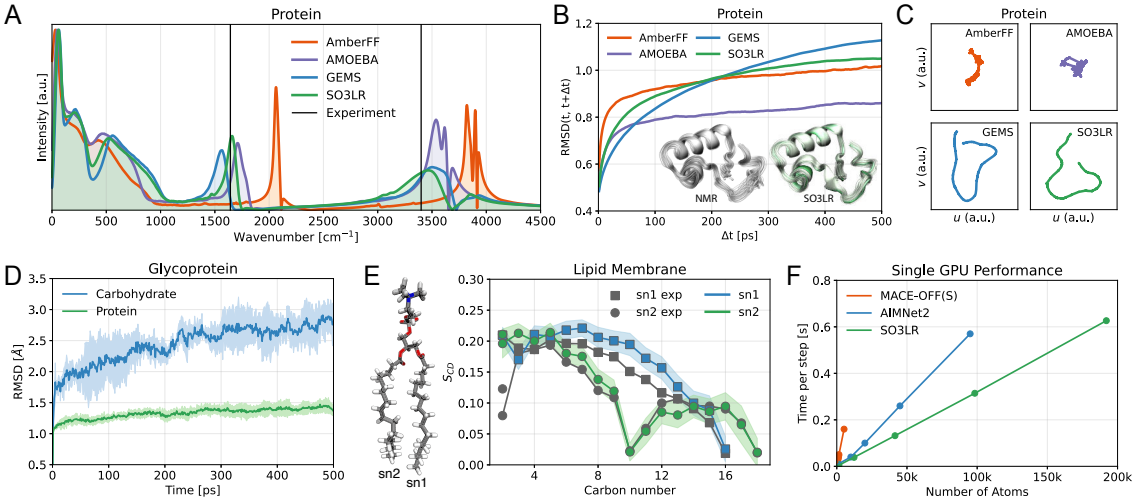

Figure 6.5: Simulations of explicitly solvated biomolecules. (A) Power spectrum of crambin in water obtained from 125 ps of dynamics. AmberFF and GEMS results are taken from Ref. 102. (B) Root mean square deviation (RMSD) of Crambin, excluding hydrogen atoms, between conformations sampled at times t and $t + \Delta t$ averaged over three 3 ns runs. The inset shows an overlay of frames from the SO3LR trajectory and NMR-derived protein structures [251]. (C) Two-dimensional Uniform Manifold Approximation and Projection (UMAP) [252] embedding of a crambin simulation trajectories. The same latent space projection is used across all subplots. (D) RMSD of protein and carbohydrate segments of glycoprotein averaged over three 500 ps runs. (E) Tail group NMR order parameters from SO3LR simulation of the 128 POPC Lipid Bilayer and from experiment [253]. The standard deviation is shown with background color. (F) Single GPU performance. SO3LR latencies were measured based on liquid water molecular dynamics using JAX-MD [227] in the NVT ensemble on an H100 80 GB GPU. The slope is 3.25×10^{-6} s/atom/step. Latencies for MACE-OFF(S) and AIMNet2 were measured on A100 and H100, respectively, and are taken from Refs. 246, 254.

agreement than GEMS, AMOEBA and AmberFF (Fig. 6.5A). We further examine the root mean square deviation RMSD(t, t+Δt) averaged over three 3 ns simulations, indicating that SO3LR shows slightly increased protein mobility on longer timescales, consistent with the GEMS model (Fig. 6.5B). We find that the overall structure stays folded during the simulation, without any indication of unfolding or bond breaking (Fig. A17). To visualize conformational space sampling, we applied the two-dimensional uniform manifold approximation and projection (UMAP) [252]. The projection of the paths reveals that SO3LR and GEMS sample the conformational space more extensively than AmberFF and AMOEBA (Fig. 6.5C), which aligns well with the high conformational variability derived from NMR measurements [251].

Table 6.2: POPC Lipid Bilayer Structural Properties: bilayer thicknesses D_{HH} (Å) and area per lipid (Å^2).

Source	D_{HH}	S/lipid
Experiment [255]	36.5	64.3 ± 1.3
Lipid21 [145]	38.50 ± 0.20	63.92 ± 0.09
CHARMM36/LJ-PME [256]	37.3 ± 0.30	65.4 ± 0.5
SO3LR	37.0	58.0 ± 0.1

For glycoprotein (48k atoms including water) we conducted a 500 ps simulation at 300 K. This system, which comprises both protein and carbohydrate segments, presented a challenge for SO3LR due to the absence of carbohydrates in its training data. Despite this, the model successfully inferred increased carbohydrate flexibility, as evidenced by the greater RMSD observed for the carbohydrate segment compared to the protein segment (Fig. 6.5D). These findings align with the results of the CHARMM force field specifically tuned for carbohydrates [257]. However, the simulation revealed limitations in sampling conformations of the N-linkage. Specifically, the Cγ-Cβ-Cα-N dihedral, located at the protein-carbohydrate junction, can adopt three conformations: $g+$ (60°), *anti* (180°), and $g-$ (300°). Our simulation only sampled the anti-conformation out of these three possible states. Longer simulations would be required to determine whether the model can explore other conformations without carbohydrate-protein linkages in the training set.

Lastly, we modeled a homogeneous POPC lipid bilayer (33k atom system consisting of 128 lipids and 5120 water molecules). We performed a 500 ps simulation at 303 K and examined the structural properties: area per lipid, bilayer thickness, and lipid tail order parameters. These properties are critical measures of the accuracy of lipid simulations and are highly sensitive to factors such as hydrophilic attraction between head groups, hydrophobic repulsion between lipid tails, and interactions with surrounding water molecules. We found that SO3LR is in good agreement with experimental data and with empirical force fields specifically fine-tuned to lipid simulations (Tab. 6.2). The 10% underestimation of the area-per-lipid likely stems from the isotropic NPT ensemble currently used in SO3LR simulations, compared to the semi-isotropic NPT used for empirical force fields. NMR lipid tail order parameters are another important quantitative measure that describe the degree of order within the acyl chains of lipids in a bilayer. The order parameters averaged over the last 250 ps suggest that the bilayer structure is in suitable agreement with NMR experiments (Fig. 6.5E) [253].

To assess the computational performance of SO3LR, we conducted NVT simulations of water boxes ranging from 1,536 to 192,000 atoms using JAX-MD on a single GPU (Fig. 6.5F). The measured scaling corresponds to a slope of 3.25×10^{-6} s/atom/step, enabling simulations of 2.6 ns/day for a 10,000-atom system using a 1 fs timestep. This performance allows nanosecond-scale simulations of large solvated systems on standard hardware. However, simulations of intricate conformational changes that occur on millisecond timescales, such as solvated protein folding, remain beyond reach at present. The current SO3LR model, with 128 features, 3 interaction layers, and a 4.5 Å local cutoff, was designed to balance accuracy and speed. These hyperparameters can be systematically adjusted to trade accuracy for performance: smaller models can be trained to accelerate simulations while maintaining chemical fidelity [258]. Notably, the presented SO3LR model was trained on a single GPU using a modest computational budget of 86 GPU hours.

6.3 Conclusion

A long-held vision in the atomistic simulation community is the development of force fields with a unified functional form that can be applied across diverse chemical spaces – such as solvents, proteins, DNA, RNA, sugars, and lipids. These force fields should closely approximate quantum-mechanical behavior while remaining efficient and scalable enough to model realistic biomolecular complexes under various conditions (e.g., pressure, temperature, and external environments). In this work, we presented significant advancements towards fulfilling these criteria through the SO3LR model, which is embedded within an openly

accessible and fully transparent framework. This framework integrates reliable and diverse quantum-mechanical datasets [56, 102, 131, 133, 158], a fast and stable SO3krates machine learning architecture [53], universal long-range interaction modules [230], a JAX-MD simulation engine [227], and robust analysis tools. Together, these components facilitate quantum-accurate molecular simulations across an extended molecular chemical space.

Our developments aim towards enabling general molecular simulations and similar goals have been pursued by the seminal efforts in the empirical force field community over many decades [30–32, 259–264]. SO3LR achieves chemical accuracy and yields an 8-fold improvement compared to AMBER for polyalanine [102] when benchmarked on the PBE0+MBD atomic forces (Force MAE of 0.9 vs 7.6 kcal/mol/Å). At the same time, SO3LR is (only) about 40 times slower on a single GPU than GROMOS [102]. Our extended assessment on energies, forces, dipoles, polarizabilities, as well as our analysis of nanosecond-long MD trajectories demonstrates that SO3LR is highly transferable throughout biochemical space and scalable to hundreds of thousands of atoms. Such transferability and scalability are achieved without the need to specify atom types, impose harmonic constraints, or introduce bespoke functional forms for interatomic interactions in different biomolecular entities. The bottom-up training on quantum mechanical data ensures that our simulations are transferable to a wider range of conditions than previously possible. This is confirmed by polyalanine simulations from 300 to 800 K, accurate structural and spectroscopic observables for high and low vibrational frequencies obtained for solvated crambin, as well as the local and global structural properties for the 1K7C glycoprotein and the POPC lipid bilayer. The toolset developed in this work complements the existing and quickly growing machinery of successful biomolecular modeling tools. Our presented advancements would not have been possible without building on a wealth of existing landmark methods, many of which were developed by the empirical force field community.

One noteworthy component of our proposed SO3LR model is a successful combination of explicit physical knowledge, such as short- and long-range force modules, coupled with a semi-local many-body potential. Importantly, all of these contributions are carefully balanced by SO3LR via learning from data. Thus, the known physical interactions do not need to be learned from data, but SO3LR can – under the correct hard-coded inductive biases (repulsion, electrostatics, and dispersion energies) – focus its nonlinear expressive power mainly on learning both: the complex many-body contributions and the appropriate balance of the diverse energy terms from Eq. 6.1.

Despite recent progress in establishing "foundation models" for atomistic systems [246, 254, 265, 266], many challenges remain in achieving truly general molecular simulations. While the SO3LR model shows broad applicability, it has several limitations that may guide future work. The model's predictive accuracy is tied to its training data, possibly leading to suboptimal performance for underrepresented chemical environments such as curved carbon systems and specific functional groups (notably dynamical simulations remain stable). Therefore, development of robust uncertainty quantification to reliably detect when the model is far in extrapolation regime is a priority. Similarly, incorporating large-scale quantum mechanical datasets covering over 80 elements (e.g., QCML [134], MPTrj [267], and OMol25 [136]), as well as expanding DFT+MBD training sets to include a broader spectrum of (bio)chemical entities such as ions, sugars, lipids, DNA, supramolecules and diverse solvents would be highly beneficial for developing more transferable models applicable to (bio)molecules, materials and their interfaces. Furthermore, future versions of SO3LR will integrate particle-mesh Ewald (PME) summation for long-range electrostatics, leveraging recent implementations in JAX-MD [268]. Several other key areas for enhancing

the SO3LR model include: (i) generating higher-level coupled cluster [64] or quantum Monte Carlo [269] reference data for small fragments, (ii) refining long-range interaction modules to effectively account for anisotropic many-body interactions [77], (iii) optimizing SO3LR for multi-GPU architectures [270], (iv) extending simulations to treat nuclear quantum effects [207, 271] beyond classical Newtonian molecular dynamics. This is a non-exhaustive list of research directions, all of which are subject of ongoing efforts in the community.

As atomistic simulations are highly sensitive to the intricacies of the underlying force fields and simulation parameters, it is imperative to establish a standardized set of benchmarks for quantum-accurate machine learning force fields. Such benchmarks will ensure reproducibility of results and enable robust modeling of experimentally relevant phenomena across realistic time and length scales.

Chapter 7

Summary and Outlook

The central challenge motivating this thesis is the decades-old compromise between accuracy and efficiency in molecular simulation, which has long forced a choice between the intractable rigor of quantum mechanics and the limited fidelity of classical approximations. This work demonstrates that this trade-off is no longer inevitable. We have shown that a new generation of machine learning force fields (MLFFs), built upon a foundation of curated data, efficient representations, and integrated physical principles, can deliver quantum accuracy to large (bio)molecules at a computationally tractable cost.

7.1 Summary

Each chapter of this thesis met one of the abovementioned challenges, together they provide a blueprint for next-generation MLFFs. The key contributions can be summarized as follows:

- In **Chapter 3**, we established the need for models that capture collective long-range interactions. This began with the introduction of MD22, a benchmark dataset comprising a handful of large, flexible molecules specifically chosen to exhibit strong nonlocal correlation effects. Using MD22 as a testing ground, we developed a large-scale training extension of the global kernel-based model sGDML that, for the first time, enabled a global MLFF to be trained on systems with hundreds of atoms.

- In **Chapter 4**, we addressed the data bottleneck hindering transferable MLFFs for biomolecules by building the QCell dataset. QCell is a comprehensive and chemically diverse quantum-mechanical database spanning the fundamental building blocks beyond proteins: nucleic acids, lipids, and carbohydrates, along with noncovalent dimers and ion-water clusters. By covering all major (bio)chemical classes, QCell provides the data to train general-purpose MLFFs applicable across the full spectrum of biomolecular systems.

- In **Chapter 5**, we addressed the scaling limitations of global models by developing the reduced Gradient-Domain Machine Learning (rGDML) method. This approach employs a novel algorithm to identify a minimal subset of critical long-range features, effectively eliminating the $\mathcal{O}(N^2)$ bottleneck of traditional global descriptors. rGDML thus preserves the physical completeness and accuracy of a global representation with a linear scaling descriptor.

- Finally, in **Chapter 6**, we introduced a unifying advance in the form of the SO3LR model, a pretrained general-purpose MLFF that combines an SO(3)-equivariant network for semi-local many-body interactions with explicit, physics-based long-range electrostatics and dispersion. SO3LR demonstrated transferability and accuracy across a wide range of biomolecular systems, including proteins, glycoproteins, lipid bilayers, enabling quantum-accurate molecular dynamics across broad chemical space.

7.2 Outlook

The framework developed in this thesis opens numerous exciting avenues for future research. While we have made significant strides in accuracy, scalability, and transferability, the journey toward a truly universal model of molecular interactions is ongoing. The following directions represent promising next steps that build upon the work described here:

- **Automated active learning workflows.** The creation of the MD22 and QCell datasets was a manual effort. The future of MLFF development lies in creating fully automated, closed-loop workflows. These "self-driving" simulation engines would use active learning to explore the vast chemical and conformational space, identifying regions where the current model is uncertain and automatically launching new quantum-mechanical calculations to improve it. This would accelerate the creation of robust and comprehensive models.
- **Multi-scale modeling.** Even with linear-scaling MLFFs, simulating cellular-level processes across biologically relevant timescales remains unattainable. A promising path forward is the development of hybrid models that seamlessly couple different levels of resolution. In such settings, regions of primary interest can be described using all-atom MLFFs, their immediate environment using coarse-grained descriptions, and the extended surroundings using implicit or continuum representations, with all components dynamically and consistently coupled. In parallel, integrating general-purpose MLFFs with generative models offers a route toward efficient exploration of complex conformational landscapes, enabling rapid access to rare or collective structural rearrangements.
- **Applications to grand biological problems.** The ultimate test of these new methods is their application to outstanding challenges in simulating complex chemical and biological systems. Promising targets include characterizing the conformational ensembles and aggregation pathways of intrinsically disordered proteins implicated in neurodegenerative diseases, elucidating complete catalytic cycles of multi-domain enzymes, modeling RNA folding pathways and ribozyme catalysis, and resolving glycan recognition in immune receptors and viral entry mechanisms.

In conclusion, the work presented in this thesis has provided a comprehensive set of solutions to the key challenges that have limited the predictive power of atomistic simulations. By addressing the problems of data generation, atomic representation, and long-range-aware architectures, we have developed a new generation of general-purpose machine learning force fields that successfully bridge the gap between quantum-mechanical accuracy and the efficiency required for large-scale simulation. These contributions help advance toward *ab initio* modeling of living matter.

Appendices

The appendices provide supplementary methodological details, extended analyses, and additional validation results for the datasets and models introduced in Chapters 3–6. Parts of the appendices reproduce or adapt supplementary material from Refs. 66, 121, 127, 157.

A1 Supplementary Information Chapter 3: MD22 & sGDML

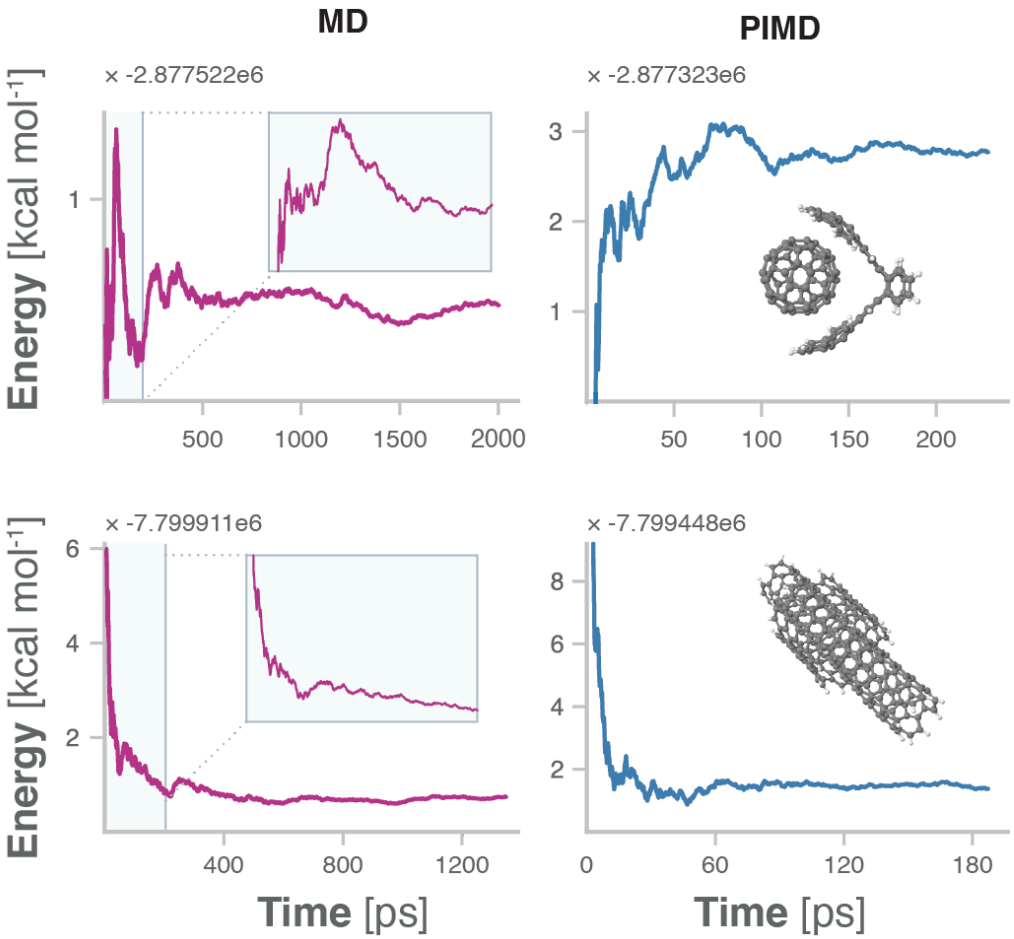

Figure A1: The cumulative potential-energy (in kcal mol^{-1}) as a function of simulation time (in ps) for the buckyball catcher and the double-walled nanotube, along classical MD and PIMD simulations obtained with sGDML FFs. The simulations were carried out until the cumulative energy remained approximately constant. The classical MD plots contain zoomed-in sections spanning a simulation time equivalent to the length of the PIMD trajectories.

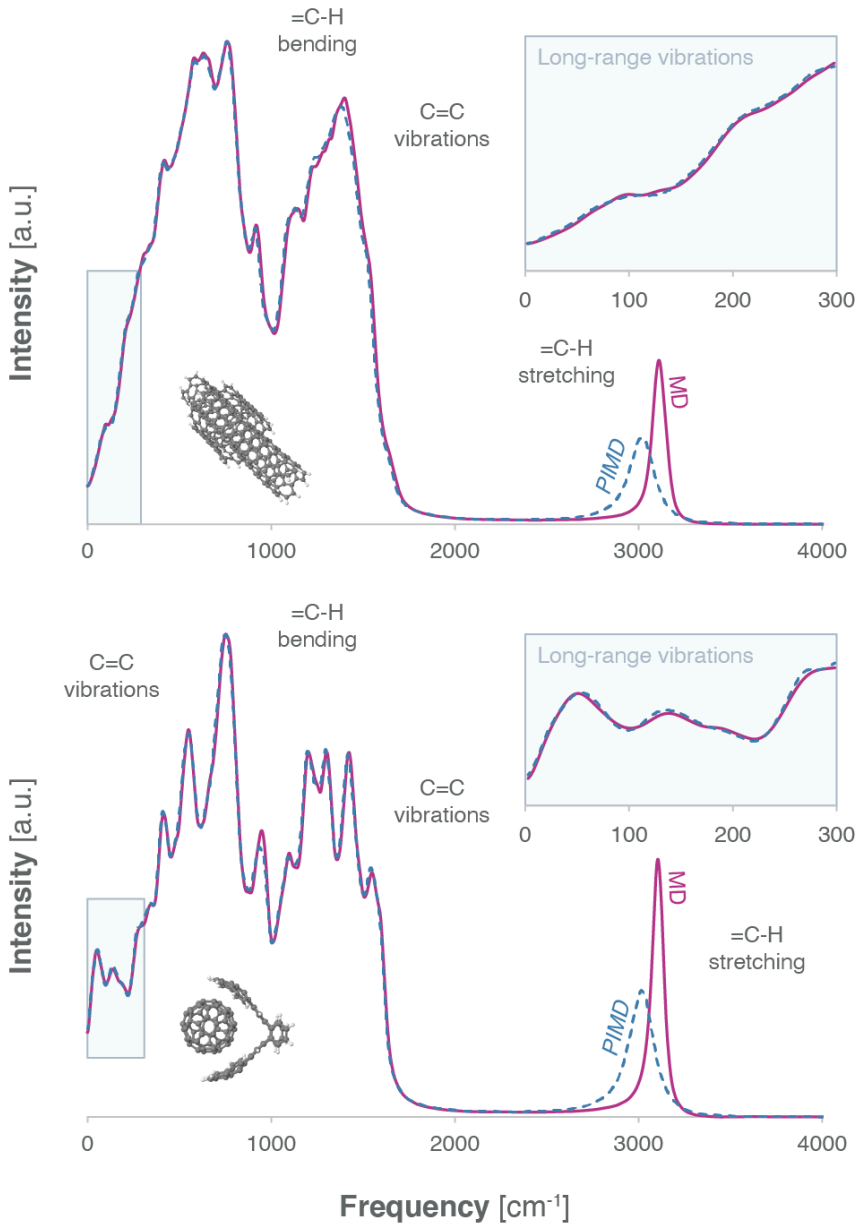

Figure A2: Molecular spectra of the double-walled nanotube and the buckyball catcher, as obtained from the velocity autocorrelation function for MD and PIMD simulations. For the sake of consistency, the MD trajectories have been cropped to the same length for this plot. The final length was determined by the total number of steps in the PIMD simulations and does not affect their convergence.

A2 Supplementary Information Chapter 4: QCell

MLFF training and parameters. The SO3LR (SO3krates with long-range terms) [53, 157] small, medium, and large models were trained using the hyperparameters listed in Tab. A1. A combined loss on forces, dipole moments, Hirshfeld ratios, and energies was employed with weighting factors of 100:10:10:1, respectively. The QCML and QM7-X subsets were sampled 10x less frequently to ensure balanced training. Training used the AdamW optimizer with an exponential learning rate decay by a factor of 0.99 every 1M steps. The global norm of gradient updates was clipped at 100, and a general robust loss with $\alpha = 1.0$ was applied [272]. A 10 Å long-range cutoff, electrostatics damping coefficient of $\sigma = 4$, and dispersion damping coefficient of $\gamma = 1.2$ were used. All models were trained on A100 for 180 GPUh.

Table A1: Training hyperparameters of the MLFF models.

Parameter	**Small**	**Medium**	**Large**
Cutoff radius (Å)	4.5	5.0	5.0
Feature dimension (H)	128	256	512
Message-passing layers (T)	2	3	3
Number of heads (h)	4	8	8
Maximal degree (L)	4	4	4
Radial basis functions (k)	32	64	128
Batch size (B)	128	128	64
Learning rate	5×10^{-4}	1×10^{-4}	1×10^{-4}

A3 Supplementary Information Chapter 5: rGDML

Interaction Heatmaps. We use corresponding GDML (ver. 0.4.11) models trained and validated on 1000 different configurations to calculate interaction heatmaps (Fig. 5.1b, d and Fig. 5.3). Heatmaps consist of pairwise contributions, F_l^k, averaged over many configurations from the dataset (3000 configurations for the $AcAla_3NHMe$ and AT-AT; 1000 configurations for the buckyball catcher).

The trained GDML force field estimator collects the contributions of the $3N$ partial derivatives (N - number of atoms) of all M training points to compile the prediction:

$$\mathbf{F}(\mathbf{x}) = \sum_{i=1}^{M} \sum_{j=1}^{3N} (\boldsymbol{\alpha}_i)_j \frac{\partial}{\partial x_j} \nabla \kappa (\mathbf{x}, \mathbf{x}_i) , \tag{A1}$$

where $\mathbf{F}(\mathbf{x})$ is a vector containing the $3N$ forces predicted for molecular geometry $\mathbf{x}$. A partial evaluation of this sum yields the contribution of a single atom k to the force prediction on all atoms (atom indices start from 0):

$$\mathbf{F}^k(\mathbf{x}) = \sum_{i=1}^{M} \sum_{j=3k+1}^{3k+3} (\boldsymbol{\alpha}_i)_j \frac{\partial}{\partial x_j} \nabla \kappa (\mathbf{x}, \mathbf{x}_i) . \tag{A2}$$

To obtain the contribution of atom k to the force prediction on atom l, F_l^k, we compute the norm of the force components of vector $\mathbf{F}^k(\mathbf{x})$ that correspond to an atom l:

$$F_l^k = \left\{ \sum_{s=1}^{3} \left(\mathbf{F}^k(\mathbf{x})_{3l+s} \right)^2 \right\}^{1/2} . \tag{A3}$$

Descriptor Reduction. The procedure starts from a pre-trained kernel-based ML model (ML_{orig}), with a default global (containing all n features) descriptor $\mathbf{x}$. Importantly, we do not require a highly accurate and thus computationally expensive ML model at this stage (see Methods for further details).

The significance of the n-th feature in the descriptor is obtained by comparing the prediction results on a subset of test configurations between the full ML_{orig} and the ML_{orig} with the n-th feature set to zero for all configurations (ML^n_{mask}). Thus, we assume separability between the features in the descriptor, but this does not imply their independence. Therefore, more advanced and computationally expensive reduction techniques can also be applied [273, 274].

This procedure is performed separately for all features in the descriptor. All other parameters of the ML_{orig} model remain unchanged when obtaining the ML^n_{mask} predictions. Therefore, the only difference between the models is in the definition of similarity between system states. The loss function

$$L_n = \sum_{i=1}^{N} \left(\mathrm{ML}_{orig}(\mathbf{x}_i) - \mathrm{ML}^n_{mask}(\mathbf{x}_i) \right)^2 , \tag{A4}$$

where N is the number of test configurations, and serves as a measure of the importance of a particular feature n in the descriptor. The descriptor features where the loss L_n is the smallest are the least important for the model and can be removed from the descriptor. As soon as our analysis is performed on a representative subset of configurations, we ensure

that we preserve all the descriptor features relevant for modeling the given PES. However, setting a threshold under which one can consider a feature as irrelevant is not trivial. The values of L_n depend on i) the predicted property, ii) the system(s) for which the model is trained, and iii) the reference data used for training. In the study, we consider every 10th percentile of all L_n values (10th to 90th). As a final step, a new ML model is trained and tested after removing from the default descriptor all the features whose corresponding L_n are below a selected percentile.

Reference Datasets. Molecular dynamics simulations at PBE+MBD level of theory with a step size of 1 fs were used to construct the reference datasets. Tab. A2 includes additional information about the datasets. Calculations were done either with i-PI [275] wrapped with FHI-aims [276] to compute forces and energies or with FHI-aims code alone.

Table A2: Settings of the MD simulations of the datasets used in the work. Temperature is given in K. PBE stands for the Perdew-Burke-Ernzerhof functional [277] and MBD stands for many-body dispersion [73, 278]. Coefficient refers to the friction coefficient (in fs) for the global Langevin thermostat, and to the effective mass (in cm^{-1}) for the Nosé-Hoover thermostat.

Molecule	**Basis set**	**Temp.**	**Thermostat**	**Coef.**
$AcAla_3NHMe$	tight	500	Global Langevin	2
AT-AT	tight	500	Global Langevin	2
Buckyball catcher	light	400	Nosé-Hoover	1700
Lactose	light	500	Nosé-Hoover	1700
Palmitic acid	light	500	Nosé-Hoover	1700

Computational Details for ML Models. The ML models were built with GDML [54, 64] and GAPs [175] with the SOAP representation [47]. GDML models were trained using a numerical solver with an initial value of 70 inducing points. All models were validated using 1000 configurations and hyperparameter search σ was performed individually for each system and training size to ensure optimal model selection (from 10 to 1000). No symmetries were considered in the models for a fair comparison between the default descriptor and those with a reduced size. GAP/SOAP models were trained using 12 radial and 6 angular functions for the descriptor. The cutoff radius was set to 5 Å. Parameter δ was set to 0.25, the atom σ was set to 0.3, and the default σs for energy and forces were set to 0.001 and 0.1, respectively. These calculations were performed with the QUIP program package [279] through the quippy python interface [280].

Molecular Dynamics Simulations. External-force DFT calculations were performed using the FHI-aims electronic structure software [162] in combination with the externalforce option in the Atomic Simulation Environment package [281]. We used the PBE and PBE+MBD [17, 29] level of theory with the *intermediate* basis set. Trajectories were generated with a resolution of 0.5 fs and sampled at 300 K using a Langevin thermostat with a friction coefficient of $1{\cdot}10^{-3}$.

To evaluate the performance of machine learning models, we utilized both the ML_{Global} and $ML_{R0.6}$ models trained on 5000 configurations with the same settings. To ensure stability during 17 parallel simulation dynamics, each averaging around 3 ns (total time of 50 ns), we utilized a time step of 0.3 fs and a Langevin thermostat with a $1{\cdot}10^{-4}$ friction coefficient. The first configuration of $AcAla_3NHMe$ from the dataset was used as a starting configuration in all calculations.

Prediction Accuracy of Models Trained with Global, Local and Reduced descriptors. Here we compare the performance of an ML model trained using an reduced descriptor to ML models trained using global and local descriptors. Fig. A3 shows distributions of force errors for default GDML models (ML_{global}), GDML models using reduced descriptors (with 40% of the original features) as obtained from the results in Fig. 5.2 (ML_{opt}), and GAP/SOAP models with a cutoff of 5 Å (ML_{SOAP}) trained on 1000 training configurations for the $AcAla_3NHMe$, the AT-AT dimer, and the buckyball catcher. Force error histograms show that the accuracy of the local ML_{SOAP} models, with respect to the ML_{global} and ML_{opt} ones, is lower with the increasing size and flexibility of the molecule. ML_{SOAP} models start with an almost equal distribution as all other models for $AcAla_3NHMe$ (Fig. A3A) but show considerably bigger errors than the ML_{global} and ML_{opt} models for the buckyball catcher (Fig. A3C). The lower accuracy of the local models is the result of neglecting nonlocal interactions that become prominent for the larger and more flexible systems. Regarding the ML_{opt} models, they present almost the same population of small force errors [under an absolute value of 1.0 kcal/mol/Å] as the ML_{global} model, while having a lower frequency of larger errors. For instance, errors above absolute values of 3.0 and 1.0 kcal/mol/Å for the AT-AT dimer (Fig. A3B) and the buckyball catcher (Fig. A3C), respectively, are more common with the ML_{global} model. Hence, ML models constructed using reduced descriptors provide more reliable predictions than typical global and local ML models when reconstructing complex PESs.

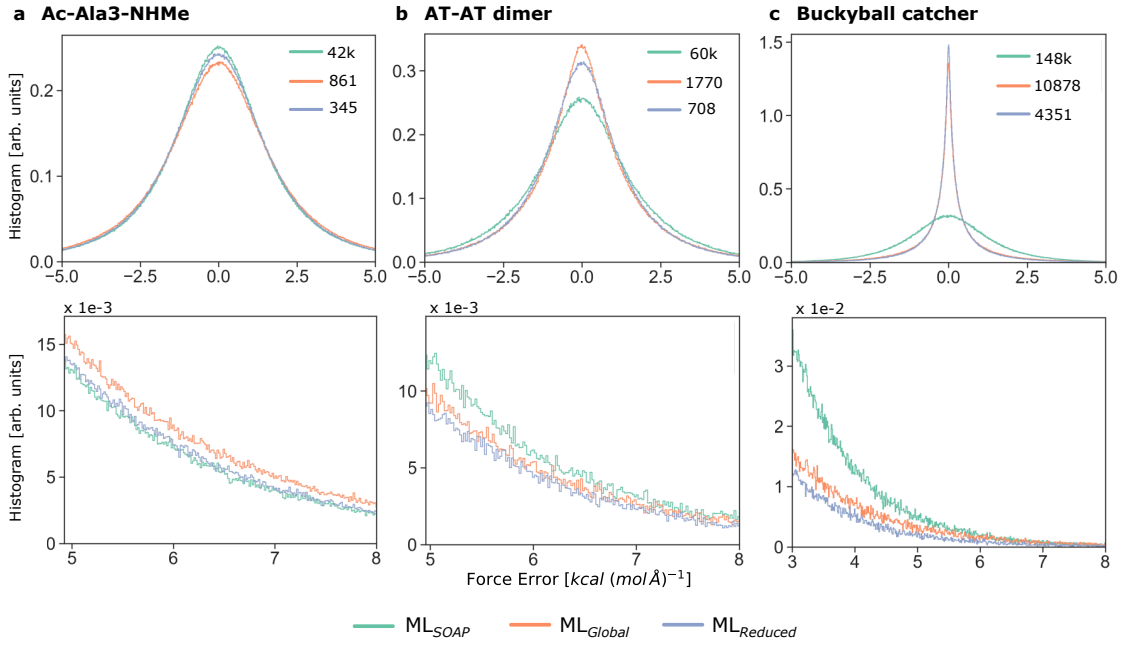

Figure A3: Histogram of force errors [in kcal/mol/Å] of the global (ML_{Global}), optimally reduced ($ML_{Reduced}$), and Gaussian Approximation Potential/Smooth Overlap of Atomic Positions (GAP/SOAP, ML_{SOAP}) models for $AcAla_3NHMe$ (**a**), the AT-AT dimer (**b**), and the buckyball catcher (**c**). The size of the descriptor of the models is given in the legend box of the figures. Upper row: section of the distribution of errors between -5 and 5 kcal/mol/Å; lower row: section of the distribution in the tail from 5 to 8 (**a, b**) and from 3 to 8 (**c**) kcal/mol/Å.

Table A3: Performance comparison of global and reduced models. Energy and Force root mean square errors (E RMSE and F RMSE) are reported in kcal/mol and kcal/mol/Å, respectively. The definitions of compact and extended training sets are explained in this subsection. The Gradient Domain Machine Learning model with default global descriptor is denoted as ML_{Global}, and the model with a descriptor reduced by 60% is denoted as $\mathrm{ML}_{R0.4}$.

Training set selected from	Model	E RMSE extended	F RMSE extended	E RMSE compact	F RMSE compact
compact	ML_{Global}	14.0	4.31	1.64	2.41
compact	$\mathrm{ML}_{R0.4}$	7.55	3.55	1.47	2.24
extended	ML_{Global}	1.74	2.44	4.72	3.09
extended	$\mathrm{ML}_{R0.4}$	1.53	2.29	3.05	2.67

Prediction of outlier data. We investigated the performance of global and reduced models trained on "extended" structures when tested on "compact" structures of the tetrapeptide. To perform the comparison, we have split the dataset based on the distance between the furthest atoms in each structure (ranges from ∼8 to 14 Å).

We selected a threshold of 12 Å which separates clusters of compact and extended structures. With this threshold, we split the dataset into dataset 1 ($max(R_{ij}) < 12$ Å, ∼80% of the initial dataset - 69k structures) and dataset 2 ($max(R_{ij}) \geq 12$ Å, 20% - 16k structures). We used 1000 points for the training and 1000 for the validation of the models from the dataset 1 (training set - compact) and used all structures from the dataset 2 for testing (E/F RMSE extended).

To check how global and reduced models trained on "extended" structures perform on "compact" structures we repeated the same procedure with a threshold of 9.5 Å, resulting in the dataset 3 ($max(R_{ij}) > 9.5$ Å, ∼80% of the initial dataset) used for training (training set - extended) and dataset 4 ($max(R_{ij}) \leq 9.5$ Å, 20%) used for testing (E/F RMSE compact).

The comparison of the Force/Energy RMSEs shows that the reduced models are more accurate than global models when dealing with "unseen" outlier structures. We can attribute such an improvement to the ability of reduced models to obtain a better description of the environments of the molecule. This means that reduced models can better identify similar structural moities between "compact" and "extended" structures while keeping the relevant information for describing nonlocal interactions.

Table A4: Relative deployment speed of the models trained on 1000 configurations. Descriptor sizes go from 1 to 0, where 1 corresponds to a default global descriptor and 0 to an empty descriptor.

Descriptor size	$AcAla_3NHMe$	AT-AT	Buckyball Catcher
1	1.00	1.00	1.00
0.9	1.06	1.04	1.10
0.8	1.19	1.17	1.23
0.7	1.32	1.28	1.33
0.6	1.48	1.47	1.50
0.5	1.68	1.72	1.77
0.4	1.93	2.06	2.10
0.3	2.22	2.56	2.63
0.2	2.62	3.22	3.83
0.1	3.17	4.23	4.90

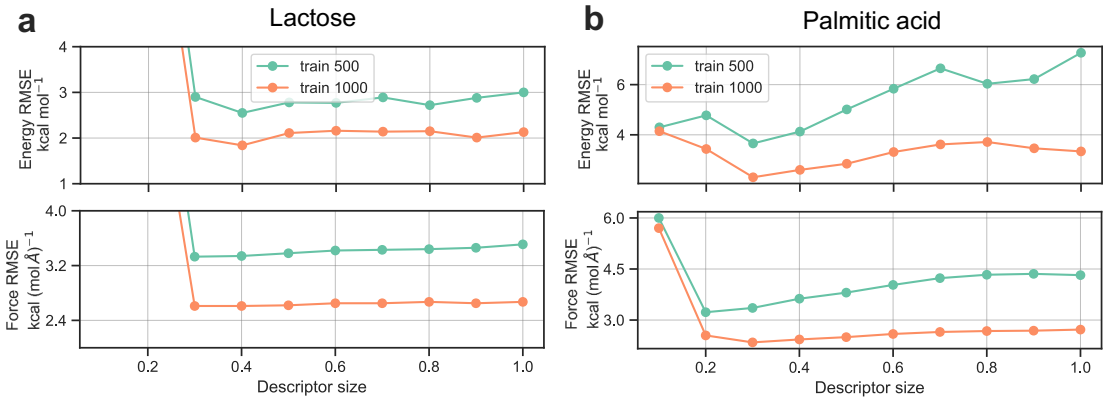

Figure A4: Accuracy of the models during the course of reduction. Energy (in kcal/mol) and force (in kcal/mol/Å) root means square errors (RMSE) as a function of the size of the descriptor for lactose (**a**) and palmitic acid (**b**) trained with 500 and 1000 training examples (green and orange colors, respectively). Descriptor sizes in x-axis go from 1 to 0, where 1 corresponds to a default global descriptor and 0 to an empty descriptor.

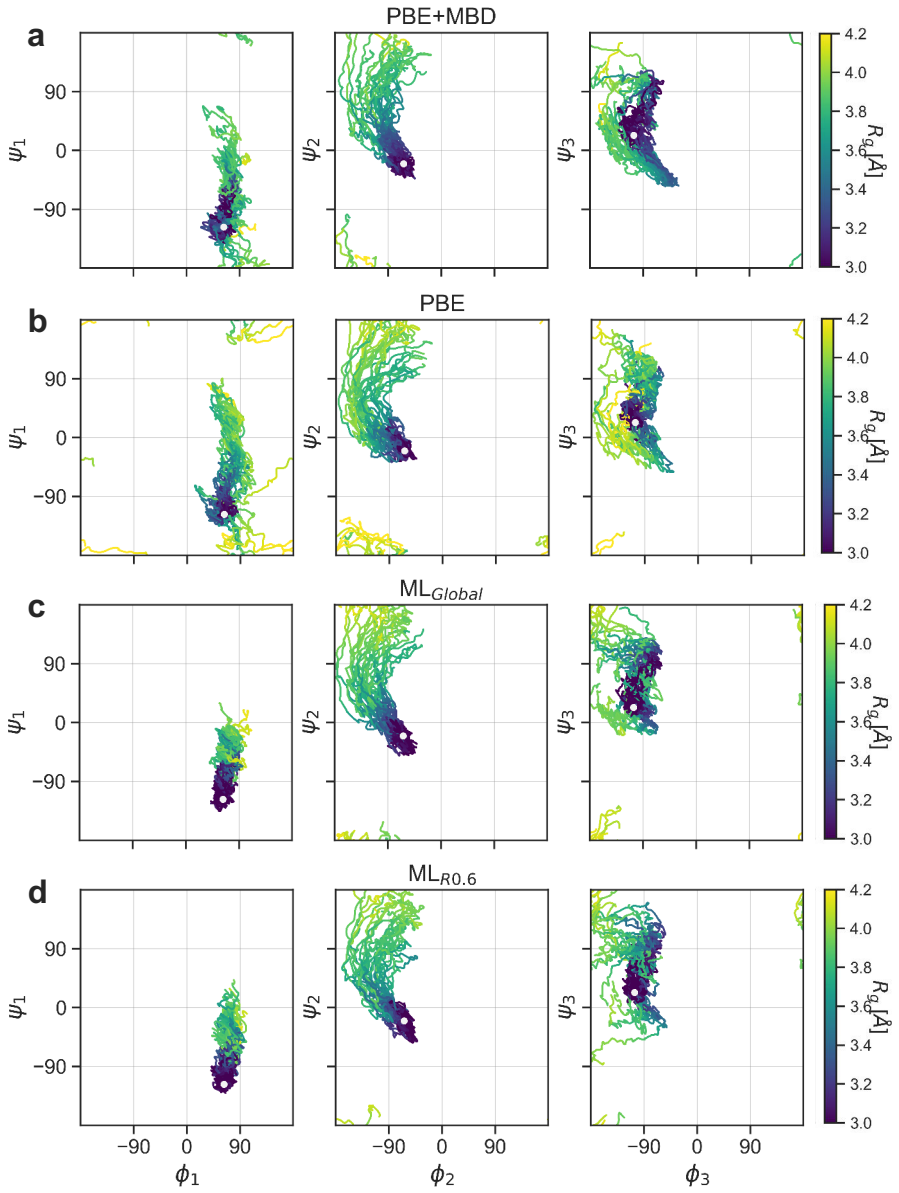

Figure A5: Ramachandran plots from 30 external-force simulations of $AcAla_3NHMe$ at 300 K employing the PBE+MBD (**a**) and PBE (**b**) level of theory, as well as ML_{Global} (**c**) and $ML_{R0.6}$ (**d**) models trained on 5000 data points. Light grey dot represent starting configuration.

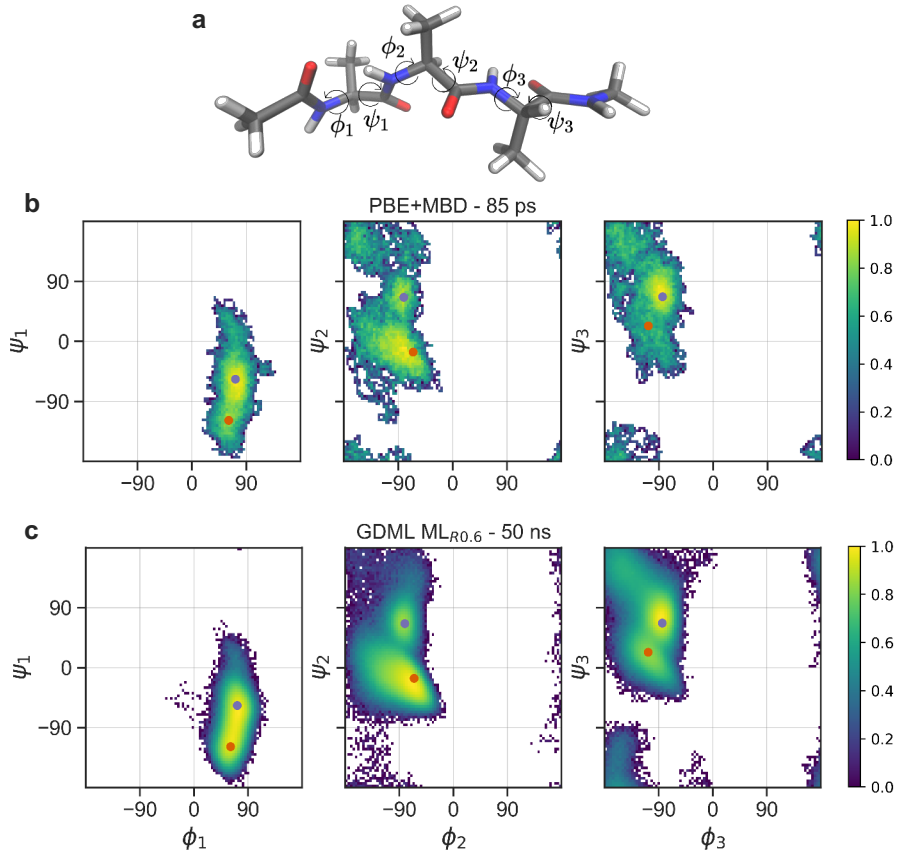

Figure A6: Ramachandran plots with angles ϕ/ψ defined in (**a**). Ramachandran plots show the initial dataset of AcAla$_3$NHMe - 85 ps, computed at the PBE+MBD level of theory (**b**), along with the resulting 50 ns dynamics obtained with the reduced ML$_{R0.6}$ model trained on 5000 data points (**c**). The color represents the population of the bins on a log scale normalized to the 0-1 range. Orange and purple dots indicate compact and extended structures, respectively.

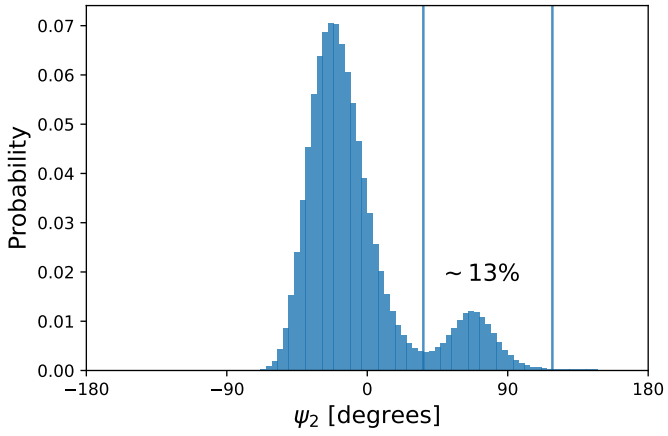

Figure A7: Distribution of the ψ_2 angle during 50 ns dynamics of AcAla$_3$NHMe. Two vertical lines mark out region of configurations in extended state.

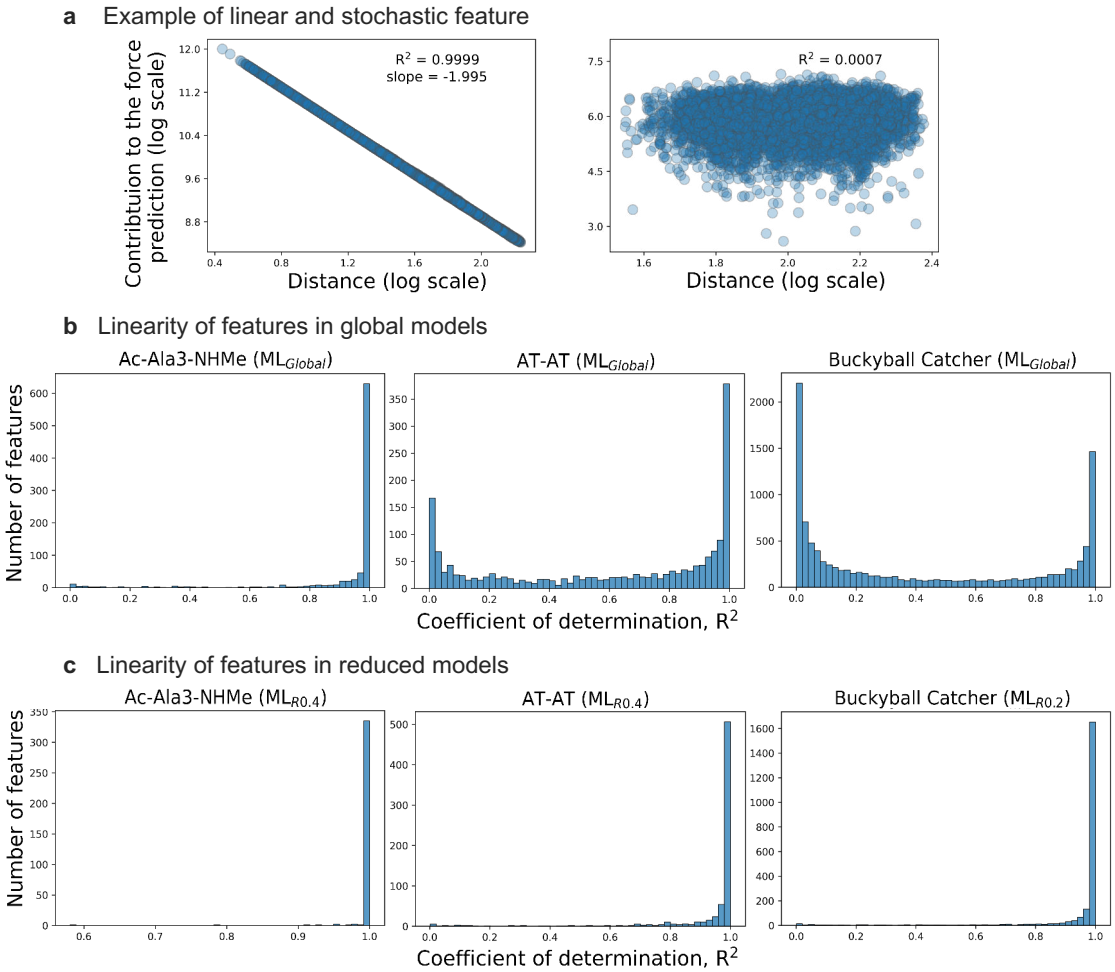

Figure A8: Analysis of the scale of the contribution with respect to distance of particular features in the global and reduced models. Example of linear and stochastic features (**a**), linearity of features in global (**b**) and reduced (**c**) models.

A4 Supplementary Information Chapter 6: SO3LR

Reference calculations. All reference calculations were performed at the PBE0+MBD level of theory using the FHI-aims code [162, 163]. "Tight" settings were applied for basis functions and integration grids. Energies were converged to 10^{-6} eV and the force accuracy was set to 10^{-5} eV/Å. The self-consistent field (SCF) optimization convergence criteria were 10^{-5} eV for the sum of eigenvalues and 10^{-6} electrons/Å^3 for the charge density. The Zenodo repository contains the relevant settings file.

Table A5: Properties present in the combined datasets.

Dataset	Size	Forces	Dipoles	Hirsh. rat.
GEMS bottom-up	2.7m	✓	✓	✗
QM7-X	1m	✓	✓	✓
AQM	60k	✓	✓	✓
SPICE Dipeptides	33k	✓	✓	✓
DES15k	15k	✓	✓	✓

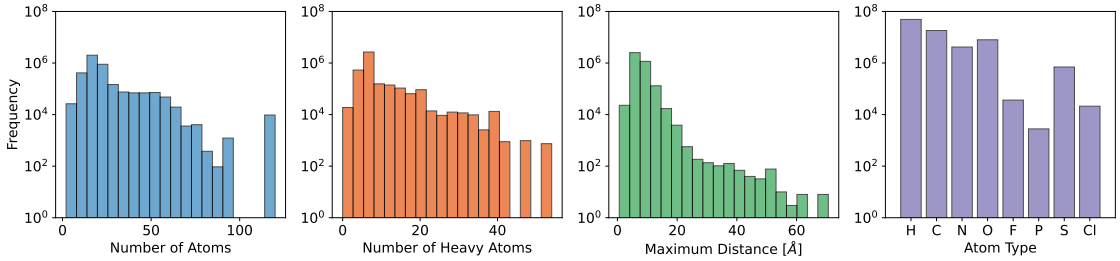

Figure A9: Statistics on a combined dataset of 3.9 million molecular fragments. Histograms of the number of atoms, number of heavy atoms, maximum distance in each fragment, and atom types.

SO3krates. The SO3krates neural network contains two sets of features: High-dimensional invariant atomic features $\mathcal{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N \,|\, \boldsymbol{h}_i \in \mathbb{R}^H\}$ and low-dimensional equivariant atomic features $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \,|\, \boldsymbol{x}_i \in \mathbb{R}^{(L+1)^2}\}$, where L denotes the maximal degree of the spherical harmonic used in the network. For a compact and formal introduction into invariance and equivariance see subsection "Symmetry and Equivariance".

Initial invariant atomic features encode information about the atomic types Z, the total charge Q and the multiplicity S of the system. Whereas the atomic types are defined for each atom in the system, the total charge and multiplicity is defined per molecule. On a high level initial features are calculated as

$$\boldsymbol{h}_i^{[0]} = \boldsymbol{e}_{i,Z} + \boldsymbol{e}_{i,Q} + \boldsymbol{e}_{i,S}, \tag{A5}$$

where each summand is a H-dimensional embedding vector for each atom i in the system. Atomic numbers are encoded as

$$\boldsymbol{e}_{i,Z} = \text{Embed}(z_i), \tag{A6}$$

where "Embed" is an embedding function that takes an atomic number $z_i \in \mathbb{N}_+$ and returns a H-dimensional embedding vector.

Following the strategy described in Ref. 71, total charge and multiplicity are encoded as

$$\begin{aligned} \boldsymbol{p}_i &= \mathrm{Embed}(Z_i), \quad \boldsymbol{k} = \begin{cases} \boldsymbol{k}_+ \text{ if } \Psi \geq 0 \\ \boldsymbol{k}_- \text{ if } \Psi < 0 \end{cases} \quad \boldsymbol{v} = \begin{cases} \boldsymbol{v}_+ \text{ if } \Psi \geq 0 \\ \boldsymbol{v}_- \text{ if } \Psi < 0 \end{cases} \\ a_i &= \frac{\Psi \ln\left(1 + \exp \boldsymbol{p}_i^\intercal \cdot \boldsymbol{k}/\sqrt{H}\right)}{\sum_{j=1}^{N} \ln\left(1 + \exp \boldsymbol{p}_j^\intercal \cdot \boldsymbol{k}/\sqrt{H}\right)}, \quad \boldsymbol{e}_{i,\Psi} = \mathrm{MLP}(a_i \boldsymbol{v}_i), \end{aligned} \tag{A7}$$

where $\boldsymbol{k}, \boldsymbol{v} \in \mathbb{R}^H$ are trainable parameters, and "Embed" is another embedding function for atomic numbers and "MLP" is a two-layered multi-layer perceptron (MLP) network. Separate parameters are used for charge $\Psi = Q$ and spin $\Psi = S$ embeddings and also for positive and negative values of Ψ, indicated by the subscripts "+" and "−", respectively. Since S is always equal or larger than zero, only the positive terms are used. Additionally, the MLP does not use any bias terms, such that $\mathrm{MLP}(a\boldsymbol{v}) = \boldsymbol{0}$ if $a\boldsymbol{v} = \boldsymbol{0}$. The described encoding procedure globally distributes the information about the total charge and the spin state, using per-atom weighting factors a_i.

The low-dimensional equivariant features are initialized to all-zeros i.e., $\boldsymbol{x} = \boldsymbol{0}$. This ensures that equivariance is preserved throughout the network. Alternative embedding schemes could be employed to embed equivariant features, i.e., an initial neighborhood scan as done in the original publications or in other equivariant MPNN approaches [52, 53].

After creating initial atomic features $(\mathcal{H}^{[t=0]}, \mathcal{X}^{[t=0]})$ they are iteratively refined via T fast equivariant MP layers as

$$\left(\mathcal{H}_i^{[t+1]}, \mathcal{X}_i^{[t+1]}\right) = \mathrm{FastEquivMP}\left[\mathcal{H}^{[t]}, \mathcal{X}^{[t]}, \mathcal{G}_\mathcal{R}\right], \tag{A8}$$

where $\mathcal{G}_\mathcal{R} = (\mathcal{R}, \mathcal{E})$ denotes a geometric graph, containing information about the atomic positions $\mathcal{R}$ and the inter-atomic connectivity via "Edges" $\mathcal{E}$. Edges are determined based on a local cutoff radius r_cut around each atom, and atoms lying within the cutoff sphere are considered a neighbor of the central atom, i.e., they share an edge. Here we use a cutoff of $r_\mathrm{cut} = 4.5\,\text{Å}$, greatly exceeding covalent bond lengths. In contrast to classical FFs, which often assume a fixed connectivity, a geometric graph is re-constructed for every set of atomic positions, such that breaking and forming of atomic bonds is handled naturally.

Each layer "FastEquivMP" layer consists of two phases. In the first phase, information from neighboring atoms is aggregated. In the second phase, the high-dimensional invariant and the low-dimensional equivariant features exchange information on a per-atom basis. This design ensures low computational cost while maintaining the benefits of equivariant feature representations. For full information the reader is referred to the original publication in Ref. 53.

The final invariant features $\boldsymbol{h}_i^{[T]} \in \mathbb{R}^H$ are used to predict the total energy of the molecule as

$$E_\mathrm{SO3k} = \sum_{i=1}^{N} \mathrm{MLP}\left(\boldsymbol{h}_i^{[T]}\right), \tag{A9}$$

with a two-layer MLP outputing a scalar energy contribution for each atom in the molecule. Forces are obtained as the gradient w.r.t. atomic positions $\vec{F}_i = -\nabla_{\vec{r}_i} E_\mathrm{SO3k}$ using automatic differentation.

Ziegler-Biersack-Littmark repulsion. The short-range repulsion between nuclei is modeled via a term inspired by the ZBL repulsion [71, 229]:

$$E_{\text{ZBL}} = k_e \sum_i \sum_{j \in \mathcal{N}_i} \frac{Z_i Z_j}{r_{ij}} f_{\text{cut}}(r_{ij}) \cdot \sum_{m=1}^{4} c_m e^{-a_m r_{ij}(Z_i^p + Z_j^p)/d}, \tag{A10}$$

where k_e is the Coulomb constant, Z_i are the atomic numbers, and a_m, c_m, p, and d are free parameters. The term $\mathcal{N}_i$ denotes the neighborhood of the i-th atom, and f_{cut} is a cutoff function that smoothly transitions between one and zero when atoms leave (or enter) the neighborhood. The ZBL term ensures a correct description of nuclear repulsion, which improves the stability of the potential for short bond-distances.

Partial charges and Dipoles. Following Ref. 71, partial charges are obtained as

$$q_i = q_{Z_i} + \tilde{q}_i + \frac{1}{N}\left(Q - \sum\nolimits_{j=1}^{N}(q_{Z_j} + \tilde{q}_j)\right), \tag{A11}$$

where $\tilde{q}_i \in \mathbb{R}$ are predicted from the final atomic representations $\boldsymbol{h}_i^{[T]} \in \mathbb{R}^H$ via a two-layered MLP network with silu nonlinearity and $q_{Z_i} \in \mathbb{R}$ is an element dependent bias. The charge correction with the total charge Q ensures charge conservation. The partial charges can be used to predict molecular dipole moments (used in the loss function, see Eq. 6.7):

$$\vec{\mu} = \sum_{i=1}^{N} q_i \vec{r}_i, \tag{A12}$$

where $\vec{r}_i \in \mathbb{R}^3$ are the atomic positions (assumed to be centered).

Long-range cutoff. For large structures with tens to hundreds of thousands of atoms, considering all pairs of atoms becomes computationally infeasible and necessitates the introduction of a long-range cutoff. Additionally, if simulations are performed in a box (e.g. with water) the largest meaningful long-range cutoff is directly connected to the box size. As such, the system under investigation and the simulation parameters, determine different values for the long-range cutoff. To account for this, we carefully designed a switching function for the long-range potentials, which allows to choose between different cutoff values up to no long-range cutoff at the time of simulation. The choice does not affect the first two terms in Eq. 6.1 or intermediate properties, partial charges and Hirshfeld ratios, which are used as inputs to calculate the last two terms.

Both the dispersion and the electrostatic potential have infinite range and take on a non-zero value at the long-range cutoff (Fig. A10). This results in a discontinuity in the forces at the cutoff value, leading to energy drift during MD simulations [282]. To ensure smoothness of the PES at the long-range cutoff we modify the pairwise electrostatic potential as

$$\tilde{u}(r) = \frac{q_i q_j k_e}{2} \cdot f_{\text{sw}}(r) \cdot u_{\text{ES}}(r) + \big(1 - f_{\text{sw}}(r)\big) \cdot u_{\text{FS}}(r), \tag{A13}$$

where $u_{\text{ES}}(r)$ is the energy-shifted potential, $u_{\text{FS}}(r)$ is the force-shifted potential (see below) and $f_{\text{sw}}(r)$ is a switching function that smoothly interpolates between 1 and 0 on a given interval from r_{on} to r_{off}. By switching between the energy- and the force-shifted terms, the potential remains unaltered within the short-range cutoff (which maintains the learned

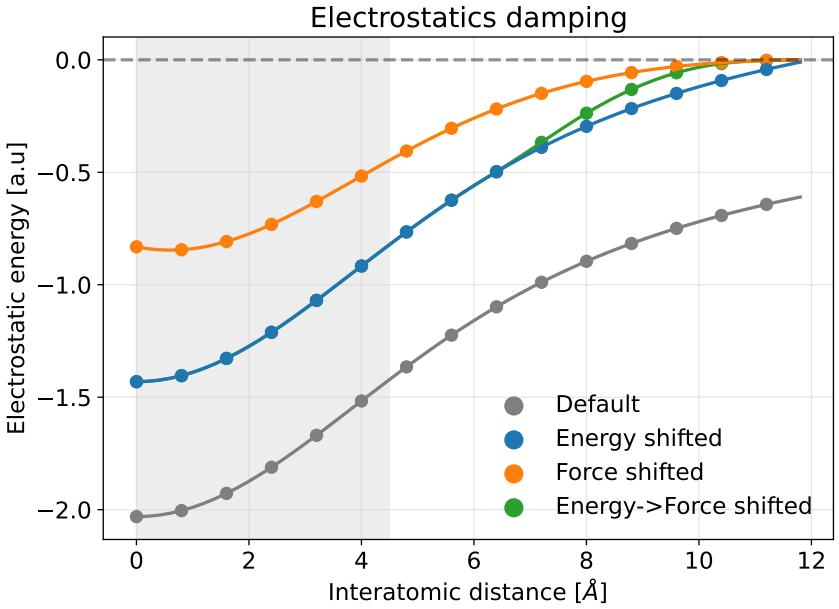

Figure A10: Switching electrostatic interactions. The model was trained with a damped $\mathrm{erf}(r_{ij}/4)/r_{ij}$ electrostatic potential on gas-phase data, with a 100 Å long-range cutoff that recovers all neighbors and a 4.5 Å short-range cutoff. In simulations with periodic boundary conditions, we employ a long-range cutoff of 12 Å to balance accuracy and computational efficiency. The potential at short-range should be the same as the one the model was trained with to maintain the learned balance between different terms. Simultaneously, the potential should smoothly transition to zero at the long-range cutoff to ensure that the potential is the exact integral of the force and to avoid introducing discontinuities in the forces. Therefore, we smoothly switch between the energy-shifted (blue curve) and force-shifted (orange curve) potentials to obtain the final potential (green curve). Dispersion interactions are smoothly energy-shifted starting 2 Å before long-range cutoff.

balance between different terms, as there was no long-range cutoff during training) while smoothly transitioning to zero at the long-range cutoff (Fig. A10). The shifted potentials are given as [282, 283]

$$u_{\mathrm{ES}}(r) = \begin{cases} u(r) - u(R_c), & r < R_c \\ 0, & r > R_c \end{cases} \tag{A14}$$

and

$$u_{\mathrm{FS}}(r) = \begin{cases} u(r) - u(R_c) - u(r - R_c) \cdot u'(R_c), & r < R_c \\ 0, & r > R_c \end{cases} \tag{A15}$$

where $u(r)$ is the unmodified pairwise electrostatic potential (Eq. 6.6) and R_c is the long-range cutoff. Dispersion interactions are smoothly switched to zero as

$$\tilde{v}(r) = f_{\mathrm{sw}}(r) \cdot v(r), \tag{A16}$$

where $v(r)$ is the pairwise potential in Eq. 6.4. The switching function parameters (r_{on}, r_{off}) were set to ($R_c \times 0.45$, R_c) for electrostatic interactions and ($R_c - 2$, R_c) for dispersion interactions. The parameters were chosen to prevent clumping artifacts at the 10 Å long-range cutoff.

Training details. SO3krates model (v1.0) was trained on a combined loss of forces, dipole moments, and Hirshfeld ratios with a weighting factor of 10:1:1, respectively. We used the AMSGrad optimizer [284] with an initial learning rate of 10^{-3} and an exponential learning rate decay every 500k steps by a factor of 0.85. The global norm of the gradient updates is clipped at 10.

The model uses a 4.5 Å cutoff, feature dimension of $H = 128$, and a maximal degree of $L = 4$ for the Euclidean variables and $T = 3$ message passing layers, electrostatics damping coefficient of $\sigma = 4$, and dispersion damping coefficient of $\gamma = 1.2$. After each attention update, a two-layered multi-layer perceptron with silu nonlinearity refines the invariant features. This increases the number of trainable parameters and thus model expressiveness, which is important in the large data regime. To stabilize training and improve gradient flow, layer normalization [285] is applied to the invariant features after the attention and the interaction block. The model was trained on a single A100 GPU for 86 h (corresponding to 5.125M gradient steps) with a batch size of $B = 200$.

Binding energy calculation. Binding energy was calculated as the difference between the bound dimer and the non-interacting monomers (separated by a distance larger than the long-range cutoff) with charges assigned for each monomer in isolation. The GitHub repository contains an example script, demonstrating the binding energy computation.

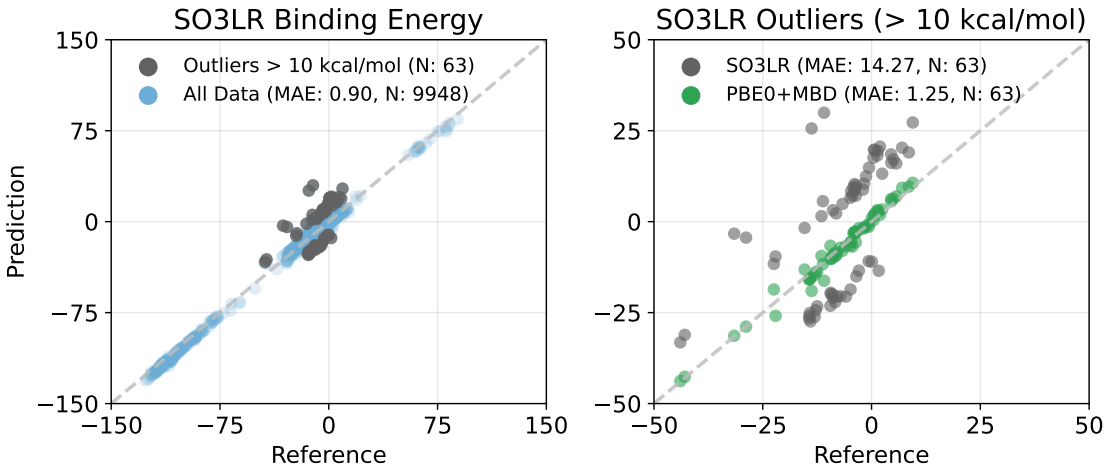

Figure A11: SAPT10k outliers analysis. Outlier structures with binding energy errors >10 kcal/mol (left) include exotic molecules such as ClF, $P(CNO)_3$, and PH_2NO_2. Recalculation of the SO3LR outliers (63 dimers) at the PBE0+MBD/tight level (right) yields a mean absolute error (MAE) of 1.25 kcal/mol, with a maximum error of 5.33 kcal/mol. These results confirm that the errors arise from the absence of these motifs in the training set, rather than from limitations of the reference theory.

Simulation details. All simulations were conducted using the NVT ensemble for gas-phase systems (with a long-range cutoff of 100 Å) and the NPT ensemble for periodic systems (with a long-range cutoff of 12 Å), with a timestep of 0.5 fs. Nosé–Hoover Chains (3 chains) were used for thermostat and barostat coupling, as implemented in JAX-MD [227], with default parameters: 1000 timesteps for the barostat and 100 timesteps for the thermostat [8–10]. Prior to simulation, all structures were pre-optimized using the FIRE algorithm [286].

MD22 molecules and polyalanines. Gas-phase simulations for stachyose and $AcAla_3NHMe$ were performed for 500 ps at 500 K, and for other MD22 molecules at 300 K. Simulations of $AcAla_{15}NHMe$ folding were carried out at 300 K (with initial velocities sampled from a Maxwell-Boltzmann distribution at 600 K). $AcAla_{15}LysH^+$ was simulated at 500–800 K with a step of 100 K, each for 500 ps. Secondary structure assignment of polyalanines was performed using the STRIDE algorithm [287].

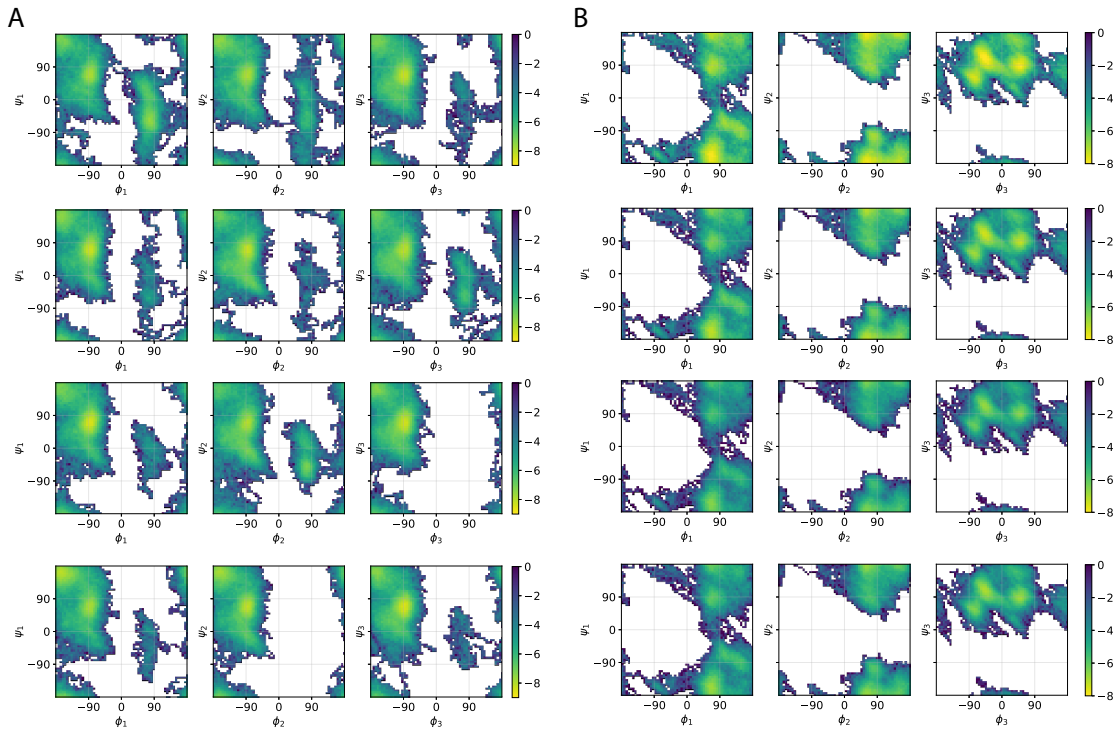

Figure A12: Simulations of small biomolecular fragments. Ramachandran plots (ϕ/ψ dihedrals) for (A) $AcAla_3NHMe$ and (B) stachyose molecules from the MD22 dataset. SO3LR simulations at 500 K for 500 ps. Trajectory is sampled every 1 fs. The Boltzmann-inverted scale is shown in kcal/mol.

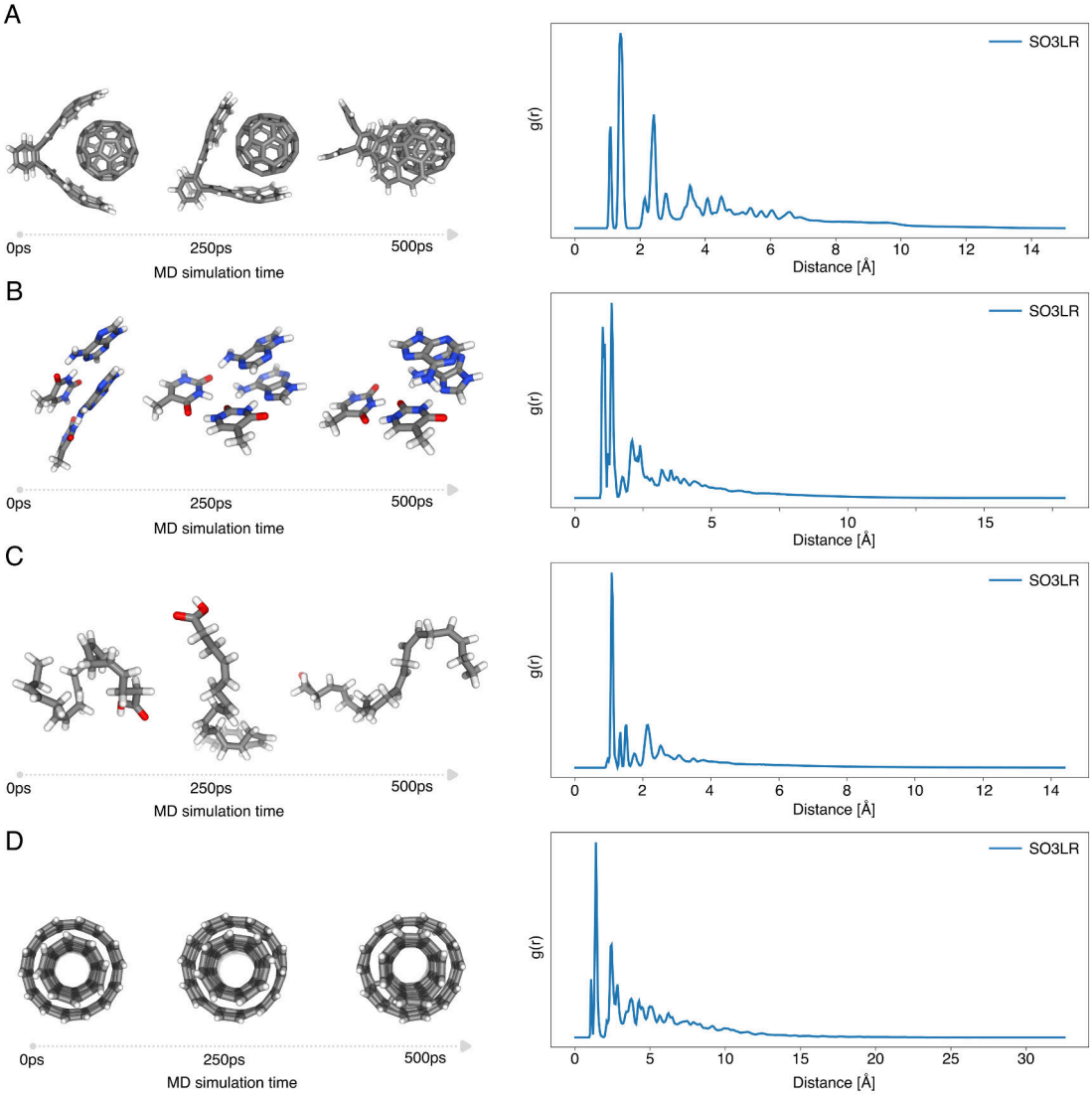

Figure A13: Simulation of structures from the MD22 dataset. Snapshots of the simulation at 0 ps, 250 ps and 500 ps (left) and the corresponding radial distribution function $g(r)$ computed over frames sampled every 1 ps (right) for the **(A)** buckyball catcher, **(B)** AT-AT, **(C)** DHA, and **(D)** double-walled nanotube. Simulations were performed for 500 ps at 300 K.

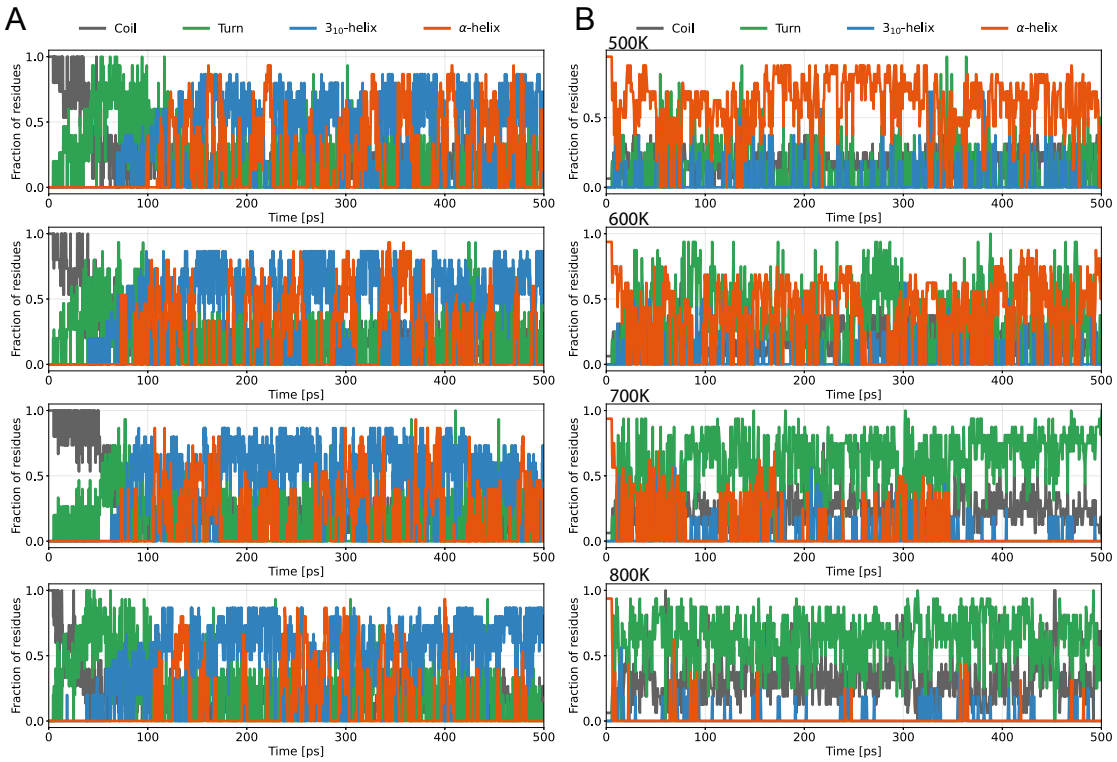

Figure A14: Polyalanine simulation. Secondary structural motifs of (A) four folding trajectories of extended $AcAla_{15}NHMe$ at 300 K in the gas phase and (B) four trajectories starting from the folded $AcAla_{15}LysH^{+}$ at 500, 600, 700, and 800 K. STRIDE was used for secondary structure assignment [287].

Water simulations. Simulations were run for 500 ps, with observables averaged over the final 300 ps. The diffusion coefficient was determined from the positions of oxygen atoms using Einstein diffusion equation [288, 289]. Double-precision was employed to enhance numerical stability.

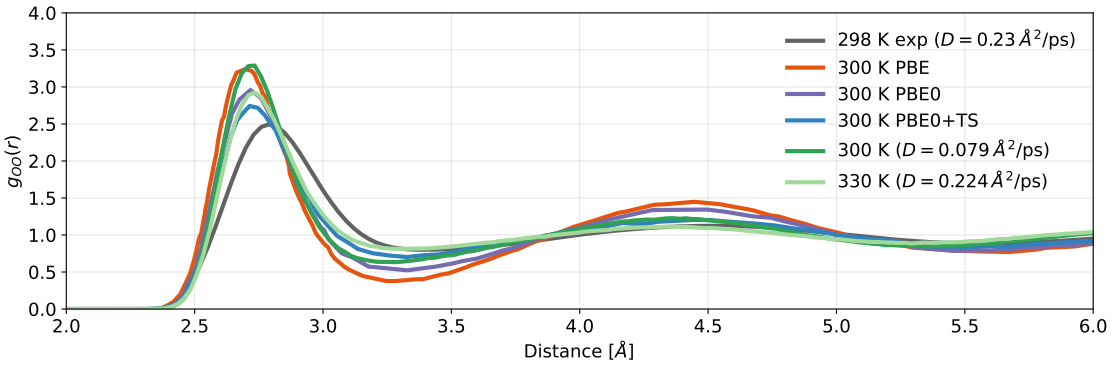

Figure A15: Oxygen-oxygen radial distribution function for bulk water. The SO3LR values were calculated from NPT molecular dynamics simulations of 4096 water molecules, run for 500 ps, with observables averaged over the final 250 ps. DFT values (PBE, PBE0, PBE0+TS) were taken from Ref. 75. The diffusion coefficients of water at 300 and 330 K, obtained using the model with a 12 Å long-range cutoff, are specified in the legend. The experimental values were taken from Refs. 169, 250.

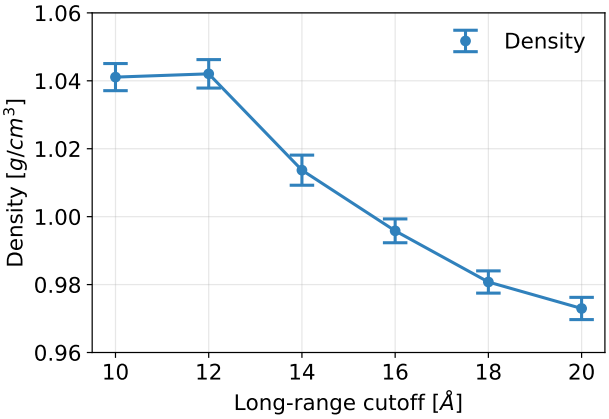

Figure A16: Dependence of the water density on long-range cutoff at 300 K. We investigated the convergence of the density as a function of the cutoff for long-range interactions (see 'long-range cutoff' subsection). The water density varies between 1.04 and 0.97 g/cm^3 for long-range cutoffs of 10–20 Å.

Crambin. The initial structure was obtained from PDB ID: 2FD7 [290], with mutated residues reverted to the wild-type sequence. The system was solvated with 8205 explicit water molecules. Simulations were performed for 3 ns at 300 K, excluding the first 0.5 ns for equilibration. The root mean square deviation of Crambin, RMSD(t, t+Δt), was obtained excluding hydrogen atoms from three 3 ns runs. Power spectra were computed from atomic velocities sampled over a 125 ps trajectory with a time resolution of 2.5 fs using schnetpack package [172, 291].

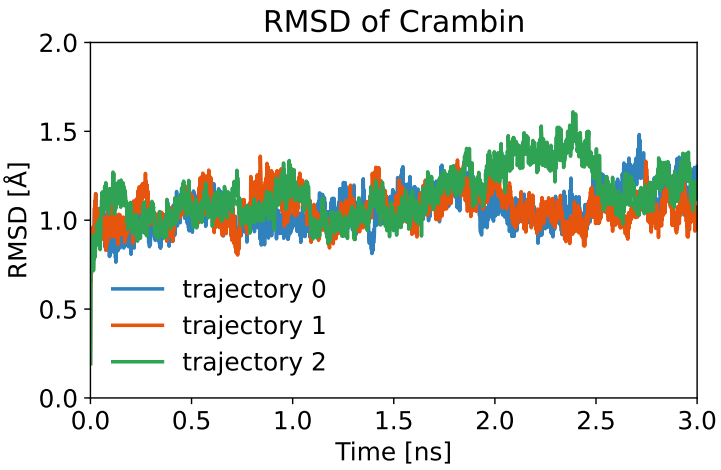

Figure A17: Crambin RMSD. Root mean square deviations of three crambin trajectories simulated with SO3LR with respect to the initial frame.

Glycoprotein. The starting structure was taken from the PDB ID: 1K7C [292]. 15008 water molecules were used for solvating the system and the pH was set to 3.7 to guarantee charge neutralization. The RMSD was calculated based on three runs of 500 ps.

POPC Lipid bilayer. The starting structure, consisting of 128 lipids and 5120 water molecules, was obtained from Ref. 145. The system was equilibrated over 250 ps using a combination of geometry relaxations and NVT simulations. Observables were then averaged over an additional 250 ps at 303 K in an isotropic NPT ensemble implemented in JAX-MD. The initial box dimensions were adjusted manually to mimic semi-isotropic NPT ensemble. The area per lipid was calculated from the simulation box dimensions. Bilayer thickness (D_{HH}), derived from electron density profiles, and NMR order parameters were both calculated using CPPTRAJ [293]. Double-precision was employed to enhance numerical stability.

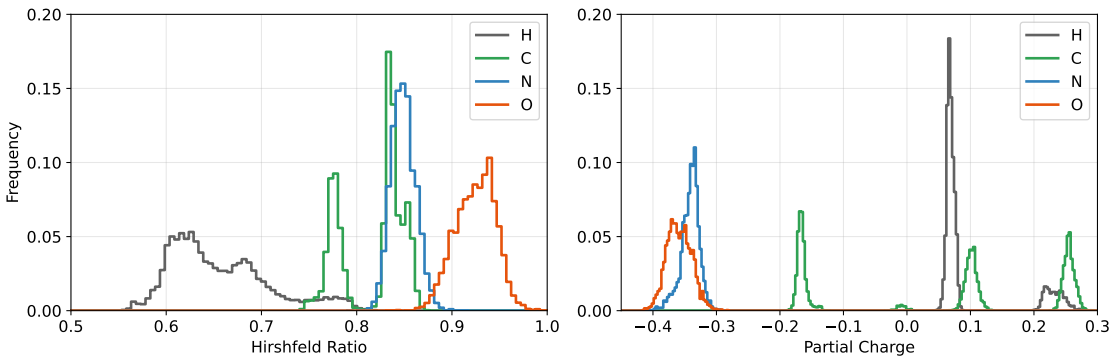

Figure A18: Hirshfeld ratio and partial charge distribution for the $AcAla_{15}NHMe$. Both quantities are dynamically changing and are predicted by the SO3LR model.

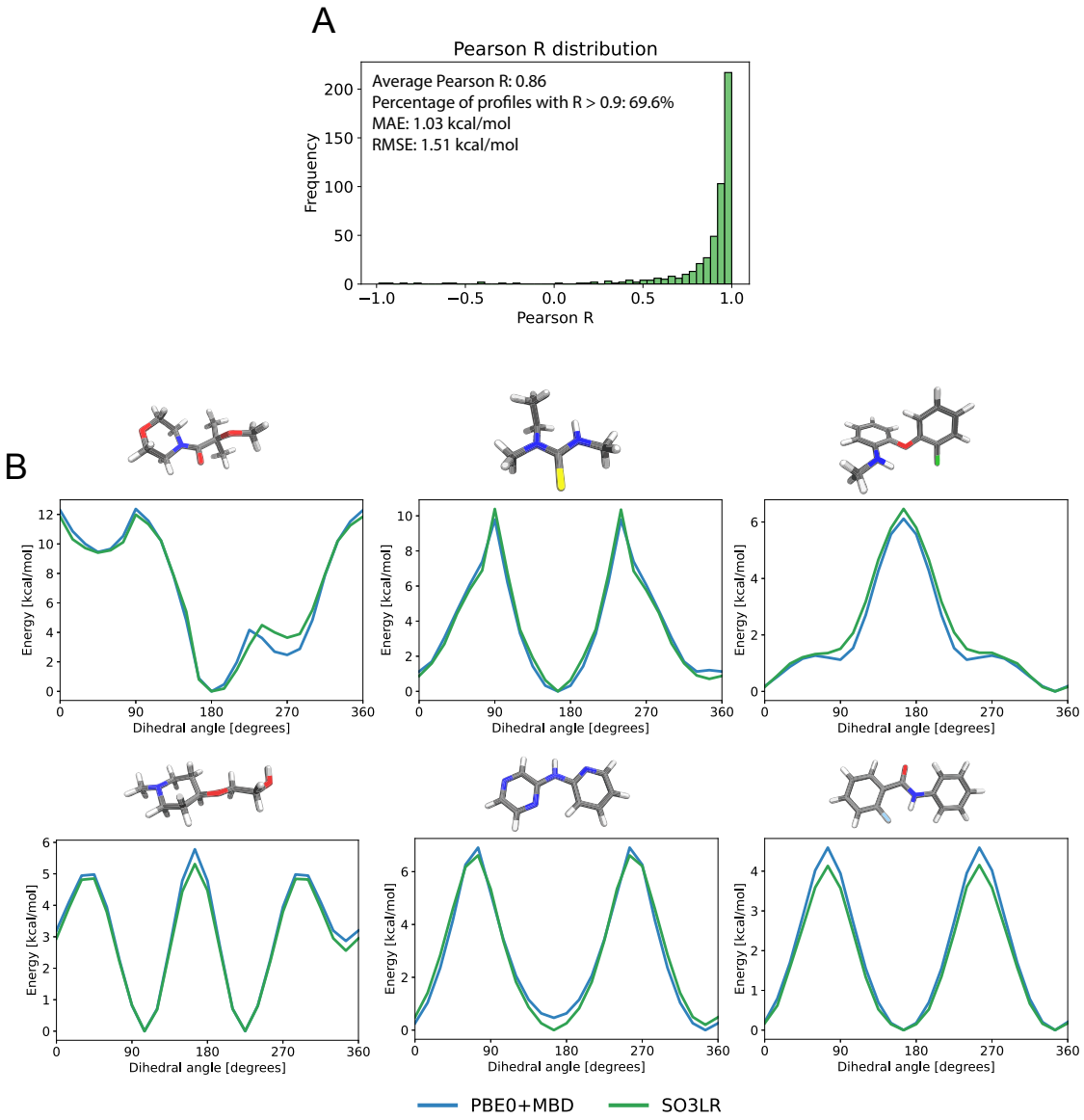

Figure A19: TorsionNet500 Benchmark. Comparison of energies predicted by SO3LR with the TorsionNet500 benchmark [242], recomputed at the PBE0+MBD level of theory. **A** Histogram of Pearson R coefficients, with additional metrics shown in the inset. **B** Torsional profiles for six molecules. The absence of certain functional groups (e.g., triazole and trifluoromethylthio groups) in the training set leads to higher average errors. In contrast, torsional profiles commonly encountered in biosimulations are predicted accurately.

Bibliography

[1] R. O. Dror, R. M. Dirks, J. Grossman, H. Xu, and D. E. Shaw, Annu. Rev. Biophys. **41**, 429 (2012).

[2] A. Singharoy, C. Maffeo, K. H. Delgado-Magnero, D. J. Swainsbury, M. Sener, U. Kleinekathöfer, J. W. Vant, J. Nguyen, A. Hitchcock, B. Isralewitz, *et al.*, Cell **179**, 1098 (2019).

[3] S. Antolínez, P. E. Jones, J. C. Phillips, and J. A. Hadden-Perilla, Biophys. J. **123**, 422a (2024).

[4] A. T. Hagler, J. Comput. Aided Mol. Des. **33**, 205 (2019).

[5] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, Chem. Rev. **121**, 10142 (2021).

[6] L. Piela, *Ideas of quantum chemistry* (Elsevier, 2013).

[7] M. E. Tuckerman, *Statistical mechanics: theory and molecular simulation* (Oxford university press, 2023).

[8] S. Nosé, J. Chem. Phys **81**, 511 (1984).

[9] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).

[10] G. J. Martyna, M. L. Klein, and M. Tuckerman, J. Chem. Phys **97**, 2635 (1992).

[11] J. Hermann, J. Spencer, K. Choo, A. Mezzacapo, W. M. C. Foulkes, D. Pfau, G. Carleo, and F. Noé, Nat. Rev. Chem. **7**, 692 (2023).

[12] C. Møller and M. S. Plesset, Phys. Rev. **46**, 618 (1934).

[13] R. J. Bartlett and M. Musiał, Rev. Mod. Phys. **79**, 291 (2007).

[14] K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, Chem. Phys. Lett. **157**, 479 (1989).

[15] B. Ladóczki, L. Gyevi-Nagy, P. R. Nagy, and M. Kállay, J. Chem. Theory Comput. **21**, 2432 (2025).

[16] J. P. Perdew and K. Schmidt, AIP Conf. Proc. **577**, 1 (2001).

[17] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[18] C. Adamo and V. Barone, J. Chem. Phys. **110**, 6158 (1999).

[19] J. Sun, A. Ruzsinszky, and J. P. Perdew, Phys. Rev. Lett. **115**, 036402 (2015).

[20] M. Stöhr, T. Van Voorhis, and A. Tkatchenko, Chem. Soc. Rev. **48**, 4118 (2019).

[21] R. A. DiStasio, V. V. Gobre, and A. Tkatchenko, J. Phys.: Condens. Matter **26**, 213202 (2014).

[22] A. M. Reilly and A. Tkatchenko, Chem. Sci. **6**, 3289 (2015).

[23] S. Grimme, A. Hansen, J. G. Brandenburg, and C. Bannwarth, Chem. Rev. **116**, 5105 (2016).

[24] A. Tkatchenko and M. Scheffler, Phys. Rev. Lett. **102**, 073005 (2009).

[25] A. D. Becke and E. R. Johnson, J. Chem. Phys. **122**, 154104 (2005).

[26] A. D. Becke and E. R. Johnson, J. Chem. Phys. **127**, 154108 (2007).

[27] J. F. Dobson, T. Gould, and G. Vignale, Phys. Rev. X **4**, 021040 (2014).

[28] V. V. Gobre and A. Tkatchenko, Nat. Commun. **4**, 2341 (2013).

[29] A. Tkatchenko, R. A. DiStasio Jr, R. Car, and M. Scheffler, Phys. Rev. Lett. **108**, 236402 (2012).

[30] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, *et al.*, J. Comput. Chem. **30**, 1545 (2009).

[31] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, J. Am. Chem. Soc. **118**, 11225 (1996).

[32] D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, J. Comput. Chem. **26**, 1668 (2005).

[33] P. Dauber-Osguthorpe and A. T. Hagler, J. Comput. Aided Mol. Des. **33**, 133 (2019).
[34] P. P. Ewald, Ann. Phys. **369**, 253 (1921).
[35] T. Darden, D. York, L. Pedersen, *et al.*, J. Chem. Phys. **98**, 10089 (1993).
[36] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).
[37] J. A. McCammon, B. R. Gelin, and M. Karplus, Nature **267**, 585 (1977).
[38] S. A. Hollingsworth and R. O. Dror, Neuron **99**, 1129 (2018).
[39] P. Ren and J. W. Ponder, J. Phys. Chem. B **107**, 5933 (2003).
[40] G. Lamoureux, A. D. MacKerell, and B. Roux, J. Chem. Phys. **119**, 5185 (2003).
[41] A. C. T. van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard, J. Phys. Chem. A **105**, 9396 (2001).
[42] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, J. Phys. Chem. B **111**, 7812 (2007).
[43] J. A. Stevens, F. Grünewald, P. van Tilburg, M. König, B. R. Gilbert, T. A. Brier, Z. R. Thornburg, Z. Luthey-Schulten, and S. J. Marrink, Front. Chem. **11**, 1106495 (2023).
[44] V. L. Deringer, M. A. Caro, and G. Csányi, Adv. Mater. **31**, 1902765 (2019).
[45] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[46] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[47] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).
[48] K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, Adv. Neural Inf. Process. Syst. **30** (2017).
[49] K. Schütt, O. Unke, and M. Gastegger, Int. Conf. on Mach. Learn. **139**, 9377 (2021).
[50] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, Nat. Commun. **13**, 1 (2022).
[51] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csányi, Adv. Neural Inf. Process. Syst. **35**, 11423 (2022).
[52] T. Frank, O. Unke, and K.-R. Müller, Adv. Neural Inf. Process. Syst. **35**, 29400 (2022).
[53] J. T. Frank, O. T. Unke, K.-R. Müller, and S. Chmiela, Nat. Commun. **15**, 6539 (2024).
[54] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Sci. Adv. **3**, e1603015 (2017).
[55] J. S. Smith, O. Isayev, and A. E. Roitberg, Sci. Data **4**, 1 (2017).
[56] J. Hoja, L. Medrano Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio Jr, and A. Tkatchenko, Sci. Data **8**, 43 (2021).
[57] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, Chem. Rev. **121**, 9759 (2021).
[58] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[59] F. A. Faber, A. S. Christensen, B. Huang, and O. A. Von Lilienfeld, J. Chem. Phys. **148**, 241717 (2018).
[60] J. Behler, J. Chem. Phys. **134**, 074106 (2011).
[61] A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. Anatole Von Lilienfeld, J. Chem. Phys. **152**, 044107 (2020).
[62] R. Drautz, Phys. Rev. B **99**, 014104 (2019).
[63] O. T. Unke and H. Maennel, arXiv preprint arXiv:2401.07595 (2024).
[64] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, Nat. Commun. **9**, 3887 (2018).
[65] S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, Comput. Phys. Commun. **240**, 38 (2019).
[66] S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko, and K.-R. Müller, Sci. Adv. **9**, eadf0873 (2023).

[67] G. Kimeldorf and G. Wahba, J. Math. Anal. Appl. **33**, 82 (1971).

[68] O. T. Unke and M. Meuwly, J. Chem. Theory Comput. **15**, 3678 (2019).

[69] J. T. Frank, *Modeling Larger Length and Time Scales in Machine Learning Force Fields*, Ph.D. thesis, Technische Universität Berlin (2025).

[70] A. Duval, S. V. Mathis, C. K. Joshi, V. Schmidt, S. Miret, F. D. Malliaros, T. Cohen, P. Lio, Y. Bengio, and M. Bronstein, arXiv preprint arXiv:2312.07511 (2023).

[71] O. T. Unke, S. Chmiela, M. Gastegger, K. T. Schütt, H. E. Sauceda, and K.-R. Müller, Nat. Commun. **12**, 1 (2021).

[72] R. H. French, V. A. Parsegian, R. Podgornik, R. F. Rajter, A. Jagota, J. Luo, D. Asthagiri, M. K. Chaudhury, Y.-m. Chiang, S. Granick, *et al.*, Rev. Mod. Phys. **82**, 1887 (2010).

[73] A. Tkatchenko, R. A. Distasio, R. Car, and M. Scheffler, Phys. Rev. Lett. **108**, 236402 (2012).

[74] J. Hermann, R. A. DiStasio Jr, and A. Tkatchenko, Chem. Rev. **117**, 4714 (2017).

[75] R. A. DiStasio, B. Santra, Z. Li, X. Wu, and R. Car, J. Chem. Phys. **141**, 084502 (2014).

[76] B. Santra, R. A. DiStasio Jr, F. Martelli, and R. Car, Mol. Phys. **113**, 2829 (2015).

[77] M. Stöhr and A. Tkatchenko, Sci. Adv. **5**, eaax0024 (2019).

[78] H. Schönherr and T. Cernak, Angew. Chem., Int. Ed. **52**, 12256 (2013).

[79] H.-X. Zhou and X. Pang, Chem. Rev. **118**, 1691 (2018).

[80] M. Patra, M. Karttunen, M. T. Hyvönen, E. Falck, P. Lindqvist, and I. Vattulainen, Biophys. J. **84**, 3636 (2003).

[81] L. Pauling, H. A. Itano, S. J. Singer, and I. C. Wells, Science **110**, 543 (1949).

[82] V. M. Ingram, Nature **178**, 792 (1956).

[83] J. D. Watson and F. H. Crick, Nature **171**, 737 (1953).

[84] J. T. Frank, S. Chmiela, K.-R. Müller, and O. T. Unke, arXiv preprint arXiv:2412.08541 (2024).

[85] Q. Li, Z. Han, and X.-M. Wu, Proc. AAAI Conf. Artif. **32** (2018).

[86] U. Alon and E. Yahav, arXiv preprint arXiv:2006.05205 (2020).

[87] M. Esders, T. Schnake, J. Lederer, A. Kabylda, G. Montavon, A. Tkatchenko, and K.-R. Müller, J. Chem. Theory Comput. **21**, 714 (2025).

[88] A. Kosmala, J. Gasteiger, N. Gao, and S. Günnemann, Int. Conf. on Mach. Learn. **202**, 17544 (2023).

[89] Y. Wang, T. Wang, S. Li, X. He, M. Li, Z. Wang, N. Zheng, B. Shao, and T.-Y. Liu, Nat. Commun. **15**, 313 (2024).

[90] N. Artrith, T. Morawietz, and J. Behler, Phys. Rev. B **83**, 153101 (2011).

[91] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, Nat. Commun. **12**, 398 (2021).

[92] A. Grisafi and M. Ceriotti, J. Chem. Phys. **151**, 204105 (2019).

[93] Y. Li, Y. Wang, L. Huang, H. Yang, X. Wei, J. Zhang, T. Wang, Z. Wang, B. Shao, and T.-Y. Liu, arXiv preprint arXiv:2304.13542 (2023).

[94] D. M. Anstine and O. Isayev, J. Phys. Chem. A **127**, 2417 (2023).

[95] J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Muller, and A. Tkatchenko, Chem. Rev. **121**, 9816 (2021).

[96] O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, Nat. Rev. Chem. **4**, 347 (2020).

[97] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, Commun. ACM **64**, 107 (2021).

[98] Z. Allen-Zhu, Y. Li, and Y. Liang, Adv. Neural Inf. Process. Syst. **32** (2019).

[99] A. Bietti and J. Mairal, Adv. Neural Inf. Process. Syst. **32** (2019).

[100] B. Chamberlain, J. Rowbottom, M. I. Gorinova, M. Bronstein, S. Webb, and E. Rossi, Int. Conf. on Mach. Learn. **139**, 1407 (2021).

[101] A. Ambrosetti, N. Ferri, R. A. DiStasio Jr, and A. Tkatchenko, Science **351**, 1171 (2016).

[102] O. T. Unke, M. Stöhr, S. Ganscha, T. Unterthiner, H. Maennel, S. Kashubin, D. Ahlin, M. Gastegger, L. Medrano Sandonas, J. T. Berryman, *et al.*, Sci. Adv. **10**, eadn4397 (2024).

[103] P. Hauseux, T.-T. Nguyen, A. Ambrosetti, K. S. Ruiz, S. P. Bordas, and A. Tkatchenko, Nat. Commun. **11**, 1651 (2020).

[104] P. Hauseux, A. Ambrosetti, S. P. Bordas, and A. Tkatchenko, Phys. Rev. Lett. **128**, 106101 (2022).

[105] T. Bereau, R. A. DiStasio, A. Tkatchenko, and O. A. Von Lilienfeld, J. Chem. Phys. **148**, 241706 (2018).

[106] K. Yao, J. E. Herr, D. W. Toth, R. Mckintyre, and J. Parkhill, Chem. Sci. **9**, 2261 (2018).

[107] S. P. Niblett, M. Galib, and D. T. Limmer, J. Chem. Phys. **155**, 164101 (2021).

[108] A. Gao and R. C. Remsing, Nat. Commun. **13**, 1 (2022).

[109] H. E. Sauceda, S. Chmiela, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, J. Chem. Phys. **150**, 114102 (2019).

[110] H. E. Sauceda, M. Gastegger, S. Chmiela, K.-R. Müller, and A. Tkatchenko, J. Chem. Phys. **153**, 124109 (2020).

[111] B. Schölkopf, A. Smola, and K.-R. Müller, Neural Comput. **10**, 1299 (1998).

[112] M. L. Braun, J. M. Buhmann, and K.-R. Müller, J. Mach. Learn. Res. **9**, 1875 (2008).

[113] H. E. Sauceda, L. E. Gálvez-González, S. Chmiela, L. O. Paz-Borbón, K.-R. Müller, and A. Tkatchenko, Nat. Commun. **13**, 1 (2022).

[114] J. R. Taylor, *An introduction to error analysis: the study of uncertainties in physical measurements* (MIT Press, 2022).

[115] S. Chmiela, H. E. Sauceda, A. Tkatchenko, and K.-R. Müller, in *Machine learning meets quantum physics* (Springer, 2020) pp. 129–154.

[116] A. M. Reilly and A. Tkatchenko, Phys. Rev. Lett. **113**, 055701 (2014).

[117] J. P. Hare, T. J. Dennis, H. W. Kroto, R. Taylor, A. W. Allaf, S. Balm, and D. R. Walton, J. Chem. Soc., Chem. Commun. **6**, 412 (1991).

[118] C. H. Choi, M. Kertesz, and L. Mihaly, J. Phys. Chem. A **104**, 102 (2000).

[119] U. Erlekam, M. Frankowski, G. Meijer, and G. von Helden, J. Chem. Phys. **124**, 171101 (2006).

[120] N. F. Schmitz, K.-R. Müller, and S. Chmiela, J. Phys. Chem. Lett. **13**, 10183 (2022).

[121] A. Kabylda, V. Vassilev-Galindo, S. Chmiela, I. Poltavsky, and A. Tkatchenko, Nat. Commun. **14**, 3562 (2023).

[122] B. Huang and O. A. von Lilienfeld, Nat. Chem. **12**, 945 (2020).

[123] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, Nat. Commun. **8**, 872 (2017).

[124] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, Nat. Commun. **8**, 13890 (2017).

[125] K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, Nat. Commun. **10**, 5024 (2019).

[126] M. Gastegger, A. McSloy, M. Luya, K. T. Schütt, and R. J. Maurer, J. Chem. Phys. **153**, 044123 (2020).

[127] A. Kabylda, S. Suárez-Dou, N. Davoine, F. N. Brünig, and A. Tkatchenko, arXiv preprint arXiv:2510.09939 (2025).

[128] I. Poltavsky, A. Charkin-Gorbulin, M. Puleva, G. Fonseca, I. Batatia, N. J. Browning, S. Chmiela, M. Cui, J. T. Frank, S. Heinen, *et al.*, Chem. Sci. **16**, 3720 (2025).

[129] I. Poltavsky, A. Charkin-Gorbulin, M. Puleva, G. Fonseca, I. Batatia, N. J. Browning, S. Chmiela, M. Cui, J. T. Frank, S. Heinen, *et al.*, Chem. Sci. **16**, 3738 (2025).

[130] S. A. Spronk, Z. L. Glick, D. P. Metcalf, C. D. Sherrill, and D. L. Cheney, Sci. Data **10**, 619 (2023).

[131] P. Eastman, P. K. Behara, D. L. Dotson, R. Galvelis, J. E. Herr, J. T. Horton, Y. Mao, J. D. Chodera, B. P. Pritchard, Y. Wang, *et al.*, Sci. Data **10**, 11 (2023).

[132] P. Eastman, B. P. Pritchard, J. D. Chodera, and T. E. Markland, J. Chem. Theory Comput. **20**, 8583 (2024).

[133] L. Medrano Sandonas, D. Van Rompaey, A. Fallani, M. Hilfiker, D. Hahn, L. Perez-Benito, J. Verhoeven, G. Tresadern, J. Kurt Wegner, H. Ceulemans, *et al.*, Sci. Data **11**, 742 (2024).

[134] S. Ganscha, O. T. Unke, D. Ahlin, H. Maennel, S. Kashubin, and K.-R. Müller, Sci. Data **12**, 406 (2025).

[135] D. M. Anstine, R. Zubatyuk, and O. Isayev, Chem. Sci. **16**, 10228 (2025).

[136] D. S. Levine, M. Shuaibi, E. W. C. Spotte-Smith, M. G. Taylor, M. R. Hasyim, K. Michel, I. Batatia, G. Csányi, M. Dzamba, P. Eastman, *et al.*, arXiv preprint arXiv:2505.08762 (2025).

[137] B. Alberts, R. Heald, A. Johnson, D. Morgan, M. Raff, K. Roberts, and P. Walter, *Molecular biology of the cell* (WW Norton & Company, 2022).

[138] S. Neidle, *Oxford Handbook of Nucleic Acid Structure* (Oxford University Press, 1999).

[139] K. Masuda, A. A. Abdullah, P. Pflughaupt, and A. B. Sahakyan, Sci. Data **11**, 911 (2024).

[140] T. J. Macke and D. A. Case, in *Molecular Modeling of Nucleic Acids* (American Chemical Society, 1997) Chap. 24, pp. 379–393.

[141] M. Zgarbová, J. Šponer, and P. Jurečka, J. Chem. Theory Comput. **17**, 6292 (2021).

[142] J. T. Berryman, A. Taghavi, F. Mazur, and A. Tkatchenko, J. Chem. Phys. **157**, 064107 (2022).

[143] C. Zhang, X. Zhang, L. Freddolino, and Y. Zhang, Nucleic Acids Res. **52**, D404 (2024).

[144] E. L. Wu, X. Cheng, S. Jo, H. Rui, K. C. Song, E. M. Dávila-Contreras, Y. Qi, J. Lee, V. Monje-Galvan, R. M. Venable, J. B. Klauda, and W. Im, J. Comput. Chem. **35**, 1997 (2014).

[145] C. J. Dickson, R. C. Walker, and I. R. Gould, J. Chem. Theory Comput. **18**, 1726 (2022).

[146] Schrödinger, LLC, Pymol molecular graphics system, version 1.8 (2015).

[147] S. Grimme, J. Chem. Theory Comput. **15**, 2847 (2019).

[148] P. Pracht, F. Bohle, and S. Grimme, Phys. Chem. Chem. Phys. **22**, 7169 (2020).

[149] P. Pracht, S. Grimme, C. Bannwarth, F. Bohle, S. Ehlert, G. Feldmann, J. Gorges, M. Müller, T. Neudecker, C. Plett, *et al.*, J. Chem. Phys. **160**, 114110 (2024).

[150] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. In't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, *et al.*, Comput. Phys. Commun. **271**, 108171 (2022).

[151] X. Zhu, M. Riera, E. F. Bull-Vulpe, and F. Paesani, J. Chem. Theory Comput. **19**, 3551 (2023).

[152] M. Riera, C. Knight, E. F. Bull-Vulpe, X. Zhu, H. Agnew, D. G. Smith, A. C. Simmonett, and F. Paesani, J. Chem. Phys. **159**, 054802 (2023).

[153] D. Beglov and B. Roux, J. Chem. Phys. **100**, 9050 (1994).

[154] P. Eastman, R. Galvelis, R. P. Peláez, C. R. Abreu, S. E. Farr, E. Gallicchio, A. Gorenko, M. M. Henry, F. Hu, J. Huang, *et al.*, J. Phys. Chem. B **128**, 109 (2024).

[155] A. Sengupta, Z. Li, L. F. Song, P. Li, and K. M. J. Merz, J. Chem. Inf. Model. **61**, 869 (2021).

[156] B. Hourahine, B. Aradi, V. Blum, F. Bonafe, A. Buccheri, C. Camacho, C. Cevallos, M. Deshaye, T. Dumitrică, A. Dominguez, *et al.*, J. Chem. Phys. **152**, 124101 (2020).

[157] A. Kabylda, J. T. Frank, S. Suárez-Dou, A. Khabibrakhmanov, L. Medrano Sandonas, O. T. Unke, S. Chmiela, K.-R. Müller, and A. Tkatchenko, J. Am. Chem. Soc. **147**, 33723 (2025).

[158] A. G. Donchev, A. G. Taube, E. Decolvenaere, C. Hargus, R. T. McGibbon, K.-H. Law, B. A. Gregersen, J.-L. Li, K. Palmo, K. Siva, *et al.*, Sci. Data **8**, 55 (2021).

[159] M. Puleva, L. Medrano Sandonas, B. D. Lőrincz, J. Charry, D. M. Rogers, P. R. Nagy, and A. Tkatchenko, Nat. Commun. **16**, 8583 (2025).

[160] Y. S. Al-Hamdani, P. R. Nagy, A. Zen, D. Barton, M. Kállay, J. G. Brandenburg, and A. Tkatchenko, Nat. Commun. **12**, 3927 (2021).

[161] J. Hermann and A. Tkatchenko, Phys. Rev. Lett. **124**, 146401 (2020).

[162] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, Comput. Phys. Commun. **180**, 2175 (2009).

[163] X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter, and M. Scheffler, New J. Phys. **14**, 053020 (2012).

[164] A. Kabylda, S. Suárez-Dou, N. Davoine, F. Brünig, and A. Tkatchenko, 10.5281/zenodo.17234182 (2025).

[165] J. F. Allemand, D. Bensimon, R. Lavery, and V. Croquette, Proc. Natl. Acad. Sci. **95**, 14152 (1998).

[166] A. Mitchell, Nature **396**, 524 (1998).

[167] Y. Marcus, Chem. Rev. **88**, 1475 (1988).

[168] F. Bruni, S. Imberti, R. Mancinelli, and M. Ricci, J. Chem. Phys **136**, 064520 (2012).

[169] A. K. Soper, Int. Sch. Res. Notices **2013**, 279463 (2013).

[170] C. J. Sahle, E. de Clermont Gallerande, J. Niskanen, A. Longo, M. Elbers, M. A. Schroer, C. Sternemann, and S. Jahn, Phys. Chem. Chem. Phys. **24**, 16075 (2022).

[171] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, J. Chem. Phys. **148**, 241722 (2018).

[172] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller, J. Chem. Theory Comput. **15**, 448 (2019).

[173] J. Behler, J. Chem. Phys. **134**, 074106 (2011).

[174] M. Herbold and J. Behler, J. Chem. Phys. **156**, 114106 (2022).

[175] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).

[176] A. P. Bartók and G. Csányi, Int. J. Quantum Chem. **115**, 1051 (2015).

[177] N. Lubbers, J. S. Smith, and K. Barros, J. Chem. Phys. **148**, 241715 (2018).

[178] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. E. Roitberg, Nat. Commun. **10**, 2903 (2019).

[179] H. Wang, L. Zhang, J. Han, and E. Weinan, Comput. Phys. Commun. **228**, 178 (2018).

[180] T. Hollebeek, T. S. Ho, and H. Rabitz, J. Chem. Phys. **106**, 7223 (1997).

[181] B. Jiang and H. Guo, J. Chem. Phys. **139**, 054112 (2013).

[182] V. L. Deringer, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, and S. R. Elliott, Nature **589**, 59–64 (2021).

[183] M. Gastegger, J. Behler, and P. Marquetand, Chem. Sci. **8**, 6924 (2017).

[184] N. Raimbault, A. Grisafi, M. Ceriotti, and M. Rossi, New J. Phys. **21**, 105001 (2019).

[185] G. M. Sommers, M. F. Calegari Andrade, L. Zhang, H. Wang, and R. Car, Phys. Chem. Chem. Phys. **22**, 10592 (2020).

[186] M. Meuwly, Chem. Rev. **121**, 10218 (2021).

[187] J. Westermayr and P. Marquetand, Chem. Rev. **121**, 9873 (2020).

[188] P. O. Dral and M. Barbatti, Nat. Rev. Chem. **5**, 388 (2021).

[189] V. Vassilev-Galindo, G. Fonseca, I. Poltavsky, and A. Tkatchenko, J. Chem. Phys. **154**, 094119 (2021).

[190] I. Poltavsky and A. Tkatchenko, J. Phys. Chem. Lett. **12**, 6551 (2021).

[191] F. Faber, A. Lindmaa, O. A. Von Lilienfeld, and R. Armiento, Int. J. Quantum Chem. **115**, 1094 (2015).

[192] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller, and A. Tkatchenko, J. Phys. Chem. Lett. **6**, 2326 (2015).

[193] H. Huo and M. Rupp, Mach. Learn.: Sci. Technol. **3**, 045017 (2022).

[194] W. Pronobis, A. Tkatchenko, and K.-R. Müller, J. Chem. Theory Comput. **14**, 2991 (2018).

[195] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. **114**, 105503 (2015).

[196] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L. M. Ghiringhelli, Phys. Rev. Mat. **2**, 083802 (2018).

[197] J. Nigam, S. Pozdnyakov, and M. Ceriotti, J. Chem. Phys. **153**, 121101 (2020).

[198] J. P. Janet and H. J. Kulik, J. Phys. Chem. A **121**, 8939 (2017).

[199] W. B. How, B. Wang, W. Chu, A. Tkatchenko, and O. V. Prezhdo, J. Phys. Chem. Lett. **12**, 12026 (2021).

[200] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, J. Chem. Phys. **148**, 241730 (2018).

[201] F. Musil, M. Veit, A. Goscinski, G. Fraux, M. J. Willatt, M. Stricker, T. Junge, and M. Ceriotti, J. Chem. Phys. **154**, 114109 (2021).

[202] R. K. Cersonsky, B. A. Helfrecht, E. A. Engel, S. Kliavinek, and M. Ceriotti, Mach. Learn.: Sci. Technol. **2**, 035038 (2021).

[203] J. P. Darby, J. R. Kermode, and G. Csányi, npj Comput. Mater. **8**, 1 (2022).

[204] L. Zhang, H. Wang, M. C. Muniz, A. Z. Panagiotopoulos, R. Car, and W. E, J. Chem. Phys. **156**, 124107 (2022).

[205] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, 2005) Chap. 8.

[206] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).

[207] H. E. Sauceda, V. Vassilev-Galindo, S. Chmiela, K.-R. Müller, and A. Tkatchenko, Nat. Commun. **12**, 442 (2021).

[208] S. Stocker, J. Gasteiger, F. Becker, S. Günnemann, and J. T. Margraf, Mach. Learn.: Sci. Technol. **3**, 045010 (2022).

[209] M. Knol, H. H. Arefi, D. Corken, J. Gardner, F. S. Tautz, R. J. Maurer, and C. Wagner, Sci. Adv. **7**, eabj9751 (2021).

[210] W. Gao and A. Tkatchenko, Phys. Rev. Lett. **114**, 096101 (2015).

[211] D. R. Bowler and T. Miyazaki, Rep. Prog. Phys. **75**, 036503 (2012).

[212] E. Schrödinger, *What is Life? The Physical Aspect of the Living Cell* (Cambridge University Press, 1974).

[213] P. A. M. Dirac, Proc. R. Soc. Lond. A. **123**, 714 (1929).

[214] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman lectures on physics, Vol. I: The new millennium edition: mainly mechanics, radiation, and heat* (Basic books, 2011).

[215] B. Huang, G. F. von Rudorff, and O. A. von Lilienfeld, Science **381**, 170 (2023).

[216] A. D. MacKerell Jr, J. Comput. Chem. **25**, 1584 (2004).

[217] W. F. Van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, P. H. Hünenberger, *et al.*, Angew. Chem. Int. Ed. **45**, 4064 (2006).

[218] T. Schlick and S. Portillo-Ledesma, Nat. Comput. Sci. **1**, 321 (2021).

[219] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, Annu. Rev. Phys. Chem. **71**, 361 (2020).

[220] K. T. Schütt, S. Chmiela, O. A. Von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller, *Machine Learning Meets Quantum Physics* (Springer, 2020).

[221] A. Illarionov, S. Sakipov, L. Pereyaslavets, I. V. Kurnikov, G. Kamath, O. Butin, E. Voronina, I. Ivahnenko, I. Leontyev, G. Nawrocki, *et al.*, J. Am. Chem. Soc. **145**, 23620 (2023).

[222] L. C. Blum and J.-L. Reymond, J. Am. Chem. Soc. **131**, 8732 (2009).

[223] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, Sci. Data **1**, 1 (2014).

[224] J. S. Smith, R. Zubatyuk, B. Nebgen, N. Lubbers, K. Barros, A. E. Roitberg, O. Isayev, and S. Tretiak, Sci. Data **7**, 134 (2020).

[225] C. Isert, K. Atz, J. Jiménez-Luna, and G. Schneider, Sci. Data **9**, 273 (2022).

[226] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, *et al.*, PLOS Comput. Biol. **13**, e1005659 (2017).

[227] S. Schoenholz and E. D. Cubuk, Adv. Neural Inf. Process. Syst. **33**, 11428 (2020).

[228] R. P. Pelaez, G. Simeon, R. Galvelis, A. Mirarchi, P. Eastman, S. Doerr, P. Thölke, T. E. Markland, and G. De Fabritiis, J. Chem. Theory Comput. **20**, 4076 (2024).

[229] J. F. Ziegler and J. P. Biersack, in *Treatise on Heavy-Ion Science: Volume 6: Astrophysics, Chemistry, and Condensed Matter* (Springer, 1985) pp. 93–129.

[230] A. Khabibrakhmanov, D. V. Fedorov, and A. Tkatchenko, J. Chem. Theory Comput. **19**, 7895 (2023).

[231] F. L. Hirshfeld, Theor. Chim. Acta **44**, 129 (1977).

[232] E. R. Johnson and A. D. Becke, J. Chem. Phys. **124**, 174104 (2006).

[233] D. V. Fedorov, M. Sadhukhan, M. Stöhr, and A. Tkatchenko, Phys. Rev. Lett. **121**, 183401 (2018).

[234] A. P. Jones, J. Crain, V. P. Sokhan, T. W. Whitfield, and G. J. Martyna, Phys. Rev. B **87**, 144103 (2013).

[235] J. Rezác, K. E. Riley, and P. Hobza, J. Chem. Theory Comput. **7**, 2427 (2011).

[236] C. Baldauf and M. Rossi, J. Condens. Matter Phys. **27**, 493002 (2015).

[237] F. Schubert, M. Rossi, C. Baldauf, K. Pagel, S. Warnke, G. von Helden, F. Filsinger, P. Kupser, G. Meijer, M. Salwiczek, *et al.*, Phys. Chem. Chem. Phys. **17**, 7373 (2015).

[238] J. Hoja, H.-Y. Ko, M. A. Neumann, R. Car, R. A. DiStasio Jr, and A. Tkatchenko, Sci. Adv. **5**, eaau3338 (2019).

[239] D. Firaha, Y. M. Liu, J. van de Streek, K. Sasikumar, H. Dietrich, J. Helfferich, L. Aerts, D. E. Braun, A. Broo, A. G. DiPasquale, *et al.*, Nature **623**, 324 (2023).

[240] A. Charkin-Gorbulin, A. Kokorin, H. E. Sauceda, S. Chmiela, C. Quarti, D. Beljonne, A. Tkatchenko, and I. Poltavsky, Mach. Learn.: Sci. Technol. **6**, 035005 (2025).

[241] C. G. Staacke, S. Wengert, C. Kunkel, G. Csányi, K. Reuter, and J. T. Margraf, Mach. Learn.: Sci. Technol. **3**, 015032 (2022).

[242] B. K. Rai, V. Sresht, Q. Yang, R. Unwalla, M. Tu, A. M. Mathiowetz, and G. A. Bakken, J. Chem. Inf. Model. **62**, 785 (2022).

[243] Y. Yang, K. U. Lao, D. M. Wilkins, A. Grisafi, M. Ceriotti, and R. A. DiStasio Jr, Sci. Data **6**, 152 (2019).

[244] C. Villot and K. U. Lao, J. Chem. Phys. **160**, 184103 (2024).

[245] S. C. Lovell, I. W. Davis, W. B. Arendall III, P. I. De Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson, Proteins Struct. Funct. Bioinform. **50**, 437 (2003).

[246] D. P. Kovács, J. H. Moore, N. J. Browning, I. Batatia, J. T. Horton, Y. Pu, V. Kapil, W. C. Witt, I.-B. Magdău, D. J. Cole, *et al.*, J. Am. Chem. Soc. **147**, 17598 (2025).

[247] G. L. Millhauser, C. J. Stenland, P. Hanson, K. A. Bolin, and F. J. van de Ven, J. Mol. Biol. **267**, 963 (1997).

[248] K. A. Bolin and G. L. Millhauser, Acc. Chem. Res. **32**, 1027 (1999).

[249] M. Kohtani, T. C. Jones, J. E. Schneider, and M. F. Jarrold, J. Am. Chem. Soc. **126**, 7420 (2004).

[250] R. Mills, J. Phys. Chem. **77**, 685 (1973).

[251] H.-C. Ahn, N. Juranić, S. Macura, and J. L. Markley, J. Am. Chem. Soc. **128**, 4398 (2006).

[252] L. McInnes, J. Healy, and J. Melville, arXiv preprint arXiv:1802.03426 (2018).

[253] T. M. Ferreira, F. Coreta-Gomes, O. S. Ollila, M. J. Moreno, W. L. Vaz, and D. Topgaard, Phys. Chem. Chem. Phys. **15**, 1976 (2013).

[254] D. M. Anstine, R. Zubatyuk, and O. Isayev, Chem. Sci. **16**, 10228 (2025).

[255] N. Kučerka, M.-P. Nieh, and J. Katsaras, Biochim. Biophys. Acta **1808**, 2761 (2011).

[256] Y. Yu, A. Kramer, R. M. Venable, B. R. Brooks, J. B. Klauda, and R. W. Pastor, J. Chem. Theory Comput. **17**, 1581 (2021).

[257] O. Guvench, S. S. Mallajosyula, E. P. Raman, E. Hatcher, K. Vanommeslaeghe, T. J. Foster, F. W. Jamison, and A. D. MacKerell Jr, J. Chem. Theory Comput. **7**, 3162 (2011).

[258] Y. Wang, K. Takaba, M. S. Chen, M. Wieder, Y. Xu, T. Zhu, J. Z. Zhang, A. Nagle, K. Yu, X. Wang, *et al.*, Appl. Phys. Rev. **12**, 021304 (2025).

[259] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, J. Comput. Chem. **25**, 1157 (2004).

[260] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, *et al.*, J. Comput. Chem. **31**, 671 (2010).

[261] X. Zhu, P. E. Lopes, and A. D. MacKerell Jr, Wiley Interdiscip. Rev. Comput. Mol. Sci. **2**, 167 (2012).

[262] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, J. Comput. Chem. **25**, 1656 (2004).

[263] N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, and W. F. Van Gunsteren, Eur. Biophys. J. **40**, 843 (2011).

[264] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio Jr, *et al.*, J. Phys. Chem. B **114**, 2549 (2010).

[265] R. Zubatyuk, J. S. Smith, J. Leszczynski, and O. Isayev, Sci. Adv. **5**, eaav6490 (2019).

[266] T. Plé, L. Lagardère, and J.-P. Piquemal, Chem. Sci. **14**, 12554 (2023).

[267] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder, Nat. Mach. Intell. **5**, 1031 (2023).

[268] P. Loche, K. K. Huguenin-Dumittan, M. Honarmand, Q. Xu, E. Rumiantsev, W. B. How, M. F. Langer, and M. Ceriotti, J. Chem. Phys. **162**, 142501 (2025).

[269] E. Slootman, I. Poltavsky, R. Shinde, J. Cocomello, S. Moroni, A. Tkatchenko, and C. Filippi, J. Chem. Theory Comput. **20**, 6020 (2024).

[270] Y. Park, J. Kim, S. Hwang, and S. Han, J. Chem. Theory Comput. **20**, 4857 (2024).

[271] F. Musil, I. Zaporozhets, F. Noé, C. Clementi, and V. Kapil, J. Chem. Phys. **157**, 181102 (2022).

[272] J. T. Barron, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (2019) pp. 4331–4339.

[273] I. Guyon and A. Elisseeff, J. Mach. Learn. Res. **3**, 1157 (2003).

[274] Y. Saeys, I. Inza, and P. Larranaga, Bioinform. **23**, 2507 (2007).

[275] V. Kapil, M. Rossi, O. Marsalek, R. Petraglia, Y. Litman, T. Spura, B. Cheng, A. Cuzzocrea, R. H. Meißner, D. M. Wilkins, *et al.*, Comp. Phys. Commun. **236**, 214 (2019).

[276] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, Comp. Phys. Commun. **180**, 2175 (2009).

[277] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[278] A. Ambrosetti, A. M. Reilly, R. A. Distasio, and A. Tkatchenko, J. Chem. Phys. **140**, 18A508 (2014).

[279] G. Csányi, S. Winfield, J. Kermode, M. Payne, A. Comisso, A. De Vita, and N. Bernstein, IoP Comput. Phys. Newsletter **1**, 1 (2007).

[280] J. R. Kermode, J. Phys. Condens. Matter **32**, 305901 (2020).

[281] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, *et al.*, J. Phys. Condens. Matter **29**, 273002 (2017).

[282] C. J. Fennell and J. D. Gezelter, J. Chem. Phys. **124**, 234104 (2006).

[283] D. Wolf, P. Keblinski, S. Phillpot, and J. Eggebrecht, J. Chem. Phys. **110**, 8254 (1999).

[284] S. J. Reddi, S. Kale, and S. Kumar, arXiv preprint arXiv:1904.09237 (2019).

[285] J. L. Ba, J. R. Kiros, and G. E. Hinton, arXiv preprint arXiv:1607.06450 (2016).

[286] E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, and P. Gumbsch, Phys. Rev. Lett. **97**, 170201 (2006).

[287] D. Frishman and P. Argos, Proteins Struct. Funct. Bioinform. **23**, 566 (1995).

[288] G. Pranami and M. H. Lamm, J. Chem. Theory Comput. **11**, 4586 (2015).

[289] E. J. Maginn, R. A. Messerly, D. J. Carlson, D. R. Roe, and J. R. Elliot, Living J. Comp. Mol. Sci. **1**, 6324 (2019).

[290] D. Bang, V. Tereshko, A. A. Kossiakoff, and S. B. H. Kent, Mol. BioSyst. **5**, 750 (2009).

[291] K. T. Schütt, S. S. P. Hessmann, N. W. A. Gebauer, J. Lederer, and M. Gastegger, J. Chem. Phys. **158**, 144801 (2023).

[292] A. Mølgaard and S. Larsen, Acta Crystallogr. D **58**, 111 (2002).

[293] D. R. Roe and T. E. Cheatham III, J. Chem. Theory Comput. **9**, 3084 (2013)