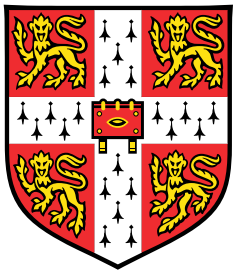
Machine Learning Force Fields for Molecular Chemistry



Dávid Péter Kovács

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

St Catharine's College

January 2024

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Dávid Péter Kovács
January 2024

Abstract

The first principles computational modelling of molecular systems is a long-standing pursuit in the scientific community. It has traditionally been tackled by developing approximate solutions to quantum mechanics. Simulations using these electronic structure based methods can be highly accurate but are limited to small system sizes or short time scales. The traditional alternative is force fields that enable fast and accurate simulations by bypassing the treatment of the electrons and describing the system solely in terms of the atomic positions. The emergence of machine learning tools has opened up the opportunity for the development of high accuracy force fields trained directly to reproduce the results of electronic structure calculations.

This thesis presents new developments that lead to improved machine learning force fields for molecular chemistry. Firstly, a set of linearly complete basis functions, called ACE, is demonstrated to yield high accuracy custom made force fields for small molecules. By recognising the symmetric tensor structure of these basis functions, the framework is extended to enable the simultaneous description of a large number of chemical elements.

Next, multi-ACE is proposed, which provides a unifying theory of most classical and machine learning force fields. Using the design space set out in this theory, a new method, called MACE is created. MACE is shown to provide simple, robust, accurate, and efficient force fields for a wide range of molecular systems. Finally, `MACE-OFF23` a new transferable force field for organic molecules is proposed and demonstrated to be capable of accurately describing not only molecules in vacuum but also in the condensed phase.

Acknowledgements

First, I would like to express my gratitude to my supervisor Prof. Gábor Csányi for his encouragement, advice, and guidance throughout my PhD. In particular, I would like to thank him for allowing me to pursue my interests freely and to allow me to choose topics that were of interest to me.

I would also like to thank the members of the Gabor group for providing an intellectually stimulating and fun environment to work in. In particular, I would like to thank Ilyes Batatia, with whom I was fortunate to work on many of my projects. I also thank Cas van der Oord for introducing me to ACE in the initial days of my PhD. I am also grateful for many fruitful discussions around the coffee machine with William Baldwin and James Darby. I would also like to thank Harry Moore for helping me with biomolecular simulations.

Furthermore, I would like to express my gratitude to the many external collaborators I had, including Chrsitoph Ortner for the guidance in the mathematical parts of my PhD, Daniel Cole for discussing with me throughout my PhD and inspiring me to work on difficult problems relevant to drug discovery, Joshua Horton for helping with the development and testing of new organic force fields, Venkat Kapil for introducing me to quantum dynamics and using ML force fields for spectroscopy, Gus Hart for providing valuable data and ideas for testing the TrACE architecture enabling force fields for many chemical elements, Gregor Simm for introducing me to equivariant neural networks, Angelos Michaelides for insightful discussions about my research, and Graeme Robb who helped me put my work into perspective from a pharmaceutical industry point of view.

I would also like to acknowledge the support of the EPSRC and Astra Zeneca for providing me with a studentship that allowed me to pursue this research. I would also like to thank the Cambridge Service for Data Driven Discovery (CSD3) for providing me with computing resources for the PhD and St Catharine's College for providing a welcoming community throughout my time in Cambridge.

Finally, I would like to thank my girlfriend Eszter for not only encouraging and motivating me to tackle difficult problems and supporting me throughout the PhD, but also for participating directly in one of the PhD projects. I am also grateful to my family for their support and encouragement to pursue my passion for research.

Table of contents

Chapter 1

Introduction

In the history of humankind, knowledge has always been advanced by the combined contribution of theoretical and experimental scientists. The scientific method requires the postulation of hypotheses that can be experimentally verified or disproved. This way of working has served as the basic foundation for rigorous academic work. As far as studying the world of atoms and molecules is concerned, quantum mechanics is a theory that is, in principle, able to predict everything with perfect accuracy. The theory was formulated in the early 20th century and the main principles have not changed over the past 100 years. Since then, the main goal of theoretical scientists has been to develop rigorous approximations of quantum mechanics that enable the calculation of interesting predictions.

Over the past decades, there has been a technological revolution that has exponentially increased the available computational resources. The availability of relatively inexpensive compute transforms what is possible to predict by theoretical science. In fact, today a third arm of science is often mentioned alongside the theoretical and experimental approach, called computational science. Computational scientists use tools derived from theories of physics and chemistry to carry out experiments on a computer simulating the conditions of real-world experiments. Computational approaches allow for very precise measurements of observables in simulations without disturbing the system compared to the real-world experiments. This can lead to accurate predictions about the behaviour of chemical systems and also to insights into the behaviour of chemicals and materials at the atomic scale.

The focus of this thesis are a set of new methods that have been developed to enable computational experiments simultaneously at a scale and accuracy that have not been possible before. In particular, the methods belong to the field of atomistic modelling of chemistry. Using the most advanced quantum mechanical methods, it is possible to accurately simulate at most tens or hundreds of atoms and for at most hundreds of picoseconds (1 picosecond = 10^{-12} second). The source of the high computational cost in these simulations is the separate

treatment of the electrons and the nuclei. In contrast, the tools discussed in this thesis build upon quantum mechanics, but themselves consider the atoms as the indivisible units of the computational experiments. This means that they bypass electrons altogether, describing the system simply as a collection of atoms. This simplification enables the simulation of 10,000-s of atoms for time scales of up to 100-s of nanoseconds (1 nanosecond = 10^{-9} second). Remarkably, this is achieved without a significant loss of accuracy compared to the quantum mechanical methods.

These developments are enabled by using machine learning methods to infer the effective interactions of atoms from high-fidelity, expensive quantum mechanical data. From the dataset of quantum mechanical calculation results, it is possible to construct force fields, also known as interatomic potentials, that describe the strength of interactions between individual atoms at a fraction of the cost of the original calculations.

There are two key contributions in this thesis. Firstly, with my co-workers, we have developed a new theoretical framework that lets us better understand the machine learning methods used for creating atomistic models. Part of the theory also enables the simulation of systems with many chemical elements, which was previously a notoriously difficult task. Secondly, we have developed a new method for creating machine learning force fields called MACE. The examples in this thesis present evidence that this new method enables the routine parameterisation of new force fields with little training data and high accuracy. The software implementation of MACE is available as an open source code that can be used by computational scientists to run their own experiments of interesting physical or chemical systems.

1.1 Outline and Key References

This thesis was written based on a number of publications. In this section, the outline of the thesis is given, with references to the papers upon which each of the chapters and sections is based. The thesis contains figures and text taken directly from these papers that are the work of the author of this thesis.

Chapter 2 introduces the necessary background to understand the thesis. It builds on a review manuscript of machine learning force fields that is under preparation. It also introduces some of the most important datasets that were used in this thesis.

Chapter 3 is based on the paper Kovács, Dávid Péter, et al. “Linear atomic cluster expansion force fields for organic molecules: beyond rmse.” Journal of chemical theory and computation 17.12 (2021): 7696-7711. It introduces the linear ACE force fields for the simulation of small molecular systems. Furthermore, in Section 3.3 the tensor-reduced

version of ACE is introduced, which enables the simulation of systems with a large number of chemical elements using machine learning force fields. This section is based on the paper Darby, James P., Kovács, Dávid Péter, et al. “Tensor-reduced atomic density representations.” Physical Review Letters 131.2 (2023): 028001.

Chapter 4 is based on the paper Batatia, Ilyes, et al. “The design space of E (3)-equivariant atom-centered interatomic potentials.” arXiv preprint arXiv:2205.06643 (2022). It extends Atomic Cluster Expansion and TrACE to equivariant outputs and describes the Multi-ACE design space of machine learning force fields.

Chapter 5 introduces the new MACE machine learning force field architecture based on the paper Batatia, Ilyes, et al. "MACE: Higher order equivariant message passing neural networks for fast and accurate force fields." Advances in Neural Information Processing Systems 35 (2022): 11423-11436. This section is also based on the follow-up paper Kovács, Dávid Péter, et al. “Evaluation of the MACE force field architecture: From medicinal chemistry to materials science.” The Journal of Chemical Physics 159.4 (2023): 044118. Finally, in Section 5.3 a pre-trained transferable force field for organic chemistry is introduced based on the paper Kovács, Dávid Péter, Moore, J. Harry, et al. “MACE-OFF23: Transferable Machine Learning Force Fields for Organic Molecules” arXiv preprint arXiv:2312.15211 (2023).

Chapter 6 contains conclusions and outlooks, in particular, about the prospect of pre-trained foundation models, like the one introduced in Ref [15].

Chapter 2

Background

In this chapter, the most important background and historical results relevant for this thesis are briefly summarised. First, the key terms from quantum mechanics and electronic structure are introduced. These serve as the foundation of the first principles or *ab initio* modelling of chemical systems. Next, the central object of computational chemistry, the Potential Energy Surface (PES), is introduced, and its significance is discussed, including the Born-Oppenheimer approximation. It is followed by a discussion of a number of important classical and machine learning approaches for parameterising the PES. Finally, a number of benchmark datasets are described that are used throughout this thesis.

2.1 Electronic Structure Methods - The Ground Truth

Quantum mechanics provides a framework in which most of materials science and chemistry can be understood. It postulates that all physical information about a system is contained in its wavefunction. The system is defined by specifying its Hamiltonian $\hat{H}$ representing the total energy operator. It is the sum of two terms that account for the kinetic energy and the potential energy of the system. The potential energy term describes the interactions of the particles (typically the electrons and nuclei in chemistry) and can also contain the contribution coming from external fields. One way of formulating quantum mechanics is via the Schrödinger-equation which the full wavefunction Ψ must satisfy.

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{r},t) = \left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r},t)\right]\Psi(\mathbf{r},t) \tag{2.1}$$

When it comes to simulating chemical systems, the Hamiltonian typically corresponds to a molecular structure. In this case, instantaneously it does not depend on time. This simplifies the equation to the time-independent Schrödinger-equation.

$$\hat{H}\Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \tag{2.2}$$

where the kinetic and potential energy terms are represented by $\hat{H}$. For a chemical system with fixed positions of the nuclei and using atomic units, the time-independent Hamiltonian can be written as

$$\hat{H} = -\frac{1}{2}\sum_i \nabla_i^2 - \sum_{iA} \frac{Z_A}{|\mathbf{r}_i - \mathbf{R}_A|} + \sum_{i>j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} + \sum_{A>B} \frac{Z_A Z_B}{|\mathbf{R}_A - \mathbf{R}_B|} \tag{2.3}$$

where i, j denote electrons with positions $\mathbf{r}_i, \mathbf{r}_j$ and A, B denote nuclei with charges Z_A, Z_B at positions $\mathbf{R}_A, \mathbf{R}_B$.

The potential energy, E from Equation (2.2), can only be computed approximately for all but the simplest systems. There are a variety of computational techniques for solving this equation, with the most accurate ones requiring the largest computing power. Quantum chemistry methods typically have the highest accuracy. These methods attempt to tackle the description of the fully correlated wavefunction as presented in Equation (2.2). Examples of quantum chemistry methods include CASSCF [177], FCI [91], and CCSD(T) [10, 164], which are capable of computing the potential energy of systems with tens of electrons. Larger systems are out of reach due to the unfavourable scaling of the computational cost with the number of electrons. For example, the cost of computing the potential energy using the CCSD(T) method scales as N^7 with the number of electrons N. Some Monte Carlo approximations of the wavefunction, such as FCIQMC [25] or diffusion Monte Carlo [237], can be scaled up to a few hundred electrons before they become untenable [129]. Recently, variational Monte Carlo methods were also combined with deep learning in new neural wavefunctions such as FermiNet [158] and PauliNet [94], which are new alternatives to more established electronic structure methods [95].

Density Functional Theory Density Functional Theory (DFT) is probably the most popular method for computing the ground state energy of chemical systems [96]. DFT bypasses the wavefunction and computes the energy as a functional of the electron density. According to the Hohenberg-Kohn theory [96] the energy can be written exactly as

$$E[\rho(\mathbf{r})] = T_e[\rho(\mathbf{r})] + V_{ee}[\rho(\mathbf{r})] + V_{ne}[\rho(\mathbf{r})] \tag{2.4}$$

where $\rho(\mathbf{r})$ denotes the electron density, T_e is the kinetic energy functional, and V_{ee} and V_{ne} represent the electron-electron and electron-nuclei interactions, respectively. From the theorem it follows that the electron density that minimises the energy corresponds to the

ground state of the system. In practise, this implies that the ground state can be determined by minimising the energy as a function of the electron density.

The most popular implementation of DFT is based on the Kohn-Sham method [114]. It assumes that the system is made up of noninteracting electrons in an effective potential $V_{eff}(\mathbf{r})$. The total wavefunction is written as a product of atomic orbital wavefunctions for each of the electrons denoted by $\psi_i(\mathbf{r})$. This reduces Equation (2.2) into an eigenvalue problem.

$$\left[\frac{\hbar^2}{2m_i}\nabla_i^2 + V_{eff}(\mathbf{r})\right]\psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r}) \tag{2.5}$$

where ε_i is the orbital energy. $V_{eff}(\mathbf{r})$ can be written as

$$V_{eff}(\mathbf{r}) = V_{ne}[\rho(\mathbf{r})] + V_{ee}[\rho(\mathbf{r})] + V_{xc}[\rho(\mathbf{r})] \tag{2.6}$$

where V_{ne} is the Coulomb term between the electrons and the nuclei, V_{ee} is the Coulomb repulsion between the electrons, also known as the Hartree term and V_{xc} is the exchange-correlation functional correcting for the error made by the noninteracting electron assumption.

This formulation reduces the computational complexity compared to the wavefunction-based quantum chemistry methods but comes at a price: The mathematical form of the exchange correlation functional is unknown; therefore, one has to rely on different approximate functionals which are not systematically improvable. There are several DFT exchange-correlation functionals whose parameters are tuned to match the electronic properties of small systems calculated exactly or at a higher fidelity level of theory or even to reproduce experimentally measured properties, such as ionisation potentials [81].

In its modern implementations, the Kohn-Sham version of density functional theory has a scaling of N^3 with the number of electrons, N, making the calculation of the energies of systems with up to 1-200 atoms relatively easily possible [115, 151]. Linear scaling implementations also exist, though these typically have a much larger prefactor to their computational cost [124].

Moving beyond DFT, more approximate semi-empirical quantum mechanics methods such as tight-binding approximations have also been developed [9, 101]. These methods still try to solve the Schrödinger-equation, but apply large simplifications to the Hamiltonian to enable the treatment of systems with a few 1,000 atoms. These simplifications typically come at the expense of decreased accuracy, making these methods useful primarily in high-throughput screening tasks where quantitative accuracy is not necessary.

2.2 The Potential Energy Surface

An important observation in theoretical chemistry is that the motion of the nuclei can be treated independently from the motion of the electrons. This means that chemical systems can be accurately simulated by decoupling the electronic and nuclear degrees of freedom, which is known as the Born-Oppenheimer approximation [28]. This leads to the definition of the potential energy surface (PES), which describes the mapping from nuclear positions to the ground state electronic potential energy of the system, $E = E(\mathbf{R})$. The intuition behind the Born-Oppenheimer approximation and the potential energy surface is that the electrons are much lighter than the nuclei; therefore, given the same amount of momentum, the electrons are moving much faster. This means that they intantenously rearrange to occupy their lowest energy state as the nuclei are moving around much slower. The approximation breaks down in cases where two potential energy surfaces are close in energy. In this thesis, only systems where that is not the case are covered; therefore, it is sufficient to parameterise the system only in its ground electronic state.

Having access to the PES, $E(\mathbf{R})$, it is possible to use it to sample the chemically relevant states of atomistic systems. There are many different sampling schemes that can be used depending on the specifics of the system and the properties of interest. In principle, since the PES is obtained via a rigorous approximation of quantum mechanics, most properties of the system can be determined from it.

The PES can be used to explore the stable states that a chemical system can occupy. This can be done in a number of different ways, for example, by employing energy landscape methods such as basin hopping [217, 218] or random structure searching [159, 160]. These methods enable the discovery of stable configurations of the system, such as the stable conformations of a protein [219] or the possible stable structures of materials [3, 191]. The PES can also be used to explore reactive systems and identify transition states and determine the heights of energy barriers corresponding to different reaction mechanisms [182].

Chemical reactions can also be simulated in real time by propagating the nuclear wave-function along the reaction coordinates. In fact, the very first potential energy surfaces were developed to analyse the quantum dynamics of small reactive systems [18, 30, 103]. The same methods can also be used to compute other observables like molecular spectra [111].

Finally, the largest use case of PESs is to simulate a chemical system using classical molecular dynamics (MD) simulations. In these simulations, the nuclei are treated as classical Newtonian point masses and the classical equations of motion are applied to them to propagate their positions [70]. The instantaneous forces acting on the nuclei are computed as the negative gradients of the PES

$$\mathbf{F} = -\nabla E(\mathbf{R}) \tag{2.7}$$

where $\mathbf{F}$ is the force vector, which has dimensions $N_{\text{atoms}} \times 3$. Without using a thermostat such a dynamics keeps the total energy of the system constant. Molecular dynamics simulations can also be coupled to thermostats and barostats to sample states of the system from constant temperature or pressure ensembles. This enables the calulation of thermodynamic observables by averaging their values over a sufficiently large set of samples of the system. Formally, the expectation value of an observable X can be computed by evaluating

$$\langle X \rangle = \frac{\sum_i \exp(-E_i/k_B T) \langle i | X | i \rangle}{\sum_i \exp(-E_i/k_B T)} \tag{2.8}$$

where the sum runs over all possible (quantum) states of the system i and $\langle X \rangle$ denotes the thermal average of the observable X and k_B is the Boltzmann constant [70]. The above expectation value gives what a measurement of X would give in real life if the system was in thermal equilibrium. In practice $\langle X \rangle$ is not computed by summing over all possible micro states, which would not be possible for more than a few degrees of freedom, but is rather estimated by taking a large enough set of samples from the above distribution. This can be achieved using thermostated molecular dynamics simulations or using Monte Carlo methods [70]. MD simulations can be useful either to make predictions from quantum mechanics on unknown systems or to rationalise known experimental results. Examples of simulations include the computation of the physical properties of materials, such as phase diagrams [8, 112, 175, 178], densities [136], or diffusivities [7, 17]. It is also possible to simulate chemical processes such as the binding of drug molecules to protein targets [44, 98], the reconstruction of surfaces [208], or the reaction mechanisms of heterogeneous catalytic systems [231].

In the next section, force fields, which are an essential tool enabling many of the aforementioned applications, will be introduced. Force fields are also known by the name interatomic potentials.

2.3 Atomistic Modelling Using Force Fields

2.3.1 What Are Force Fields?

Force fields (FFs), are simplified representations of the PES. A force field is a set of mathematical functions that describe the energy of a system as a function of the positions of atoms and their chemical elements. Typically, they parameterise the PES directly, rather than as

the solution of a complicated equation, as is the case in electronic structure methods. This approach allows for the calculation of forces acting on atoms in a computationally efficient way, enabling the simulation of larger systems where a full quantum mechanical treatment of the PES would be computationally prohibitive. By approximating the PES, force fields enable the study of molecular behaviour over longer time scales and larger spatial dimensions than would otherwise be feasible, thus playing a crucial role in extending our understanding of molecular chemistry beyond the quantum scale.

Empirical force fields have been used for atomistic simulations for several decades [130]. These models typically assumed a very simple functional form with the parameters fitted both to quantum mechanical and real experimental data [176]. With recent advances in compute hardware and machine learning over the past decade, machine learning potentials emerged that also directly map atomic positions to the potential energy. The key difference from empirical force fields lies in the functional form and the parameterisation relying on quantum mechanical data only. These models are capable of approximating the quantum mechanical PES to much greater accuracy, with the newest methods even retaining the accuracy far from the training set used for the parameterisation [16].

In this section, the most important terminology and requirements that guide the design of new force field functional forms are introduced.

2.3.2 Symmetries of Force Fields

There is a well-defined set of symmetries that a force field energy expression should obey, which are illustrated in Figure 2.1. The potential energy is invariant with respect to translations of the system, which is trivially incorporated into the functional form by either expressing the energy as a function of internal coordinates or by using an atom-centred local coordinate system.

Next, the potential energy should be invariant with respect to the rigid rotations of the system, also commonly referred to as the actions of the $SO(3)$ group.

The final symmetry of the PES is the invariance with respect to permutation of the like atoms, or the actions of the permutation group S_n. This means that the energy should be exactly unchanged if the atoms of the same chemical element are swapped with each other in the force field energy expression.

Different models address these symmetry constraints in different ways. This has been reviewed in Ref [147] and is further elaborated in Section 2.4 for a selection of classical and machine learning force fields.

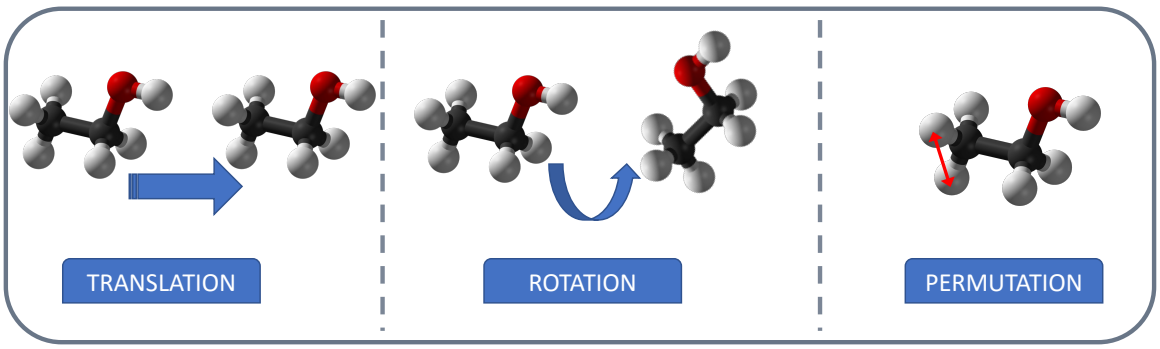


Fig. 2.1 **The main symmetries of force fields** The figure illustrates the symmetries a potential energy function should obey. The energy is invariant with respect to translations, rotations and permutations of the like atoms.

2.3.3 Body Order

The body order expansion is a systematic method for approximating high-dimensional functions in terms of lower-dimensional ones. The general form of the expansion for an N dimensional function Ψ is

$$\Psi(x_1,...,x_N) = \Psi^{(0)} + \sum_{i=1}^{N} \Psi^{(1)}(x_i) + \sum_{i,j} \Psi^{(2)}(x_i,x_j) + \cdot + \Psi^{(N)}(x_1,...,x_N). \tag{2.9}$$

where $\Psi^{(i)}$ denotes the i-body term, depending on the simultaneous values of i coordinates. Such an approximation is useful if it can be truncated at a maximum body order much smaller than the total number of variables N. The quantum mechanical PES is an N-body function, which means that solving the full Schrödinger equation of an N-particle system results in an energy expression that simultaneously depends on the position of all N nuclei.

Writing the PES in terms of low body order terms and truncating early (around 3-5) is an approximation that underpins directly or indirectly many of the force field models developed in the past. It builds on the intuition that the complex potential energy surface can be described in terms of lower-dimensional transferable terms that describe the interactions of pairs of atoms (distances), triplets of atoms (angles), etc. This approximation has theoretical foundations in quantum mechanical tight-binding theory [194] and is generally found to work well when describing the ground state potential energy surface.

To compare the different force field models in terms of their body order, it is useful to have a precise definition which captures the expressivity of the functional form. The body order of an atomic representation $\phi(r_1,...,r_N)$ of N atoms in interaction is the largest

integer $\mathscr{T}$ such that the descriptor can be be made complete on the space of $\mathscr{T}$ atoms. Here completeness loosely refers to being able to approximate any smooth function of $\mathscr{T}$ atoms to arbitrary accuracy, as explained in detail in Ref [63]. Based on this definition, the body order of a force field model expresses the maximum number of atoms N for which it is always possible to find parameters approximating a smooth energy function of the N atoms to arbitrary accuracy.

2.3.4 Topology - transferability and reactivity

A force field parameterisation is called transferable if the same functional form is able to describe different chemical systems without changing the parameters.

A nontransferable force field, such as sGDML [38] or PIP-s force fields for molecules [30], can only describe the one molecule for which they were parameterised. The input to these models is the entire structure either with a canonical ordering of the atoms or in a permutationally symmetrised form. These models are inherently reactive, meaning that they can be parameterised to be able to describe arbitrary (potentially reactive) rearrangements of the atoms of that one system.

On the other hand, most classical empirical force fields and many of the machine learning force fields are transferable. This means that they are made up of functions that can simultaneously parameterise the interatomic interactions of many chemical systems. It is also possible to fit them to energies and forces of small systems and evaluate them on larger ones.

There are a number of different ways in which a chemical system can be specified as the input to a force field model. Classical empirical force fields typically require the topology of the system in addition to the coordinates of the atoms. The topology comprises the chemical connectivity of the atoms together with the assignment of so-called atom types. These are determined by using a set of rules that encode all information about an atom's local environment. They are then used to retrieve the appropriate parameters for the force field functional form (bond, angle, torsion parameters) for the simulated system. The topology and the force field parameters together define the PES of the system and are fixed at the beginning of a simulation. This makes these models non-reactive. By using parameter sets that depend on the local connectivity of the system, the force field parameterisations are transferable to a variety of systems of different sizes and different chemistries as long as the atom-types can be assigned accurately.

Most local machine learning force fields are not only transferable, but also reactive. They usually decompose the total energy into atomic site contributions which depend on the local environment only. Since neighbour atoms are allowed to enter or leave the local environment, which is determined on-the-fly, the models are reactive. The transferability of these models

is achieved by only using local terms, as discussed in detail in the next subsection (2.3.5). Transferability in practise means that the model remains applicable for any system that is locally similar to the training set used for parameterisation [13, 20].

2.3.5 Local or Global Terms

As introduced in Section 2.3.1 force fields provide a simplified representation of the PES compared to computing it by solving the Schrödinger equation for each new arrangement of the nuclei. Such a simplified functional form can have many different building blocks. Locality is one of the key concepts that can be used to classify the terms in a force field energy expression.

Early work on parameterizing potential energy surfaces relied on a fully global description. This means that all interatomic interactions in the system were taken into account [30]. Such a functional form has an evaluation cost that increases combinatorially as the body order of the terms is increased. There are $\binom{N}{2}$ pair terms, $\binom{N}{3}$ 3-body terms, etc. This makes the global methods feasible for simulations of systems with up to dozens of atoms.

Most transferable force field models build the energy expression from local terms. For example, many of the most successful machine learning force fields decompose the total energy of an N atom system into the sum of atomic site energies.

$$E_{\text{tot}} = \sum_{i}^{N} E_i \tag{2.10}$$

where E_i is the energy contribution of atom i and the sum is taken over all atoms in the system. An energy expression is called local if E_i depends only on a subset of all atoms j that are in the local environment of i [13, 20]. To be precise, the local environment of an atom i is made up of all atoms j for which $||\mathbf{r}_i - \mathbf{r}_j|| \leq r_{\text{cut}}$, where $r_{\text{cut}} < \infty$ is the predefined cutoff distance. Therefore, a term in a force field expression E_i is called local if for all set of atoms $\{\mathbf{r}_j\}$ the value of E_i is constant if any of the atoms j is not in the local neighbourhood of atom i.

If a potential energy surface is parameterised using local terms only, it is called a local force field and is valid under the locality assumption. This assumption can be formally justified by the nearsightedness of quantum mechanics [43, 205] and can be tested emprically using locality tests [56]. It has also been shown recently that for condensed phase systems, a short-range (local) model can give an accurate description of bulk properties even if the true underlying function is long-range (non-local) [45]. The use of local terms was originally introduced with the bonded terms (bond, angle, dihedral) of classical empirical (Lifson-type) force fields as early as the 1970s [89, 220]. Building force fields from local terms has a

number of favourable computational properties. Their evaluation is scalable to large systems because the computational cost scales linearly with the system size. Furthermore, under the assumption of the system being well represented by local terms, the functional form also implies a constant per-atom error compared to the true quantum mechanical PES. The evaluation is also trivially parallelisable. Finally, local terms can be transferable across many different systems that are made up of similar local environments [144].

For small systems, building a model from global terms can have advantages. Global terms parameterise the interactions between all atoms of the system regardless of their distance, ensuring that no interaction is excluded from the model. For example, when it comes to parameterising the PES of individual small molecules or a small number of atoms for a reactive system, the traditional approach is to use global models. The prime example of such potential energy surfaces are the Permutation Invariant Polynomials (PIPs) that have been used to fit the PES of several small molecules with very high accuracy [150]. The disadvantages of global models are that they have to be fitted uniquely for each system studied and that they are typically not scalable to more than 10s of atoms. Recent work has demonstrated that it is possible to scale up these models to up to a few hundred atoms by filtering out some of the features, thereby preventing the explosion in model size [40, 165], but it is still far from the large-scale (up to millions of atoms) simulations possible with local models.

There is a third class of terms in force field expressions that are global, but rather than being general functions, they have a simple physically motivated functional form. These are usually pair potentials describing Van der Waals dispersion and electrostatic interactions. For these global long-range terms, there are several efficient algorithms such as Particle Mesh Ewald [48] making the evaluation of these terms in large-scale simulations with hundreds of thousands of atoms still feasible.

2.4 Overview of Force Field Methods for Molecular Chemistry

In this section, a brief overview of different force field functional forms is given. First classical empirical force fields are described followed by an introduction to machine learning force fields. The strength and limitation of these methods are also briefly discussed.

2.4.1 Classical Empirical Force Fields

Functional form and parametrisation

The functional form of the major empirical force fields dates back to 1969 [130]. These models were developed for running large-scale simulations specifically targeting biochemical applications, such as the simulation of proteins, DNA and RNA. The energy expression of the so-called Lifson-type force fields, which includes AMBER [180, 226], CHARMM [33, 215], GROMOS [188] and OPLS [107, 108] can be written as

$$\begin{aligned} E(\mathbf{r}) = E_{\text{bond}}(\mathbf{r}) + E_{\text{angle}}(\mathbf{r}) + \\ + E_{\text{torsion}}(\mathbf{r}) + E_{\text{improper torsion}}(\mathbf{r}) + E_{\text{electrostatics}}(\mathbf{r}) + E_{\text{vdW}}(\mathbf{r}) \end{aligned} \tag{2.11}$$

where $E(\mathbf{r})$ is the potential energy of the system with the position of the atoms described by $\mathbf{r}$ [50, 88, 176]. The terms can be divided into bonded and non-bonded contributions depending on whether they express contributions from atoms that are covalently bonded or not. The bonding graph of the system is predefined in the topology and is kept fixed throughout the simulations. The bonding interactions usually have a harmonic functional form; for example, a typical bond term between atoms i and j is of the form

$$E_{\text{bond}}(\mathbf{r}_i, \mathbf{r}_j) = \frac{1}{2} K_{ij} [\mathbf{r}_i - \mathbf{r}_j]^2 \tag{2.12}$$

where K_{ij} is the harmonic force constant of the bond between atom i and j. Its value is usually stored in a look-up table and is determined by the atom types of i and j. Similar simple functional forms apply for the angle and torsional terms. These terms represent a body order expansion, bonded terms being 2-body, angles 3-body, and torsions 4-body. It is important to note that the terms in this functional form do not represent general N-body terms, meaning that they can only approximate a small subset of all possible N-body functions. This restricted functional form is the key limitation of the accuracy of Lifson-type force fields.

The parameters of the bond and angle terms in empirical force fields are typically fitted to reproduce experimental vibrational frequencies, crystallographic, and microwave spectroscopy data [176]. Parameterisation of dihedral torsion terms was initially also done using experimental data, but more recently it has been increasingly carried out using reference quantum mechanical calculations [99]. More modern classical force fields also include cross terms such as bond-bond interactions, or angle-angle interactions. These can also be parameterised directly from quantum-mechanical data [67, 140, 202].

The non-bonded terms describe long-range interactions and act between atoms that are not connected by a covalent bond. These global terms are usually 2-body and have a physics-derived functional form. The bonded terms can be viewed as the remaining short-range part of the quantum mechanical energy after subtracting the long-range electrostatic and dispersion interactions. For example, the electrostatic energy has exactly the functional form of the potential energy of two fixed point charges q_i and q_j separated by a distance r_{ij}

$$E_{\text{electrostatics}}(\mathbf{r}) = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_1}\frac{1}{r_{ij}} \tag{2.13}$$

where ε_1 is the background dielectric permittivity. Due to the very special decaying 2-body functional form it is possible to implement highly optimised algorithms such as the particle mesh Ewald method [48] which scale as $N\log(N)$ with the system size, making simulations of millions of atoms feasible [156]. The partial charges can be determined using quantum chemistry or by fitting to experimental data such as liquid and hydration properties or a combination of the two.

Classical force fields and machine learning

Several avenues of development have been pursued to improve the accuracy of empirical force fields using machine learning. One approach is to reparameterise the force field for the small molecule parts of simulations, which is especially relevant for protein-ligand binding free energy calculations. For example, the QueBeKit method automatically refits the bonded terms using reference Quantum Mechanical calculations. This improves the description of torsion barriers, whilst retaining the parameters of the generic force fields for the large biomolecules (the protein part of the system) where they work best [99, 100].

An alternative is the Espaloma approach which aims to machine learn the parameters of classical force fields using graph neural networks [223]. The key idea is to relax the discrete atom-types of classical force fields and replace them with smooth graph neural networks. The output of these neural networks is atomic environment-dependent force field parameters, thus retaining the computationally efficient simple classical force field functional form whilst enabling increased flexibility and fitting to experimental or QM data thanks to the continuous atom typing. The parameters are determined once at the beginning of the simulation and then kept fixed. This way the force fields have identical computational cost compared to the previously described classical force fields.

Current capabilities, challenges and recent developments

The typical total energy errors of Lifson-type force fields for small molecules compared to quantum mechanical energies are on the order of 2-10 kcal / mol (80-400 meV), although it can vary widely depending on how close the structure is to its equilibrium geometry. Furthermore, large differences in accuracy can be observed for different functional groups, for example, strained (3-4 member) rings being particularly challenging for these force fields [26, 99, 119].

Current capabilities in terms of computational performance are rapidly improving thanks to developments in both software algorithms and available hardware. Single-GPU performance of up to 800 ns / day with a 2 fs time step for a system of 24,000 atoms is possible using even relatively old GPUs. For larger systems of up to 1 million atoms using multiple GPUs, it is possible to achieve simulation performance over 100 ns / day with a 2 fs time step [125, 156].

The current limit of possible achievable performance is set by D E Shaw's custom built Anton 3 supercomputer which achieves up to 100 microseconds / day using a 2.5 fs time step for 1 million atoms [190].

2.4.2 Machine Learning Force Fields

The key limitation of the classical empirical force fields is their relatively high error in the PES. This error can be reduced significantly by relaxing the restricted functional form of classical force fields. Several new machine learning (ML) based functional forms have been proposed over the past 15 years. ML in the context of the approximation of potential energy surfaces can be regarded as a tool for accurately regressing high-dimensional functions.

The first methods that used machine learning to parameterize potential energy surfaces were all global methods. As discussed in Section 2.3.5 these methods are only applicable for small systems. The topic of the rest of the thesis is local force fields that potentially have a number of long-range terms.

In the following, a brief description of the most important machine learning force fields architectures is given. These methods served as the foundation and inspiration for the new methods proposed in this thesis in Chapter 3 and Chapter 5.

Symmetry-function based neural network force fields

The first class of widely adapted machine learning force fields is the symmetry-function based neural network potentials proposed by Ref [20] in 2007. The atom centred symmetry function (ACSF) based force fields rely on the locality assumption already introduced in

Section 2.3.5. They decompose the total energy of the system into atomic site energies following Equation (2.10). Site energies are parameterised as a function of the geometry of the atomic environment and are characterised by a set of fixed symmetric features, the ACSFs [19]. The ACSFs respect all relevant physical symmetries of the systems; i.e. they are invariant with respect to rigid rotations and translations of the environment and to permutations of like atoms in the environment. The ACSF descriptors are fed through a learnable nonlinear feedforward neural network to obtain the atomic site energies. In the following, the precise form of the ACSF descriptors as well as neural network regression are briefly reviewed.

Symmetry Function Descriptors The symmetry function descriptors serve the purpose of representing the local chemical environment of an atom by an array of numbers that are invariant with respect to the symmetries described in Section 2.3.2. Invariance with respect to rotations and translations is achieved by using internal coordinates (distances and angles) to define the symmetry functions. The radial symmetry functions are the products of Gaussian-type terms and a smooth cutoff function:

$$G_{i,\mu z}^{\text{rad}} = \sum_{\substack{j \neq i \\ r_{ij} \leq r_{\text{cut}}}} e^{-\eta_\mu (r_{ij} - r_\mu)^2} f_{\text{cut}}(r_{ij}) \delta_{z z_j} \tag{2.14}$$

where the sum is taken over all atoms j within the local environment of atom i as defined in Section 2.3.5. $f_c(r_{ij})$ is a cutoff envelope function that ensures that the value of the symmetry function goes smoothly to 0 at r_{cut} distance from the central atom. A separate set of radial symmetry functions are used for each central atom chemical element μ and neighbour atom chemical element z. A usual choice for the cutoff function is

$$f_{\text{cut}}(r_{ij}) = \begin{cases} 0.5 \times [\cos\left(\frac{\pi r_{ij}}{r_{\text{cut}}}\right) + 1] & \text{for } r_{ij} \leq r_{\text{cut}} \\ 0 & \text{for } r_{ij} > r_{\text{cut}} \end{cases} \tag{2.15}$$

The parameters η_μ and r_μ control the rate and position of decay of the symmetry functions. Using only radial symmetry functions would not be a sufficiently expressive representation of an atom's environment, as even trivially different environments like a square and a tetrahedron could not be distinguished if they have the same set of distances from the central atom. That is why a set of angular symmetry functions was also introduced.

$$G^{\text{ang}}_{i,\mu zz'} = 2^{1-\zeta} \sum_{\substack{j,k\neq i \\ r_{ij},r_{ik}\leq r_{\text{cut}}}} (1+\lambda_\mu \cos\theta_{ijk})^\zeta \, e^{-\eta_\mu (r_{ij}^2+r_{ik}^2+r_{jk}^2)} f_{\text{cut}}(r_{ij}) f_{\text{cut}}(r_{ik}) f_{\text{cut}}(r_{jk}) \delta_{zz_j} \delta_{z'z_k} \tag{2.16}$$

where the sum is taken over all neighbour pairs and θ_{ijk} is the angle between neighbour atoms j and k and central atom i. A set of different ζ values are used to control the angular resolution of the features. The parameter λ takes the values $+1$ or -1 and is used to center the maxima of the functions at $\theta_{ijk} = 0^\circ$ or at $\theta_{ijk} = 180^\circ$. In some variants of the symmetry functions the cutoff envelope between the two neighbour atoms j and k is omitted, leading to a sum over all possible triplets centred at atom i.

The final descriptor of the environment is formed by evaluating the radial and angular symmetry functions on all neighbours and pairs of neighbours and concatenating them into an array of numbers, where each entry corresponds to a distinct symmetry function. Note that the above is not the only valid choice of symmetry functions, and several other sets have been introduced since. A notable one is the ANI symmetry functions, which modified the functional form of the angular part [197]. It is also possible to create symmetry functions that have learnable parameters which are optimised based on the training set, as done by the DeepMD models [239].

Feedforward Neural Network Regression The descriptors introduced above are typically fed into a relatively small feedforward neural network, which is introduced next. Feedforward neural networks are probably the simplest examples of machine learning algorithms that can be used for regression or classification tasks and that also serve as a key component of many of the more advanced architectures. They take as input a number of variables and apply a series of linear and non-linear transformations to them to form the output.

More precisely, a layer of a neural network has the form

$$\mathbf{a}^{(t)} = f(\mathbf{W}^{(t)}\mathbf{a}^{(t-1)} + \mathbf{b}^{(t)}) \tag{2.17}$$

where f is a non-linear function, like sigmoid [23], ReLU [90] or SiLU [65]. The argument to the pointwise non-linearity is a linear transformation of the previous layer values. The width of layer t is the size of the array $\mathbf{a}^{(t)}$ and the depth of the network is the number of layers, indexed by t above.

$\mathbf{W}^{(t)}$ and $\mathbf{b}^{(t)}$ are the free parameters of the neural network that are estimated during training to minimise the value of a loss function $\mathscr{L}$, given a set of labelled training examples. In the case of fitting potential energy surfaces, the training examples contain quantum

mechanical energies and forces (gradients) of a number of arrangement of atoms. Neural networks are usually trained using the stochastic gradient descent algorithm [29]. It is an optimisation method that iteratively updates the weights of the neural network using gradient decent based on a random subset (batch) of the data. The parameters θ of a neural network are updated as

$$\theta = \theta - \eta \cdot \nabla_\theta \mathscr{L}(\theta; x^{(i)}, y^{(i)}) \tag{2.18}$$

where η is a parameter called the learning rate that determines the step-size during the optimisation, $\mathscr{L}(\theta; x^{(i)}, y^{(i)})$ is a loss function computed on a set of randomly selected datapoints labelled by i and ∇_θ means the gradient of the loss function with respect to the model parameters. In regression tasks, such as fitting the PES, the mean squared error is the commonly used loss function.

$$\mathscr{L}(\theta; x^{(i)}, y^{(i)}) = \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - \hat{y}^{(i)})^2 \tag{2.19}$$

where the sum is taken over the batch with N data points, $\hat{y}^{(i)}$ are the predictions of the neural network and $y^{(i)}$ are the ground truth labels.

ACSF based neural network potentials in practice Usually the ACSF based neural network potentials use small feedforward neural networks with 2-5 hidden layers. This makes fitting them relatively straightforward without the need for large-scale specialised hardware [20]. Recently, methods for automated parameterisation of neural network potentials have also been developed using active learning based on a committee of potentials [184].

Although symmetry functions provide a simple and intuitive way to describe the chemical environment, the framework has several limitations. The first key limitation is that most ACSF sets only use up to 3-body terms in the characterisation of the environment, but there are physical arrangements of atoms that cannot be distinguished by 3-body terms only [163]. It is possible to define 4-body ACSFs, but their evaluation cost gets prohibitively expensive. A further limitation is the quadratic scaling of the number of symmetry functions with the number of different chemical elements, making the models substantially slower for systems with many different chemical elements. This can be overcome by using element-specific weighting of ACSFs rather than indexing them by the chemical elements [75] which is a special case of a general class of tensor-reduced descriptors introduced in Section 3.3 [47]. A further limitation of the method is the need to explicitly sum over all triplets in the environment to construct the angular symmetry functions. This can become a computational bottleneck, especially in simulations of condensed phase materials where each atom can have

a large number of neighbours, on the order of 50 to 100, in a typical 6 Å neighbourhood. For example, in the case of a typical water model based on ACSF-s with a 6 Å cutoff the average number of neighbours is around 80, resulting in $\binom{80}{2} = 3,160$ triplets to sum over for each central atom.

ACSF based potentials have been used successfully over the years, for example, to study the 2D phase diagram of nanoconfined water [112] and to create a general purpose transferable organic force field [58]. The method has also been used to simulate reactive phenomena such as proton transfer at interfaces between ZnO and water [166] or the decomposition of urea in water [230].

SOAP-GAP Force Fields

In parallel with the development of ACSFs, an alternative representation of the atomic environment was also developed. The goal was to create a local Gaussian process regression based potential, also called Gaussian Approximation Potential (GAP) [13]. This required a rotationally and permutationally invariant kernel that quantifies the similarity of atomic environments. The most successful such kernel was based on the Smooth Overlap of Atomic Positions (SOAP) descriptor [12].

The construction of the SOAP descriptor incorporates the permutations symmetry first, followed by symmetrisation with respect to rotations. This is in contrast to ACSFs, which start by incorporating rotational symmetry first through using internal coordinates, and incorporate permutation symmetry second by summing the functions of internal coordinates over the neighbours.

In more detail, the SOAP framework is derived by first constructing a smooth atomic neighbourhood density, ρ. For a central atom i, of chemical element μ and with neighbours indexed by j, the neighbourhood density is written as

$$\rho^{i,\mu} = \sum_j \delta_{\mu\mu_j} \exp\left(\frac{-|\mathbf{r}-\mathbf{r}_i|^2}{2\sigma_\mu^2}\right) f_{\text{cut}}(r_{ij}) \tag{2.20}$$

The smooth cutoff function f_{cut} ensures that only atoms within a given radius contribute to the neighbourhood density. This density is a smooth and continuous function of atomic positions and is invariant to the permutation of like atoms. The parameter σ is a hyperparameter.

In the second step, the atomic density is expanded in a set of radial basis functions, R_n, and spherical harmonics, Y_{lm}, giving expansion coefficients with radial and angular indices $c^{i,\mu}_{nlm}$:

$$c^{i,\mu}_{nlm} = \int d\mathbf{r}\, R_n(\mathbf{r}) Y^*_{lm}(\mathbf{r}) \rho^{i,\mu}(\mathbf{r}) \tag{2.21}$$

Finally, a rotation invariant descriptor $p_{nn'l}^{i,\mu\mu'}$ called the power spectrum of the atomic environment is constructed from the expansion coefficients

$$p_{nn'l}^{i,\mu\mu'} = \frac{1}{\sqrt{2l+1}} \sum_m (c_{nlm}^{i,\mu})^\star c_{n'lm}^{i,\mu'} \tag{2.22}$$

The SOAP descriptor encodes the symmetries of permutation, translation and rotation invariance and by using a smooth cutoff in Equation (2.20), it varies smoothly with atomic positions.

In principle, a surrogate model of the PES could be constructed as any function of the power spectrum coefficients. The most widely used models fit the PES through kernel regression [23]. In this case, the potential energy of an environment, described by its power spectrum $\mathbf{p}^i$ is written in terms of a kernel function on the power spectrum $k(\mathbf{p}, \mathbf{p}')$ and a set of reference environments $\{\mathbf{p}^j\}_{i=1}^N$:

$$E(\mathbf{p}^i) = \sum_{i=1}^{N} w_j k(\mathbf{p}^i, \mathbf{p}^j) \tag{2.23}$$

This formulation explains the name 'Smooth Overlap of Atomic Positions'. If a linear dot product kernel is used ($k(\mathbf{p}^i, \mathbf{p}^j) := \mathbf{p}^i \cdot \mathbf{p}^j$), then it can be rewritten as a symmetrised overlap integral between two atomic neighbourhood densities ρ^i and ρ^j [53].

The emergence of symmetrised atomic density-based descriptors provided a simple recipe for constructing invariant or equivariant functions of local environments [147]. Coupled with the machinery of Gaussian processes in GAP [13], this method allowed an accurate description of complex materials with relatively straightforward parameterisation. The method has been applied to a wide range of materials science and chemistry problems, as recently reviewed in Ref. [53]. Examples are force fields for pure elements such as carbon [179], phosphorus [55], boron [57] and silicon [11, 54] as well as numerous functional materials [242], heterogeneous catalysis [182] and battery electrolytes [137].

Notably, the critical ingredients of the modern machine learning potentials discussed in this thesis were already present in SOAP. The use of a product of a radial basis and spherical harmonics as the set of 2-body functions still largely prevails to this date. The density trick is used in most models to achieve permutation invariance at no cost. Finally, the tensor products and the Clebsch-Gordan coefficients are the key elements of modern equivariant architectures such as NequIP [17] and MACE that will be introduced in Chapter 5 [16].

The main limitations of the SOAP GAP framework are similar to those of ACSF. It also scales poorly with the number of chemical elements and suffers from the same problem of incompleteness of 3-body descriptors [163]. These limitations are probably best observed

when the method is applied to approximate the PES of molecular systems, where it cannot achieve the accuracy of newer methods [119].

Message Passing Neural Network Force Fields

Message Passing Neural Networks (MPNN) are a class of machine learning methods that are designed to learn functions on graph structured data [79]. Chemical structures can naturally be represented as graphs embedded in 3D space. The nodes of the graph are the atoms, and there is an edge between two atoms if their Euclidean distance is less than a cutoff distance r_{cut}.

In the following, the precise form of MPNNs is summarised. This is a crucial background for both the unified design space of machine learning force fields presented in Chapter 4 and the MACE model presented in Chapter 5.

Invariant Message Passing Potentials MPNNs have four sub-parts. In the first part, a node state $\sigma_i^{(t)}$ is defined on each atom as a collection of three quantities,

$$\sigma_i^{(t)} = (\mathbf{r}_i, z_i, \mathbf{h}_i^{(t)}), \tag{2.24}$$

with $\mathbf{r}_i$ the position of atom i, z_i the atomic number of i, and h_i^t a collections of learnable features or descriptors of the environment of atom i. In the 0-th layer, these features are usually initialised as the one-hot embedding of the chemical element. A one-hot embedding is a technique often used in machine learning to turn categorical features into numerical representations.

$$\mathbf{h}_i^{(0)} = \sum_z W_{kz} \delta_{zz_i} \tag{2.25}$$

where the learnable embedding matrix W has dimensions $k \times Z$, with k being the length of the embedding and Z being the number of different chemical elements of the model.

In the next step, a general learnable function, usually a feedforward neural network, M_t is applied to the states of each pair of neighbouring atoms i and j to form the edge features. This function M depends on the states of the atoms but is otherwise the same for the entire chemical structure, ensuring that the model is extensible and transferable to new unseen graphs. To achieve permutation invariance, a “message" is constructed by applying a pooling operation $\bigoplus_{j \in \mathcal{N}(i)}$ to the set of edge features $M_t(\sigma_i^{(t)}, \sigma_j^{(t)})$. This pooling can in principle be any permutation invariant operation, such as `min` or `max`, but in practice it is almost always

chosen to be the sum of the edge features

$$\mathbf{m}_i^{(t)} = \bigoplus_{j \in \mathcal{N}(i)} M_t(\boldsymbol{\sigma}_i^{(t)}, \boldsymbol{\sigma}_j^{(t)}), \tag{2.26}$$

where $\mathbf{m}_i^{(t)}$ represents the message on atom i in layer t of the network. The resulting message contains information about the neighbourhood of the central atom i. An MPNN force field is called invariant if the message $\mathbf{m}_i^{(t)}$ is invariant with respect to the rigid rotations of the structure. Next, the message is used to update the state of the central atom to form a set of new features,

$$\mathbf{h}_i^{(t+1)} = U_t(\boldsymbol{\sigma}_i^{(t)}, \mathbf{m}_i^{(t)}), \tag{2.27}$$

with U_t being a linear or non-linear learnable update function. The steps in Equations (2.24) - (2.27) represent a layer of an MPNN.

Each iteration of the message passing operation leads to an increase in the body order. The features h_i are multiplied by one additional term leading to an iterative increase in body order. A side effect of this mechanism is that the receptive field of the representation increases in each iteration by r_{cut}. Many typical MPNN models use up to 5 layers, leading to a receptive field of 25-30 Å.

In the final readout phase, the states of the atoms are mapped to the output quantity, usually the site energy, by a readout function,

$$E_i = \sum_{t=1}^{T} \mathscr{R}_t(\boldsymbol{\sigma}_i^{(t)}). \tag{2.28}$$

This step is analogous to fitting a model on a set of features in ACSFs or SOAP GAP, with the difference that the features themselves already contain a large number of free parameters as they were constructed by stacking MPNN layers. Although most of the message functions, M_t, of MPNNs are two body functions, meaning that they depend simultaneously on the state of just two atoms; DimeNet [76] and GemNet [113] expanded them to invariant three and four body functional forms by considering triplets or quadruplets within the neighbourhoods of a central atom. The mechanism for this is analogous to how ACSFs create higher body order features by incorporating functions of angles between two neighbours.

The most well-known example of invariant message passing force fields is SchNet [187] which uses a graph convolutional architecture. In SchNet a learnable convolution filter is iteratively applied to the chemical graph with the filter being a function of interatomic distances. This means that M_t in Equation (2.26) is only a function of the distance r_{ij} and the node features of the neighbour atom $\mathbf{h}_j^{(t)}$, but it is not a function of the features of the

central atom i. The use of distances only is analogous to forming a one-particle basis in SOAP, but taking only the $l = 0$ spherical harmonics, i.e. only using radial features. The formal connection between MPNNs and graph convolutional neural networks like Schnet has been discussed at length in Ref. [32].

Due to its relative simplicity, the SchNet architecture gained popularity as a building block for creating more sophisticated networks. When used as a force field though it is not able to achieve the same accuracy as the kernel and linear models [42, 119]. One of the key limitations of SchNet is similar to the problems with ACSF and SOAP-GAP models, there are atomic configurations that a SchNet model, even with infinitely many layers cannot differentiate [162].

Equivariant Message Passing Potentials

There are two ways to improve the expressiveness of MPNN models. One way is to increase the body order of the messages. If done naively by summing over triplets and quadruplets, this significantly increases the computational cost of the models [113]. The other way is to allow for the messages to be of higher order geometric tensors describing the atomic environment. MPNNs that use messages which are vectors or tensors are called equivariant MPNN models. Cormorant [4] and Tensor Field Networks [206] were the first two examples of MPNNs that used tensor products of spherical features to construct rotationally equivariant messages. A message $\mathbf{m}_{i,L}$ is rotationally equivariant (with symmetry label L) if it transforms according to the irreducible representation L of the SO(3) symmetry group,

$$\mathbf{m}_{i,L}\left(Q\cdot(\mathbf{r}_1,...,\mathbf{r}_N)\right) = \mathbf{D}^L(Q)\mathbf{m}_{i,L}(\mathbf{r}_1,...,\mathbf{r}_N), \quad \forall Q \in \mathrm{O}(3) \tag{2.29}$$

where $Q\cdot(\mathbf{r}_1,...,\mathbf{r}_N)$ denotes the action of an arbitrary rotation on the set of atomic positions $(\mathbf{r}_1,...,\mathbf{r}_N)$ and $\mathbf{D}^L(Q)$ is the corresponding Wigner D-matrix. Hence, a message indexed by L transforms like the spherical harmonic Y_{LM} under rotations. By using equivariant messages, the model becomes capable of constructing a more comprehensive set of invariant features in the subsequent layers and at the readout phase [186].

The NequIP model demonstrated that this architecture is capable of improving the accuracy of invariant models on several benchmarks [17]. Similarly to Cormorant, NequIP also uses tensor products of spherical features to construct equivariant messages, but achieves significantly higher accuracy by using better normalisation of the weights, a better radial basis, readouts, and updates [14].

In NequIP the message of Equation (2.26) is constructed from edge features defined as a tensor product of a radial part, spherical harmonics and the equivariant node features of the

neighbour atoms $h^{(t)}_{j,kl_2m_2}$,

$$m^{(t)}_{i,kl_3m_3} = \frac{1}{\lambda} \sum_{l_1m_1,l_2m_2} C^{l_3m_3}_{l_1m_1,l_2m_2} \sum_{j \in \mathcal{N}(i)} R^{(t)}_{k,l_1l_2l_3}(r_{ij}) Y_{l_1m_1}(\mathbf{r}_{ij}) h^{(t)}_{j,kl_2m_2} \tag{2.30}$$

One key innovation of NequIP was the use of different radial basis for each quadruplet of channels and angular momenta kl_1, l_2, l_3, which greatly increased the flexibility of the model. The message of NeuquIP is then used to update the node features, and the process is repeated for typically 4-5 iterations. Finally, the site energy of the atoms is predicted from the invariant part of the last layer features using a feedforward neural network.

NeuqIP demonstrated the power of equivariant MPNN models for parameterising potential energy surfaces. There remain two limitations of NequIP: it still uses only 2-body messages, and it has an inherently large computational cost. There are two reasons for the relatively high computational cost. On the one hand, the equivariant tensor products that are repeated in each layer, altogether 4 or 5 times. On the other hand, the very large receptive field makes the model very difficult to parallelise across multiple GPUs. Both of these limitations will be addressed by the MACE model introduced in Chapter 5.

2.5 Benchmark Datasets for Comparing Molecular Force Fields

Benchmark datasets have played a very important role in the development of new force field functional forms. Typically benchmark datasets contain a number of different atomistic structures, which are labelled with quantum mechanically calculated energies and forces. The task is to fit a model of the potential energy on a subset of the dataset, the training set, and to evaluate the accuracy of the fitted PES on an unseen test set. These datasets are best viewed as proxies: the models and PESs parameterised on these datasets are not useful for any scientific purpose. However, they can be a good tool for guiding model development. Empirically, it has been found that the models that are simple to fit and perform well on the benchmark datasets are also the ones that are the most useful in real scientific applications requiring high accuracy potential energy surfaces.

In this section, some of the most important benchmark datasets are summarised. They are used in the later part of the thesis to compare the performance of the newly developed force field functional forms with other relevant models in the literature.

There is also a different set of benchmark datasets that can be used purely for the evaluation of already fitted PES-s rather than for functional forms. These do not provide a

training set, only a set of test structures or tasks. These are useful for evaluating potential energy surfaces that are designed to be useful for scientific applications, such as a transferable organic force field. Some of these tests will be discussed in Section 5.3 which tests the MACE-OFF transferable organic force field.

2.5.1 Pre-existing Benchmark Datasets

QM9 The QM9 dataset [171] is one of the oldest and most widely used benchmark datasets in atomistic machine learning. It can be used to validate ML models for chemistry in general and not just for force field fitting. Most of the machine learning architectures published over the past 10 years have reported their results on QM9. The dataset contains about 130,000 equilibrium molecular geometries that are made up of chemical elements H, C, N, O, and F and contain up to 9 heavy atoms. The molecules were generated by enumerating all possible such compounds using cheminformatics rules. Once the equilibrium geometries were found using geometry optimisation, 12 different properties of the molecules were calculated which include the potential energy, but also other intensive quantities such as heat capacities and band gap energies. The ML models are typically fitted independently for the 12 properties [120].

rMD17 The original MD17 benchmark dataset consists of configurations of 10 small organic molecules in vacuum sampled from density functional theory (DFT) molecular dynamics simulations at 500 K [39]. It has been recognised that some of the calculations in the original dataset did not properly converge; in particular, many of the forces are noisy. A subset of the full dataset was recomputed with very tight SCF convergence settings and is called the rMD17 (revised MD17) dataset published in Ref. [41]. This new version of the dataset is used throughout this thesis, including the five train-test splits as originally reported. These revised training sets consist of 1,000 configurations to avoid the problem of correlated training and test sets: When more than 1,000 configurations are used from the full published trajectory, some of the test set configurations will necessarily fall between two neighbouring training set data points that are separated by a much smaller time difference than the decorrelation time of the trajectory, resulting in an underestimate of the generalisation error [41, 119].

MD22 The MD22 dataset was designed to be challenging for short-range (local) machine learning models of the PES [40]. The dataset includes large molecules and molecular assemblies containing hundreds of atoms with complex intermolecular interactions. Similarly to MD17, it was created by running ab initio molecular dynamics simulations at elevated

temperatures to better sample the configuration space of the systems. The size of the training set was determined so that the total energy error of the sGDML model [110] is below chemical accuracy, 1 kcal/mol ($\approx$ 43 meV). In this thesis, the same sizes of the training set are used to allow a fair comparison of the different models [120].

Water The water dataset consists of 1593 liquid water configurations, with 64 molecules in each [36]. This dataset contains a wide range of liquid water structures with about a third of them being generated by path integral molecular dynamics (PIMD) simulations. The energy and force labels were computed using the CP2K software [123] at the revPBE0-D3 level of density functional theory which is known to give a reasonably good description of the structure and dynamics of water at a variety of pressures and temperatures [141].

2.5.2 Benchmark Datasets Developed in This Thesis

3BPA The 3BPA dataset contains snapshots of a large flexible druglike organic molecule, 3-(benzyloxy)pyridin-2-amine (3BPA), sampled from different temperature molecular dynamics trajectories [119] generated using the ANI-1x force field [198].

To ensure a good coverage of the full energy landscape in the training dataset, first a grid of the three dihedral angles (α, β and γ) was defined and the configurations with atom overlap were removed. From each of the configurations corresponding to the grid points, short (0.5 ps) MD simulations were performed using the ANI-1x force field [198]. This time scale is sufficient to perturb the structures towards lower potential energies, but is not enough to significantly equilibrate them. In this way, a set of 7,000 configurations was obtained, as shown in the left panel of Figure 2.2.

From the configurations obtained, five different densely populated pockets were identified in the space of the three dihedral angles. One random configuration was selected from each of the five pockets, and a long 25 ps MD simulation was performed at three different temperatures (300 K, 600 K, 1,200 K) using the Langevin thermostat and 1 fs time step. 460 configurations were sampled from each of the trajectories starting after a delay of 2 ps. This protocol resulted in the final data set of 2,300 molecular geometries. The configurations were re-evaluated using ORCA [151] at the DFT level of theory using the ωB97X exchange correlation functional [34] and the 6-31G(d) basis set. These settings are similar to those used in the creation of the ANI-1x data set [201]. The benchmark contains two training sets, one is created by randomly selecting 500 geometries from the 300 K set, and the other, labelled "mixed-T", is constructed by selecting 133 random configurations from each of the trajectories at the three temperatures. The rest of the data not present in any of the training sets make up the three test sets, each corresponding to a different temperature. The right-hand

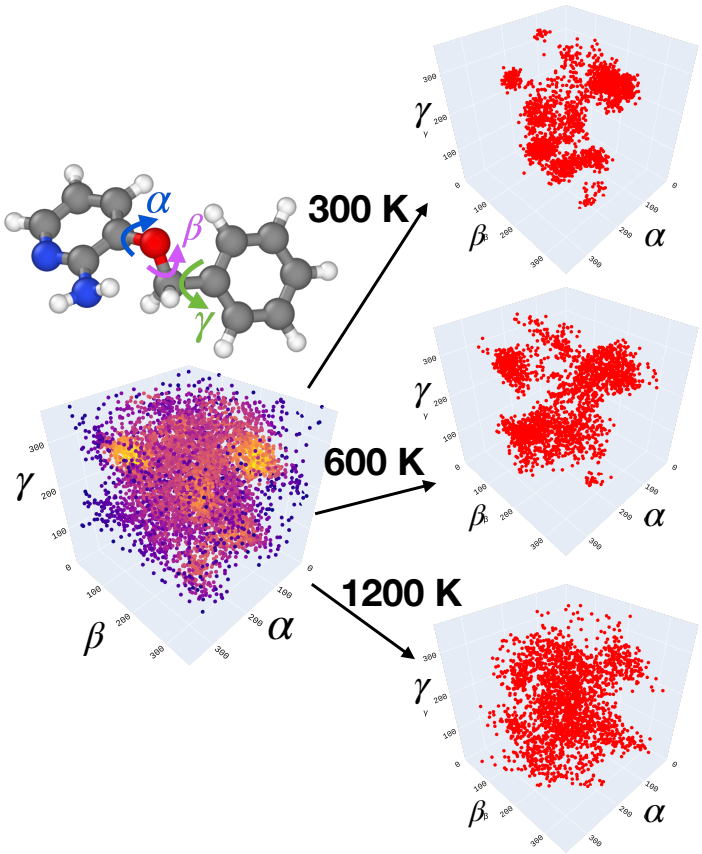


Fig. 2.2 **3BPA data set.** The three freely rotating angles of the 3BPA molecule together with a characterization of the three different data sets sampled at different temperatures showing how the phase space sample increases significantly with temperature.

panels of Figure 2.2 show the distribution of dihedral angles in the test sets. At 300 K the stable pockets of the configuration space are sampled individually, whereas at 1200 K the distribution widens significantly, and the sampling connects the stable pockets.

This benchmark dataset is significantly different from those introduced in Section 2.5.1. It directly probes the smoothness and extrapolation of the fitted PESs. In particular, in the case of the 300K training set, the 300K test configurations measure in-domain accuracy, whereas the 600K and 1,200K test sets measure the extrapolation accuracy of the PES. Performance on out-of-distribution samples is crucial because it correlates well with the usefulness of the model in actual applications. When using the models to run MD simulations, they are likely to encounter configurations far from the training set. Models that achieve low errors on this extrapolation benchmark are typically the ones that perform best in molecular dynamics applications. These are also the models that require the least amount of iterative retraining to obtain a model capable of running stable MD simulations [15–17, 120, 121].

Chapter 3

Linear ACE Force Fields for Small Molecules

In this chapter, first, the Atomic Cluster Expansion (ACE) introduced by Ref. [61] is presented. This method was developed as a general framework for atomistic modelling and PES fitting in the context of materials science. Following Ref [119], the ACE method is demonstrated to be capable of parameterising high-fidelity PESs of small molecules even if it is used as a regularised linear model. Finally, following Ref [47], a new general method is introduced that enables the simulation of systems with a very large number of chemical elements using atom centred descriptor based models like ACE and SOAP GAP.

3.1 Atomic Cluster Expansion (ACE)

The ACE model[61, 63] uses the body order expansion defined in Section 2.3.3. One of the key features of ACE is that it reduces the computational cost of evaluating such an expansion compared to models like PIPs that explicitly sum over the many-body clusters. This is accomplished by projecting the atomic neighbour density onto isometry invariant basis functions. This idea, detailed in the following, is referred to as the "density trick", and was introduced originally to construct the power spectrum (also known as SOAP) and bispectrum descriptors [13, 12] which are, in fact, equivalent to the 3- and 4-body terms in ACE, respectively. The ACE features can be considered to be a generalisation of SOAP and the bispectrum to an arbitrary body order.

ACE features, or basis functions, can be derived similarly to how SOAP is constructed in Section 2.4.2. First, the neighbourhood density of an atom i is defined as

$$\rho_i^z(\mathbf{r}) = \sum_j \delta_{zz_j}\delta(\mathbf{r}-\mathbf{r}_{ji}); \tag{3.1}$$

where ρ_i^z denotes the density of atoms of element z in the neighborhood of atom i. Note that in SOAP the neighbourhood density was constructed from Gaussian basis functions centred at the atomic positions, whereas in ACE it is made up of Dirac delta functions. This density is projected onto a set of 1-particle basis functions, $\phi_{nlm}^{z_i z_j}$, which are chosen to be the product of a radial basis, $R_{nl}^{z_i z_j}$, and real spherical harmonics, Y_{lm}.

$$\phi_{nlm}^{z_i z_j}(\mathbf{r}) = R_{nl}^{z_i z_j}(r)Y_{lm}(\hat{\mathbf{r}}). \tag{3.2}$$

Here, "1-particle" refers to the fact that the value of these basis functions depends on the position of one neighbour particle j. There is considerable flexibility in the choice of the radial basis; the specifics for this work are documented at the end of this section. Next, the atomic base is defined as the projection of the neighbourhood density onto the 1-particle basis functions.

$$A_{z_i,znlm} = \langle \rho_i^z | \phi_{nlm}^{z_i z} \rangle = \sum_{\substack{j \\ \text{where } z_j = z}} \phi_{nlm}^{z_i z}(\mathbf{r}_{ji}) \tag{3.3}$$

where the index z_i refers to the chemical element of atom i. For notational convenience, the rest of the 1-particle basis indices are collected into a multi-index,

$$(znlm) \equiv \nu. \tag{3.4}$$

From the permutation invariant atomic base $A_{z_i\nu}$, also called the "A-basis", many-body basis functions can be formed by taking the products,

$$\mathbf{A}_{z_i\mathbf{v}} = \prod_{t=1}^{\nu} A_{z_i\nu_t}, \quad \mathbf{v} = (\nu_1, \ldots, \nu_\nu). \tag{3.5}$$

where the bold $\mathbf{A}_{z_i\mathbf{v}}$ is called the product basis. These basis functions are $(\nu+1)$ body because the product containing ν factors gives a basis function that is the sum of terms, each of which depends on the coordinates of at most ν neighbours. These functions are also sometimes referred to as ν-correlations. A graphical illustration of this construction is shown in Figure 3.1 for the special case where the two factors are the same. For many (different) factors, taking products of the atomic base (left side of Figure 3.1) takes much less time to

evaluate than the explicit sum of all possible products (right side of Figure 3.1). This is the key step of the density trick.

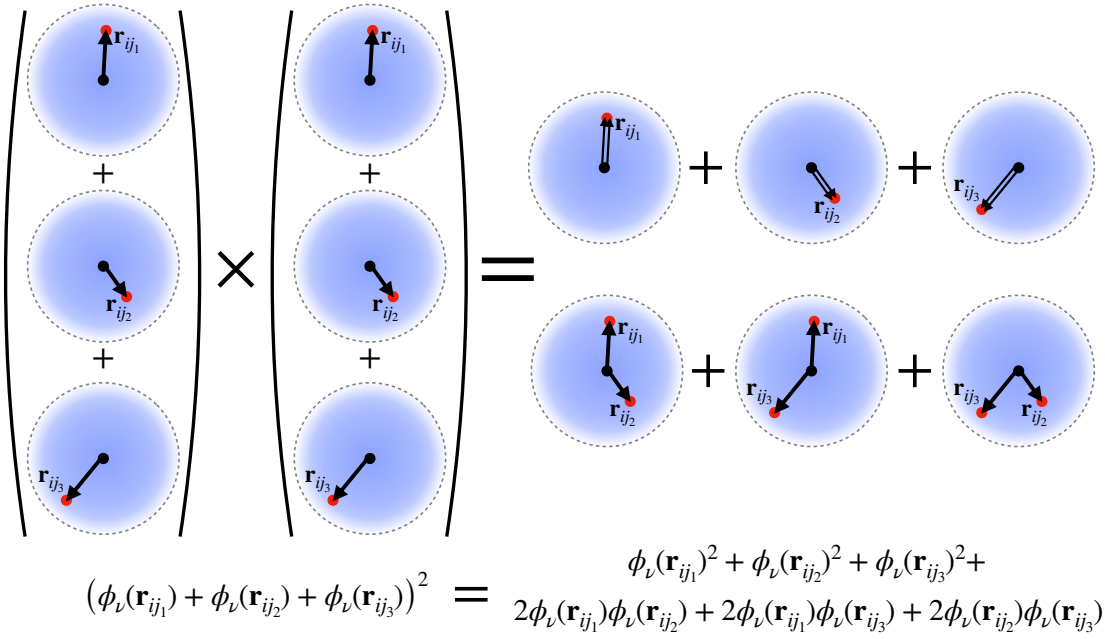


Fig. 3.1 **Construction of high body order invariant basis functions.** A graphical illustration showing how higher body-order basis functions can be constructed as products of the projected neighborhood density. The evaluation cost of the basis functions scales linearly with the number of neighbours rather than exponentially by performing the density projection first and then taking the products to obtain higher order basis functions. The figure (and expression) also makes explicit the occurrence of self-interaction terms in the ACE basis. They are automatically corrected through the inclusion of lower-order correlations in the basis [63].

The product basis is not rotationally invariant. A fully permutation and isometry-invariant overcomplete set of functions called the B-basis (technically not a basis but a spanning set), can be constructed by averaging the $\mathbf{A}$-basis over the three dimensional rotation group, $O(3)$,

$$B_{z_i\mathbf{v}} := \int_{\hat{R}\in O(3)} \prod_{t=1}^{\nu} A_{z_i\nu_t}(\{\hat{R}\mathbf{r}_{ij}\})\, d\hat{R} = \sum_{\mathbf{v}'} C_{\mathbf{v}\mathbf{v}'}\mathbf{A}_{z_i\mathbf{v}'}, \tag{3.6}$$

where on the right the integral was rewritten as a linear projection expressed by the matrix of generalised Clebsch-Gordan coupling coefficients $C_{\mathbf{v}\mathbf{v}'}$. The integral can be rewritten in such a simple form because the angular dependence of $\mathbf{A}_{z_i\mathbf{v}'}$ is expressed using the spherical harmonic basis [63]. Many of the resulting basis functions will be linearly dependent (or even

zero), but it is relatively straightforward to remove these dependencies in a pre-processing step to arrive at an actual basis set. The details of the procedure used are outlined in Ref [63].

The B-basis in Equation (3.6) is complete in the sense that any smooth and continuous function of the neighbouring atoms that is invariant to permutations and rotations can be expanded as a linear combination of the basis functions. The atomic site energy of linear ACE can therefore be written as

$$E_i = \sum c_{z_i\mathbf{v}} B_{z_i\mathbf{v}} = \mathbf{c}\cdot\mathbf{B}. \tag{3.7}$$

The above equation makes it clear that the model is linear in its free parameters, the $\mathbf{c}$ coefficients. The B-basis functions are polynomials of the atomic coordinates, and to show that the explicit body ordering has been retained, the site energy can be rewritten in terms of the A-basis (with the product explicitly written out),

$$E_i = \sum_{v} \tilde{c}^{(1)}_{z_i v} A_{z_i v} + \sum_{v_1 v_2}^{v_1 \geq v_2} \tilde{c}^{(2)}_{z_i v_1 v_2} A_{z_i v_1} A_{z_i v_2} + \sum_{v_1 v_2 v_3}^{v_1 \geq v_2 \geq v_3} \tilde{c}^{(3)}_{z_i v_1 v_2 v_3} A_{z_i v_1} A_{z_i v_2} A_{z_i v_3} + \ldots \tag{3.8}$$

where the $\tilde{c}$ are a linear combinations of the c coefficients appearing in Equation (3.7), using the transformation defined in Equation (3.6). Now, the body ordering is readily identifiable. Each term corresponds precisely to a sum of ν-correlations, that is, $(\nu+1)$-body terms as in the traditional body-order expansion. In practice, a recursive scheme can be used to efficiently evaluate basis functions, which leads to an evaluation cost that is $O(1)$ per basis function, independent of body order [63]. The number of basis functions increases with body order, at a rate that has an exponent ν.

The construction outlined so far yields infinitely many polynomials $B_{z_i\mathbf{v}}$, which can be characterised by their correlation order ν, and their (modified) polynomial degree $D = \sum_t^{\nu} n_t + w_Y l_t$, where n_t and l_t come from the multi-index v_t and the weight w_Y is used to trade off the radial and angular resolution of the basis set. When it comes to defining a model in practice, the expansion is truncated both in the body order and in the maximum polynomial degree at each body order.

3.1.1 Choice of Radial Basis

In the models in this chapter a simplified radial basis is used where $R_{nl}^{z_i z_j}(r) = R_n(r)$, such that

$$R_n(r) = p_n(x(r)) f_{\text{cut}}(x), \tag{3.9}$$

$r \mapsto x(r)$ is a one dimensional radial transformation, f_{cut} is a cutoff or envelope function and p_n are orthogonal polynomials. The radial transform is chosen to be

$$x(r) = \frac{1}{(1+r/r_0)^2}, \tag{3.10}$$

which amplifies the effect of neighbours closer to the central atom. For the cutoff function, both the inner and outer cutoffs are specified, $r_{\text{in}} < r_{\text{out}}$, and the envelope is defined as

$$f_{\text{cut}}(x) = (x - x(r_{\text{in}}))^2 (x - x(r_{\text{out}}))^2, \tag{3.11}$$

The polynomials p_n are defined recursively by specifying that $p_0(x) = 1, p_1(x) = x$, and the orthogonality requirement

$$\int_{x(r_{\text{in}})}^{x(r_{\text{out}})} R_n(r(x)) R_{n'}(r(x)) x^2 dx = \delta_{nn'}, \tag{3.12}$$

where the inverse of the radial transform is used, $x \mapsto r(x)$. Equation (3.12) implies that the radial basis R_n and not the polynomials p_n are orthonormal in x-coordinates.

The introduction of an inner cutoff is necessary to prevent wildly oscillating behaviour of the many-body basis functions in regions of configuration space where pairs of atoms are very close to one another and little or no training data is available. Alternatively, one could introduce such training data, but that would unnecessarily complicate the construction of training data sets, and this inner cutoff mechanism is sufficient. To ensure short-range repulsion, the large multi-body ACE basis is augmented by a small auxiliary basis set, consisting only of low-polynomial-degree two body functions. The construction is exactly the same as before, but the cutoff function is changed to

$$f_{\text{cut}}^{\text{rep}} = (x - x(r_{\text{out}}))^2. \tag{3.13}$$

3.1.2 Basis Selection

Before the linear ACE force field can be parameterised, a specific finite basis set should be chosen from the complete ACE basis. There are three approximation parameters in linear ACE:

- cutoff radius ($r_{\text{cut}} = r_{\text{out}}$)
- maximum correlation order $\nu^{\max}$
- maximum polynomial degrees $D_\nu^{\max}$ corresponding to order ν basis functions

The cutoff radius was already defined in the definition of the radial basis in Equation (3.9). The basis is then chosen as (a linearly independent subset of) all possible basis functions $B_{i\mathbf{v}}$ with correlation order at most $\nu^{\max}$ and polynomial degree at most $D_{\nu}^{\max}$.

In all models for molecules with three or fewer distinct elements, the maximum correlation $\nu^{\max} = 4$ was used, which corresponds to a general 5-body potential. In models for molecules with four or more distinct chemical elements, the correlation order was reduced to $\nu^{\max} = 3$ (4-body potential). The weight w_Y specifies the relative importance of the radial and angular basis components; here $w_Y = 2$ was used. The maximum polynomial degrees $D_{\nu}^{\max}$ can be adjusted to balance the size of the basis set against the accuracy of the fit and the evaluation cost of the force field.

3.2 Regularised Linear ACE Models for Small Molecules

In this section, the parameterisation of linear ACE force fields for small molecules is summarised followed by a number of selected results benchmarking the model against many previously used methods developed to create small-molecule force fields.

3.2.1 Parameterisation of the Linear ACE Force Fields

The total energy of a linear ACE model with parameters $\mathbf{c}$ corresponding to a spatial configuration of atoms (denoted by X, e.g. a molecule in a particular configuration) is defined as the sum of site energies,

$$E(\mathbf{c};X) = \sum_{i \in X} E_i(\mathbf{c}) \tag{3.14}$$

where E_i is a site energy defined in Equation (3.7). Optimal parameters are obtained by minimising the loss function given a set of labelled training configurations

$$L(\mathbf{c}) = \sum_{X} \left(w_X^E \left| E(\mathbf{c};X) - E_{\mathrm{QM}}(X) \right|^2 + w_X^F \left| F(\mathbf{c};X) - F_{\mathrm{QM}}(X) \right|^2 \right), \tag{3.15}$$

where the E_{QM} and F_{QM} are energies and forces, respectively, in the training data, obtained from electronic structure calculations and E and F are the energies and forces predicted by the ACE model. The sum is taken over all configurations in the training set, and w_X^E, w_X^F are weights specifying the relative importance of the energies and forces. Since the predicted energies and forces are both linear in the free parameters, the loss can be written in a linear least squares form.

$$L(\mathbf{c}) = \|\Psi\mathbf{c} - \mathbf{t}\|^2, \tag{3.16}$$

where the vector $\mathbf{t}$ contains the QM energy and force observations, and the design matrix Ψ contains the values and gradients of the basis evaluated at the training geometries. Ψ has a number of rows equal to the total number of observations (energies and force components) in the training set and a number of columns equal to the total number of basis functions.

The least squares problem has to be regularised, especially when the basis contains high degree polynomials [214]. One option is to apply Tychonov regularisation, where the loss function is modified as

$$L(\mathbf{c}) = \|\Psi\mathbf{c} - \mathbf{t}\|^2 + \lambda\|\Gamma\mathbf{c}\|^2. \tag{3.17}$$

This is widely used to regularise linear regression, often taking Γ as just the identity matrix, or alternatively in the case of kernel ridge regression (and Gaussian process regression) as the square root of the kernel matrix [173]. In the present case, a diagonal Γ is used with entries corresponding to a rough estimate for the p-th derivative of the basis functions,

$$\|\nabla^p B_{z\mathbf{v}}\|_2 \approx \sum_{t=1}^{\mathrm{len}(\mathbf{v})} (n_t)^p + (l_t)^p, \tag{3.18}$$

where n_t and l_t are part of the elements of the multi-index vector $\mathbf{v}$ (cf. Equation (3.4)). This regularisation scales down high degree basis functions, encouraging a smooth potential, which is crucial for extrapolation, and is loosely analogous to the smooth Gaussian prior of Gaussian process regression. The actual solutions are then found using the standard iterative LSQR solver[155].

In the other approach used for solving the least squares problem the same Γ matrix is introduced, but without a Tychonov term,

$$L(\mathbf{c}) = \|\left(\Psi\Gamma^{-1}\right)\left(\Gamma\mathbf{c}\right) - \mathbf{t}\|^2, \tag{3.19}$$

and the solution is found using the rank revealing QR factorisation [97] (RRQR), in which a QR factorization of the scaled design matrix $\Psi\Gamma^{-1}$ is performed. The algorithm truncates the factorisation so that small singular values below some tolerance parameter λ are omitted. For more details on the exact implementation, see Refs. [2, 97]. When the linear system is not underdetermined, RRQR is found to give solutions with lower error than LSQR.

The last modelling choice that needs to be made is the 1-body term, that is, the energies of the isolated atoms of each element in the model. The energy of the isolated atoms evaluated with the reference electronic structure method can be used, which ensures the correct behaviour of the model in the dissociation limit. In other words, the force field is modelling the interaction energy of the atoms. An alternative approach, often used in the ML fitting of molecular energies, is to take the average energy of the training set, divided

by the number of atoms in the molecule, and assign the result to each element. In this case, the fitted model has zero mean energy. This usually improves the fit accuracy slightly by normalising the target data. A third option is to use no reference potential for the fit, but only use the forces. Once the coefficients are determined, the potential can be shifted by a constant energy chosen to minimise the energy error of the training set. In this section, all three strategies are evaluated for linear ACE. It is found that using the isolated atom energies for the 1-body term gives slightly higher root mean squared (RMS) errors in domain but leads to superior extrapolation. The other two strategies (using the average energy for the 1-body term and fitting only to forces) result in similar somewhat lower test set errors but inferior physical extrapolation properties.

3.2.2 Experimental Results

The purpose of this section is to demonstrate the performance of linear ACE force fields for small organic molecules. First, the performance on the rMD17 [39, 41] benchmark data set is evaluated. It is also important to go beyond the RMSE or mean absolute error (MAE) of energies and forces (the typical target of the loss function in the fit), because practically useful force fields have other desirable properties too: chemically sensible extrapolation, good description of vibrational modes, and accuracy on trajectories self-generated with the force field, just to name a few. The insufficient nature of mean error metrics has been pointed out before [69, 118, 232]. In addition, linear ACE is also fitted on the slightly larger and significantly more flexible 3BPA molecule that is more representative of the needs of medicinal chemistry applications.

rMD17

Energy and Force Accuracy First, the ability of linear ACE to fit small molecular potential energy surfaces was assessed using the rMD17 benchmark dataset introduced in Section 2.5. Table 3.1 shows the MAE of the different force field models trained on 1,000 configurations. For comparison, in Table 3.1, a wide selection of models is included from the various classes of force field fitting approaches that were discussed in Section 2.3. They include ML approaches such as feedforward neural network architectures (ANI [58, 73], GMsNN [236]), Gaussian process regression models (sGDML [38], FCHL [42], GAP [13]) and graph neural network based models (DimeNet [76], PaiNN [186]). The models on the left of Table 3.1 were retrained for this evaluation using the exact train test splits of rMD17 (except for FCHL which has published numbers available), while the models on the right of the solid vertical

Table 3.1 **Mean Absolute Error of MD17 molecules.** Energy (meV) and force (meV/Å) errors of different models trained on 1,000 samples. Models on the left were trained and tested using the same rMD17 train-test splits, whereas models on the right use MD17. The best models for each molecule (on the left and right) are shown in bold font. The average energy MAE is calculated per atom. For reference 43 meV $\approx$ 1 kcal / mol.

		ACE	**sGDML**	**FCHL**[41]	**GAP**	**ANI**	**FF**	**PaiNN**[186]	**GMsNN**[236]	**DimeNet**[76]
Aspirin	E	**6.1**	7.2	6.2	17.7	16.6	93.2	6.9	16.5	8.8
	F	17.9	31.8	20.9	44.9	40.6	260	**16.1**	29.9	21.6
Azobenzene	E	3.6	4.3	**2.8**	8.5	15.9	112	-	-	-
	F	10.9	19.2	**10.8**	24.5	35.4	246	-	-	-
Benzene	E	**0.04**	0.06	0.35	0.75	3.3	13.2	-	3.5	3.4
	F	**0.5**	0.8	2.6	6.0	10.0	105	-	9.1	8.1
Ethanol	E	1.2	2.4	**0.9**	3.5	2.5	42.1	2.7	4.3	2.8
	F	7.3	16.0	**6.2**	18.1	13.4	208	10.0	14.3	10.0
Malonaldehyde	E	1.7	3.1	**1.5**	4.8	4.6	45.9	3.9	5.2	4.5
	F	11.1	18.8	**10.3**	26.4	24.5	234	13.8	19.5	16.6
Naphthalene	E	0.9	**0.8**	1.2	3.8	11.3	65.3	5.1	7.4	5.3
	F	5.1	5.4	6.5	16.5	29.2	292	**3.6**	15.6	9.3
Paracetamol	E	4.0	5.0	**2.9**	8.5	11.5	93.9	-	-	-
	F	12.7	23.3	**12.3**	28.9	30.4	248	-	-	-
Salicylic acid	E	**1.8**	2.1	**1.8**	5.6	9.2	68.4	4.9	8.2	5.8
	F	9.3	12.8	9.5	24.7	29.7	263	**9.1**	21.2	16.2
Toluene	E	1.1	**1.0**	1.7	4.0	7.7	36.9	4.2	6.5	4.4
	F	6.5	6.3	8.8	17.8	24.3	183	**4.4**	14.7	9.4
Uracil	E	1.1	1.4	**0.6**	3.0	5.1	43.3	4.5	5.2	5.0
	F	6.6	10.4	**4.2**	17.6	21.4	233	6.1	14.3	13.1
Average MAE	E*	**0.12**	0.16	**0.12**	0.37	0.50	3.9	**0.33**	0.49	0.36
	F	**8.0**	12.8	8.6	22.5	24.1	227	**8.0**	17.3	13.0

line are from the literature and were trained on the original MD17 data set using different train test splits.

Of the descriptor based models, sGDML, FCHL and linear ACE have the lowest MAE for some molecules. Overall, based on the per-atom energy and force errors, the ACE model achieves the lowest errors averaged across the entire data set. It is interesting to note that of the neural network models, the PaiNN equivariant neural network achieves very low force errors, but its energy errors are almost three times higher compared to ACE and FCHL. In the second half of the thesis, the newer generations of equivariant neural network force fields will be discussed and assessed in detail showing that they are capable of also fitting the energy to very high accuracy. It is important to note that it is crucial for a model to have a low energy error, especially in the relative energy of different molecular geometries, because even though a molecular dynamics trajectory is only affected directly by the forces, the stationary probability distribution that MD is used to sample is solely a function of the energy through the Boltzmann weight, and so errors in predicted energy translate into errors of the stationary distribution and thus of all equilibrium observables.

The ANI model in this table refers to a fine-tuning of the published ANI-2x model [58]. This means that the weights of the neural network were initialised as the weights of the

ANI-2x model. Using the pre-trained model was crucial for achieving the errors shown. When the weights were randomly initialised, the errors were higher by a factor of 2. The GAP model, using SOAP features, which are similar 3-body objects to ANI's symmetry functions, achieves comparable errors to the ANI model with pre-training. The fact that ANI is only competitive with GAP if it is pre-trained can be rationalised by the relative sample efficiency of kernel models compared to feedforward neural networks. Just like SOAP and ANI, FCHL kernel models also use 2- and 3-body features, which have been designed and tuned for molecular systems and are therefore able to achieve very low errors [42].

The classical force field (FF) refers to a reparametrization of the GAFF functional form [220, 221] using the ForceBalance program[221, 222] and the rMD17 training set. This model gives at least an order of magnitude higher errors compared to the ML force fields. This is not a surprise given the very simple harmonic functional form. These results provide a quantitative characterisation of the limitations of the functional form.

Learning Curves The first property to consider beyond the raw energy and force errors is the learning curve, which shows how a model's performance improves with additional training data. For kernel models such as FCHL and sGDML, the "kernel basis" grows precisely together with the training data. In contrast, for the linear ACE potentials the size of the training set and the size of the basis are decoupled. This has the advantage that the evaluation cost is independent of the size of the training set. By choosing a finite basis set through the specification of the maximum body order and polynomial degree, the linear ACE potentials can be tuned to trade off computational cost and accuracy. To motivate the particular choices made here, Figure 3.2 shows the force accuracy of linear ACE as a function of basis set size and the corresponding evaluation time, trained on 1,000 azobenzene configurations, the largest molecule in MD17.

The timings were obtained using a single 2.3 GHz Intel Xeon Gold 5218 CPU. For context, the accuracy and evaluation time of some of the other ML models is also shown, each called in their native environment: ACE in Julia, GAP via the fortran executable, and sGDML and ANI directly from their respective Python packages (TorchANI in the case of ANI [73]). Note that in the case of ANI considerable speed up could be achieved using a GPU when multiple molecules are evaluated simultaneously. Recently, further optimisation of the implementation of ANI has been reported that led to increased computational performance [72]. The solid part of the ACE curve corresponds to 4-body potentials ($\nu = 3$) where only the polynomial degree is varied, while for the last point (dashed), the body order was increased to 5, because the 4-body part of the curve showed saturating accuracy. Increasing the body order further is likely to bring the error down even more; however, the cost of evaluation would also grow

unacceptably if all basis functions for the given body and polynomial degree are retained. Effective sparsification strategies can be developed that allow the inclusion of some high body order basis functions without the concomitant very large increase of the overall basis set size. One such approach, generating the high body order basis functions via message passing is the subject of Chapter 5. For the comparisons in this section, for each molecule in MD17 a basis set size was selected such that the evaluation cost was roughly comparable to the other ML models. Note, however, that in a real ML force field application, one might very well choose a much smaller basis to take advantage of sub-millisecond evaluation times.

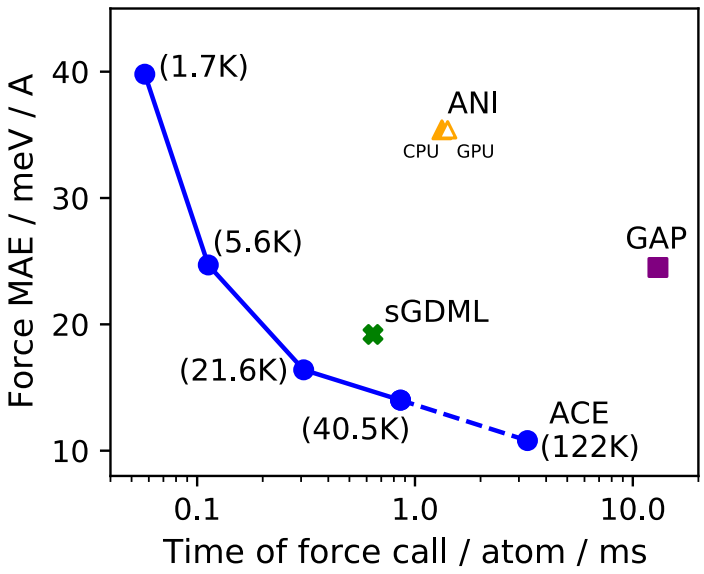


Fig. 3.2 **Force evaluation times.** The timing of force calls per atom for the azobenzene molecule. In the case of ACE the number of basis functions is shown in parentheses. The classical force field has a timing of about 1 μs, which would not fit on this scale. For the ANI architecture both the CPU and GPU timings are shown obtained from TorchANI [73]. Better performance is likely possible using the NNPOPS package and OpenMM [72].

In Figure 3.3 the learning curves for linear ACE and sGDML (the best models trained from Table 3.1) are shown and compared to the literature results of FCHL [41]. The low body order linear ACE is equal or better than the other many-body kernel models in the low data limit, but with additional training data the kernel models overtake ACE in several cases. The latter also saturates, showing the limitations of the relatively low body order model. The energy and force learning curves show a broadly similar trend. Interestingly, the force errors show a less pronounced saturation for ACE.

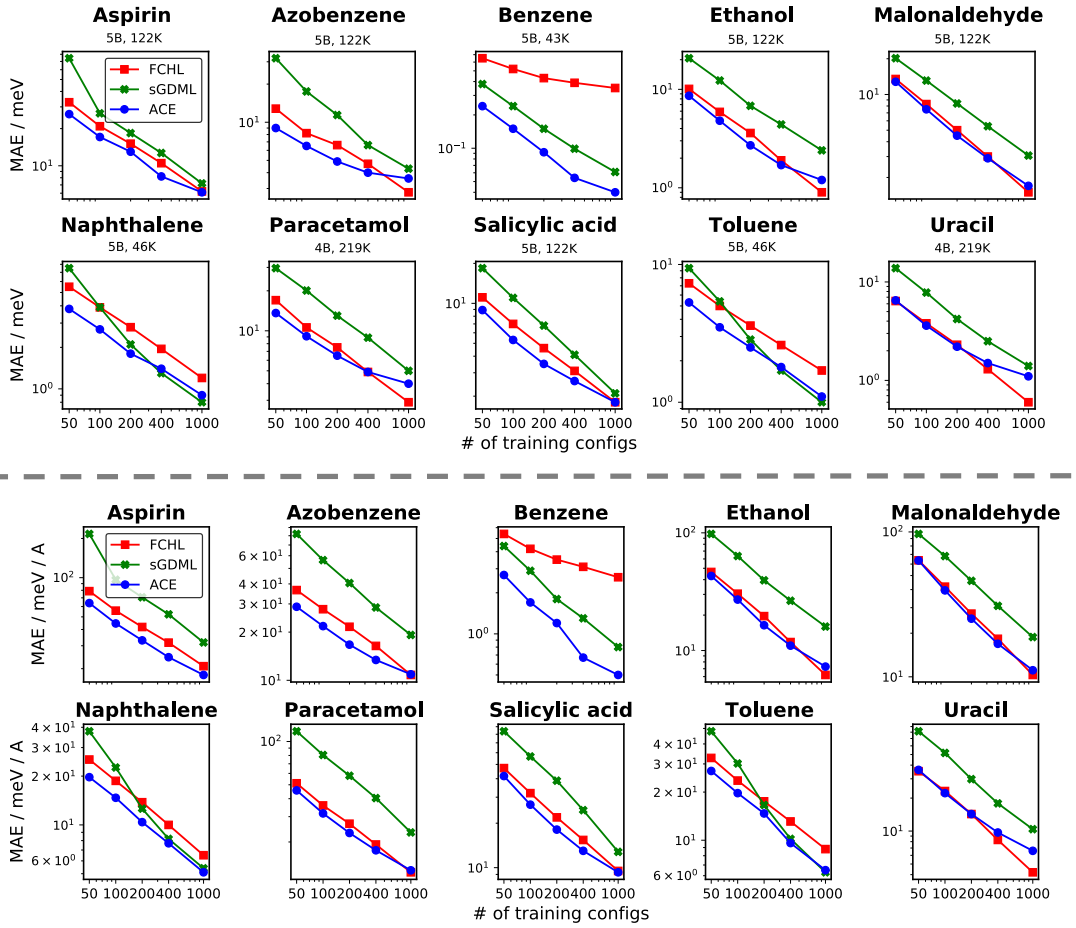


Fig. 3.3 **rMD17 learning curves.** The learning curves of the best performing models on the rMD17 data set. The body order and basis set size for the ACE models are given under the title of each panel. The top panel shows the energy errors, whereas the bottom panel shows the force errors.

Normal Model Analysis The normal modes and their corresponding vibrational frequencies characterise the potential energy surface near equilibrium. This is interesting in the context of the MD17 models because their training set contains geometries sampled at 500 K, which means that they are, in general, far from the equilibrium geometry. This test evaluates the ability of the models to describe the minima of the PES, even if it is not in the training set.

To test how well the different models infer the normal modes, the DFT optimised geometries were taken of each of the 10 molecules and re-relaxed with the force field models. At the force field minima, vibrational analysis was carried out to find the normal modes and their corresponding vibrational frequencies. They were computed as the eigenvectors and eigenvalues of the Hessian matrix obtained by numerical differentiation.

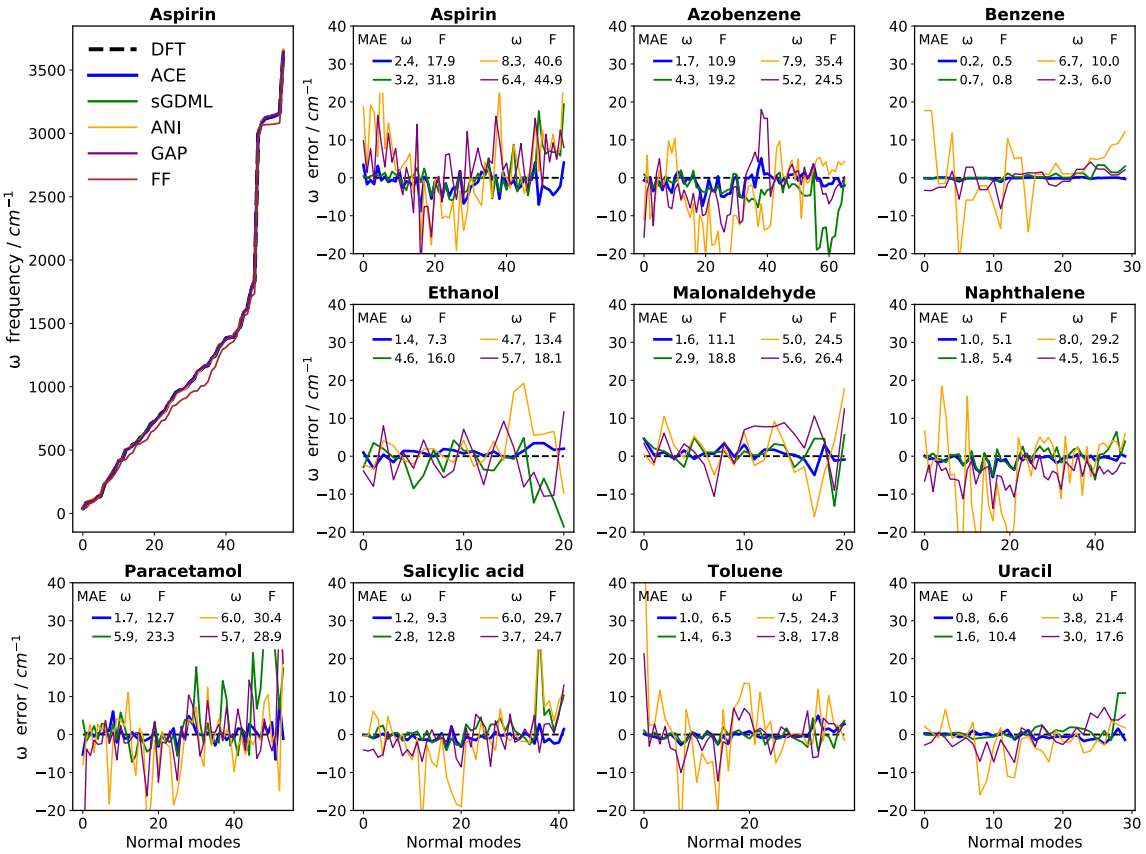


Fig. 3.4 **Normal mode frequency test.** The frequency error of the normal modes of each of the MD17 molecules. The legend shows the frequency (ω) MAE (in cm^{-1}) and also the force (F) MAE (in meV/Å) from Table 3.1 for each model

Fig 3.4 shows the errors in the predicted normal mode vibrational frequencies for each of the 10 MD17 molecules. The ACE model achieves the lowest error for all 10 molecules, surprisingly even for those for which sGDML has slightly lower errors based on the 500 K MD test set of Table 3.1. For example, for toluene, sGDML has both lower energy and force errors, but at the same time the ACE model has significantly lower errors in predicting the vibrational frequencies, achieving a MAE of 1.0 cm^{-1} compared to sGDML with an error of 1.4 cm^{-1}. Observing the individual molecules in Fig 3.4 it is notable that the ACE model has the lowest fluctuation in the errors of the normal modes, achieving nearly uniform accuracy across the entire spectrum. The case of benzene also shows the limitations of characterising the models by the force MAE alone. The linear ACE model has only slightly lower force MAE than sGDML (0.5 meV/Å compared to 0.8 meV/Å) but the normal mode frequency prediction is more than 3 times more accurate: 0.2 cm^{-1} compared to 0.7 cm^{-1}. The linear

ACE model has very low errors for all normal modes, whereas sGDML has much higher errors for the high frequency modes.

The ML models are also compared to the classical force field. The normal mode frequency errors of the empirical FF are about 10 times higher than the errors of the ML force fields. These errors do not fit on the scale of Figure 3.4.

Extrapolation test To test the extrapolation properties of the different models, two further tests were carried out probing the torsional profile of azobenzene and O-H bond breaking in ethanol. Both of these tests probe how far away the models can smoothly extrapolate from the training data.

These tests were carried out with several different versions of the linear ACE models differing in the definition of their 1-body terms because this choice is expected to influence how chemically reasonable the fitted models are far from the training set. The energy and force errors of these models are compared in Table 3.2

ACE models fitted only using force data are denoted by `ACE F`. This has the lowest force error on the test set. For the other two ACE models, energies were also included in the training. They differ only in the 1-body term, the model using average per-atom training set energy is denoted as `ACE AVG`, while the model using isolated atom energies as the 1-body term is denoted as `ACE E0`. The third option is the natural choice, as this ensures that if all atoms are separated from each other, the predicted energy will correctly correspond to the sum of the isolated atom energies.

Figure 3.5(a) shows the torsional energy profile of the azobenzene molecule. The `ACE E0` model with the isolated atom 1-body term is able to extrapolate farthest, somewhat overestimating the energy, while the ANI and sGDML models also extrapolate smoothly, but slightly underestimate the energy. The linear ACE model with the average energy 1-body term and the GAP model fail to extrapolate and predict a nonphysical drop in energy for smaller values of the dihedral angle.

Figure 3.5(b) shows the energy profile as the O-H distance is varied starting from the equilibrium geometry of ethanol. The only force field that shows qualitative agreement with DFT is the `ACE E0` model. Note that none of the fitted models is expected to quantitatively reproduce the DFT energy profile, even when the isolated H atom is described correctly by design, because the $C_2H_5O^\bullet$ radical is not. The smooth extrapolation of the linear ACE model can probably be attributed to careful regularisation - as was the case in a similar test for other polynomial models [2]. Figure 3.5(c) shows a detailed comparison of the different ACE models together with their test set MAE value. This shows that having the lowest possible test set error does not coincide with the most physically reasonable model, and

Table 3.2 **Comparison of different 1-body terms** Comparison of the mean absolute error of energies (meV) and forces (meV / Å) of ACE models trained on energies and forces using the average energy shift, the isolated atom QM energy shift and trained on forces only and then shifted to minimize training energy error.

		ACE with iso E0	**ACE with average E0**	**ACE with forces only**
Aspirin	Energy	6.1	**5.9**	6.1
	Force	19.1	18.7	**17.9**
Azobenzene	Energy	**3.4**	3.5	3.6
	Force	12.8	14.8	**10.9**
Benzene	Energy	0.04	0.04	0.04
	Force	0.5	0.5	0.5
Ethanol	Energy	1.4	1.4	**1.2**
	Force	8.4	8.2	**7.3**
Malonaldehyde	Energy	1.9	1.8	**1.7**
	Force	12.0	11.5	**11.1**
Naphthalene	Energy	0.9	0.9	0.9
	Force	**5.1**	5.2	**5.1**
Paracetamol	Energy	4.0	**3.8**	4.0
	Force	14.9	13.8	**12.7**
Salicylic acid	Energy	2.3	2.2	**1.8**
	Force	11.2	10.5	**9.3**
Toluene	Energy	1.1	1.1	1.1
	Force	6.7	6.7	**6.5**
Uracil	Energy	1.4	1.2	**1.1**
	Force	8.7	7.8	**6.6**

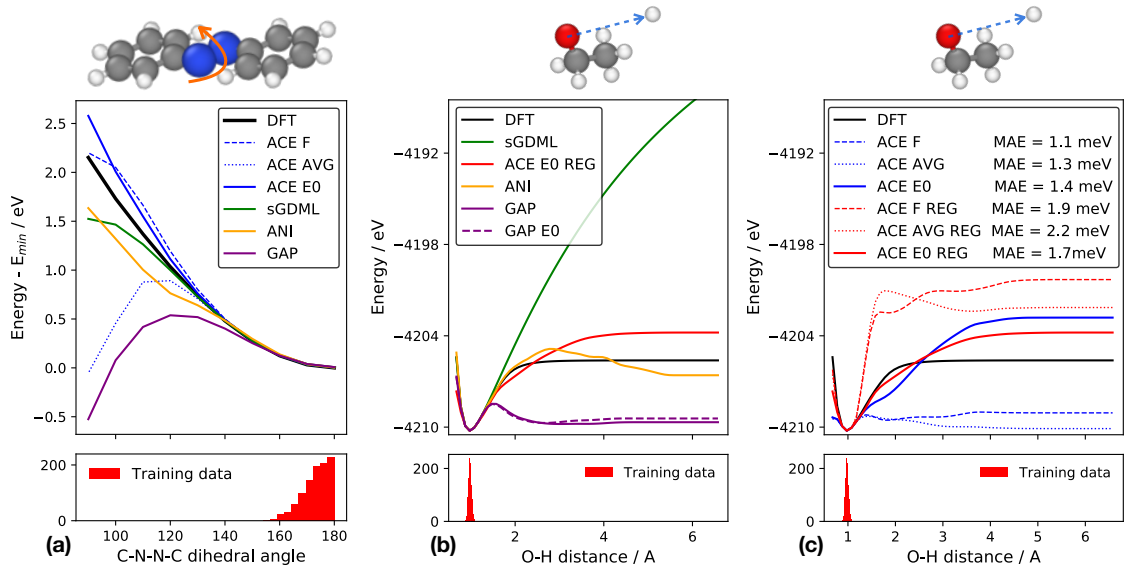


Fig. 3.5 **Extrapolation test far away from training data.** The bottom panels show the histogram of the variables in the training set. **(a)** the energy change predicted by the different models as the C-N-N-C dihedral angle of azobenzene is decreased from the equilibrium 180 degrees. `ACE F` refers to training on forces only, `ACE AVG` refers to using average per atom energy as the 1-body term and `ACE E0` refers to using the isolated atom energy as the 1-body term. **(b)** change in energy as the O-H bond distance of ethanol is extended from equilibrium, as predicted by the different models. **(c)** comparison of ACE models with (i) lowest force MAE and (ii) with stronger regularisation, the latter indicated with the `REG` label.

using stronger regularisation can lead to much smoother extrapolation. The more strongly regularised ACE models with relatively higher force error are still significantly more accurate than sGDML, ANI, GAP or the classical force field.

Interestingly, having the isolated atom as the 1-body term is not sufficient for good extrapolation. This is shown by the two different GAP models in Figure 3.5(b), which show essentially no difference to the extrapolation, presumably due to the very poor description of the radical.

3BPA

As a final test of the linear ACE method, it is evaluated on the newly introduced 3BPA benchmark described in Section 2.5. Noting that all MD17 molecules are rather rigid, this test can assess the capabilities of the different force field models on a more challenging system that has relevance for medicinal chemistry applications. Although smaller than many drug molecules, with a molecular weight of 200, this molecule has three consecutive rotatable

Table 3.3 **Comparison of errors on the 3BPA benchmark** Root mean squared error of the energy (meV) and force (meV/Å) predictions of different models of the flexible 3BPA molecule.

		ACE	sGDML	GAP	FF	ANI	ANI-2x
Fit to 300K							
300 K	E	**7.1**	9.1	22.8	60.8	23.5	38.6
	F	**27.1**	46.2	87.3	302.8	42.8	84.4
600 K	E	**24.0**	484.8	61.4	136.8	37.8	54.5
	F	**64.3**	439.2	151.9	407.9	71.7	102.8
1200 K	E	85.3	774.5	166.8	325.5	**76.8**	88.8
	F	187.0	711.1	305.5	670.9	**129.6**	139.6
Fit to mixed-T							
300 K	E	**9.3**	11.7	27.7	86.0	21.6	38.6
	F	**30.5**	55.1	85.7	307.8	56.4	84.4
600 K	E	**19.7**	25.6	50.2	115.6	40.3	54.5
	F	**54.4**	94.3	123.8	392.0	81.4	102.8
1200 K	E	**47.1**	78.2	103.0	268.8	77.6	88.8
	F	**118.4**	177.1	217.8	634.3	131.0	139.6

bonds, as was shown in Figure 2.2. This leads to a complex dihedral potential energy surface with many local minima, which can be challenging to approximate using classical or ML force fields [216].

Comparison of force fields models Linear ACE, sGDML, ANI and GAP force fields were trained, and the bonded terms of a classical force field (FF) were reparameterised, using the 300 K and the mixed-T training sets. Table 3.3 shows the energy and force RMSEs of the different models along with the transferable pre-trained ANI-2x force field errors on the same configurations. Just as before, the ANI label refers to a fine tuned version of the ANI-2x force field.

For the case of training on the 300 K configurations the linear ACE and sGDML models are able to achieve very low errors when tested at the same temperature, but the ACE model shows significantly better extrapolation properties to the configurations sampled at higher temperatures. The model extrapolating most accurately to 1200 K is the fine tuned ANI force field. Just as for the smaller molecules, the fitted empirical force field shows much higher errors, about a factor of 2-4 for energies and a factor of 4 for forces compared with the ANI-2x machine learning force field.

Training on the mixed-T training set leads to a significant drop in the errors for the higher temperature test sets for all ML models, but not for the empirical force field. The linear ACE model achieves the lowest error in all cases, showing an approximately 40% decrease in the

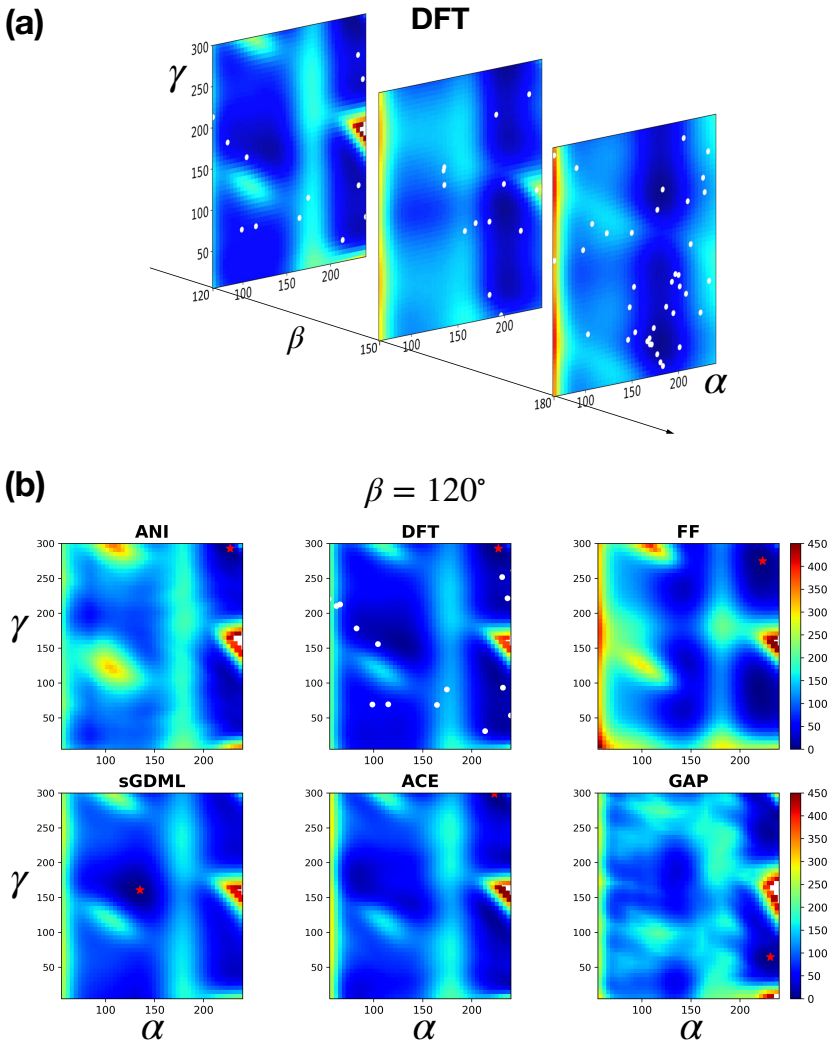


Fig. 3.6 **Dihedral PES of 3BPA.** (a) The dihedral potential energy landscape of 3BPA for different fixed β, as predicted by DFT. (b) The $\beta = 120^o$ section of the PES for the different force field models. The white dots on the DFT PES correspond to configurations from the training set that lie within $\pm 10^o$ of the planes considered here and the red star shows the position of the energy minimum on each slice.

error for the high temperature test set. The other ML models also improve, by even bigger factors because their extrapolation was rather poor before. Custom force fields outperform the transferable ANI-2x model by nearly a factor of two in energies for all three test sets. This shows that there is scope for much improvement in transferable organic force fields for small molecules. The errors in the empirical force field are mostly unchanged, quantifying the limitations of the simple functional form when describing the anharmonic high energy parts of the potential energy surface.

To look beyond the energy and force RMSE, dihedral torsional scans were carried out using the different force field models and DFT. The complex energy landscape is visualised in Figure 3.6. The Greek letters denote the three rotatable dihedral angles as introdcued on Figure 2.2. Figure 3.6(a) visualises the PES at three different fixed values of β, in the α-γ plane, limiting the range to avoid overlapping atoms. Figure 3.6(b) shows a comparison of ML and empirical force fields with DFT for the case of $\beta = 120^\circ$, the plane with the fewest training data points. The energy landscape of the empirical force field has most of the features of the DFT landscape and even correctly predicts the position of the lowest energy minimum in the $\beta = 120^\circ$ plane. However, some of the potential energies on this plane are clearly too high. On the other hand, the landscape of the GAP model is quite irregular; some of the most basic features are either missing or blurred together. The fine-tuned ANI landscape is somewhat less irregular than GAP, but some of the high energy peaks are too high and too broad. This is an example where the fixed functional form of the classical force field gives better extrapolation behaviour to parts of the configuration space where there is little training data. The RMSE results clearly do not give a full characterisation of these models.

The ACE and sGDML models reproduce the landscape much more closely, and indeed these are the models with the lowest RMSE. Some differences include the sGDML getting the position of the lowest energy minimum wrong and ACE having too high a peak at $\alpha = 230^\circ$, $\gamma = 150^\circ$.

3.2.3 Conclusions About Linear ACE

This section has demonstrated great potential in using higher body order models built using the ACE framework for the simulation of molecular chemistry. The large improvements compared to the SOAP-GAP and ACSF feedforward neural network type ML models served as a guide to building even more accurate force fields in the next chapters of the thesis.

There are three key limitations or outstanding challenges that need to be addressed. First, the ACE model, similarly to SOAP and ACSF-s, cannot be scaled for a large number of chemical elements as the size of the model increases steeply. This is addressed in a general new framework in the next section. Second, the fitting of linear ACE models can be quite fiddly, as the polynomials can be highly oscillatory, requiring carefully tuned regularisation for each new system. Finally, though the accuracy of the linear ACE is a large improvement compared to the methods before, there is still great scope to improve the accuracy of these models, as will be demonstrated later in Chapter 5.

3.3 Tensor Reduced Atomic Cluster Expansion

In this section, the first limitation of linear ACE models is addressed: the unfavourable scaling of the number of features with the number of different chemical elements, S. This limitation is not a special feature of ACE; indeed, most of the different classes of atomic representations used to create machine learning force fields have this same problem with the exception of graph neural network based models. In the case of ACE, the number of features scales as S^ν for basis functions with correlation order ν (i.e., a body order of $\nu+1$). This poor scaling severely restricts the use of these representations in many applications. For example, when using the representations to build machine learning force fields for systems with many (more than 5) different chemical elements, the large size of the models results in memory limitations being reached during parameter estimation, as well as significantly reducing the evaluation speed.

Multiple strategies have been proposed to address this scaling problem for the environment descriptor based models. This includes element weighting [6, 75] or embedding elements into a fixed dimensional space [228], directly reducing element-sensitive correlation order [46], low-rank tensor-train approximations for lattice models [117] and data-driven approaches for selecting the most relevant subset or combination of the original features for a given dataset [82, 152, 238].

A different class of machine learning methods are graph neural networks or MPNNs that were introduced in Section 2.4.2 [79, 187]. Instead of constructing full tensor products, these models embed chemical element information into a fixed size latent space, using a learnable transformation $\mathbb{R}^S \to \mathbb{R}^K$ where K is the dimension of the latent space. This construction avoids the poor scaling with the number of chemical elements. As will be demonstrated in Chapter 5, MPNN methods can achieve very high accuracy [14, 16, 17], strongly suggesting that the true complexity of the relevant chemical element space does not grow as S^ν.

In this section, two approaches, that together can be referred to as Tensor-reduced ACE or TrACE, are introduced for significantly reducing the scaling of atomic density-based representations like SOAP and ACE. The key idea is to exploit the tensor structure of these descriptors to either i) approximate the parameters of a linear model with a tensor decomposition or ii) use tensor sketching [229] to compress the features. Both of the resulting approximations are systematically improvable and can be converged to the original full descriptor limit. This is supported by a number of numerical experiments on real data in Section 3.3.2. It is also possible to generalise this scheme to compress not only the chemical element information but also the radial degrees of freedom, yielding an even more compact representation. When fitting interatomic potentials for organic molecules a ten-fold reduction in the number of features required is demonstrated for the linear ACE model. In Ref. [47]

the method is also demonstrated for SOAP GAP models using a high entropy alloy dataset showing a similar reduction in the size of the models.

3.3.1 Methods

All many-body density based descriptors can be understood in terms of the Atomic Cluster Expansion [61]. To derive the new compressed descriptors, it is easiest to rewrite the ACE equations emphasising the chemical element dependence. The one-particle basis $\phi_{nlm}^{z_i z_j}(\mathbf{r})$ from Equation 3.2 can be rewritten to $\phi_{znlm}(\mathbf{r}_{ij}, Z_j)$ as shown in Equation (3.20), where $\mathbf{r}_{ij}$ denotes the relative position of the central atom i and the neighbour j, while z and Z_j denote the atomic number of atoms i and j respectively. Following the ACE recipe, permutation invariance is introduced by summing over the neighbour atoms in Equation (3.21) after which $(\nu+1)$-body features are formed in Equation (3.22) by taking tensor products of the atomic basis $A_{i,znlm}$ with itself ν times. Finally, Equation (3.23) shows how the product basis $\mathbf{A}_{i,\mathbf{znlm}}$ is rotationally symmetrised using the generalised Clebsch-Gordon coefficients $C_{\mathbf{m}}^{\mathbf{l}\eta}$, where η enumerates all possible symmetric couplings [61, 63, 152].

$$\phi_{znlm}(\mathbf{r}_{ij}, Z_j) = R_n(r_{ij}) Y_l^m(\hat{\mathbf{r}}_{ij}) \delta_{zZ_j}, \tag{3.20}$$

$$A_{i,znlm} = \sum_{j \in \mathcal{N}(i)} \phi_{znlm}(\mathbf{r}_{ij}, Z_j), \tag{3.21}$$

$$\mathbf{A}_{i,\mathbf{znlm}} = \prod_{t=1}^{\nu} A_{i,z_t n_t l_t m_t} \tag{3.22}$$

$$\mathbf{B}_{i,\mathbf{znl}\eta} = \sum_{\mathbf{m}} C_{\mathbf{m}}^{\mathbf{l}\eta} \mathbf{A}_{i,\mathbf{znlm}} \tag{3.23}$$

Rewriting Equation (3.7) to display the indices of the parameter tensor c, a linear ACE model of the atomic site energy E_i is

$$E_i = \sum_{\mathbf{znl}\eta} c_{\mathbf{znl}\eta} \mathbf{B}_{i,\mathbf{znl}\eta} \tag{3.24}$$

where $c_{\mathbf{znl}\eta}$ are the model parameters and in practice the expansion is truncated using the maximum correlation order $\nu_{\max}$, and polynomial degrees $l_{\max}$ and $n_{\max} = N$. Note that since $\mathbf{B}_{i,\mathbf{znl}\eta}$ is invariant under $(z_a, n_a, l_a) \leftrightarrow (z_b, n_b, l_b)$, symmetrically equivalent terms are usually omitted from Equation (3.24).

The tensor product in Equation (3.22) causes the number of features and therefore the number of model parameters to grow rapidly as $\mathcal{O}(N^\nu S^\nu)$. Previous work has reduced this

scaling by first embedding the chemical and radial information into K channels as shown in Equation (3.25), then taking a full tensor product across the $\bar{\mathbf{A}}_{i,\mathbf{klm}}$ in Equation (3.26). This leads to a $\mathcal{O}(K^\nu)$ scaling. The embedding weights are optimised either before or during fitting, the latter causing the models to be non-linear [228].

$$\bar{A}_{i,klm} = \sum_{zn} W^k_{zn} A_{i,znlm} \tag{3.25}$$

$$\bar{\mathbf{A}}_{i,\mathbf{klm}} = \prod_{t=1}^{\nu} \bar{A}_{i,k_t l_t m_t} \tag{3.26}$$

In this section, two principled approaches are proposed to further reduce the size of the basis to $\mathcal{O}(K)$. In the first approach, the model parameters $\mathbf{c}_\eta \equiv c_{\mathbf{znl}\eta}$ in Equation (3.24) are identified as elements of a symmetric tensor, invariant under $(z_a, n_a, l_a) \leftrightarrow (z_b, n_b, l_b)$, which can be expanded as a sum of products of rank-1 tensors as,

$$\mathbf{c}_\eta = \sum_{k=1}^{K} \lambda_{k\eta} \underbrace{\mathbf{w}_k \otimes \mathbf{w}_k \cdots \otimes \mathbf{w}_k}_{\nu \text{ times}}, \tag{3.27}$$

or equivalently in component form

$$c_{\mathbf{znl}\eta} = \sum_{k}^{K} \lambda_{k\eta} \prod_{t=1}^{\nu} W^k_{z_t n_t l_t} \tag{3.28}$$

where $W^k_{z_t n_t l_t}$ are the components of $\mathbf{w}_k$. This expansion is exact for finite K, as $\mathbf{c}$ is finite due to basis truncation, and is equivalent to eigenvalue decomposition of a symmetric matrix when $\nu = 2$. Note that the same weights $W^k_{z_t n_t l_t}$ were chosen to be used for all ν and η, which significantly reduces the number of free parameters of this model. In practice, one can choose to expand over the **zn** (or **z**) indices only, and then substitute the expansion into Equation (3.24) as

$$E_i \approx \sum_{k\mathbf{l}\eta} \lambda_{k\mathbf{l}\eta} \left[\sum_{\mathbf{m}} C^{\mathbf{l}\eta}_{\mathbf{m}} \sum_{\mathbf{zn}} \prod_{t=1}^{\nu} W^{kl_t}_{z_t n_t l_t} A_{i,z_t n_t l_t m_t} \right] \tag{3.29}$$

$$= \sum_{k\mathbf{l}\eta} \lambda_{k\mathbf{l}\eta} \left[\sum_{\mathbf{m}} C^{\mathbf{l}\eta}_{\mathbf{m}} \prod_{t=1}^{\nu} \tilde{A}_{i,kl_t m_t} \right] \tag{3.30}$$

$$= \sum_{k\mathbf{l}\eta} \lambda_{k\mathbf{l}\eta} \tilde{\mathbf{B}}_{i,k\mathbf{l}\eta} \tag{3.31}$$

Table 3.4 The density projection $\vec{A}_i$ are viewed as vectors with a composite index (z,n,l,m) whereas the embedded density projections $\bar{A}_i = (\mathbf{W}\vec{A}_i)$ etc. are indexed by (k,l,m). The symbol $\otimes_k^{lm}$ means full tensor product across l and m but element-wise product across k whereas $\otimes$ indicates a full tensor product across all indices.

Name	Product Basis	Index Notation	Basis size
ACE	$\mathbf{A}_i = \vec{A}_i \otimes \vec{A}_i \cdots \otimes \vec{A}_i$	$\mathbf{A}_{i,\mathbf{znlm}} = \prod_{t=1}^{\nu} A_{i,z_t n_t l_t m_t}$	$\mathcal{O}(NS)^\nu$
Embedding	$\bar{\mathbf{A}}_i = (\mathbf{W}\vec{A}_i) \otimes (\mathbf{W}\vec{A}_i) \cdots \otimes (\mathbf{W}\vec{A}_i)$	$\bar{\mathbf{A}}_{i,\mathbf{klm}} = \prod_{t=1}^{\nu} \bar{A}_{i,k_t l_t m_t}$ $\bar{A}_{i,k_t l_t m_t} = \sum_{zn} W_{zn}^{k_t} A_{i,znl_t m_t}$	$\mathcal{O}(K^\nu)$
Tensor decomposition	$\tilde{\mathbf{A}}_i = (\mathbf{W}\vec{A}_i) \otimes_k^{lm} (\mathbf{W}\vec{A}_i) \ldots \otimes_k^{lm} (\mathbf{W}\vec{A}_i)$	$\tilde{\mathbf{A}}_{i,k\mathbf{lm}} = \prod_{t=1}^{\nu} \tilde{A}_{i,kl_t m_t}$ $\tilde{A}_{i,kl_t m_t} = \sum_{zn} W_{zn}^{kl_t} A_{i,znl_t m_t}$	$\mathcal{O}(K)$
Tensor sketch	$\hat{\mathbf{A}}_i = (\mathbf{W}^1\vec{A}_i) \otimes_k^{lm} (\mathbf{W}^2\vec{A}_i) \ldots \otimes_k^{lm} (\mathbf{W}^\nu\vec{A}_i)$	$\hat{\mathbf{A}}_{i,k\mathbf{lm}} = \prod_{t=1}^{\nu} \hat{A}_{it,kl_t m_t}$ $\hat{A}_{it,kl_t m_t} = \sum_{zn} W_{zn}^{tk} A_{i,znl_t m_t}$	$\mathcal{O}(K)$

where $\tilde{\mathbf{B}}_{i,k\mathbf{l}\eta}$ are the new tensor reduced features and the approximation arises because in practice the tensor decomposition can be truncated early. The key novelty of this approach is that element-wise products are taken across the k index of the embedded channels $\tilde{A}_{i,kl_t m_t}$, rather than a full tensor product. For a detailed comparison of the different approaches, see Table 3.4.

There are multiple natural strategies for specifying the embedding weights W_{zn}^{kl}, including approximating a precomputed $c_{\mathbf{znl}\eta}$ or treating the weights as model parameters to be estimated during the training process, as is done in MACE (see Chapter 5). In this section, the use of random weights is investigated, which is a simpler alternative. This ensures that Equation 3.31 remains a linear model and allows $\tilde{\mathbf{B}}_{i,k\mathbf{l}\eta}$ to be used directly in other, not force field fitting tasks, such as data visualisation.

The second strategy developed by my co-author in Ref. [46] to reduce the scaling with the number of elements is to compress the ACE features $\mathbf{B}_{i,\mathbf{znl}\eta}$ directly. Random Projection (RP) [22, 49] is an established technique where high dimensional feature vectors $\{\vec{x}_1,\ldots,\vec{x}_N\} \subset \mathbb{R}^d$ are compressed as $\tilde{x}_i = \mathbf{W}\vec{x}_i \in \mathbb{R}^K$, with the entries of the matrix $\mathbf{W}$ being normally distributed. This simple approach offers a tuneable level of compression and is underpinned by the Johnson-Lindenstrauss Lemma [106] which bounds the fractional error made in approximating $\vec{x}_i^T\vec{x}_j$ by $\tilde{x}_i^T\tilde{x}_j$. This is used in compressed linear regression [1, 109, 139] where features are replaced by their projections, thus reducing the number of model parameters. The drawback of RP is that it requires the full feature vector to be constructed, which means that applying RP to ACE would not avoid the unfavourable $\mathcal{O}(N^\nu S^\nu)$ scaling. Tensor sketching can be used instead of RP to avoid this problem. For vectors with tensor

structure $\vec{x} = \vec{y} \otimes \vec{z}$ where $\vec{x} \in \mathbb{R}^{d_1 d_2}$, $\vec{y} \in \mathbb{R}^{d_1}$ and $\vec{z} \in \mathbb{R}^{d_2}$, the Random Projection $\mathbf{W}\vec{x}$ can be efficiently computed directly from $\vec{y}$ and $\vec{z}$ as

$$\mathbf{W}\vec{x} = \mathbf{W}'\vec{y} \circ \mathbf{W}''\vec{z}. \tag{3.32}$$

where $\circ$ denotes the element-wise (Hadamard) product. Similarly, the ACE product basis can be tensor sketched, across the **zn** indices, as

$$\hat{\mathbf{A}}_i = (\mathbf{W}^1\vec{A}_i) \otimes_k^{lm} (\mathbf{W}^2\vec{A}_i) \dots \otimes_k^{lm} (\mathbf{W}^\nu\vec{A}_i) \tag{3.33}$$

where $\otimes_k^{lm}$ denotes taking the tensor product over the upper indices lm and the element-wise product over the lower index k and $\mathbf{W}^1, \mathbf{W}^2$ etc. are independent random matrices. The $\hat{\mathbf{A}}_{i,k\mathbf{lm}}$ can then be symmetrised as in Equation 3.23 yielding

$$\hat{\mathbf{B}}_{i,k\mathbf{l}\eta} = \sum_m C_{\mathbf{m}}^{\mathbf{l}\eta} \prod_{t=1}^{\nu} \hat{A}_{it,kl_tm_t} \tag{3.34}$$

where $\hat{A}_{it,kl_tm_t}$ is defined in Table 3.4. A summary comparing standard ACE, embedded ACE and the two new proposed schemes, tensor decomposition and tensor sketching, is given in Table 3.4.

3.3.2 Numerical Experiments

Linear ACE design matrix rank First it is demonstrated that the tensor-reduced features are able to efficiently and completely describe a many-element training set. This is achieved by considering a dataset comprised of all symmetry inequivalent fcc structures made up of 5 elements with up to 6 atoms per unit cell [93]. A set of features is considered complete on this dataset if the design matrix for a linear model fit to total energies has full (numerical) row rank, where each row corresponds to a different training configuration.

Figure 3.7 shows the numerical rank of the design matrix as a function of the basis set. At a given correlation order, the standard ACE basis set is grown by increasing the polynomial degree, and the tensor-reduced basis set is enlarged by increasing K, the number of independent channels. In both cases, once the rank stops increasing at the given correlation order ν is incremented. The colours in Figure 3.7 correspond to three different geometrical variations: blue contains on-lattice configurations only, whilst in magenta and red the atomic positions have been perturbed by a random Gaussian displacement with mean 0 and standard deviation of 0.025 and 0.25 Å, respectively. The dotted lines correspond to the standard ACE basis, whereas the solid lines corresponds to the tensor-reduced version from Equation (3.31).

Although the standard ACE basis can always achieve full row rank since it is a complete linear basis, it does this very inefficiently. In contrast, the row rank using the tensor-reduced basis grows almost linearly. This demonstrates that the tensor-reduced basis, having removed unnecessary redundancies, still retains the expressive power of the full basis.

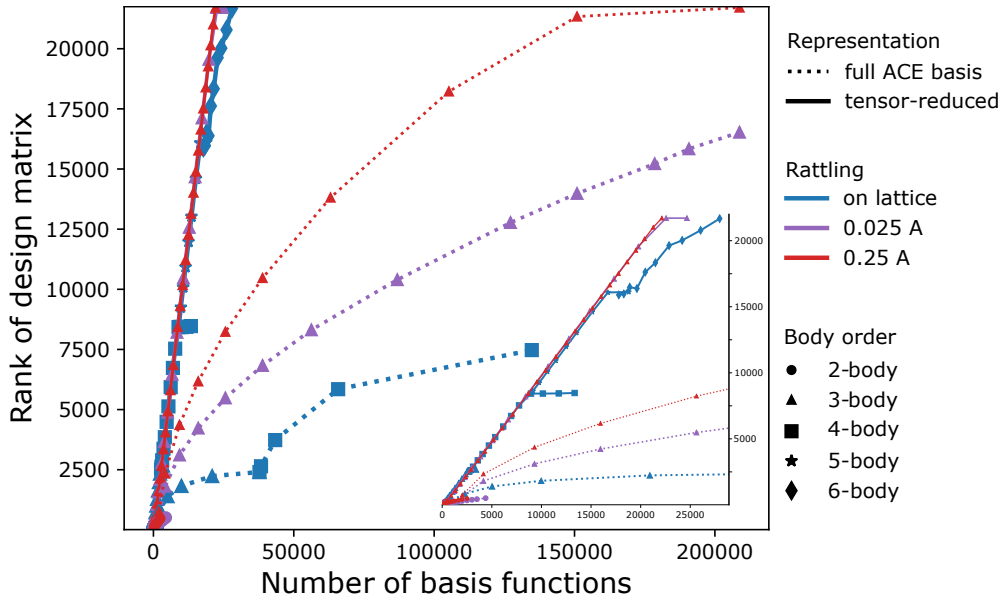


Fig. 3.7 **Design matrix rank** The row rank of the design matrix as a function of basis set size on a dataset of all symmetry inequivalent fcc lattices of 5 chemical elements and unit cell sizes of up to 6 atoms. The inset zooms in on the $x = y$ region.

Organic molecules with 10 elements Next, a linear ACE [119] model is fitted on a training set of 400 different organic molecules of size 19-168 atoms, randomly selected from the QMugs dataset [104]. The molecules are made up of 10 different chemical elements (H, C, N, O, F, P, S, Cl, Br, I). The conformers were created by running 800 K NVT molecular dynamics for 1 ps starting from a published minimum energy structure using the semi-empirical GFN2-xTB method [9]. The test set is composed of 1,000 different molecules sampled the same way. This is a small-data regime task that is particularly challenging because of the chemical and conformational diversity. Figure 3.8 shows the convergence of the energy error with the number of basis functions for the fully coupled and the tensor-reduced ACE models, both using $\nu_{\max} = 3$ (4-body). By increasing the number of uncoupled channels K the accuracy can be converged to the previous level, whilst reducing the size of the model by a factor of 10.

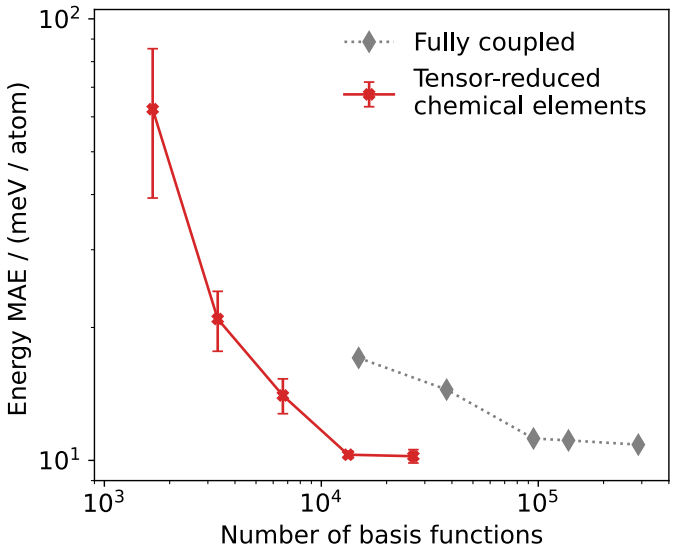


Fig. 3.8 **Energy error convergence of TrACE to ACE** Convergence of the energy errors on the independent test set with respect to the number of basis functions is shown for a linear tensor-reduced ACE model and a standard linear ACE model. Error bars show the standard error in the mean, computed across 5 fits using independently chosen random weights

3.3.3 Conclusions About TrACE

In this section a new theoretical method was proposed for reducing the scaling of the model size with respect to the number of chemical elements and radial functions for atom centred density based ML force fields. The proposed methods achieve this whilst retaining the rigorous mathematical results of linear completeness of the ACE basis. The canonical tensor decomposition was also analysed numerically and shown to be a promising route to create linear models for systems with many elements. There are several other tensor decomposition schemes that could be tried and that might have advantages in certain situations [210]. In our paper on this topic, the tensor sketching is also shown to work well in materials science examples, and an implementation for the SOAP kernel is also provided by my co-authors [46].

Importantly, the tensor-decomposed ACE product basis is used in Chapter 5 as a key building block of the MACE model. By combining TrACE with equivariant message passing, the remaining two limitations of ACE can also be overcome. MACE models are very easy to fit and are robust with respect to the choice of the hyperparameters, and MACE also greatly improves on the accuracy of ACE as will be demonstrated later.

The TrACE ideas combined with message passing also enables the development of transferable foundation models. Recently, a paper on organic chemistry describing 10 chemical elements [121] (see section 5.3) and another on materials chemistry describing 89

chemical elements [15] were published. These would not have been possible without the techniques described in this section.

Chapter 4

Multi-ACE: The Design Space of Machine Learning Force Fields

In Chapter 2 several different approaches were introduced to parameterise potential energy surfaces. In particular, machine learning approaches could be divided into two main categories. There were the atom-centred descriptor based models and the message passing graph neural network models. In Chapter 3 the Atomic Cluster Expansion descriptors are described in detail. As they provide a linearly complete basis of symmetric functions of local atomic environments, all other atom-centred descriptors can be expressed in terms of ACE [61, 63, 147].

In this chapter, the multi-ACE theory is introduced. This theory provides a framework for understanding the design space of machine learning interatomic potentials. Crucially, it brings the message passing neural networks, including the more recent equivariant neural networks and the atom-centred machine learning force fields, to the same footing. The connection between MPNN models and atom-centred models was also investigated simultaneously with our work in Ref. [153]. This work was motivated by a number of papers showing the great potential of equivariant MPNN force fields, significantly improving on the accuracy of the linear ACE models in molecular chemistry benchmarks. By understanding these models in a unified framework, it is possible to rigorously analyse the architectures and propose new ones that can further improve both accuracy and computational efficiency. This led to the development of a new model called MACE that will be discussed in detail in the next chapter.

4.1 The Multi-ACE layer: Equivariance and Continuous Embedding

In the following, a version of the ACE formalism is presented for deriving $E(3)$-invariant and equivariant basis functions that incorporates a continuous embedding of chemical elements. This is closely related to TrACE and will serve as the main building block of the Multi-ACE framework.

One-particle basis As before in Chapter 3, the first step is to specify the form of the one-particle basis as introduced in Equation (3.2). It is used to describe the spatial arrangement of atoms j around the central atom i:

$$\phi_{nlm}^{z_i z_j}(\mathbf{r}_{ji}) = R_{nl}^{z_i z_j}(r_{ji}) Y_l^m(\hat{\mathbf{r}}_{ji}), \tag{4.1}$$

where the index z_i and z_j refer to the chemical elements of atoms i and j. The one-particle basis functions are formed as the product of a set of orthogonal radial basis functions R_{nl} and spherical harmonics Y_l^m. The positional argument $\mathbf{r}_{ji}$ in Equation (4.1) can be obtained from the states of the atoms introduced in Equation (2.24) $(\sigma_i^{(t)}, \sigma_j^{(t)})$, thus making the value of the one-particle basis function depend on the states of two atoms.

The formulation in Equation (4.1) uses discrete chemical element labels. The drawback of this approach is that the number of different basis functions rapidly increases with the number of chemical elements in the system, as discussed in detail in Section 3.3. MPNNs typically leverage a learnable mapping from discrete chemical element labels to a continuous fixed-length representation. Using such an embedding with ACE eliminates the scaling of the number of basis functions with the number of chemical elements. This embedding is closely related to the tensor decomposition in TrACE, but can be expressed in a more general form by introducing two new indices as explained below.

$$\phi_{kv}(\sigma_i, \sigma_j) = R_{kcl}(r_{ji}) Y_l^m(\hat{\mathbf{r}}_{ji}) T_{kc}(\theta_i, \theta_j), \tag{4.2}$$

where T_{kc} is a generic function of the chemical attributes θ_i and θ_j and is endowed with two indices, k and c, and the radial basis likewise. Of these, c, together with l and m, will be coupled together when the many-body basis functions are formed (see Equation (4.4) below). These coupled indices are collected into a single multi-index $(v \equiv lmc)$ for ease of notation. On the other hand, k will be referred to as the uncoupled index.

Beyond chemical element labels, T_{kc} can account for the dependence of the one-particle basis functions on other attributes of the atoms, such as charge, magnetic moment [62],

or learnable features. Furthermore, the output of T_{kc} can be invariant or equivariant to rotations. In the case of equivariant outputs, the indices k (in the uncoupled case) or c (in the coupled case) will themselves be multi-indices that contain additional indices (e.g., l' and m') describing the transformation properties of these outputs.

To recover Equation (4.1) with the discrete element labels, θ_i, θ_j can be set to z_i, z_j and it can be assumed that $k \equiv 1$, i.e., there are no uncoupled indices. Furthermore, c is chosen to be a multi-index, $(c \equiv nz_iz_j)$, with T_{kc} being an index selector, $T_{kc} = T_{1nz_iz_j} = \delta_{z_i\theta_i}\delta_{z_j\theta_j}$. In this case, the index n of the radial basis $R_{nlz_iz_j}$ in Equation (4.1) is also part of the "coupled" multi-index c.

In the language of MPNNs, the values of the one-particle basis functions would be thought of as edge features of a graph neural network model. This graph would be directed since the one-particle basis functions are not symmetric with respect to the swapping of the central atom i and the neighbour atom j.

Higher-order basis functions The next step of the ACE construction is analogous to traditional message passing: the values of the one-particle basis functions evaluated on the neighbours are summed to form the atomic- or A-basis. This corresponds to a projection of the atomic density onto the one-particle basis. Therefore, in the atomic environment representation literature, this step is often referred to as the density projection [147],

$$A_{i,kv} = \sum_{j \in \mathcal{N}(i)} \phi_{kv}(\sigma_i, \sigma_j). \tag{4.3}$$

The A-basis is invariant with respect to the permutation of the neighbour atoms, and its elements are 2-body functions. This means that it can represent functions that depend on all neighbours' positions but can be decomposed into a sum of 2-body terms.

To create basis functions with higher body-order, tensor products of the A-basis functions are formed to obtain the product basis, $\mathbf{A}_{i,k\mathbf{v}}$:

$$\mathbf{A}_{i,k\mathbf{v}} = \prod_{\xi=1}^{\nu} A_{i,kv_\xi}, \quad \mathbf{v} = (v_1, ..., v_\nu), \tag{4.4}$$

where ν denotes the correlation order and the array index $\mathbf{v}$ collects the multi-indices of the individual A-basis functions, representing a ν-tuple.

Taking the product of ν A-basis functions results in basis functions of correlation order ν, which thus have body-order $\nu + 1$, on account of the central atom. In the language of density-based representations, these tensor products correspond to ν-correlations of the density of atoms in the atomic neighbourhood [153].

For example, the $\nu = 3$, four-body basis functions have the form

$$A_{i,kv} = A_{i,kv_1} A_{i,kv_2} A_{i,kv_3}, \tag{4.5}$$

where $v = (v_1 v_2 v_3)$. This illustrates the difference between the uncoupled k channels and the coupled v channels - tensor products were not formed with respect to the indices collected in k. This corresponds to the tensor decomposition introduced in TrACE. In the Multi-ACE framework, there is freedom to apply tensor decompositions across any number of different indices, including chemical elements or the radial features.

Symmetrisation of basis functions The product basis linearly spans the space of permutationally and translationally invariant functions but does not account for rotational invariance or equivariance of predicted properties or intermediate features. To create rotationally invariant or equivariant basis functions, the product basis must be symmetrised with respect to the rotation group $O(3)$ or $SO(3)$. Symmetrisation takes its most general form as an averaging over all possible rotations of the neighbourhood. In the case of rotationally invariant basis functions, this averaging is expressed as an integral of the product basis over rotated local environments, as was already introduced in Equation (3.6). In the more general notation introduced in this chapter, the symmetrisation can be written as

$$B_{i,k\mathbf{v}} := \int_{O(3)} \mathbf{A}_{i,k\mathbf{v}}\Big(\big\{Q \cdot (\sigma_i, \sigma_j)\big\}_{j \in \mathcal{N}(i)}\Big)\, dQ, \tag{4.6}$$

where the dependence of the product basis on the atomic states is written out explicitly, and $Q \cdot (\sigma_i, \sigma_j) = (Q \cdot \sigma_i, Q \cdot \sigma_j)$ denotes the action of the rotation on a pair of atomic states. The above integral is purely formal. To explicitly create a spanning set of the symmetric B functions above, one can instead use tensor contractions, as the angular dependence of the product basis is expressed using products of spherical harmonics (see Equation (4.9) below).

The construction of Equation (4.6) is readily generalised if equivariant features are required [62, 152, 240]. If the action of a rotation Q on a feature $\mathbf{h}$ is represented by a matrix $\mathbf{D}(Q)$, then following Equation (2.29) the equivariance constraint can be written as

$$\mathbf{D}(Q)^{-1}\mathbf{h}\Big(\big\{Q \cdot (\sigma_i, \sigma_j)\big\}_{j \in \mathcal{N}(i)}\Big) = \mathbf{h}\Big(\big\{\sigma_i, \sigma_j\big\}_{j \in \mathcal{N}(i)}\Big). \tag{4.7}$$

To linearly expand the feature $\mathbf{h}$, the basis functions must satisfy the same symmetries, which is achieved by defining the symmetrised basis as

$$B_{i,k\mathbf{v},\alpha} = \int_{O(3)} (\mathbf{D}(Q)^{-1} e_\alpha) \mathbf{A}_{i,k\mathbf{v}}\Big(\big\{Q \cdot (\sigma_i, \sigma_j)\big\}_{j \in \mathcal{N}(i)}\Big)\, dQ, \tag{4.8}$$

where the e_α are a basis of the feature space $\mathbf{h}$. This approach can be applied to parameterise tensors of any order, both in Cartesian and spherical coordinates. For example, if $\mathbf{h}$ represents a Euclidean 3-vector, the e_α denotes simply the three Cartesian unit vectors, $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$.

Going forward, features with spherical L-equivariance are discussed. They are labelled accordingly as $\mathbf{h}_L$ and the corresponding basis functions as $B_{i,k\alpha LM}$. The matrices $\mathbf{D}(Q)$ become the Wigner-D matrices, i.e., $\mathbf{D}^L(Q)$.

The integration over the rotations can be reduced to recursions of products of Wigner D-matrices and carried out explicitly as a tensor contraction [63, 152]. It is then possible to create a spanning set of L-equivariant features of integrals of the types of equations (4.6) and (4.8) using linear operations. This can be done by introducing the generalized coupling coefficients:

$$B_{i,k\eta,LM} = \sum_{\mathbf{v}} C^{LM}_{\eta,\mathbf{v}} \mathbf{A}_{i,k\mathbf{v}}, \tag{4.9}$$

where $C^{LM}_{\eta,\mathbf{v}}$ are the coupling coefficients corresponding to correlation order ν and imposed equivariance L. The output index η enumerates the different possible combinations of $\mathbf{A}_{i,k\mathbf{v}}$ that have equivariance L.

An additional degree of freedom is to have a different $\mathbf{A}_{i,k\mathbf{v}L}$ product basis for each symmetry L (e.g., by choosing different one-particle basis functions depending on L) which is a choice made for example in the NequIP model [17].

The values of the B functions can be combined into an output $m_{i,kLM}$ on each atom i and each channel k via a learnable linear transformation

$$m_{i,kLM} = \sum_{\eta} w_{k\eta L} B_{i,k\eta,LM}. \tag{4.10}$$

Finally, to generate the target output of the layer for the atom i, the uncoupled channels k can be mixed via a learnable (linear or non-linear) function $\Phi_{i,L} = \mathscr{F}(\mathbf{m}_{i,L})$.

4.2 A General Framework of Many-Body Equivariant Interatomic Potentials

In this section, multiple equivariant ACE layers from the previous section are combined to build a message passing model. The resulting framework encompasses most equivariant MPNN-based interatomic potentials. In the case of using a single message passing layer, the framework can be reduced to linear ACE or the other atom-centred descriptor-based models.

To create a Multi-ACE model, it needs to be specified how the output of one ACE layer is used in the next layer. This is done by updating the state of the atoms, assigning the output

Table 4.1 Different machine learning potentials in the framework of MPNNs. We identify SchNet, NequIP, and ACE as examples of MPNNs and exhibit their explicit components in the design space: the message, symmetric pooling, and update functions. Note that in NequIP, the choice of non-linearity is not fixed, and we have chosen a normed activation with tanh to be shown here. In each case, learnable parameters (weights) are shown as W and biases as b.

	SchNet	NequIP	Linear ACE
Message function M_t	$R_k^{(t)}\left(\|\|r_j - r_i\|\|\right) h_{j,k}^{(t)}$	$R^{(t)}_{kl_1l_2L}(r_{ji}) Y_{l_1}^{m_1}(\hat{\mathbf{r}}_{\mathbf{ji}}) h^{(t)}_{j,kl_2m_2}$	$R_n(r_{ji}) Y_l^m(\hat{\mathbf{r}}_{ji}) \delta_{z_i\theta_i}\delta_{z_j\theta_j}$
Symmetric pooling $\bigoplus_{j\in\mathcal{N}(i)}$	$\sum_{j\in\mathcal{N}(i)}$	$\sum_{l_1m_1l_2m_2} \mathscr{C}^{LM}_{l_1m_1l_2m_2} \sum_{j\in\mathcal{N}(i)}$	$\sum_\eta w_\eta \sum_{\mathbf{v}} \mathscr{C}^{00}_{\eta,\mathbf{v}} \prod_{\xi=1}^{\nu} \sum_{j\in\mathcal{N}(i)}$
Update function U_t	$h_i^{(t)} + \tanh\left(W^{(t)} m_i^{(t)} + b^{(t)}\right)$	$h_i^{(t)} + \tanh\left(\|W^{(t)} m_i^{(t)}\|^2\right) W^{(t)} m_i^{(t)}$	-

of the previous layer to the feature $\mathbf{h}_i^{(t+1)}$:

$$\begin{aligned} \boldsymbol{\sigma}_i^{(t+1)} &= (\mathbf{r}_i, \theta_i, \mathbf{h}_i^{(t+1)}), \\ \mathbf{h}_i^{(t+1)} &= U_t(\boldsymbol{\sigma}_i^{(t)}, \mathbf{m}_i^{(t)}), \end{aligned} \tag{4.11}$$

where $\mathbf{m}_i^{(t)}$ is a set of messages at iteration t as defined in Equation (4.10) and U_t is the update function for each layer (see Equation (2.27)). In most MPNNs, the k channel of the message corresponds to the dimension of the learned embedding of the chemical elements [187]. Next, Equation (4.2) needs to be extended further to incorporate the dependence on the output of the previous ACE layer into the one-particle basis. This can be achieved by making it an argument of the T_{kc} functions

$$\phi^{(t)}_{kvL}(\boldsymbol{\sigma}_i^{(t)}, \boldsymbol{\sigma}_j^{(t)}) = R^{(t)}_{kcl_1L}(r_{ji}) Y_{l_1}^{m_1}(\hat{\mathbf{r}}_{ji}) T^{(t)}_{kcL}(\mathbf{h}_j^{(t)}, \theta_i, \theta_j), \tag{4.12}$$

where $(v \equiv l_1 m_1 c)$. The index L is also added to the one-particle basis to enable having a different set of one-particle basis functions for messages $m_{i,kLM}$ with different symmetry L.

These equations can now be directly related to the general MPNN framework introduced in Section 2.4.2. First, the message function M_t can be identified with the one-particle basis of Equation (4.12):

$$M_t(\boldsymbol{\sigma}_i^{(t)}, \boldsymbol{\sigma}_j^{(t)}) := M^{(t)}_{kvL}(\boldsymbol{\sigma}_i^{(t)}, \boldsymbol{\sigma}_j^{(t)}) = \phi^{(t)}_{kvL}(\boldsymbol{\sigma}_j^{(t)}, \boldsymbol{\sigma}_i^{(t)}). \tag{4.13}$$

Next, the permutation invariant pooling operation $\bigoplus_{j\in\mathcal{N}(i)}$ of Equation (2.26) should be defined. To obtain a symmetric many-body message $m^{(t)}_{i,kLM}$ of correlation order ν, the pooling operation must map the one-particle basis, which are two-body, to a set of many-

body symmetric features. These can be combined in a learnable way to form the message on each node. This is what the ACE formalism of Section 4.1 achieves. The central equation of Multi-ACE can then be written as:

$$m_{i,kLM}^{(t)} = \bigoplus_{j\in\mathscr{N}(i)} M_t(\sigma_i^{(t)}, \sigma_j^{(t)}) = \sum_{\eta} w_{i,k\eta L}^{(t)} \sum_{\mathbf{v}} C_{\eta,\mathbf{v}}^{LM} \prod_{\xi=1}^{\nu} \sum_{j\in\mathscr{N}(i)} \phi_{kv_\xi L}^{(t)}(\sigma_i^{(t)}, \sigma_j^{(t)}), \tag{4.14}$$

where $w_{i,k\eta L}^{(t)}$ are learnable weights, and ν is the maximum correlation order. $C_{\eta,\mathbf{v}}^{LM}$ denotes the generalised Clebsch-Gordan coefficients defined in Equation (4.9).

The update function U_t from Equation (2.27) corresponds to a learnable linear combination of the uncoupled channels of the symmetrised message. U_t can be written as

$$h_{i,kLM}^{(t+1)} = U_t(\sigma_i^{(t)}, \mathbf{m}_i^{(t)}) = \sum_{\tilde{k}} W_{k\tilde{k}L}^{(t)} m_{i,\tilde{k}LM}^{(t)} \tag{4.15}$$

with $W^{(t)}$ being a block diagonal weight array of dimension $[N_{\text{channels}} \times N_{\text{channels}} \times L_{\max}]$, N_{channels} is the number of uncoupled k channels in the message and $L_{\max}$ is the maximum order of symmetry in the message that is passed from one layer to the next. U_t can also depend on the attributes (e.g., the chemical element) of the central atom via a so-called "self-connection" [14]. In general, the update functions acting on equivariant features can also be non-linear, but for that, it has to have a particular gated form [225].

After the T-th layer, a learnable (linear or non-linear) readout function that can depend on the final message or all previous ones gives the site energy of the atom i.

4.2.1 Interpreting Models as Multi-ACE

The Multi-ACE framework includes many of the previously published equivariant message passing networks. The most basic specification of a multi-ACE model considers the following:

- the number of layers T
- the correlation order of each layer ν
- the internal order of the spherical harmonic expansion within the layer in the one-particle basis $l_{\max}$
- the order of the spherical harmonics in the message passing phase after symmetrisation, $L_{\max}$

Table 4.2 **ML force fields in the multi-ACE framework** Different choices in the Multi-ACE formalism lead to different models in the literature. The internal $l_{\max}$ specifies the angular information contained in the messaging function M_t indexed by the highest weights of the irreducible representations of $O(3)$. The update $L_{\max}$ specifies the angular information in the update function. The local correlation order is the correlation order of the first message $m_i^{(0)}$. The total correlation order corresponds to the correlation order of the entire model as a function of individual atoms. The models above the separation line correspond to spherical equivariant interatomic potentials and the models under to Cartesian equivariant interatomic potentials.

	$l_{\max}$	Update $L_{\max}$	Local correlation order (ν)	Number of layers (T)	Total correlation order	$T_{kc}^{(t)}\left(\mathbf{h}_j^{(t)}, \theta_i, \theta_j\right)$	Coupling (ν)
SOAP [12]	≥ 3	0	2	1	≥ 3	$\delta_{z_i\theta_i}\delta_{z_j\theta_j}$	nlm
Linear ACE [119]	≥ 1	0	≥ 1	1	≥ 3	$\delta_{z_i\theta_i}\delta_{z_j\theta_j}$	nlm
SchNet [187]	0	0	1	T ≥ 2	T	$h^{(t)}_{j,kl=0}$ (Scalars)	$\emptyset$
DimeNet [76]	0	0	2	T ≥ 2	2T	$h^{(t)}_{j,l=0}$ (Scalars)	$\emptyset$
Cormorant [4]	≥ 1	≥ 1	1	T ≥ 2	T	$h^{(t)}_{j,klm}$ (Spherical Vec.)	lm
NequIP [17]	≥ 1	≥ 1	1	T ≥ 2	T	$h^{(t)}_{j,klm}$ (Spherical Vec.)	$l_1m_1l_2m_2$
GemNet [113]	≥ 1	≥ 1	3	T ≥ 2	T	$h^{(t)}_{j,klm}$ (Spherical Vec.)	$l_1m_1l_2m_2$
NewtonNet [86]	1	1	1	T ≥ 2	T	Cartesian Vectors	-
EGNN [181]	1	1	1	T ≥ 2	T	Cartesian Vectors	-
PaINN [186]	1	1	1	T ≥ 2	T	Cartesian Vectors	-
TorchMD-Net [204]	1	1	1	T ≥ 2	T	Cartesian Vectors	-

Other choices include the type of features (Cartesian or spherical basis) and the type of dependence of the radial basis on the indices kcl in Equation (4.12). Note that the point-wise nonlinearities present in some of those models affect both the local correlation and the total correlation, but do not change the expressivity of the architecture [14]. For simplicity, they are not considered for the following discussion. A comparison of the design choices of many different models is summarised in Table 4.2.

For example, the SchNet network uses $T \geq 2$ layers and a 2-body invariant convolution, meaning $\nu = 1$, $L = 0$, and $l_{\max} = 0$. The DimeNet invariant message passing network includes higher correlation order messages (more precisely, 3-body messages by incorporating angular information), which means that $T \geq 2$, $\nu = 2$, $L_{\max} = 0$, and $l_{\max} = 5$. The equivariant MPNN NequIP uses several layers, $T \geq 2$, each of them being 2-body, $\nu = 1$ and having a higher order symmetry, $L_{\max} \geq 1$, and $l_{\max} = L_{\max}$. In this case the symmetrisation of Equation (4.14) can be simplified

$$m^{(t)}_{i,kLM} = \sum_{l_1m_1l_2m_2} C^{LM}_{l_1m_1,l_2m_2} \sum_{j\in\mathcal{N}(i)} R^{(t)}_{kl_1l_2L}(r_{ji}) Y^{m_1}_{l_1}(\hat{\mathbf{r}}_{ji}) h^{(t)}_{j,kl_2m_2} \tag{4.16}$$

The models in the lower part of the table do not use a spherical harmonic expansion, but work with Cartesian tensors. However, they fit into this framework by considering the equivalence of vectors and $l = 1$ spherical tensors. The coordinate displacements present in

EGNN [181] and NewtonNet [86] can be rewritten as an $l = 1$ spherical expansion of the environment through a change of basis.

Based on the models presented in Table 4.2, the Multi-ACE framework presents two main routes that have been taken thus far in building interatomic potentials. The models have either few layers and high local correlation order, like linear ACE (and other descriptor-based models), or many layers and low local correlation order, such as NequIP. In Chapter 5 MACE will be introduced, which is a new model that combines high local correlation order with equivariant message passing.

4.2.2 Message Passing as a Chemically Inspired Sparsification

A central aspect of message passing models is the treatment of semi-local information: while in approaches such as ACE, the atomic energy is only influenced by neighbouring atoms within the local cut-off sphere (panel (a) of Figure 4.1), the message passing formalism iteratively propagates information, allowing for semi-local information to be communicated. Equivariant MPNNs update the atom states based on a tensor product between the edge features and neighbouring atoms' states, which leads to "chain-like" information propagation.

In particular, considering a much simplified message passing architecture with a single channel k and an update U which is just the identity

$$h^{(t+1)}_{i,LM} = \sum_{l_1 m_1 l_2 m_2} C^{LM}_{l_1 m_1, l_2 m_2} \sum_{j \in \mathcal{N}(i)} R^{(t)}_{l_1 l_2 L}(r_{ji}) Y^{m_1}_{l_1}(\hat{\mathbf{r}}_{ji}) h^{(t)}_{j,l_2 m_2} \tag{4.17}$$

leads to a simple two-layer update that can be written explicitly as

$$\begin{aligned} h^{(2)}_{i,LM} &= \sum_{l_1 m_1 l_2 m_2} C^{LM}_{l_1 m_1, l_2 m_2} \sum_{j_1 \in \mathcal{N}(i)} R^{(1)}_{l_1 l_2 L}(r_{ji}) Y^{m_1}_{l_1}(\hat{\mathbf{r}}_{ji}) h^{(1)}_{j_1, l_2 m_2} \\ &= \sum_{l_1 m_1 l_2 m_2} C^{LM}_{l_1 m_1, l_2 m_2} \sum_{j_1 \in \mathcal{N}(i)} R^{(1)}_{l_1 l_2 L}(r_{ji}) Y^{m_1}_{l_1}(\hat{\mathbf{r}}_{ji}) \sum_{j_2 \in \mathcal{N}(j_1)} R^{(0)}_{l_2}(r_{ji}) Y^{m_2}_{l_2}(\hat{\mathbf{r}}_{ji}) h^{(0)}_{j_2} \end{aligned} \tag{4.18}$$

where $h^{(0)}_{j_2}$ is assumed to be a scalar, learnable embedding of the chemical elements, such that it does not possess the l index. The 2-layer message passing is illustrated graphically in panel (b) of Figure 4.1.

This mechanism defines a pattern of information flow in which the state of j_2 is first passed onto the atom j_1, resulting in the (j_2, j_1)-correlation being captured. This is then passed onto atom i, which encodes the 3-body interaction between atoms (i, j_1, j_2) on atom i. This scheme induces a chain-wise propagation mechanism ($j_2 \to j_1 \to i$), which is different from local models such as ACE, in which the three-body correlation on atom i arises from an

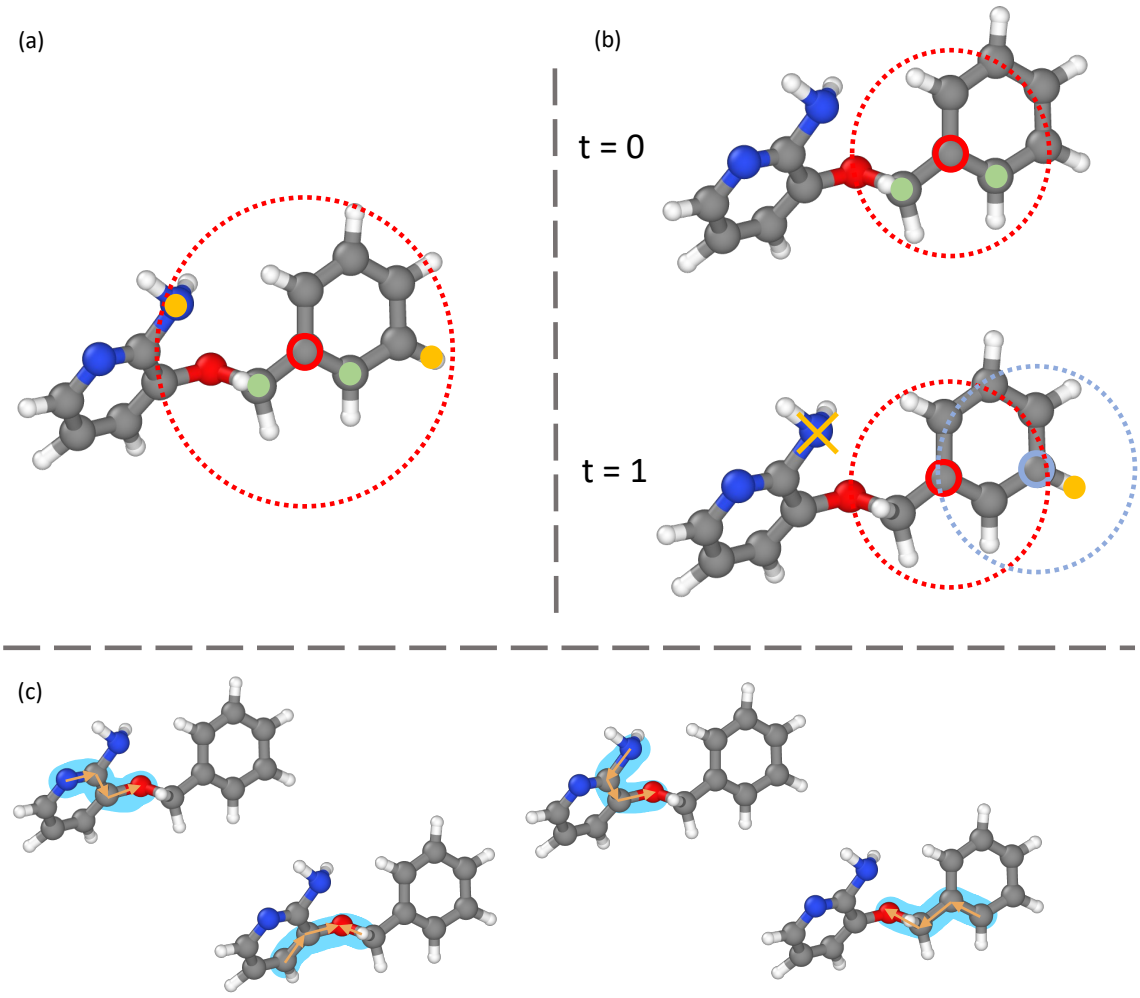


Fig. 4.1 Panel (a) shows a green and a yellow 3-body clusters of an atom centred descriptor model. Panel (b) shows a 2-layer MPNN with the same cutoff showing the green cluster present in the MPNN, but the yellow not. Panel (c) illustrates the chain-like clusters of the MPNN architecture.

interaction between (i, j_1) and (i, j_2). Some of these chain-like clusters are illustrated in the bottom panel (c) of Figure 4.1.

One can then, under the assumption of linearity, view equivariant MPNNs as a sparsification of an equivalent one-layer ACE model but which has a larger cutoff radius $r_{\text{cut,ACE}} = T \times r_{\text{cut,MPNN}}$, where T denotes the number of message passing steps and $r_{\text{cut,ACE}}$ is the maximal distance of atoms that can see each other in a T layer MPNN. While in a one-layer ACE, all clusters with central atom i would be considered, the MPNN formalism sparsifies this to only include walks along the graph (the topology of which is induced by local cutoffs) of length T that end on atom i.

In practice, for typical settings of T, r_{cut}, and ν, a local model like ACE with a cutoff of $T \times r_{\text{cut}}$ would be impractical due to the large number of atoms in the neighbourhood. Furthermore, the clusters created by atom-centred representations for an equivalent cutoff to MPNNs are less physical, as illustrated by the atom highlighted with a cross in panel (b) of Figure 4.1. Most physical interactions in chemistry are short-ranged and semi-local information propagates in a chain-like mechanism, thus causing the message passing sparsification to roughly correspond to the chemical bonding topology. A more in-depth discussion on the relationship between message passing and semi-local information can be found in Refs. [24, 153].

4.3 Conclusions about the Multi-ACE Design Space

In this chapter, Multi-ACE was introduced, which is a framework that enables the understanding of many previously published $O(3)$-equivariant (or invariant) machine-learning interatomic potentials in a common footing. Using this framework, a large design space can be identified. In Ref [14] we have systematically studied how the different choices made by the different models affect the accuracy, smoothness, and extrapolation of the fitted interatomic potentials.

Using this framework, one can identify the choices made by existing ML force fields: some use invariant 2 and 3-body features and nonlinear regression (SOAP-GAP, BPNN, etc.), and others use higher body-order features and linear regression (linear ACE, MTP [189]) whereas most message-passing models use 2-body features locally but increase the body-order via nonlinear activations and applying multiple messages passing layers. A yet unexplored part of the design space was the use of locally many-body features in a message-passing model, and it is the subject of the next chapter.

Chapter 5

MACE: Higher Order Equivariant Message Passing Force Fields

Developing fast and accurate force fields for molecular systems was the main goal of this PhD research project. Chapter 3 focused on the Atomic Cluster Expansion, which provided a mathematical framework as well as a practical recipe to create features that can be used to parameterise force fields. As discussed before, these models had several limitations. They have very poor scaling with the number of different chemical elements modelled, which was addressed by the tensor-reduced version introduced in Section 3.3. Further limitations became apparent with the publication of a new generation of force fields based on equivariant message passing neural networks [17, 186]. These models improved on the accuracy of linear ACE substantially, whilst also being simpler to train, not requiring the careful regularisation necessary for the linear models. They also had limitations; to achieve high accuracy, they required 4-5 layers resulting in large computational cost and poor scalability to large system. By having several layers, the receptive field of the models increased to 20-30Å, making them impractical to parallelise on large computers.

Seeing the strengths and limitations of each of the methods motivated the new theoretical work described in Chapter 4, unifying the language of equivariant deep learning with the more traditional formalism used in the machine learning force field community. This work led to the definition of the design space of machine learning force fields.

In this chapter, a new machine learning force field architecture called MACE is introduced [16, 120]. It was designed to address the limitations of previous equivariant MPNN force fields by directly combining them with the mechanisms of TrACE for creating higher body order messages.

First, the MACE architecture is described in detail, followed by a number of benchmark experiments that show the excellent accuracy of MACE on a wide range of tasks. Remarkably,

these results are achieved without changing or tuning the hyperparameters of the model for each new system.

Finally, `MACE-OFF23`, a new transferable organic force field, is introduced based on the MACE architecture [121]. This force field model is able to describe any closed shell, neutral organic molecule at a hybrid DFT level of accuracy without any further training. The capabilities of `MACE-OFF23` are demonstrated on a number of different tasks both in the gas phase and in the condensed phase.

5.1 MACE Architecture

In this section, the MACE higher order equivariant message passing force field architecture is introduced. This is followed by a brief discussion of a general training strategy relying on a loss scheduler. The body order of MACE models is also highlighted and its implications for smoothness and extrapolation are briefly discussed.

5.1.1 Higher Order Equivariant Message Passing

The MACE model parametrises the mapping from the positions and chemical elements of the atoms to their potential energy. This is achieved by decomposing the total potential energy of the system into site energies (atomic contributions). The site energy of each atom depends on symmetric features that describe its chemical environment. MACE parametrises these features using a many-body expansion. To construct the features on the atoms first the local environment $\mathcal{N}(i)$ of atom i has to be defined. As before, $\mathcal{N}(i)$ is the set of all atoms j in the system for which $|\mathbf{r}_{ij}| \leq r_{\text{cut}}$, where $\mathbf{r}_{ij}$ denotes the vector from atom i to atom j and r_{cut} is a predefined cutoff. Defining a local neighbourhood allows the construction of a graph from the geometry where the nodes are the atoms, and the edges connect atoms in each other's local environment. The array of features of node (atom) i is denoted by $\mathbf{h}_i$ and is expressed in the spherical harmonic basis. Therefore, the elements of the features are always indexed by l and m. The superscript on $\mathbf{h}^{(t)}$ indicates the iteration steps (corresponding to "layers" of message passing in the parlance of graph neural networks), and the number of layers in the model is denoted by T. Equivariance of the model is achieved by utilising the transformation properties of spherical harmonics Y_{lm} under 3D rotations, which are inherited by the node features with the corresponding indices.

In the following, the MACE model is described as it was introduced in Refs. [16] and [120]. The description of the model focuses on showing both the key equivariant operations and also on identifying all free parameters of the model that are optimised during training.

All free ("learnable") parameters are denoted by the letter W, and all occurrences below correspond to different blocks of free parameters - the shapes of these parameter arrays are indicated by the use of different indices. Note that in addition to the explicitly shown free parameters W, the fully connected multi-layer perceptrons (MLPs) also contain internally further learnable parameters.

First, the node features $\mathbf{h}_i^{(0)}$ are initialised as a (learnable) embedding of the chemical elements with atomic number z_i into k learnable channels.

$$h_{i,k00}^{(0)} = \sum_z W_{kz} \delta_{zz_i} \tag{5.1}$$

This kind of mapping has been used extensively for graph neural networks [17, 76, 186, 187] and elsewhere [85, 228] and has been shown to lead to some transferability between molecules with different elements [14]. The zeros for the lm indices correspond to these initial features being scalars, i.e. rotationally invariant. The higher order elements of $\mathbf{h}_i^{(0)}$ with nonzero lm indices are implicitly initialised to zero. At the beginning of each subsequent iteration, the node features (both scalars and higher order) are linearly mixed together, resulting in $\bar{\mathbf{h}}_j$.

$$\bar{h}_{i,kl_2m_2}^{(t)} = \sum_{\tilde{k}} W_{k\tilde{k}l_2}^{(t)} h_{i,\tilde{k}l_2m_2}^{(t)} \tag{5.2}$$

Next, the features of each of the neighbouring atoms (edge in the graph) are combined with the interatomic displacement vectors pointing to them from the central atom. The interatomic displacement vector is expressed using radial and spherical harmonic basis. This is analogous to the construction of the one-particle basis of neighbour density representations, such as SOAP [12] and ACE [14, 61] introduced in earlier chapters. The construction is also closely related to the Cormorant [4] and NequIP [17] equivariant neural networks. The relationship between the different approaches is discussed in Chapter 4. The radial basis set is constructed using the first spherical Bessel function j_0^n for different wavenumbers, n, up to some small maximum (typically 8) as proposed in Ref. [76].

$$j_0^n(r_{ij}) = \sqrt{\frac{2}{r_{\text{cut}}}} \frac{\sin\left(n\pi \frac{r_{ij}}{r_{\text{cut}}}\right)}{r_{ij}} f_{\text{cut}}(r_{ij}) \tag{5.3}$$

The Bessel functions are multiplied by a polynomial cutoff function $f_{\text{cut}}(r_{ij})$ that goes smoothly to zero at $r_{ij} = r_{\text{cut}}$. The radial information expressed in this basis is then passed through an MLP that typically has 3 hidden layers of width 64. This learnable function has

many outputs, indexed by (η_1, l_1, l_2, l_3), effectively referring to a large number of learnable radial functions, but with substantial weight sharing between them.

$$R^{(t)}_{k\eta_1 l_1 l_2 l_3}(r_{ij}) = \mathrm{MLP}\left(\left\{j^n_0(r_{ij})\right\}_n\right) \tag{5.4}$$

This additional degree of freedom becomes relevant when combining the positional information (itself an equivariant that transforms under rotation like a vector) with equivariant node features. This is achieved by using the spherical tensor product formalism of angular momentum addition [227]. All possible combinations of equivariants are constructed using the appropriate Clebsch-Gordan coefficients. This operation creates the one-particle basis, $\phi^{(t)}_{ij,k\eta_1 l_3 m_3}$, analogously to ACE.

$$\phi^{(t)}_{ij,k\eta_1 l_3 m_3} = \sum_{l_1 l_2 m_1 m_2} C^{l_3 m_3}_{\eta_1, l_1 m_1 l_2 m_2} R^{(t)}_{k\eta_1 l_1 l_2 l_3}(r_{ij}) \times Y_{l_1 m_1}(\hat{r}_{ij})\, \bar{h}^{(t)}_{j,kl_2 m_2} \tag{5.5}$$

This operation is implemented using the `e3nn` library [78]. There are multiple ways of constructing an equivariant combination with a given symmetry corresponding to (l_3, m_3), and these multiplicities are enumerated by the index η_1.

The one-particle basis ϕ is summed over the neighbourhood to form the permutation invariant 2-body descriptors. Note that the identity of chemical elements has already been embedded, and hence this sum is over all atoms, regardless of their atomic number. A linear mixing of k channels with learnable weights yields the atomic basis, A_i (c.f. ACE [14, 61] in Section 3.1).

$$A^{(t)}_{i,kl_3 m_3} = \sum_{\tilde{k},\eta_1} W^{(t)}_{k\tilde{k}\eta_1 l_3} \sum_{j \in \mathcal{N}(i)} \phi^{(t)}_{ij,\tilde{k}\eta_1 l_3 m_3} \tag{5.6}$$

Further following the ACE and TrACE recipe, many-body symmetric features are created on each atom by taking the tensor product of the atomic basis, A, with itself ν times, yielding the "product basis", $\mathbf{A}_i$.

$$\mathbf{A}^{(t),\nu}_{i,k\mathbf{lm}} = \prod_{\xi=1}^{\nu} A^{(t)}_{i,kl_\xi m_\xi} \tag{5.7}$$

Note that in forming the tensor product, each k channel is treated independently. This method of tensor decomposition has been widely used in signal processing [192] and was formally shown not to degrade the expressibility of many-body models while substantially reducing computational cost compared to a full tensor product [47] (Section 3.3).

Next, the product basis is contracted to yield the fully symmetric basis, $\mathbf{B}_i$, using the generalised Clebsch-Gordan coefficients, C, where again there are multiple ways to arrive at a given output symmetry and these are enumerated by η_ν.

$$\mathbf{B}^{(t),\nu}_{i,\eta_\nu kLM} = \sum_{\mathbf{lm}} C^{LM}_{\eta_\nu \mathbf{lm}} \mathbf{A}^{(t),\nu}_{i,k\mathbf{lm}} \tag{5.8}$$

The bold **lm** signify that these are multi-indices, and the bold styles of **A** and **B** are a reminder that these are many-body features. Maximum body order is controlled by limiting ν. The tensor product and contraction of Equation (5.7)-(5.8) is implemented using an efficient loop tensor contraction algorithm [16].

Finally, a "message" m_i is formed on each atom as a learnable linear combination of the symmetrised many-body features of the neighbours.

$$m^{(t)}_{i,kLM} = \sum_{\nu} \sum_{\eta_\nu} W^{(t),\nu}_{z_i \eta_\nu kL} \mathbf{B}^{(t),\nu}_{i,\eta_\nu kLM} \tag{5.9}$$

To form the node features of the next layer, this message is added to the atoms' (nodes') features from the previous iteration using weights that depend explicitly on the atoms' chemical element (z_i).

$$h^{(t+1)}_{i,kLM} = \sum_{\tilde{k}} W^{(t)}_{kL,\tilde{k}} m^{(t)}_{i,\tilde{k}LM} + \sum_{\tilde{k}} W^{(t)}_{kz_iL,\tilde{k}} h^{(t)}_{i,\tilde{k}LM} \tag{5.10}$$

Note that because the initial node features $\mathbf{h}^{(0)}$ are solely functions of the chemical element corresponding to the node, in the first layer the second sum of Equation (5.10) is omitted. This allows the setting and fixing of the energy of isolated atoms (i.e. those with no neighbours in their environment) [14], which is often desirable [53] as also discussed in Section 3.2.

Equations (5.2)-(5.10) comprise a MACE layer, and multiple layers are built by iteration. This means that the effective receptive field of the model (the region around an atom from which information is used to determine the site energy of the atom) is approximately a sphere of radius $T \times r_{\text{cut}}$. More precisely, a neighbouring atom contributes to the site energy if it can be reached in T hops on the graph defined above. The selected clusters form a sparse set of all possible clusters in $T \times r_{\text{cut}}$. Therefore, message passing methods can be viewed as sparsifications of larger atom-centred methods, as discussed in Section 4.2.2. In practice, almost always two layers of MACE are used, so $T = 2$.

The output of the model, the site energy, is a learnable combination of the rotationally invariant part of the node features,

$$E_i = \sum_{t=1}^{2} E_i^{(t)} = \sum_{t=1}^{2} \mathscr{R}^{(t)}\left(\mathbf{h}_i^{(t)}\right), \tag{5.11}$$

where $\mathscr{R}$ is a linear map for the first layer features and a shallow one hidden layer MLP for the second layer features,

$$\mathscr{R}^{(t)}\left(h_i^{(t)}\right) = \begin{cases} \sum_k W_k^{(t)} h_{i,k00}^{(t)} & \text{if } t = 1 \\ \text{MLP}\left(\left\{h_{i,k00}^{(t)}\right\}_k\right) & \text{if } t = 2 \end{cases} \tag{5.12}$$

Keeping the readout function linear for all but the last layer helps preserve the body ordered nature of the model; see Section 5.1.2 and Ref. [14]. The forces on the atoms are determined as usual by taking analytical derivatives of the total potential energy, using auto-differentiation tools,

$$\mathbf{F} = -\nabla \sum_i E_i \tag{5.13}$$

Neural network models such as MACE have a very large number of free parameters, especially compared with linear or kernel based models previously applied to the same tasks. This overparameterisation is believed to help training when using stochastic gradient descent and brings accuracy and regularity [77, 105]. The trade-off between the size (and therefore computational cost) of the model and its accuracy needs to be controllable. The key controls for model size in MACE are the following:

- number of embedding channels k
- highest order L_{max} of the symmetric features $\mathbf{B}_{i,\eta_\nu kLM}^{(t),\nu}$

In this thesis, the different sized MACE models will be characterised using these two numbers; for example, the invariant MACE model (corresponding to $L_{\text{max}} = 0$) with 64 channels is denoted `MACE 64-0`.

Computational cost of MACE The computational cost of running a MACE model depends on the details of the systems studied (e.g. the average number of neighbours) and the hyperparameters that control the size of the model. The latency of the models (i.e., the shortest time to calculate forces on small systems, excluding calculation of the neighbour list and other book-keeping necessary for running MD) ranges from 0.7-20 ms, corresponding to

small (`MACE 64-0`) or larger (`MACE 256-2`) models as measured on an Nvidia A100 GPU. A comprehensive assessment of MD performance needs to consider the relationship between model size, accuracy, and execution speed. It is also useful to report the fastest possible MD performance, which can be taken to be that of the smallest model able to run stable MD for an arbitrary (in practice very large) number of steps. At present the performance of MACE in MD simulations is between 2-50M steps / day (a million steps correspond to a nanosecond with 1 femtosecond time step), depending on model size and the details of the full software stack, with the highest performance achieved using the JAX version of MACE using JAX-MD [183]. A comprehensive analysis of the computational speed and scaling of MACE models can be found in Ref. [121], showing execution times in the OpenMM and LAMMPS MD packages. The training times of the MACE models also vary vastly depending on the dataset size and the model size, and typically range from 15 minutes to 1 day on a single A100 GPU for the models shown here.

The computational performance of MACE compares very favourably with other alternative equivariant neural network potentials. This is primarily the result of MACE forming the high body order features very efficiently via the symmetric tensor products carried out on the nodes. In comparison, the Allegro model [146], which is one of the other leading architectures, which uses tensor products between the features on the edges of the graph rather than nodes, has a fastest reported speed of 18M steps per day [146]. However, it should be noted that the Allegro model has already been demonstrated to scale well to many millions of atoms on many GPUs [122]. This is in principle possible for MACE, which is also a strictly local model with a similarly moderate size receptive field, 5-10 Å, though the necessary LAMMPS domain decomposition implementation that would make it efficient is ongoing work.

5.1.2 The Body Order of MACE Models

The node features $\mathbf{h}_i$ of MACE can be considered as body ordered descriptors of atomic environments. In the limit of complete basis of radial functions and spherical harmonics, these descriptors, or basis functions, fully linearly span the space of symmetric functions over chain-like clusters with hops of size r_{cut}, up to the maximum body order of the features [14, 47, 63]. In most MACE models with the default hyperparameters, each layer has a body order of 4 (corresponding to $\nu = 3$ in Equation (5.7)). Therefore, the node features of the first layer are 4-body functions. They are then expanded in the second layer in the second one-particle basis, which are $4+1=5$-body functions, since each application of the one-particle basis adds one to the body order on account of the central atom. Then, taking the tensor products of these features with themselves three times, $3\times 4+1=13$-body features are obtained. Note

that the 5-body terms of the first layer all share the same central atom. This mechanism of efficiently forming very high body order, linearly complete features is unique to the MACE architecture, and it might be one of the reasons underpinning its excellent performance in the low-data regime and extrapolation.

5.1.3 Loss Scheduler

MACE models are trained to reproduce the energies, forces and, if available, the stresses of atomistic structures based on labelled reference data coming from quantum mechanical calculations. The training loss, $\mathscr{L}$, is typically computed as the weighted sum of the mean squared errors of the total energy, the force components (and virials or stresses if available for periodic systems),

$$\mathscr{L} = \frac{\lambda_E}{B} \sum_{b=1}^{B} \left(\frac{E_b - \hat{E}_b}{N_b} \right)^2 + \frac{\lambda_F}{3B} \sum_{b=1}^{B} \sum_{i_b,\alpha=1}^{N_b,3} \left(-\frac{\partial E_b}{\partial r_{i_b,\alpha}} - \hat{F}_{i_b,\alpha} \right)^2, \tag{5.14}$$

where the sum is taken over the atomic configurations b in the current batch with batch size B, E_b is the model's prediction of the total energy, $\hat{E}_b$ and $\hat{F}_b$ are the training data corresponding to total energy and force, respectively. The number of atoms in the configuration is N_b, atoms are indexed by i_b with elements of their position vector denoted by $r_{i_b,\alpha}$, in the Cartesian direction α. The weights of energies and forces are controlled by λ_E and λ_F.

To obtain the best performance from the MACE model, particular care has to be taken with the weight factors in the loss function. The most accurate models are obtained when, during the training, a higher weight is placed on the forces compared to the other properties. A possible reason for this could be that the forces are local quantities and contain information about the dependence of the total energy on each atom. However, having very accurate force predictions does not necessarily result in accurate energy predictions. This is especially the case in systems where the training set is heterogeneous, meaning that it is composed of a wide variety of different systems well separated in atomic configuration space, e.g. different molecules or phases of a solid with a very different structure. In this case, the model can accurately learn the local potential energy surface of each system individually, but it might not learn their relative energies correctly, resulting in large absolute energy errors. These appear as shifts in the predicted energy - true energy plots.

To reduce absolute energy errors, a loss scheduler can be used. For about 60% of the total training time, $\lambda_F > \lambda_E$ is used. For the second part of the training, the weights are switched so that $\lambda_E > \lambda_F$, and the learning rate is decreased by a factor of 10. In this way, the absolute energy errors can be reduced while keeping the force predictions' high accuracy.

An example of the energy validation error dramatically improving as the loss is changed is shown on Figure 5.1.

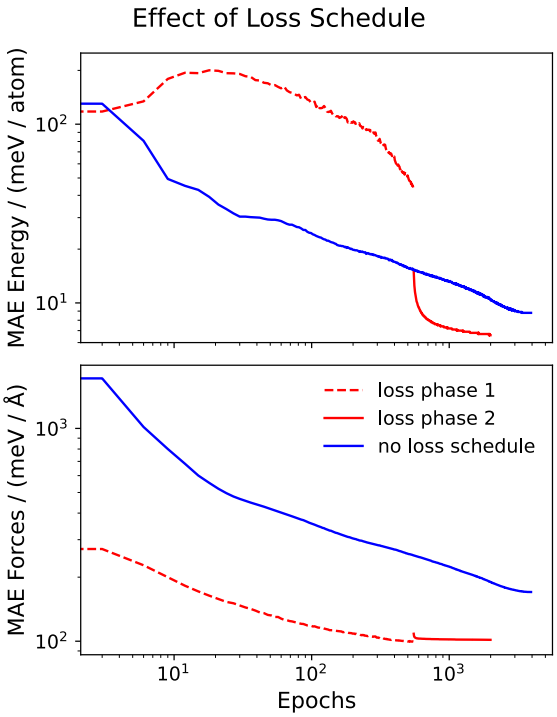


Fig. 5.1 **The effect of using a loss schedule** The figure shows that the two phase training not only accelerates the training of the model but also results in overall decreased errors.

The figure also compares the two phase training with using the weights of the second phase from the beginning and shows that the two phase protocol not only accelerates the convergence, but also results in overall decreased errors.

5.2 Selected MACE Applications

In this section a number of MACE applications are presented that showcase the capabilities of this new machine learning force field architecture on a wide range of chemical systems.

5.2.1 QM9 Benchmark

First, the MACE architecture is evaluated on a general molecular machine learning benchmark dataset. QM9 was introduced in Section 2.5, and contains 133,000 relaxed molecular geometries and 12 corresponding quantum mechanical observables.

	Gap meV	**Homo** meV	**Lumo** meV	$\mathbf{C_V}$ cal/mol K	μ D	**ZPVE** meV	$\mathbf{R^2}$ α_0^2	α α_0^3	**G** meV	**H** meV	**U** meV	$\mathbf{U_0}$ meV
NMP [79]	69	43	38	0.040	0.030	1.50	0.180	0.092	19	17	20	20
SchNet [187]	63	41	34	0.033	0.033	1.70	0.073	0.235	14	14	19	14
Cormorant [4]	61	34	38	0.026	0.038	2.03	0.961	0.085	20	21	21	22
LieConv [68]	49	30	25	0.038	0.032	2.28	0.800	0.084	22	24	19	19
DimeNet++ [76]	33	25	20	0.023	0.030	1.21	0.331	0.044	7.6	6.5	6.3	6.3
EGNN [181]	48	29	25	0.031	0.029	1.55	0.106	0.071	12	12	12	11
PaiNN [186]	46	28	20	0.024	0.012	1.28	0.066	0.045	7.4	6.0	5.8	5.9
TorchMD-NET [203]	36	20	18	0.026	**0.011**	1.84	**0.033**	0.059	7.6	6.2	6.4	6.2
SphereNet [133]	32	23	18	0.022	0.026	1.12	0.292	0.046	7.8	6.3	6.4	6.3
SEGNN [31]	42	24	21	0.031	0.023	1.62	0.660	0.060	15	16	13	15
EQGAT [128]	32	20	16	0.024	**0.011**	2.00	0.382	0.053	23	24	25	25
Equiformer [132]	30	**15**	**14**	0.023	**0.011**	1.26	0.251	0.046	7.6	6.6	6.7	6.6
MGCN [134]	64	42	57	0.038	0.056	1.12	0.110	**0.030**	15	16	14	13
Allegro [146]	-	-	-	-	-	-	-	-	5.7	4.4	4.4	4.7
NoisyNodes [80]	29	20	19	0.025	0.025	1.16	0.700	0.052	8.3	7.4	7.6	7.3
GNS-TAT+NN [233]	**26**	17	17	0.022	0.021	**1.08**	0.65	0.047	7.4	6.4	6.4	6.4
Wigner Kernels [21]	-	-	-	-	-	-	-	-	-	-	-	4.3
TensorNet [193]	-	-	-	-	-	-	-	-	6.0	**4.3**	4.3	4.3
MACE	42	22	19	**0.021**	0.015	1.23	0.210	0.038	**5.5**	4.7	**4.1**	**4.1**

Table 5.1 **Performance of MACE on QM9** Mean absolute error (MAE) of various models on the QM9 dataset demonstrating that MACE improves on the state of the art in several tasks. The bold model has the lowest error, while the underline indicates models with error within 10% of the best model for each task.

For all 12 tasks, the results are summarised for a large number of different machine learning architectures in Table 5.1. MACE achieves state-of-the-art results on 3 of the 4 energy related tasks (G, H, U, U_0) and also improves the state-of-the-art in one other non-energy related task.

To achieve the best result on intensive properties (gap, homo, lumo, μ, C_V and zpve), the readout function of the MACE model had to be modified. This was necessary because the simple sum of the site output in Equation (5.11) that is used in the MACE force fields results in a size-extensive overall prediction. The modified readout is a non-linear pooling operation which first applies both sum, mean and standard deviation pooling, followed by the application of an attention mechanism to form the final output [145]. Such a readout function led to up to two-fold improvement compared to the simple linear size-extensive pooling operation.

Learning curve on Potential Energies Since MACE was designed to predict potential energies, it is interesting to examine the (U_0) task in more detail. On the left panel of Figure 5.2, the learning curves of 4 different MACE models are compared. The blue models use hyperparameters of MACE that are similar to those used throughout the thesis (`256-2`).

In comparison, the models denoted with red are larger versions of MACE, which use a deeper and wider MLP in the learnable radial basis (see Eq. (5.4)) which has been shown to help for QM9 for models like Allegro [146]. This extra flexibility results in higher accuracy, especially in the very high data regime. A further interesting point is the comparison of the models trained using energies only (the usual practice for QM9, denoted by the solid line) with the models whose loss function also included forces exploiting the knowledge that QM9 is made up of equilibrium geometries with zero forces on all the atoms. This extra information (but needing no new QM calculations) increases the accuracy of both the small and large MACE models.

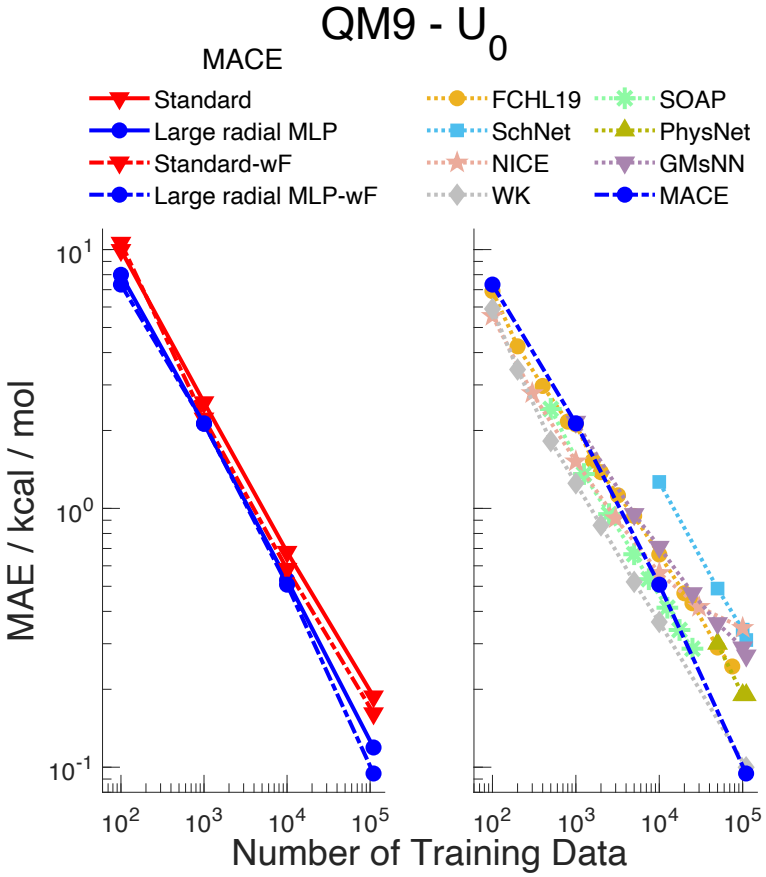


Fig. 5.2 **QM9 Learning Curves** The left panel compares the standard large MACE model (256-2) to a modified version for QM9 where the size of the MLP in the radial basis was increased. The figure also shows that including the 0 force information (labelled -wF) in the training leads to lower errors. The right panel compares the best MACE with a number of other machine learning models.

In the right panel of Figure 5.2, the learning curve of the best MACE model is compared to the learning curve of several other very good models. It shows that kernel models such as FCHL [42], SOAP [12], the linear NICE model [152], the feed-forward neural network

Table 5.2 **Root-mean-square errors on the 3BPA dataset.** Energy (E, meV) and force (F, meV/Å) errors of models trained and tested on configurations collected at 300 K of the flexible drug-like molecule 3-(benzyloxy)pyridin-2-amine (3BPA). Standard deviations are computed over three runs and shown in brackets if available.

		Linear ACE	Allegro (L=3)	NequIP (L=3)	BOTNet (L=3)	MACE (L=0)	MACE (L=1)	MACE (L=2)
300 K	E	7.1	3.84 (0.1)	3.3 (0.1)	3.1 (0.1)	4.5 (0.3)	3.4 (0.2)	**3.0** (0.2)
	F	27.1	12.98 (0.2)	10.8 (0.2)	11.0 (0.1)	14.6 (0.5)	10.3 (0.3)	**8.8** (0.3)
600 K	E	24.0	12.07 (0.5)	11.2 (0.1)	11.5 (0.6)	13.7 (0.2)	9.9 (0.8)	**9.7** (0.5)
	F	64.3	29.17 (0.2)	26.4 (0.1)	26.7 (0.3)	33.3 (1.4)	24.6 (1.1)	**21.8** (0.6)
1200 K	E	85.3	42.57 (1.5)	38.5 (1.6)	39.1 (1.1)	37.1 (0.8)	31.7 (0.5)	**29.8** (1.0)
	F	187.0	82.96 (1.8)	76.2 (1.1)	81.1 (1.5)	81.6 (3.9)	67.8 (1.8)	**62.0** (0.7)

GMsNN [235] and the message passing SchNet [187] and PhysNet [212] models all achieve comparable errors. The models that can surpass this significantly are Allegro [146], Wigner kernels [21] and MACE, all of which use higher body-order features. This strongly hints that high body order is a crucial property of the most successful atomistic ML models.

5.2.2 3BPA Benchmark

The 3BPA dataset is introduced in Section 2.5 and is used to evaluate the linear ACE model in Section 3.2. This benchmark tests a model's extrapolation capabilities. Its training set contains 500 geometries sampled from 300 K molecular dynamics simulation of the large and flexible drug-like molecule 3BPA. The three test sets contain geometries sampled at 300 K, 600 K, and 1200 K to assess in- and out-of-domain accuracy.

The root-mean-squared errors (RMSE) on the energies and forces for linear ACE (Section 3.2), three of the best equivariant neural networks Allegro [146], NequIP [17] and BOTNet [14] as well as a series of MACE models are shown in Table 5.2. Remarkably, MACE not only outperforms the linear ACE model by a factor of 2-3, but also achieves lower errors than the other equivariant models by a significant margin. In particular, when extrapolating to 1200 K data, MACE with $L = 2$ reduces the error of the NequIP and Allegro models by approximately 30%. Even the MACE model with invariant messages ($L = 0$) often nearly matches or exceeds the performance of competitive equivariant models, whilst improving on linear ACE by a very significant margin.

The need for equivariant features It might seem counterintuitive that MACE uses equivariant features to predict an invariant property, the potential energy. From Ref. [63] it is known that the invariant linear ACE features provide a complete basis for invariant functions. However, Table 5.2 shows that the MACE model improves significantly when the invariant ACE features are complemented with the equivariant ones. The reason for this might be

the fact that in all practical scenarios the models are very far from the complete basis limit. Therefore, in the pre-asymptotic regime, the extra flexibility in the functional form coming from the equivariant features appears to be beneficial for these models. The need for equivariant features is easier to see in models such as NequIP, whose functional form is formally not complete when only invariant features are used [162].

Dihedral torsions To further test the extrapolation capabilities of the models, Figure 5.3 compares BOTNet, NequIP, and MACE ($L = 2$) by inspecting their energy profile for three dihedral torsion scans. Overall, it can be seen that all models produce smooth energy profiles and that, in general, MACE comes closest to the ground truth. The fact that MACE

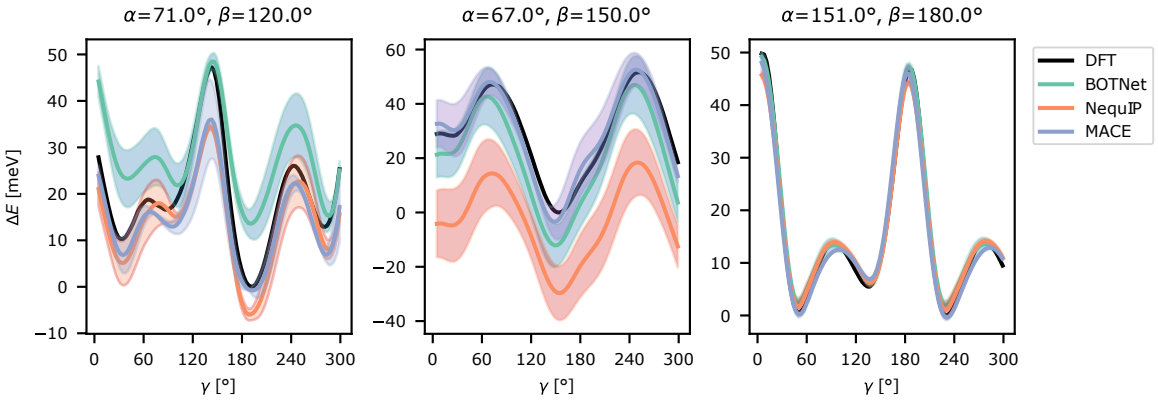


Fig. 5.3 Energy predictions on three cuts through the potential energy surface of the 3-(benzyloxy)pyridin-2-amine (3BPA) molecule by BOTNet, NequIP, and MACE ($L = 2$). The ground-truth energy (DFT) is shown in black. For each cut, the curves have been shifted vertically so that the lowest ground-truth energy is zero.

outperforms the other methods in the middle panel, which contains geometries furthest from the training dataset [14], suggests superior extrapolation capabilities.

5.2.3 Vibrational Spectrum from 50 Coupled Cluster Calculations

In this subsection, the performance of MACE in the low-data regime is evaluated. Data efficiency is crucial in atomistic machine learning as it can save time and computational cost or enable the models to be trained with more accurate reference data. Recent work has found that the majority of machine learning potentials are not able to run stable molecular dynamics simulations without iterative fitting, even when they are trained on thousands of configurations of a system [71]. Furthermore, it was found that very low energy and force RMSE did not correlate with the ability to perform stable simulations.

To demonstrate the excellent data efficiency and extrapolation capabilities of MACE, the relatively large 256-2 MACE models published in Ref. [16] trained using rMD17 [41] can be used. The models are trained on 50 and 1000 small molecule geometries sampled randomly from the dataset.

Remarkably, 50 configurations are sufficient to train a MACE model that has sub-kcal/mol total energy error on a large test set. For three selected systems, ethanol, paracetamol and salicylic acid, 50 ps long NVT molecular dynamics simulations were performed and found to be stable for both training set sizes, obviating the need for any further active learning or iterative fitting.

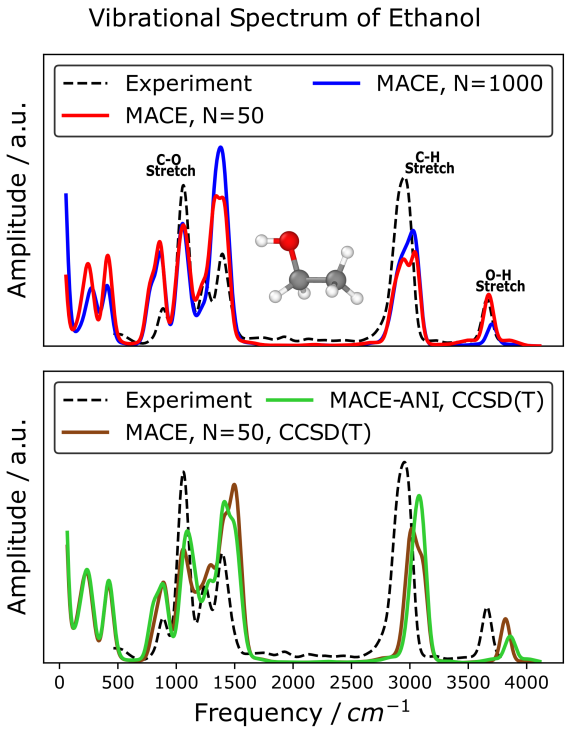


Fig. 5.4 **Vibrational spectrum of ethanol** The figure illustrates molecular vibrational spectrum computed from the velocity-autocorrelation function. The top panel compares the MACE model fitted to 50 (red) and 1,000 (blue) random revMD17 ethanol geometries. The bottom panel compares the spectrum of the small transferable MACE-COMP6-CC model from Section 5.2.5 (brown) and a MACE model fitted to the same 50 random revMD17 geometries (green) recomputed using CCSD(T) level of theory.

The molecular vibrational spectrum can be used to validate the dynamics of the model trained using 50 training points. The top panel of Figure 5.4 shows the spectra of ethanol. The two MACE models are in excellent agreement, and are able to reproduce the peaks of the experimental gas-phase IR spectrum of ethanol [66].

To demonstrate that these findings are applicable to larger and more complex molecules compared to ethanol, the same protocol of training to 50 and 1,000 reference geometries and running MD calculations to produce vibrational spectra was repeated for the salicylic acid and paracetamol molecules. The results are shown on Figure 5.5 and indicate that MACE is capable of learning a stable and accurate force field from 50 random training geometries even for these more complex molecules.

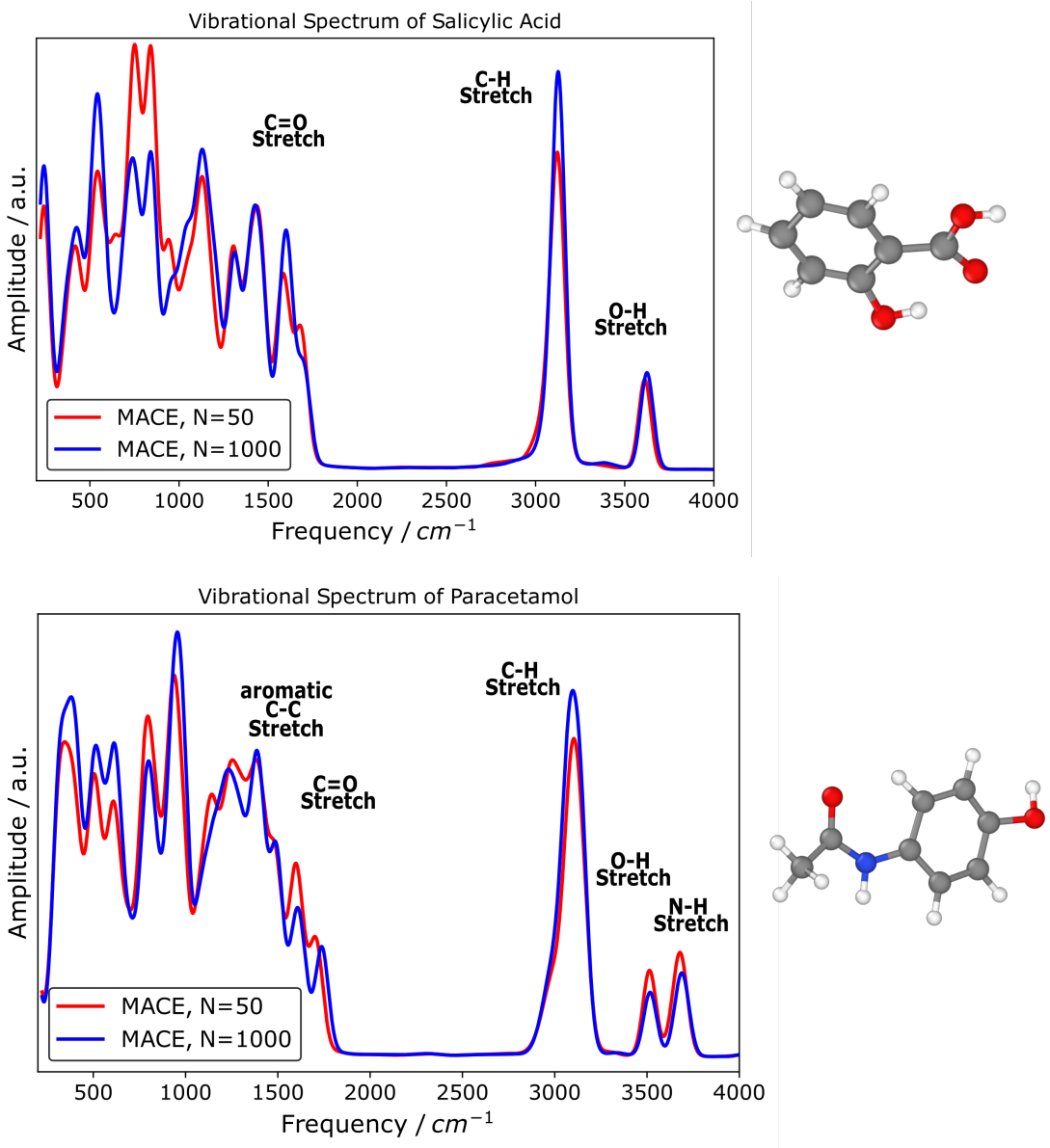


Fig. 5.5 **Vibrational spectrum of larger molecules** The figure illustrates molecular vibrational spectrum computed from the velocity-autocorrelation function. The top panel shows the spectrum of salicylic acid, whereas the bottom panel shows the spectrum of paracetamol.

Being able to train accurate force field models from as few as 50 molecular geometries unlocks the possibility of using much more expensive, higher levels of theory to create the training data. To demonstrate this, the 50 training configurations from rMD17 were recomputed using CCSD(T) level of theory, including the forces [151]. A new MACE model was fitted on this data directly, without transfer learning. The spectrum is also compared to the smallest (`64-0`) transferable organic MACE model that was transfer learned to coupled-cluster data from Section 5.2.5. Custom trained and transferable MACE models are producing very similar spectra as shown in the bottom panel of Figure 5.4. Interestingly, the peaks corresponding to stretching modes involving hydrogens are blue-shifted compared to the experimental positions. This is the result of nuclear quantum effects, and the experimental spectrum can be recovered by including these effects through techniques that treat the nuclei as quantum particles [35]. Note how these shifts are smaller for the original rMD17-derived models, an example of cancellation of errors from missing correlation in DFT and quantum nuclear effects.

5.2.4 MD22 - Large Molecules

In this subsection, the effect of the locality assumption used by MACE is evaluated. The local MACE model is compared to a global machine learning force field, sGDML, which does not use a cutoff to decompose the energy into local site energy terms [37]. In particular, MACE is assessed against the recent improved version of sGDML which uses a reduced set of global descriptors to allow the fitting of systems with hundreds of atoms [110]. Furthermore, MACE is also evaluated compared to the VisNet-LSRM model, which employs a mixed short-range long-range description of the potential energy surface by combining local message passing with long-range message passing between larger fragments [131].

As explained in Section 5.1 the effective cutoff in MACE is the number of layers times the cutoff distance in each layer. In this section, the extent to which information is transferred between MACE layers is also investigated by looking at how property predictions of large molecular systems differ between a two-layer and a single-layer MACE models with the same receptive fields.

For this study, the MD22 dataset is used, which was designed to be challenging for short-range models [40] and was introduced in Section 2.5. The dataset includes large molecules and molecular assemblies containing hundreds of atoms with complex intermolecular interactions. The task is to parameterise a custom force field for each of the different systems using the geometries and the corresponding energy and force labels.

Table 5.3 **Global vs Long-range vs Local models on MD22 dataset of large molecules.** Energy (E, meV/atom) and force (F, meV/Å) mean absolute errors (MAE) of models. The approximate diameter of the system is denoted by d.

	d (Å)		**MACE** 256-2	**MACE** 256-0	**MACE** 256-2	**VisNet-LSRM**	**sGDML**
Cutoff distance			2×3Å	1×6Å	2×5Å	Long-range	Global
Tetrapeptide	∼12	E	0.608	0.345	**0.064**	0.080	0.40
		F	7.6	17.0	**3.8**	5.7	34
Fatty acid	∼16	E	0.446	0.399	0.102	**0.058**	1.0
		F	6.2	23.5	**2.8**	3.6	33
Tetrasaccharide	∼14	E	0.252	0.357	0.062	**0.044**	2.0
		F	6.8	27.0	**3.8**	5.0	29
Nucleic acid (AT-AT)	∼22	E	0.902	0.155	0.079	**0.055**	0.52
		F	13.3	14.9	**4.3**	5.2	30
Nucleic acid (AT-AT-CG-CG)	∼24	E	0.603	0.166	0.058	**0.049**	0.52
		F	16.3	20.1	**5.0**	8.3	31
Buckyball catcher	∼15	E	0.476	0.171	0.141	**0.124**	0.34
		F	13.1	22.2	**3.7**	11.6	29
Double-walled nanotube	∼33	E	0.207	0.231	0.194	**0.117**	0.47
		F	17.9	39.6	**12.0**	28.7	23

Effect of locality on energy and force errors

In order to test the influence of the receptive field of local models on their accuracy, a series of three MACE models with different combinations of cutoff distances and number of layers were trained. The typical MACE model with two layers and 5 Å cutoff at each layer is compared to a one layer MACE with 6 Å cutoff and a two layer MACE with both layers having 3 Å cutoffs.

The best performing MACE model, which employs a 2×5 Å cutoff, improves the errors of the sGDML model by up to a factor of 10. Crucially, even this 10 Å receptive field is considerably smaller than the diameter of the systems in the dataset, so the MACE model is effectively local. Even the 2×3 Å and 1×6 Å models outperform sGDML for all systems. This is likely because the strength of local intramolecular (covalent) interactions is much higher than that of long-range intermolecular interactions. Therefore, a better overall accuracy can be achieved by only improving the short-range description, which MACE evidently does.

In Figure 5.6 the vibrational spectrum of the tetrapeptide is displayed. Again, all models were stable for molecular dynamics simulations without any iterative fitting, and they are showing excellent agreement between them. Even the low frequency part of the spectrum aligns, which corresponds to the larger scale bending modes of the peptide. This plot indicates the validity of the locality assumption for this system.

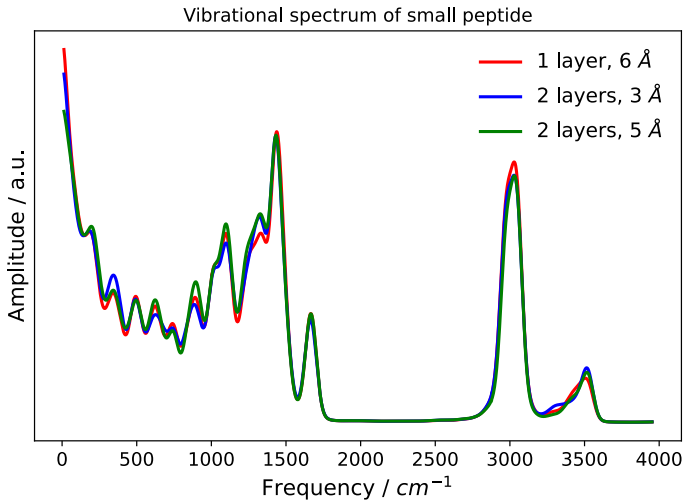


Fig. 5.6 **Vibrational spectrum of a tetrapeptide** The figure illustrates the vibrational spectrum of the tetrapeptide system from the MD22 dataset. It shows that all three MACE models, with different numbers of layers and receptive field, are capable of capturing the vibrational modes accurately.

When comparing MACE to VisNet-LSRM, the best model reported to date, the strictly local MACE model has significantly lower force errors, but in most cases the long-range model has lower energy errors, though it should be noted that the energy errors of both models are on the order of 0.1 meV / atom or lower. This small but consistent improvement in the energies could result from a more accurate description of the remaining small long-range interactions beyond the 10 Å cutoff of the longest range MACE model, or from different relative force and energy weights used during training. It is difficult to come up with benchmarks that specifically test the description of long range interactions and might pose a greater challenge for short range models.

Typical intermolecular interactions appear to be well captured at the 5-6 Å range. The evidence for this is in the energy errors for the nucleic acid and the Bucky-ball catcher systems. In these systems, the intermolecular interactions contribute considerably to the total energy, and the MACE model with 3 Å cutoff in each layer cannot describe them well. In contrast, the two longer-range MACE models have significantly lower energy errors. It is

not a coincidence that most short range ML force field models in the literature use cutoff distances between 5 and 10 Å.

Effect of locality beyond RMSE - the dynamics of the bucky-ball catcher

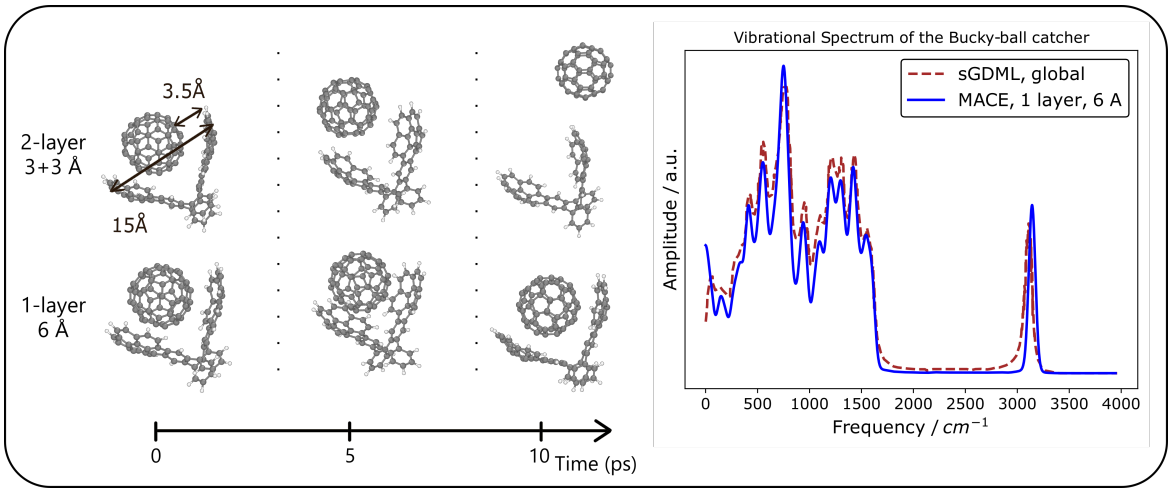


Fig. 5.7 **Bucky-ball catcher MD** The left panel illustrates that the Bucky-ball catcher dissociates for short cutoff MACE model, but stays together with longer cutoff. The right panel compares the molecular vibrational spectrum computed using local MACE and the global sGDML model

Next, the Bucky-ball catcher system is investigated in more detail because this is a system where intermolecular interactions play a crucial role in the dynamics. Following the experiment in Ref. [40] a 200 ps, NVT molecular dynamics simulation was performed with each of the three MACE models introduced in the previous subsection. As shown in the left panel of Figure 5.7 for the 2×3 Å MACE model, the system dissociated into the Bucky-ball and the catcher within 10 ps. The model, which on the whole has excellent accuracy (just 0.5 meV/atom for energies and 13 meV/Å for forces), is unable to fit the attractive dispersion between the two sub-parts of the system because they are typically further than 3 Å apart. On the other hand, the two slightly longer ranged MACE models provide qualitatively correct dynamics. The molecular vibrational spectrum was also computed by taking the Fourier transform of the velocity-velocity auto-correlation function to analyse the dynamics quantitatively. In the right panel of Figure 5.7, the spectrum of the 1-layer 6 Å MACE model is compared to the spectrum published in Ref. [40] computed using sGDML. The figure shows excellent agreement including for low frequencies, demonstrating that a strictly local model can simulate the dynamics of systems even when intermolecular interactions are important.

Table 5.4 **Mean Absolute Errors on the COMP6 benchmark dataset** Total energies are given in kcal/mol, forces in kcal/mol/ Å. Note, that the ANI-1x model was trained on 10× more data than the other models. *For NewtonNet, the decomposition of errors for the subsets was not published and conformations of molecules whose energies were outside a 100 kcal/mol energy range were omitted from the testing.

		ANI-1x	GM-NN	NewtonNet	TensorNet	MACE 64-0	MACE 96-1	MACE 192-2
ANI-MD	E	3.40	3.83	-	**1.61**	10.3	2.81	3.25
	F	2.68	1.43	-	0.82	1.92	0.89	**0.62**
DrugBank	E	2.65	2.78	-	0.98	1.81	1.04	**0.73**
	F	2.86	1.69	-	0.75	1.20	0.70	**0.47**
GDB 7 - 9	E	1.04	1.22	-	0.32	0.77	0.40	**0.21**
	F	2.43	1.41	-	0.53	0.96	0.54	**0.34**
GDB 10 - 13	E	2.30	2.29	-	0.83	1.54	0.88	**0.53**
	F	3.67	2.25	-	0.97	1.52	0.92	**0.62**
S66x8	E	2.06	2.95	-	0.62	1.17	0.69	**0.39**
	F	1.60	0.93	-	0.33	0.65	0.33	**0.22**
Tripeptides	E	2.92	3.06	-	0.92	2.10	1.18	**0.79**
	F	2.49	1.48	-	0.62	1.09	0.66	**0.44**
COMP6 total	E	1.93	2.03	1.45*	-	1.47	0.76	**0.48**
	F	3.09	1.85	1.79*	-	1.31	0.77	**0.52**

5.2.5 COMP6 - Benchmark of Transferable Small Molecule Force Fields

In this subsection, the MACE architecture is evaluated for training transferable organic force fields on large datasets. This task is rather different from the previous ones presented in the thesis, which focused on parameterising custom force fields for each chemical system. To train the model, a subset of the ANI-1x dataset was used, which contains coupled cluster calculations [201]. A series of MACE models was trained, going from a small (`64-0`) MACE model to a medium (`96-1`) and a large (`192-2`) model.

First, the models were trained using DFT energies and forces and tested using the COMP6 benchmark suite [198]. The results are summarised in Table 5.4. Even the smallest MACE model outperforms most previously published models [198, 234, 86]. The large MACE model improves on the previous state-of-the-art by about a factor of 5, achieving an overall error well below 0.5 kcal/mol. The only model with performance comparable to at least the medium MACE model is another equivariant neural network, TensorNet [193].

The ANI-MD subset is one where the MACE total energy errors are relatively high compared with the other subsets. This subset is comprised of configurations of 14 different molecules sampled from ANI-MD trajectories. The MACE errors are relatively low on 12 of

the 14 molecules. However, on the two largest ones (Chignolin - 149 atoms and TrpCage - 312 atoms), the MACE energy is shifted by a constant value in comparison to the DFT reference energy. The energy-energy correlation plot for all molecules in this subset is shown in Figure 5.8. It can be seen that the MACE model predicts the relative energies of the molecular conformers very accurately, but in the case of the two largest molecules at the bottom two panels, the MACE energies are systematically shifted compared to the reference energies. Predicting the energy of the larger molecules is made particularly challenging by the fact that the training set contains very few molecules with more than 50 atoms; hence, when testing on larger systems there can be an accumulation of small errors that is not controlled in the training. A simple test of this hypothesis could be the addition of a small number of larger molecules to the training set. Another possible reason for this systematic error is that it results from missing long-range non-bonded contributions to the energy. This is contradicted by the fact that the shift is going down during training, albeit very slowly towards the end.

Transfer learning to the coupled cluster level of theory was also performed using the models. This step only involved the energy labels, as the coupled cluster forces were not available, following Ref. [234]. One of the coupled cluster accurate models was used in the ethanol example in Section 5.2.3. The resulting six pre-trained MACE models are published and available to download from the MACE GitHub page [120].

Biaryl torsion benchmark

Next, the coupled cluster transfer learned versions of the above MACE organic force fields are evaluated on the challenging biaryl dihedral torsion dataset introduced in Ref. [126]. This dataset consists of 88 small drug-like molecules with different biaryl dihedral torsional profiles. Such a benchmark is of particular interest in connection with small molecule drug discovery. The description of torsional barriers in small molecules is a typical task in which classical empirical force fields cannot provide sufficiently accurate descriptions of the potential energy surface [176, 99].

MACE dihedral torsional scans were carried out on the subset of the full data set that contains only H, C, N and O chemical elements (78 different molecules). In the following the largest, most accurate MACE model (`192-2`) transfer learned to CCSD(T) level of theory is compared to the ANI-1ccx model which was trained on the same dataset. The MACE model achieves a mean absolute barrier height error compared to CCSD(T) of 0.36 kcal/mol, improving significantly on the ANI-1ccx model, which was identified as the best in Ref. [126] and has an error of 0.78 kcal/mol. The upper panels of Figure 5.9 show examples of torsional profiles identified in Ref. [126] as particularly challenging. The MACE models are able

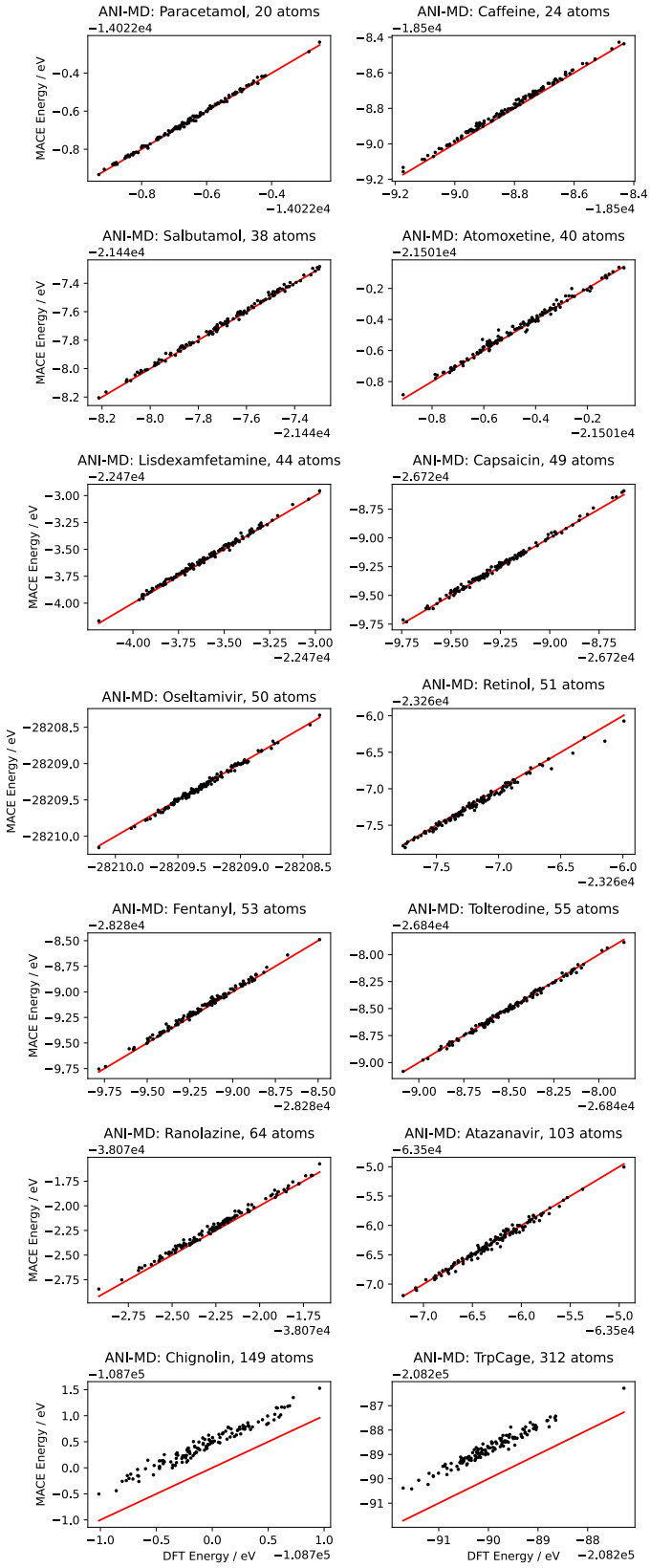


Fig. 5.8 **ANI-MD subset energies** The Figure illustrates the ANI-MD subset of COMP6 showing MACE vs DFT energy correlation.

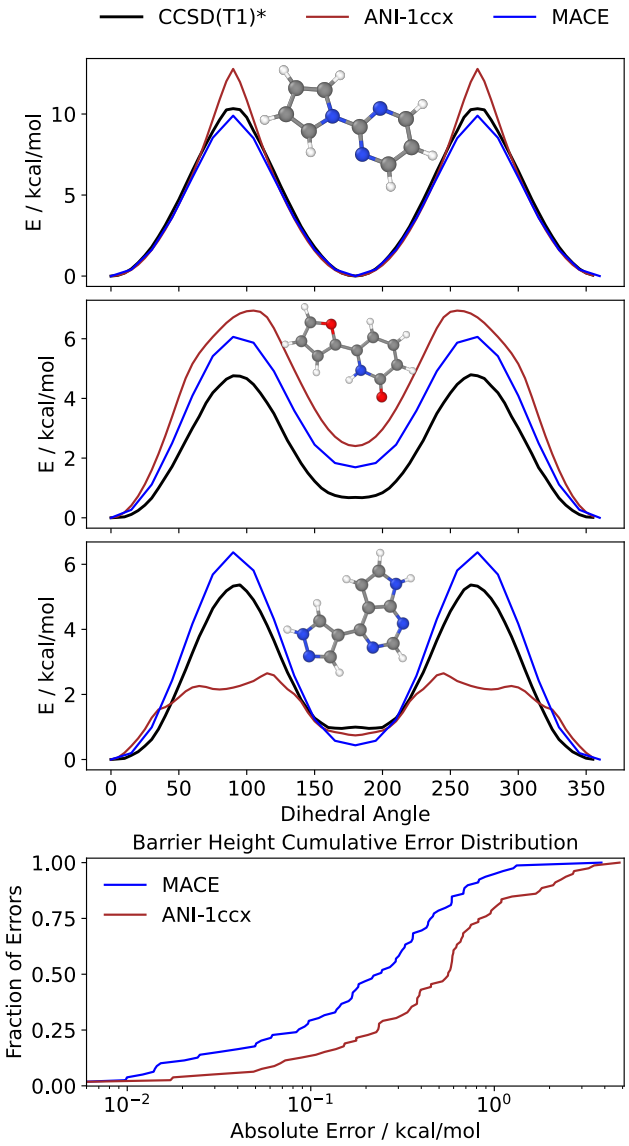


Fig. 5.9 **Dihedral scans** The top three panels illustrates a selection of challenging dihedral torsional scans computed using MACE 192-2 from Ref. [126] where the ANI force field has particularly large errors. The bottom panel compares MACE 192-2 and ANI-1ccx torsional barrier height errors.

to describe these profiles quantitatively accurately for the first and third cases and match qualitatively in the second case. In the bottom panel of Figure 5.9 the cumulative error distribution of MACE and ANI is displayed. The figure shows that MACE has very few molecules for which it makes an error close to or larger than 1 kcal/mol. Notably, even the medium and small MACE models perform better or close to the same as the ANI model with average barrier height errors of 0.56 and 0.83 kcal / mol.

Table 5.5 **Energy and Force errors on liquid water dataset** Ref. [36]

	BP-NN [36]	**REANN** [241]	**NequIP** [17] (L=2)	**MACE** 64-0	**MACE** 192-2
E RMSE (meV / H_2O)	7.0	2.4	-	**1.9**	**1.9**
F RMSE (meV / Å)	120	53.2	-	37.1	**36.2**
E MAE (meV / H_2O)	-	-	2.5	**1.2**	**1.2**
F MAE (meV / Å)	-	-	21	20.7	**18.5**

Finally, it is important to remark that the MACE potential energy surfaces are found to be smooth, resulting in rapidly converging constrained geometry optimisations across the entire set. This is in contrast to the experience of slow convergence of geometry optimisations with previous generations of machine learning force fields [92].

5.2.6 Water Structure and Dynamics

To assess the ability of MACE to describe complex molecular liquids, it can be tested in the description of water. MACE models were trained on a dataset of 1593 liquid water configurations, made up of 64 molecules each [36]. The QM labels of the dataset were computed using the CP2K software [123] at the revPBE0-D3 level of density functional theory which is known to give a reasonably good description of the structure and dynamics of water at a variety of pressures and temperatures [141].

Table 5.5 compares the energy and force errors of the MACE model with other machine learning force fields trained on the same dataset, but using different train test splits. The 3-body Atom-Centred Symmetry Function based feed-forward neural network model (BPNN) has the highest errors. The 3-body invariant message passing model REANN [241] and the 2-body equivariant message passing model NequIP [17] significantly improve the errors compared to the BPNN model. A further improvement is achieved by the many-body equivariant MACE model. Interestingly a relatively small MACE model using invariant messages, but having an overall body order of 13 achieves already lower errors than the other best models and the larger MACE model only slightly improves on the force errors. This suggests that the model might be getting close to the inherent noise of the training and test labels.

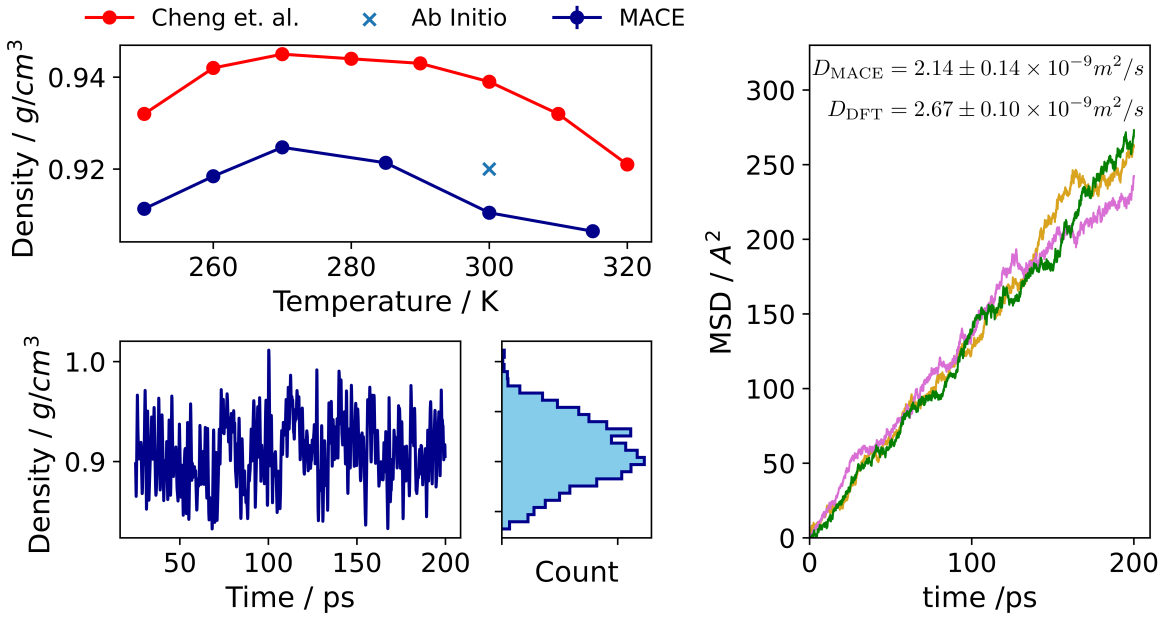


Fig. 5.10 **Thermodynamic properties of liquid water** The top left panel shows the liquid water density isobar at $p = 1.0$ bar pressure and compares the isobar from from Cheng et. al. [36] and the *ab initio* density [154] to MACE `64-0`. The bottom panel shows an example of the density distribution in the NPT simulations. The right panel shows the mean squared displacement of the water molecules in 3 independent NVT simulations at the equilibrium density (0.91 g / cm^3) of MACE. The corresponding DFT value was obtained from a simulation at the DFT equilibrium density of 0.92 g / cm^3.

Thermodynamics and kinetics of liquid water

To characterise the smaller and faster MACE water model 200 ps NPT simulations were performed at a range of temperatures from 250 K to 315 K using a combined Nose-Hoover and Parrinello-Rahman barostat [143, 142]. As shown on Figure 5.10 the average density was calculated after an initial equilibration period at each temperature. The top-left panel shows the water density isobar, showing the characteristic density maximum at around 270 K. This is somewhat lower than the experimental value, but is consistent with previous DFT based studies of water [36]. At 300 K the density obtained using the MACE model can be compared to the available ab initio value from Ref. [154] showing very good agreement.

Finally, a dynamic property, the diffusivity of water can also be investigated. This is a property that is notoriously difficult to model accurately [136]. To obtain the diffusivity, a water configuration from the NPT simulation was taken with the equilibrium density of 0.91 g / cm^3. It was used as the initial structure for 3 independent 200 ps long NVT simulations. The mean squared displacements from these simulations are shown in the right panel of Figure 5.10. By fitting a linear function on the diffusive part of the MSD, the value of the

diffusion coefficient can be computed. It is found to be $2.14 \pm 0.14 \times 10^{-9} m^2/s$ which is in reasonably good agreement with the ab initio value estimated from much smaller simulations in Ref. [141].

5.3 MACE-OFF23: Transferable Organic Force Field

5.3.1 Motivation for Transferable Organic Force Fields

Machine learning force fields have undergone major improvements in their accuracy, robustness, and computational speed, as exemplified in Chapter 3 and Chapter 5 [16, 17, 120, 135]. They are now routinely used in materials chemistry simulations, where density functional theory was previously the method of choice. In these applications, available empirical force fields, such as the embedded-atom method [52], do not provide sufficient accuracy and transferability to describe many scientifically interesting and challenging phenomena.

In contrast, simulating bio-organic systems entails a different set of trade-offs, with greater emphasis on simulating systems over long timescales. This has meant that empirical force fields, which sacrifice accuracy for computational speed, continue to be used routinely for studying molecular liquids, crystals, biological systems, and drug-like molecules [27, 51, 74, 87].

Two alternatives to empirical force fields are available in molecular chemistry applications. The first is semi-empirical quantum mechanics, such as the series of extended tight-binding models [9], which represents a low-cost solution for small molecules. The method is limited by its moderate accuracy compared to quantum chemistry methods, its restriction to modelling non-periodic systems, and its cubic scaling with system size.

Second, a number of transferable machine learning force fields have also been developed for organic chemistry. The most notable are the series of ANI [58, 197–199] and AIMNet potentials [5, 243, 244]. ANI potentials pioneered the use of local symmetry function-based feedforward neural networks [20] trained on a large dataset of organic molecular geometries [196, 200] to create transferable force fields. The ANI-2x model became the most widely adopted ML force field and therefore serves as one of the primary points of comparison in this section. The ANI-2x model was recently combined with a polarisable electrostatic model [102] in a hybrid ML/MM simulation setting, and also with a neural network based dispersion correction [209]. The AIMNet models apply a message passing architecture [79], where the initial embeddings are the ANI symmetry functions. Compared to ANI, AIMNet extends the applicability of the models to a larger set of chemical elements, as well as to charged species. These models relax the locality assumptions by incorporating electrostatic

and dispersion interactions. The PhysNet model uses a message passing architecture and in addition to the semi-local terms also includes long-range electrostatic and dispersion interactions [212].

Further recent (bio)organic force fields include the FENNIX model, which combines a local equivariant machine learning model with a physical long-range functional form for electrostatics and dispersion [161]. The model was trained to reproduce the CCSD(T)/CBS energies of small molecules and molecular dimers. It was shown that such ML force fields can be used to run stable dynamics of liquid water, the solvated alanine dipeptide, and an entire protein in the gas phase. However, wider benchmarking is required to assess the accuracies of the intramolecular potential outside the training set and of condensed phase molecular dynamics simulations.

Similarly, the ANA2B potential employs a short-ranged two molecular body ML potential, with long-ranged, classical multipolar electrostatics, polarisation and dispersion interactions [207]. Although this long-ranged model shows promising accuracy for condensed phase properties and crystal structure ranking, its accuracy and computational performance has not yet been demonstrated for larger biomolecules. Finally, the GEMS model [213], built on the SpookyNet architecture [211], is another recent ML force field for biomolecular simulations. Although SpookyNet has demonstrated application to more challenging condensed phase simulations, including protein dynamics, these required a significant number of additional reference quantum chemistry calculations on relevant large peptide fragments for each new simulation to obtain a stable model. Therefore, this model is not a transferable force field.

In this section, MACE-OFF23, a new transferable force field for organic molecules is introduced. It is based on the MACE force field architecture and was trained to reproduce first-principles reference data computed with a high level of quantum mechanical theory. MACE-OFF23 demonstrates the remarkable capabilities of local, short-range models by accurately predicting a wide variety of gas and condensed phase properties of molecular systems. It produces accurate, easy-to-converge dihedral torsion scans of unseen molecules as well as reliable descriptions of molecular crystals and liquids, including quantum nuclear effects. In Ref. [121], the capabilities of the models are further demonstrated for the determination of free energy surfaces in explicit solvent, as well as the for the folding dynamics of peptides. The model is also shown to be capable of simulating a fully solvated small protein, observing accurate secondary structure and vibrational spectrum. These developments enable first-principles simulations of molecular systems for the broader chemistry community at higher accuracy and lower computational cost compared to the previously available methods.

In this section, the training details of the MACE-OFF23 model and a subset of the experimental validation results are presented. The rest of the test examples were carried out by co-authors and can be found in Ref [121].

5.3.2 Training Data

The core of the MACE-OFF23 training set is the SPICE dataset [64]. Table 5.6 provides a summary of both the training and test sets. The MACE-OFF23 model is trained to reproduce the energies and forces computed at the ωB97M-D3(BJ)/def2-TZVPPD level of quantum mechanics [83, 84, 149, 172, 224], as implemented in the PSI4 software [195]. A subset of SPICE was selected that contains the ten chemical elements H, C, N, O, F, P, S, Cl, Br, and I and has a neutral formal charge excluding the ion pair subset. Approximately 85% of the SPICE dataset was retained. The geometries within the SPICE dataset were generated by running molecular dynamics simulations using classical force fields [138] and sampling maximally different conformations from the trajectories [64].

Table 5.6 Summary of training and test sets

	PubChem	DES370K Monomers	DES370K Dimers [60]	Dipeptides	Solvated Amino Acids	Water	QMugs [104]	Tripeptides
Chemical elements	H, C, N, O, F, P, S, Cl, Br, I	H, C, N, O, F, P, S, Cl, Br, I	H, C, N, O, F, P, S, Cl, Br, I	H, C, N, O, S	H, C, N, O, S	H, O	H, C, N, O, F, P, S, Cl, Br, I	H, C, N, O
System size	3-50	3-22	4-34	26-60	79-96	3-150	51-90	30-69
# Train	646821	16861	263065	19773	948	1597	2748	0
# Test	33884	889	13896	1025	52	84	144	898

The SPICE dataset only contains small molecules of up to 50 atoms. To facilitate the learning of intramolecular non-bonded interactions, the dataset was augmented with larger 50–90 atom molecules randomly chosen from the QMugs dataset [104]. These geometries were generated by running molecular dynamics simulations using GFN2-xTB [9] similarly to the protocol described in Ref. [47]. The energies and forces were re-calculated at the level of QM theory used in SPICE. Finally, to obtain a better description of water, the dataset was further augmented with a number of water clusters carved out of molecular dynamics simulations of liquid water [185], with sizes of up to 50 water molecules. 95% of each subset of the the final dataset was used for training and validation, and 5% for testing.

In addition, part of the COMP6 tripeptide geometry dataset [198] was also recomputed at the SPICE level of theory and used as part of the test evaluation.

After training the small MACE model and observing the training errors, the presence of outliers was observed in the dataset. This is probably caused by errors in the underlying

electronic structure calculations, some of which have been documented on the SPICE GitHub repository (https://github.com/openmm/spice-dataset). To address this, the configurations that had a maximum force error greater than 2 eV/Å were removed from the training set. This meant the removal of just 808 configurations. Many of the configurations contained heavy elements, in particular phosphorus and iodine, which might have a more challenging electronic structure. The outliers with the largest errors were re-evaluated, using the level of DFT used in SPICE, and found that for about a third of the configurations, the recomputed energies and forces agreed well with the MACE prediction and not with the original DFT labels. This observation further validates the hypothesis that the outliers were failed DFT calculations.

5.3.3 Training Details

The MACE model has parameters that enable systematic control of model expressivity (accuracy) against computational cost, as discussed in Section 5.1. In the following, three variants of the MACE-OFF23 model are presented, a small, a medium, and a large one, denoted in the text as MACE-OFF23(S), MACE-OFF23(M) and MACE-OFF23(L) respectively. The hyperparameters of the models are displayed in Table 5.7.

Table 5.7 Hyperparameters of the three MACE-OFF23 models

	Small	Medium	Large
Cutoff radius (Å)	4.5	5.0	5.0
Chemical channels k (Eq. (5.1))	96	128	224
max L (Eq. (5.9))	0	1	2

The MACE-OFF23(S) model has 96 channels, $L = 0$ invariant messages and a cutoff of 4.5 Å. The MACE-OFF23(M) model has 128 channels and $L = 1$ messages, and finally the MACE-OFF23(L) has 224 channels and $L = 2$ equivariant messages. The medium and large models have a cutoff of 5.0 Å in each layer. All models used two layers and a body order of 4 in each layer. The models were trained using the PyTorch [157] implementation of MACE, available on https://github.com/ACEsuit/mace.

During training, initially, the force weight in the loss in Equation (5.14) was set to 1,000 and the energy weight to 40. The learning rate was 0.01 and Adam optimiser with Amsgrad was used. The exponential moving average of the weights was taken in each training step. When the force error converged, the second phase of the training was started with force weight 10 and energy weight 1000 and the learning rate was reduced to 0.00025. Finally,

the training was terminated when the energy error also stopped decreasing significantly. All models were trained on a single Nvidia A100 GPU. Training the small model took about 6 days, the medium about 10 days, and the large model 14 days.

5.3.4 Results

Extended SPICE Test Set

First, the pointwise errors of the energy and force predictions are evaluated on a held-out test set for each of the three `MACE-OFF23` models. Figure 5.11 shows the per-atom energy and force mean absolute errors. As the size of the model increases, the models gradually become more accurate, with the large model achieving errors of around 0.5 meV/atom and 15 meV/Å, well below the 1 kcal/mol (43 meV) chemical accuracy limit for the organic molecules studied here.

The last column of Figure 5.11 looks at an extrapolation task, the training set contains only dipeptides, and this test set looks at tripeptides, indicating that the models are able to extrapolate to larger fragments with more complex interactions.

Dihedral scans

Next, the performance of the `MACE-OFF23` model is evaluated on dihedral scans of drug-like molecules. This task is routinely carried out using quantum mechanical methods to create reference data for re-parameterising classical empirical force field dihedral terms [100]. The task is particularly challenging, as constrained geometry optimisations can be difficult to converge if the potential energy surface is not sufficiently smooth [92]. This has been observed in particular for the ANI family of potentials, which are known to have convergence difficulties in geometry optimisation tasks, as also discussed in Section 5.2.5, where the MACE and ANI architectures were compared when they are trained on the same training data.

TorsionNet-500 The top panel of Figure 5.12 summarises the results for the TorsionNet-500 dataset [167]. This dataset contains torsion drives of 500 different molecules, selected to cover a wide range of pharmaceutically relevant chemical space. The original data set was reported at the B3LYP/6-31G(d) level of DFT theory. For consistency with the SPICE dataset, the torsion profiles were recalculated using the DFT setting of SPICE, which is a higher level of theory.

The first panel shows an example of a torsion drive, indicating the complex energy profile that the MACE models are able to capture closely, including geometries far from equilibrium.

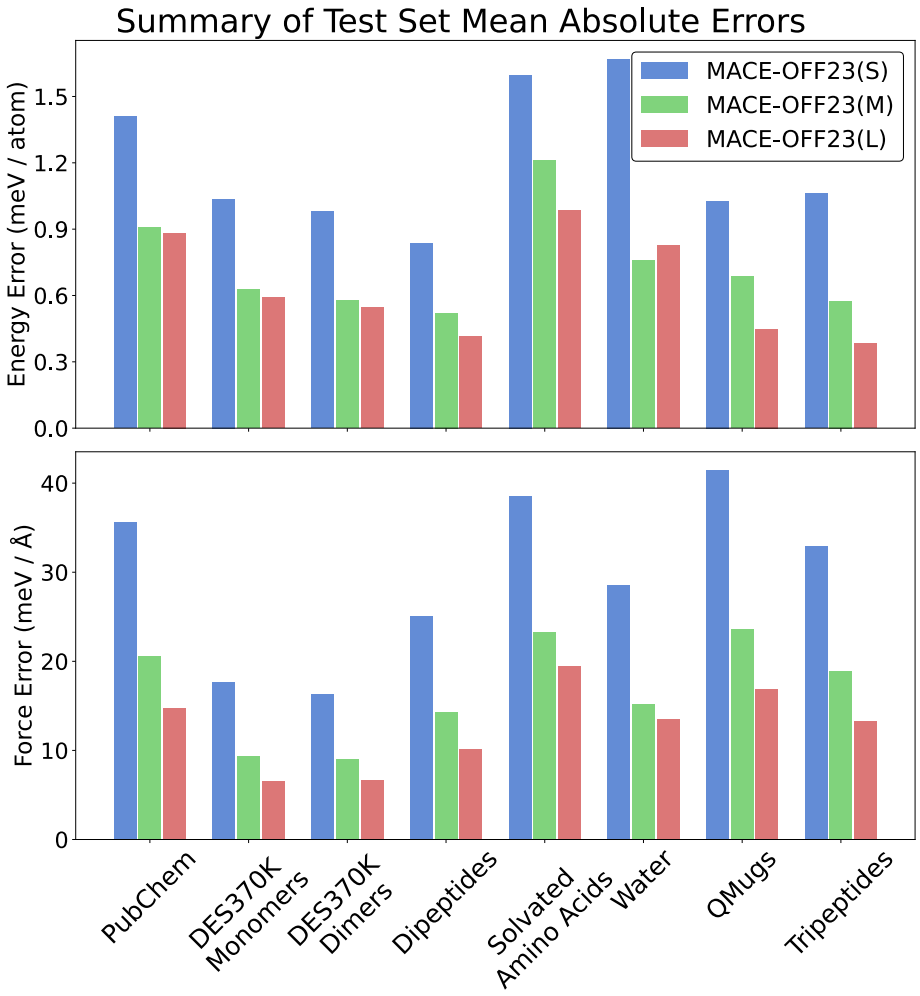


Fig. 5.11 **Test set mean absolute error.** Errors in the MACE-OFF23 models compared to the underlying DFT reference data, highlighting the relative accuracy of the three models.

The central panel shows the mean barrier height error of a number of representative models, comparing the Sage classical empirical force field [27], a semi-empirical quantum mechanical method GFN2-xTB [9], a recent transferable machine learning force field AIMNet-2 [5], and the three MACE-OFF23 models. Again, systematic improvements in accuracy with the size of the MACE model are observed, with medium and large models, in particular, achieving errors of around 0.25 kcal/mol compared to the reference method. The AIMNet-2 model achieves comparable accuracy to the small MACE-OFF23 model, a significant improvement compared to the previous generation ANI models. A similar conclusion can be drawn from the comparison of the molecular geometries by looking at the root mean squared deviation of the atomic positions averaged over the full torsion scans, as indicated by the top right panel of Figure 5.12. It shows that MACE optimised geometries have a deviation of about 0.025 Å,

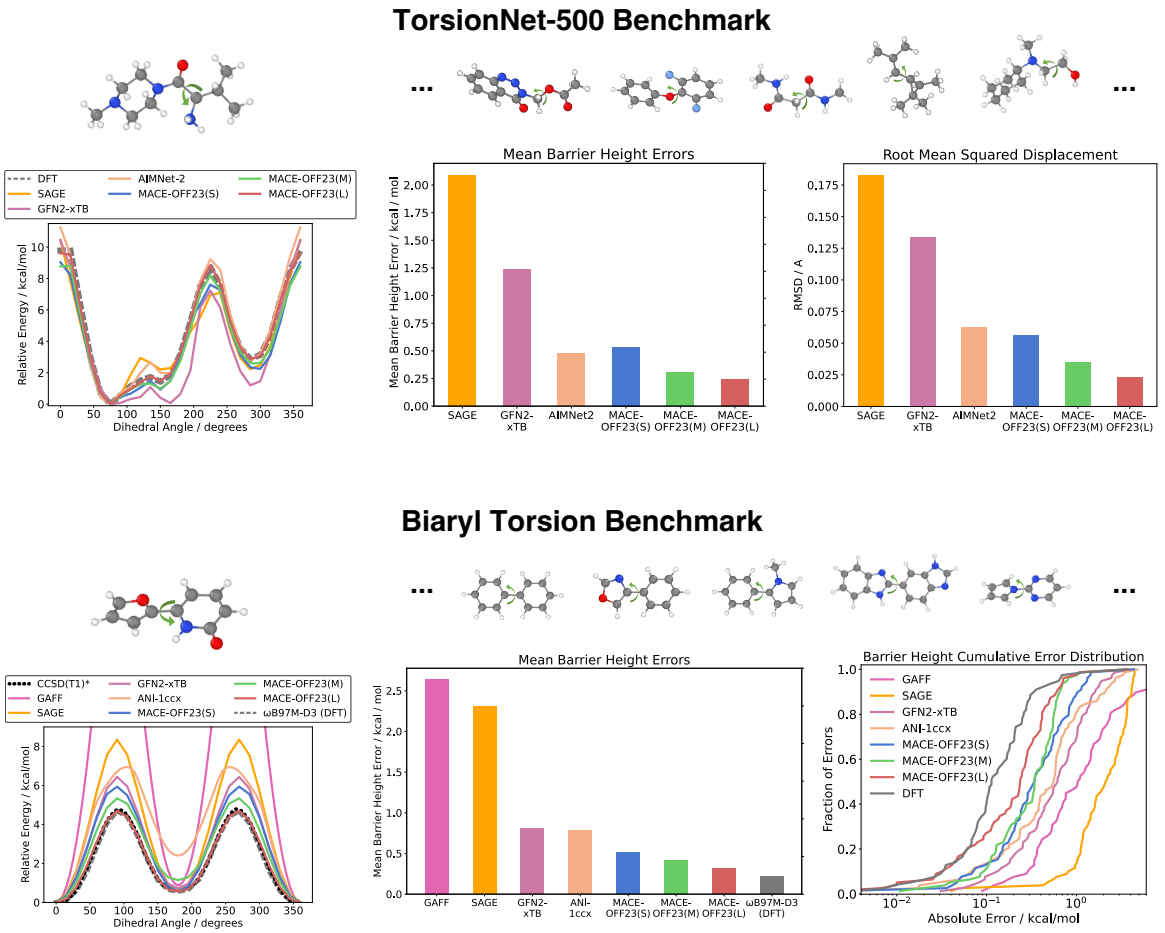


Fig. 5.12 **Dihedral benchmark scans.** The top panel shows torsion drive data for the TorsionNet-500 dataset [167], which has a wide chemical diversity (five example molecules are shown). The bottom panel focusses on the torsion angle between two aromatic rings in the biaryl torsion benchmark [127] which contains 78 molecules (five examples are shown).

meaning they are almost indistinguishable from DFT optimised structures. It is important to note that the different models were trained to different levels of DFT, which might also contribute to the observed differences.

Biaryl fragments The biaryl torsion benchmark was introduced in Section 5.2.5, here the same subset of 78 molecules is used to evaluate the `MACE-OFF23` models.

In the bottom panel of Figure 5.12, the results of the torsion drives are compared for empirical force fields, semi-empirical QM methods, and the ANI-1ccx machine learning force field [199] to the `MACE-OFF23` models. The DFT potential energy surfaces (using the

SPICE level of theory) are also displayed alongside the published gold standard coupled cluster data.

The DFT torsion drives achieve a mean barrier height error of 0.2 kcal/mol, which is therefore the best theoretically possible result using the MACE-OFF23 models. The MACE-OFF23(L) model comes close to this, with a mean absolute error of 0.3 kcal/mol. The medium and small MACE models have barrier height errors of 0.4 and 0.5 kcal/mol, respectively. Remarkably, the small MACE model is significantly more accurate than the next-best non-DFT methods, ANI-1ccx and GFN2-xTB, as illustrated in the bottom-centre plot of Figure 5.12. In particular, unlike MACE-OFF23, coupled cluster reference calculations were used to parameterise the ANI-1ccx and GFN2 models. In the bottom right, the cumulative error distributions are also shown to verify that the MACE-OFF23 barrier height errors are not only accurate on average but are also robust, having essentially no outliers.

These examples showcased the potential of MACE-OFF23 for the quick and easy calculation of dihedral torsion profiles for drug-like molecules.

Molecular crystals

In the following, the ability of the MACE-OFF23 force fields to simulate the vibrational and thermal properties of molecular crystals is demonstrated. This is an out of domain test, as the model was trained only on individual molecules and molecular dimers in vacuum, but not on periodic molecular crystals.

Vibrational spectroscopy of paracetamol Raman spectroscopy is one of the most widely used techniques for characterising molecular crystals. Unlike IR spectroscopy, which only detects vibrational modes that distort dipoles, Raman spectroscopy is more sensitive to collective modes governed by weak, non-bonded interactions in a broad range of molecular materials. The low-frequency region of the Raman spectrum (e.g., the THz regime) gives a vibrational fingerprint of the intermolecular interactions. Thus, it is widely used to differentiate between polymorphs of molecular crystals. Meanwhile, the high-frequency Raman spectrum probes intramolecular vibrational modes and their coupling to low-frequency modes.

Here, the MACE-OFF23(S) model is tested for the prediction of the Raman spectrum of the “Form II" polymorph of paracetamol. To compare MACE-OFF23(S) directly with experiments, quantum nuclear effects are rigorously incorporated using a recently introduced ML-aided framework [111]. In particular, a bespoke effective potential energy surface [148] is fitted also using the MACE architecture to calculate quantum nuclear corrections to the MACE-OFF23(S) model within the path-integral coarse-grained simulations (PIGS) method.

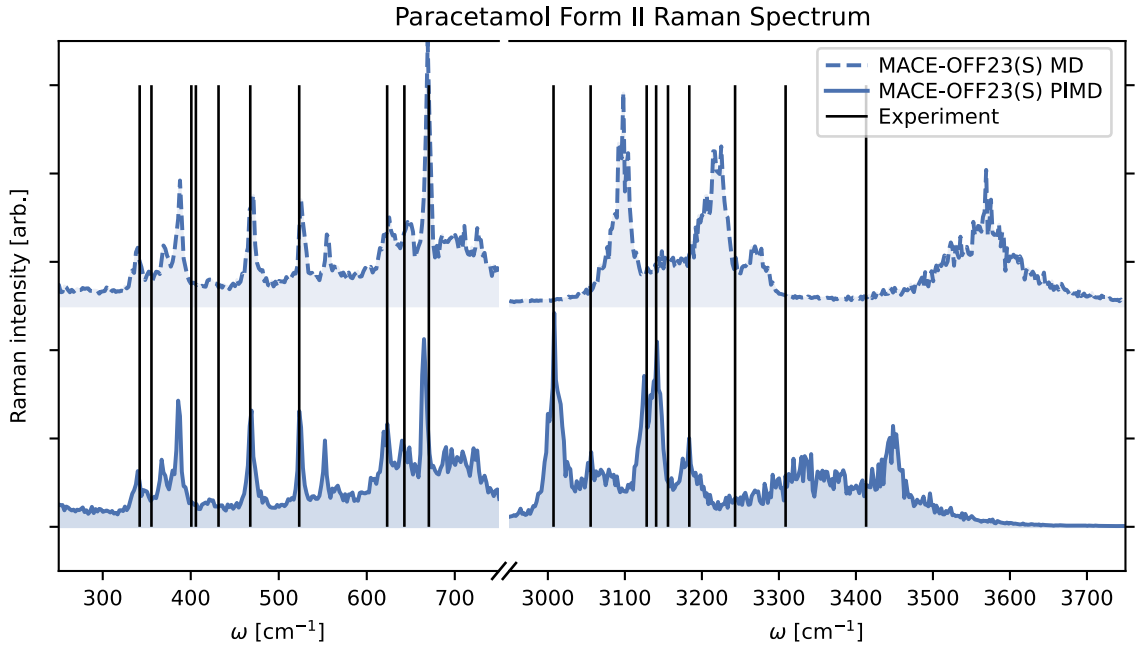


Fig. 5.13 **Powder Raman spectrum of paracetamol form II [168].** Spectrum computed at ambient conditions using the `MACE-OFF23(S)` model for the potential energy surface, a MACE model of the polarizability, and a MACE model that incorporates quantum nuclear effects on the potential energy surface using the PIGS approach [148]. The black lines represent experimentally determined band positions [116].

A separate MACE model with equivariant readout function is also fitted to the first-principles polarisability of paracetamol polymorphs (data taken from Refs. 169, 170). The remaining steps to produce the spectra involve a classical calculation involving a NVE molecular dynamics simulation on the effective PIGS PES, prediction of the isotropic and anisotropic components of the polarisability tensor, and calculation of their time correlation functions following Ref. [111].

As shown in Figure 5.13, both the high- and low-frequency regions of the Raman spectrum of paracetamol form II can be predicted with overall good agreement with the experimental band positions [116]. Since the experiment captures the Raman spectra along different crystal directions while the computational experiment estimates the "powder" Raman spectrum [170], the band intensities cannot be directly compared. It is interesting to note that the predictions based on MACE classical MD are consistently shifted with respect to the experiment. At the same time, quantum nuclear predictions encoded by the PIGS method play an essential role in improving the agreement between theory and experiments. Moreover, a broad

band at around 3300 cm^{-1} is only captured at the level incorporating quantum nuclear effects.

Overall, this example provides evidence that the MACE-OFF23 transferable force fields are capable of accurately simulating the vibrational spectrum of molecular crystals, which only a few years ago was only possible with custom made machine learning or classical force field models [170].

Lattice enthalphies Next, the MACE-OFF23 force fields are used to describe the structure and stability of molecular crystals. The enthalpies of sublimation were computed for a range of 23 representative small molecular crystals [174] following the protocol of Ref. [59].

Figure 5.14 compares the predicted sublimation enthalpies with the experimentally measured ones. This task is often used to test various DFT functionals. Since the ωB97M-D3(BJ) functional used to parameterize MACE-OFF23 does not have a periodic implementation, an estimate of the highest achievable accuracy is not available. The figure shows that the three MACE models are capable of capturing trends and have higher accuracy than the ANI-2x model, which was not designed for tasks involving molecular crystals. The MACE-OFF23(L) model achieves a mean error of just 1.7 kcal/mol, which is comparable to the errors of several different dispersion-corrected density functionals for a fraction of the computational cost [174].

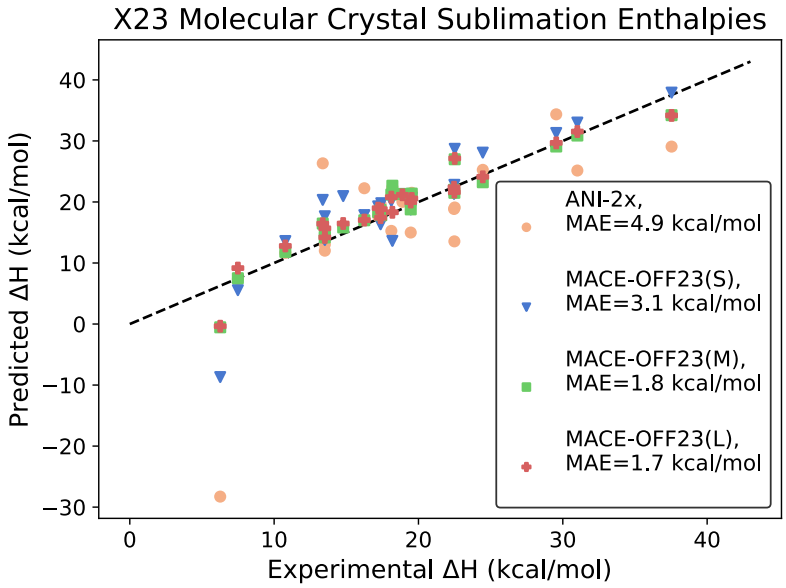


Fig. 5.14 **Sublimation enthalphies of molecular crystals.** Comparison between predicted sublimation enthalphies of the MACE-OFF23 and ANI models and experiment.

The relaxed unit cell vectors of the MACE-OFF23 models were also compared to the values measured experimentally as shown in Table 5.8. The MACE-OFF23(L) model has a relative error as low as 5%, which is in close agreement with the experimental values.

Table 5.8 **Crystal structure geometries** The table compares the MACE-OFF23(L) and experimental lattice vectors for the X23 set of organic molecular crystals.

Molecule	T_{exp}		Experiment	MACE-OFF23(L)
Acetic acid	40	a	13.151	13.359
		b	3.923	3.809
		c	5.762	5.542
		V	297.27	282.01
Ammonia	2	a	5.048	4.946
		V	128.63	120.98
Benzene	4	a	7.351	6.59
		b	9.364	9.303
		c	6.695	6.851
		V	460.84	420.51
Naphthalene	10	a	8.0846	7.821
		b	5.9375	5.836
		c	8.6335	8.430
		β	124.67	125.26
		V	340.83	314.23
Pyrazine	184	a	9.325	9.351
		b	5.850	5.508
		c	3.733	3.545
		V	203.64	182.57
Urea	40	a	5.565	5.330
		c	4.684	4.670
		V	145.06	123.68

5.3.5 Conclusions

In Ref. [121] there are several further example use-cases of the MACE-OFF23 models. These include several biochemical applications, such as the simulation of a small protein in explicit solvent. Thanks to using a purely local short-ranged functional form, the models are capable of simulating 10-s of thousands of atoms, whilst still keeping the computational cost relatively low. This is further aided by a series of custom-made accelerated CUDA kernels. A detailed analysis of the computational speed of the MACE-OFF23 models can also be found in Ref [121].

The accuracy, extrapolation capabilities, and computational speed demonstrated above make the MACE-OFF23 models a good starting point for many molecular chemistry applications and projects. There are a number of limitations, though, that will be addressed in its next version. The lack of explicit long-range interactions limits the domain of applicability of the present model to neutral, non-radical, and non-reactive systems. This is something that the recently published AIMNet-2 model addresses by extending the ANI models to include charged species and long-range interactions [5] These models were also trained on a more extensive training set than ANI-2x or the MACE-OFF23 models, with a training set size of about 20 million molecules.

The next generation of the MACE-OFF23 models will similarly include an explicit description of charges, enabling the description of amino acids with different protonation states, charged nucleic acids, and counter-ions. This will pave the way towards obtaining an accurate quantum mechanical transferable machine learning force field for simulating a wide range of biologically relevant systems.

Chapter 6

Conclusions and Outlook

6.1 Summary

This thesis presents theoretical, methodological, and practical advances that improve molecular simulations. The machine learning force field methods developed enable the simple parameterisation of accurate and computationally efficient potential energy surfaces in a systematically improvable way.

In particular, linear ACE gives the basic framework for this work, providing a rigorous linearly complete basis for fitting functions of atomic environments. These basis functions can be used to successfully fit custom force fields for small molecules. Linear ACE force fields were competitive in accuracy with other state-of-the-art methods at the time of publication. However, they have a number of limitations. Firstly, the size of the models scales poorly with the number of different chemical elements in the system. Secondly, the fitting of the force field can be sensitive to the regularisation hyperparameters. Finally, shortly after publishing linear ACE, a number of new methods appeared based on equivariant message passing neural networks, which significantly improved the accuracy compared to linear ACE.

The subsequent sections of this thesis detail the methodological developments to overcome the aforementioned limitations of linear ACE. By exploiting the symmetric tensor structure of ACE and applying tensor decompositions, the TrACE method reduces the scaling of the model size with the number of elements to be effectively constant, with a single convergence parameter. The method also comes with theoretical guarantees of convergence.

Next, the multi-ACE framework is presented which provides a unifying framework of atom-centred and message-passing machine learning force fields. With the help of this framework, it is possible to select different points of the design space to systematically design more powerful models.

Bringing together the tensor decomposed ACE descriptors of TrACE, and the understanding of their relationship to equivariant neural networks via the multi-ACE method, led to the creation of MACE, a new machine learning force field. The robustness and ease of fitting of the MACE models, evidenced by the examples in Chapter 5 marks a significant leap forward from linear ACE. One of the highlights of the capabilities of MACE is the fact that it can be fitted even at the very low data regime, to obtain a stable force field for molecular dynamics simulations without the need for hyperparameter tuning or iterative fitting. This is particularly significant because it greatly reduces the time and effort required by users to obtain a working force field for a given scientific problem. Importantly, it is possible to fit MACE models without having to have a deep understanding of the actual fitting process or the details of the machine learning architecture. MACE is released as an open source software and has an active user community with several contributors.

Finally, the MACE architecture was used to fit a series of short-range transferable organic force fields released under the name `MACE-OFF23`. The models are shown to be accurate for a wide variety of chemical systems, including gas phase and condensed phase use cases.

6.2 Outlook

The development of simple and accurate machine learning force fields holds the potential of transforming the workflows of computational scientists. The new ML force fields enable the use of ab initio methods in scenarios previously deemed impractical or even impossible. For researchers reliant on ab initio calculations, the methods presented in this thesis allow exploration of larger system sizes and longer time-scales, far exceeding previous limitations. The classical empirical force field community can also benefit from these methods, gaining the accuracy and transferability typically associated with ab initio simulations.

The remaining challenges are partially scientific and partially engineering. Despite efficient implementations, these machine learning methods are still at least one order of magnitude slower than their classical counterparts. This is likely to improve with new hardware and software engineering efforts.

One of the key future development directions is the creation of foundation models such as `MACE-OFF23` and the recently published MACE-MP models [15]. These models can be a useful starting point for almost any atomistic simulation research project. In particular, the MACE-MP models were parameterised for 89 different chemical elements. This means that it is possible to run geometry optimisation or molecular dynamics simulation for almost any chemical or material system at quantum mechanical accuracy and force field speed. This transforms the workflow of computational scientists. It is now possible to set up the

simulations and obtain initial results and chemically reasonable structures with the foundation model straight away. If the results are not sufficiently accurate, due to the level of quantum mechanics or the accuracy for the particular system is not sufficient then the foundation model can be fine tuned with little additional data. This is made easier by the availability of self-generated trajectories. Future research will focus on the development of robust fine-tuning protocols.

These recent developments have the potential to accelerate both academic and industry research. This technology not only paves the way for breakthroughs in fundamental sciences but also enables the discovery of novel materials and pharmaceuticals, promising substantial societal benefits.

References

[1] Ahmed, S. E. (2017). *Big and complex data analysis: methodologies and applications*. Springer.

[2] Allen, A. E. A., Dusson, G., Ortner, C., and Csányi, G. (2021). Atomic permutationally invariant polynomials for fitting molecular force fields. *Machine Learning: Science and Technology*, 2(2):025017.

[3] Altman, A. B., Tamerius, A. D., Koocher, N. Z., Meng, Y., Pickard, C. J., Walsh, J. P., Rondinelli, J. M., Jacobsen, S. D., and Freedman, D. E. (2020). Computationally directed discovery of mobi2. *Journal of the American Chemical Society*, 143(1):214–222.

[4] Anderson, B., Hy, T. S., and Kondor, R. (2019). Cormorant: Covariant molecular neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[5] Anstine, D., Zubatyuk, R., and Isayev, O. (2023). Aimnet2: A neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *ChemRxiv*.

[6] Artrith, N., Urban, A., and Ceder, G. (2017). Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species. *Physical Review B*, 96(1):014112.

[7] Atsango, A. O., Morawietz, T., Marsalek, O., and Markland, T. E. (2023). Developing machine-learned potentials to simultaneously capture the dynamics of excess protons and hydroxide ions in classical and path integral simulations. *The Journal of Chemical Physics*, 159(7).

[8] Baldock, R. J., Pártay, L. B., Bartók, A. P., Payne, M. C., and Csányi, G. (2016). Determining pressure-temperature phase diagrams of materials. *Physical Review B*, 93(17):174108.

[9] Bannwarth, C., Ehlert, S., and Grimme, S. (2019). Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation*, 15(3):1652–1671.

[10] Bartlett, R. J. and Musiał, M. (2007). Coupled-cluster theory in quantum chemistry. *Reviews of Modern Physics*, 79(1):291.

[11] Bartók, A. P., Kermode, J., Bernstein, N., and Csányi, G. (2018). Machine learning a general-purpose interatomic potential for silicon. *Physical Review X*, 8(4):041048.

[12] Bartók, A. P., Kondor, R., and Csányi, G. (2013). On representing chemical environments. *Physical Review B*, 87(18):184115.

[13] Bartók, A. P., Payne, M. C., Kondor, R., and Csányi, G. (2010). Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 104(13).

[14] Batatia, I., Batzner, S., Kovács, D. P., Musaelian, A., Simm, G. N. C., Drautz, R., Ortner, C., Kozinsky, B., and Csanyi, G. (2022a). The design space of e(3)-equivariant atom-centered interatomic potentials.

[15] Batatia, I., Benner, P., Chiang, Y., Elena, A. M., Kovács, D. P., Riebesell, J., Advincula, X. R., Asta, M., Baldwin, W. J., Bernstein, N., Bhowmik, A., Blau, S. M., Cărare, V., Darby, J. P., De, S., Pia, F. D., Deringer, V. L., Elijošius, R., El-Machachi, Z., Fako, E., Ferrari, A. C., Genreith-Schriever, A., George, J., Goodall, R. E. A., Grey, C. P., Han, S., Handley, W., Heenen, H. H., Hermansson, K., Holm, C., Jaafar, J., Hofmann, S., Jakob, K. S., Jung, H., Kapil, V., Kaplan, A. D., Karimitari, N., Kroupa, N., Kullgren, J., Kuner, M. C., Kuryla, D., Liepuoniute, G., Margraf, J. T., Magdău, I.-B., Michaelides, A., Moore, J. H., Naik, A. A., Niblett, S. P., Norwood, S. W., O'Neill, N., Ortner, C., Persson, K. A., Reuter, K., Rosen, A. S., Schaaf, L. L., Schran, C., Sivonxay, E., Stenczel, T. K., Svahn, V., Sutton, C., van der Oord, C., Varga-Umbrich, E., Vegge, T., Vondrák, M., Wang, Y., Witt, W. C., Zills, F., and Csányi, G. (2023). A foundation model for atomistic materials chemistry.

[16] Batatia, I., Kovacs, D. P., Simm, G., Ortner, C., and Csányi, G. (2022b). Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 35:11423–11436.

[17] Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., and Kozinsky, B. (2022). E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453.

[18] Beck, M. H., Jäckle, A., Worth, G. A., and Meyer, H.-D. (2000). The multiconfiguration time-dependent hartree (mctdh) method: a highly efficient algorithm for propagating wavepackets. *Physics reports*, 324(1):1–105.

[19] Behler, J. (2021). Four generations of high-dimensional neural network potentials. *Chemical Reviews*, 121(16):10037–10072.

[20] Behler, J. and Parrinello, M. (2007). Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14).

[21] Bigi, F., Pozdnyakov, S. N., and Ceriotti, M. (2023). Wigner kernels: body-ordered equivariant machine learning without a basis. *arXiv preprint arXiv:2303.04124*.

[22] Bingham, E. and Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250.

[23] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[24] Bochkarev, A., Lysogorskiy, Y., Ortner, C., Csányi, G., and Drautz, R. (2022). Multilayer atomic cluster expansion for semilocal interactions. *Phys. Rev. Res.*, 4:L042019.

[25] Booth, G. H., Thom, A. J., and Alavi, A. (2009). Fermion monte carlo without fixed nodes: A game of life, death, and annihilation in slater determinant space. *The Journal of chemical physics*, 131(5):054106.

[26] Boothroyd, S., Behara, P. K., Madin, O. C., Hahn, D. F., Jang, H., Gapsys, V., Wagner, J. R., Horton, J. T., Dotson, D. L., Thompson, M. W., Maat, J., Gokey, T., Wang, L.-P., Cole, D. J., Gilson, M. K., Chodera, J. D., Bayly, C. I., Shirts, M. R., and Mobley, D. L. (2023a). Development and benchmarking of open force field 2.0.0: The sage small molecule force field. *Journal of Chemical Theory and Computation*, 19(11):3251–3275. PMID: 37167319.

[27] Boothroyd, S., Behara, P. K., Madin, O. C., Hahn, D. F., Jang, H., Gapsys, V., Wagner, J. R., Horton, J. T., Dotson, D. L., Thompson, M. W., Maat, J., Gokey, T., Wang, L.-P., Cole, D. J., Gilson, M. K., Chodera, J. D., Bayly, C. I., Shirts, M. R., and Mobley, D. L. (2023b). Development and benchmarking of open force field 2.0.0: The sage small molecule force field. *J. Chem. Theory Comput.*, 0(0):null. PMID: 37167319.

[28] Born;, M. and Oppenheimer, J. R. (1927). Zur Quantentheorie der Molekeln. *Annalen der Physik*, 389(20):457–484.

[29] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer.

[30] Braams, B. J. and Bowman, J. M. (2009). Permutationally invariant potential energy surfaces in high dimensionality. *International Reviews in Physical Chemistry*, 28(4):577–606.

[31] Brandstetter, J., Hesselink, R., van der Pol, E., Bekkers, E. J., and Welling, M. (2022). Geometric and physical quantities improve e(3) equivariant message passing. In *International Conference on Learning Representations*.

[32] Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges.

[33] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. a., and Karplus, M. (1983). Charmm: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry*, 4(2):187–217.

[34] Chai, J. D. and Head-Gordon, M. (2008). Systematic optimization of long-range corrected hybrid density functionals. *Journal of Chemical Physics*, 128(8).

[35] Chen, Z. and Yang, Y. (2023). Incorporating nuclear quantum effects in molecular dynamics with a constrained minimized energy surface. *The Journal of Physical Chemistry Letters*, 14:279–286.

[36] Cheng, B., Engel, E. A., Behler, J., Dellago, C., and Ceriotti, M. (2019). Ab initio thermodynamics of liquid and solid water. *Proceedings of the National Academy of Sciences*, 116(4):1110–1115.

[37] Chmiela, S., Sauceda, H. E., Müller, K.-R., and Tkatchenko, A. (2018). Towards exact molecular dynamics simulations with machine-learned force fields. *Nature communications*, 9(1):3887.

[38] Chmiela, S., Sauceda, H. E., Poltavsky, I., Müller, K. R., and Tkatchenko, A. (2019). sGDML: Constructing accurate and data efficient molecular force fields using machine learning. *Computer Physics Communications*, 240:38–45.

[39] Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T., and Müller, K. R. (2017). Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5).

[40] Chmiela, S., Vassilev-Galindo, V., Unke, O. T., Kabylda, A., Sauceda, H. E., Tkatchenko, A., and Müller, K.-R. (2023). Accurate global machine learning force fields for molecules with hundreds of atoms. *Science Advances*, 9(2):eadf0873.

[41] Christensen, A. S. and Anatole von Lilienfeld, O. (2020). On the role of gradients for machine learning of molecular energies and forces. *Machine Learning: Science and Technology*, 1.

[42] Christensen, A. S., Bratholm, L. A., Faber, F. A., and Anatole von Lilienfeld, O. (2020). Fchl revisited: Faster and more accurate quantum machine learning. *The Journal of chemical physics*, 152(4):044107.

[43] Combes, J.-M. and Thomas, L. (1973). Asymptotic behaviour of eigenfunctions for multiparticle schrödinger operators. *Communications in Mathematical Physics*, 34(4):251–270.

[44] Cournia, Z., Allen, B., and Sherman, W. (2017). Relative binding free energy calculations in drug discovery: recent advances and practical considerations. *Journal of chemical information and modeling*, 57(12):2911–2937.

[45] Cox, S. J. (2020). Dielectric response with short-ranged electrostatics. *Proceedings of the National Academy of Sciences*, 117(33):19746–19752.

[46] Darby, J. P., Kermode, J. R., and Csányi, G. (2022). Compressing local atomic neighbourhood descriptors. *npj Computational Materials*, 8(1):1–13.

[47] Darby, J. P., Kovács, D. P., Batatia, I., Caro, M. A., Hart, G. L. W., Ortner, C., and Csányi, G. (2023). Tensor-reduced atomic density representations. *Phys. Rev. Lett.*, 131:028001.

[48] Darden, T., York, D., and Pedersen, L. (1993). Particle mesh ewald: An n log (n) method for ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092.

[49] Dasgupta, S. (2013). Experiments with random projection. *arXiv preprint arXiv:1301.3849*.

[50] Dauber-Osguthorpe, P. and Hagler, A. T. (2019a). Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there? *Journal of computer-aided molecular design*, 33(2):133–203.

[51] Dauber-Osguthorpe, P. and Hagler, A. T. (2019b). Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there? *J Comput Aided Mol Des*, 33:133–203.

[52] Daw, M. S. and Baskes, M. I. (1984). Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals. *Physical Review B*, 29(12):6443.

[53] Deringer, V. L., Bartók, A. P., Bernstein, N., Wilkins, D. M., Ceriotti, M., and Csányi, G. (2021a). Gaussian process regression for materials and molecules. *Chemical Reviews*, 121(16):10073–10141.

[54] Deringer, V. L., Bernstein, N., Csányi, G., Ben Mahmoud, C., Ceriotti, M., Wilson, M., Drabold, D. A., and Elliott, S. R. (2021b). Origins of structural and electronic transitions in disordered silicon. *Nature*, 589(7840):59–64.

[55] Deringer, V. L., Caro, M. A., and Csányi, G. (2020). A general-purpose machine-learning force field for bulk and nanostructured phosphorus. *Nature communications*, 11(1):5461.

[56] Deringer, V. L. and Csányi, G. (2017). Machine learning based interatomic potential for amorphous carbon. *Physical Review B*, 95(9):094203.

[57] Deringer, V. L., Pickard, C. J., and Csányi, G. (2018). Data-driven learning of total and local energies in elemental boron. *Physical review letters*, 120(15):156001.

[58] Devereux, C., Smith, J. S., Huddleston, K. K., Barros, K., Zubatyuk, R., Isayev, O., and Roitberg, A. E. (2020). Extending the applicability of the ani deep learning molecular potential to sulfur and halogens. *Journal of Chemical Theory and Computation*, 16(7):4192–4202.

[59] Dolgonos, G. A., Hoja, J., and Boese, A. D. (2019). Revised values for the x23 benchmark set of molecular crystals. *Physical Chemistry Chemical Physics*, 21(44):24333–24344.

[60] Donchev, A. G., Taube, A. G., Decolvenaere, E., Hargus, C., McGibbon, R. T., Law, K.-H., Gregersen, B. A., Li, J.-L., Palmo, K., Siva, K., et al. (2021). Quantum chemical benchmark databases of gold-standard dimer interaction energies. *Scientific data*, 8(1):55.

[61] Drautz, R. (2019). Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, 99(1).

[62] Drautz, R. (2020). Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer. *Phys. Rev. B*, 102:024104.

[63] Dusson, G., Bachmayr, M., Csányi, G., Drautz, R., Etter, S., van der Oord, C., and Ortner, C. (2022). Atomic cluster expansion: Completeness, efficiency and stability. *Journal of Computational Physics*, 454:110946.

[64] Eastman, P., Behara, P. K., Dotson, D. L., Galvelis, R., Herr, J. E., Horton, J. T., Mao, Y., Chodera, J. D., Pritchard, B. P., Wang, Y., et al. (2023). Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(1):11.

[65] Elfwing, S., Uchibe, E., and Doya, K. (2018). Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11.

[66] EthanolIR2021 (1964). Ethanol ir spectrum. https://webbook.nist.gov/cgi/cbook.cgi?ID=C64175&Type=IR-SPEC&Index=2.

[67] Ewig, C. S., Berry, R., Dinur, U., Hill, J.-R., Hwang, M.-J., Li, H., Liang, C., Maple, J., Peng, Z., Stockfisch, T. P., Thacher, T. S., Yan, L., Ni, X., and Hagler, A. T. (2001). Derivation of class II force fields. VIII. derivation of a general quantum mechanical force field for organic compounds. *Journal of Computational Chemistry*, 22(15):1782–1800.

[68] Finzi, M., Stanton, S., Izmailov, P., and Wilson, A. G. (2020). Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, pages 3165–3176. PMLR.

[69] Fonseca, G., Poltavsky, I., Vassilev-Galindo, V., and Tkatchenko, A. (2021). Improving molecular force fields across configurational space by combining supervised and unsupervised machine learning. *The Journal of Chemical Physics*, 154(12):124102.

[70] Frenkel, D. and Smit, B. (2023). *Understanding Molecular Simulation: From Algorithms to Applications*. Elsevier, 3 edition.

[71] Fu, X., Wu, Z., Wang, W., Xie, T., Keten, S., Gomez-Bombarelli, R., and Jaakkola, T. S. (2023). Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *Transactions on Machine Learning Research*.

[72] Galvelis, R., Varela-Rial, A., Doerr, S., Fino, R., Eastman, P., Markland, T. E., Chodera, J. D., and De Fabritiis, G. (2023). Nnp/mm: Accelerating molecular dynamics simulations with machine learning potentials and molecular mechanics. *Journal of chemical information and modeling*, 63(18):5701–5708.

[73] Gao, X., Ramezanghorbani, F., Isayev, O., Smith, J. S., and Roitberg, A. E. (2020). Torchani: A free and open source pytorch-based deep learning implementation of the ani neural network potentials. *Journal of chemical information and modeling*, 60(7):3408–3415.

[74] Gapsys, V., Pérez-Benito, L., Aldeghi, M., Seeliger, D., Van Vlijmen, H., Tresadern, G., and De Groot, B. L. (2020). Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chem. Sci.*, 11(4):1140–1152.

[75] Gastegger, M., Schwiedrzik, L., Bittermann, M., Berzsenyi, F., and Marquetand, P. (2018). wacsf—weighted atom-centered symmetry functions as descriptors in machine learning potentials. *The Journal of chemical physics*, 148(24):241709.

[76] Gasteiger, J., Groß, J., and Günnemann, S. (2020). Directional message passing for molecular graphs. In *International Conference on Learning Representations*.

[77] Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. (2020). Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401.

[78] Geiger, M. and Smidt, T. (2022). e3nn: Euclidean neural networks.

[79] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry.

[80] Godwin, J., Schaarschmidt, M., Gaunt, A., Sanchez-Gonzalez, A., Rubanova, Y., Veličković, P., Kirkpatrick, J., and Battaglia, P. (2021). Simple gnn regularisation for 3d molecular property prediction & beyond. *arXiv preprint arXiv:2106.07971*.

[81] Goerigk, L., Hansen, A., Bauer, C., Ehrlich, S., Najibi, A., and Grimme, S. (2017). A look at the density functional theory zoo with the advanced gmtkn55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Physical Chemistry Chemical Physics*, 19(48):32184–32215.

[82] Goscinski, A., Musil, F., Pozdnyakov, S., Nigam, J., and Ceriotti, M. (2021). Optimal radial basis for density-based atomic representations. *The Journal of Chemical Physics*, 155(10):104106.

[83] Grimme, S., Antony, J., Ehrlich, S., and Krieg, H. (2010). A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *The Journal of chemical physics*, 132(15).

[84] Grimme, S., Ehrlich, S., and Goerigk, L. (2011). Effect of the damping function in dispersion corrected density functional theory. *Journal of computational chemistry*, 32(7):1456–1465.

[85] Gubaev, K., Podryabinkin, E. V., Hart, G. L., and Shapeev, A. V. (2019). Accelerating high-throughput searches for new alloys with active learning of interatomic potentials. *Computational Materials Science*, 156:148–156.

[86] Haghighatlari, M., Li, J., Guan, X., Zhang, O., Das, A., Stein, C. J., Heidar-Zadeh, F., Liu, M., Head-Gordon, M., Bertels, L., et al. (2022). Newtonnet: A newtonian message passing network for deep learning of interatomic potentials and forces. *Digital Discovery*, 1(3):333–343.

[87] Hagler, A. T. (2019a). Force field development phase ii: Relaxation of physics-based criteria... or inclusion of more rigorous physics into the representation of molecular energetics. *J Comput Aided Mol Des*, 33:205–264.

[88] Hagler, A. T. (2019b). Force field development phase ii: Relaxation of physics-based criteria… or inclusion of more rigorous physics into the representation of molecular energetics. *Journal of computer-aided molecular design*, 33(2):205–264.

[89] Hagler, A. T., Huler, E., and Lifson, S. (1974). Energy Functions for Peptides and Proteins. I. Derivation of a Consistent Force Field Including the Hydrogen Bond from Amide Crystals. *Journal of the American Chemical Society*, 96(17):305319–5327.

[90] Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., and Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *nature*, 405(6789):947–951.

[91] Handy, N. C. (1980). Multi-root configuration interaction calculations. *Chemical Physics Letters*, 74(2):280–283.

[92] Hao, D., He, X., Roitberg, A. E., Zhang, S., and Wang, J. (2022). Development and evaluation of geometry optimization algorithms in conjunction with ani potentials. *Journal of Chemical Theory and Computation*, 18(2):978–991.

[93] Hart, G. L. and Forcade, R. W. (2008). Algorithm for generating derivative structures. *Physical Review B*, 77(22):224115.

[94] Hermann, J., Schätzle, Z., and Noé, F. (2020). Deep-neural-network solution of the electronic schrödinger equation. *Nature Chemistry*, 12(10):891–897.

[95] Hermann, J., Spencer, J., Choo, K., Mezzacapo, A., Foulkes, W. M. C., Pfau, D., Carleo, G., and Noé, F. (2023). Ab initio quantum chemistry with neural-network wavefunctions. *Nature reviews. Chemistry*, 7(10):692–709.

[96] Hohenberg, P. and Kohn, W. (1964). Inhomogeneous electron gas. *Physical review*, 136(3B):B864.

[97] Hong, Y. P. and Pan, C.-T. (1992). Rank-Revealing QR factorizations and the singular value decomposition. *Mathematics of Computation*, 58(197):213–232.

[98] Horton, J. T., Allen, A. E., and Cole, D. J. (2020). Modelling flexible protein–ligand binding in p38α map kinase using the qube force field. *Chemical Communications*, 56(6):932–935.

[99] Horton, J. T., Allen, A. E., Dodda, L. S., and Cole, D. J. (2019). Qubekit: Automating the derivation of force field parameters from quantum mechanics. *Journal of chemical information and modeling*, 59(4):1366–1381.

[100] Horton, J. T., Boothroyd, S., Wagner, J., Mitchell, J. A., Gokey, T., Dotson, D. L., Behara, P. K., Ramaswamy, V. K., Mackey, M., Chodera, J. D., et al. (2022). Open force field bespokefit: Automating bespoke torsion parametrization at scale. *Journal of Chemical Information and Modeling*, 62(22):5622–5633.

[101] Hourahine, B., Aradi, B., Blum, V., Bonafé, F., Buccheri, A., Camacho, C., Cevallos, C., Deshaye, M., Dumitrică, T., Dominguez, A., et al. (2020). Dftb+, a software package for efficient approximate density functional theory based atomistic simulations. *The Journal of chemical physics*, 152(12).

[102] Inizan, T. J., Plé, T., Adjoua, O., Ren, P., Gökcan, H., Isayev, O., Lagardère, L., and Piquemal, J.-P. (2023). Scalable hybrid deep neural networks/polarizable potentials biomolecular simulations including long-range effects. *Chemical Science*, 14(20):5438–5452.

[103] Ischtwan, J. and Collins, M. A. (1994). Molecular potential energy surfaces by interpolation. *The Journal of chemical physics*, 100(11):8080–8088.

[104] Isert, C., Atz, K., Jiménez-Luna, J., and Schneider, G. (2022). Qmugs, quantum mechanical properties of drug-like molecules. *Scientific Data*, 9(1):1–11.

[105] Jiang*, Y., Neyshabur*, B., Mobahi, H., Krishnan, D., and Bengio, S. (2020). Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*.

[106] Johnson, W. B. (1984). Extensions of lipschitz mappings into a hilbert space. *Contemp. Math.*, 26:189–206.

[107] Jorgensen, W. L., Maxwell, D. S., and Tirado-Rives, J. (1996). Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236.

[108] Jorgensen, W. L. and Tirado-Rives, J. (1988). The opls [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666. PMID: 27557051.

[109] Kabán, A. (2013). A new look at compressed ordinary least squares. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 482–488. IEEE.

[110] Kabylda, A., Vassilev-Galindo, V., Chmiela, S., Poltavsky, I., and Tkatchenko, A. (2023). Efficient interatomic descriptors for accurate machine learning force fields of extended molecules. *Nature Communications*, 14(1):3562.

[111] Kapil, V., Kovács, D. P., Csányi, G., and Michaelides, A. (2023). First-principles spectroscopy of aqueous interfaces using machine-learned electronic and quantum nuclear effects. *Faraday Discussions*.

[112] Kapil, V., Schran, C., Zen, A., Chen, J., Pickard, C. J., and Michaelides, A. (2022). The first-principles phase diagram of monolayer nanoconfined water. *Nature*, 609(7927):512–516.

[113] Klicpera, J., Becker, F., and Günnemann, S. (2021). Gemnet: Universal directional graph neural networks for molecules. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

[114] Kohn, W. and Sham, L. J. (1965a). Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138.

[115] Kohn, W. and Sham, L. J. (1965b). Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133.

[116] Kolesov, B. A., Mikhailenko, M. A., and Boldyreva, E. V. (2011). Dynamics of the intermolecular hydrogen bonds in the polymorphs of paracetamol in relation to crystal packing and conformational transitions: a variable-temperature polarized raman spectroscopy study. *Physical Chemistry Chemical Physics*, 13(31):14243.

[117] Kostiuchenko, T., Körmann, F., Neugebauer, J., and Shapeev, A. (2019). Impact of lattice relaxations on phase transitions in a high-entropy alloy studied by machine-learning potentials. *npj Computational Materials*, 5(1):1–7.

[118] Kovács, D. P., McCorkindale, W., and Lee, A. A. (2021). Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. *Nature Communications*, 12(1):1–9.

[119] Kovács, D. P., Oord, C. v. d., Kucera, J., Allen, A. E., Cole, D. J., Ortner, C., and Csányi, G. (2021). Linear atomic cluster expansion force fields for organic molecules: beyond rmse. *Journal of chemical theory and computation*, 17(12):7696–7711.

[120] Kovács, D. P., Batatia, I., Arany, E. S., and Csányi, G. (2023a). Evaluation of the MACE force field architecture: From medicinal chemistry to materials science. *The Journal of Chemical Physics*, 159(4):044118.

[121] Kovács, D. P., Moore, J. H., Browning, N. J., Batatia, I., Horton, J. T., Kapil, V., Witt, W. C., Magdău, I.-B., Cole, D. J., and Csányi, G. (2023b). Mace-off23: Transferable machine learning force fields for organic molecules.

[122] Kozinsky, B., Musaelian, A., Johansson, A., and Batzner, S. (2023). Scaling the leading accuracy of deep equivariant models to biomolecular simulations of realistic size. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–12.

[123] Kühne, T. D., Iannuzzi, M., Del Ben, M., Rybkin, V. V., Seewald, P., Stein, F., Laino, T., Khaliullin, R. Z., Schütt, O., Schiffmann, F., et al. (2020). Cp2k: An electronic structure and molecular dynamics software package-quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics*, 152(19):194103.

[124] Kussmann, J., Beer, M., and Ochsenfeld, C. (2013). Linear-scaling self-consistent field methods for large molecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(6):614–636.

[125] Kutzner, C., Pall, S., Fechner, M., Esztermann, A., de Groot, B. L., and Grubmueller, H. (2015). Best bang for your buck: Gpu nodes for gromacs biomolecular simulations. *Journal of Computational Chemistry*, 36(26):1990–2008.

[126] Lahey, S.-L. J., Thien Phuc, T. N., and Rowley, C. N. (2020a). Benchmarking force field and the ani neural network potentials for the torsional potential energy surface of biaryl drug fragments. *Journal of Chemical Information and Modeling*, 60(12):6258–6268.

[127] Lahey, S.-L. J., Thien Phuc, T. N., and Rowley, C. N. (2020b). Benchmarking force field and the ani neural network potentials for the torsional potential energy surface of biaryl drug fragments. *Journal of Chemical Information and Modeling*, 60(12):6258–6268.

[128] Le, T., Noé, F., and Clevert, D.-A. (2022). Equivariant graph attention networks for molecular property prediction. *arXiv preprint arXiv:2202.09891*.

[129] Levine, D. S., Hait, D., Tubman, N. M., Lehtola, S., Whaley, K. B., and Head-Gordon, M. (2020). Casscf with extremely large active spaces using the adaptive sampling configuration interaction method. *Journal of chemical theory and computation*, 16(4):2340–2354.

[130] Levitt, M. and Lifson, S. (1969). Refinement of protein conformations using a macromolecular energy minimization procedure. *Journal of molecular biology*, 46(2):269–279.

[131] Li, Y., Wang, Y., Huang, L., Yang, H., Wei, X., Zhang, J., Wang, T., Wang, Z., Shao, B., and Liu, T.-Y. (2023). Long-short-range message-passing: A physics-informed framework to capture non-local interaction for scalable molecular dynamics simulation. *arXiv preprint arXiv:2304.13542*.

[132] Liao, Y.-L. and Smidt, T. (2023). Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations*.

[133] Liu, Y., Wang, L., Liu, M., Lin, Y., Zhang, X., Oztekin, B., and Ji, S. (2022). Spherical message passing for 3d molecular graphs. In *International Conference on Learning Representations (ICLR)*.

[134] Lu, C., Liu, Q., Wang, C., Huang, Z., Lin, P., and He, L. (2019). Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1052–1060.

[135] Lysogorskiy, Y., Oord, C. v. d., Bochkarev, A., Menon, S., Rinaldi, M., Hammerschmidt, T., Mrovec, M., Thompson, A., Csányi, G., Ortner, C., et al. (2021). Performant implementation of the atomic cluster expansion (pace) and application to copper and silicon. *npj Computational Materials*, 7(1):1–12.

[136] Magdău, I.-B., Arismendi-Arrieta, D. J., Smith, H. E., Grey, C. P., Hermansson, K., and Csányi, G. (2023a). Machine learning force fields for molecular liquids: Ethylene carbonate/ethyl methyl carbonate binary solvent. *npj Computational Materials*, 9(1):146.

[137] Magdău, I.-B., Arismendi-Arrieta, D. J., Smith, H. E., Grey, C. P., Hermansson, K., and Csányi, G. (2023b). Machine learning force fields for molecular liquids: Ethylene carbonate/ethyl methyl carbonate binary solvent. *npj Computational Materials*, 9(1):146.

[138] Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of chemical theory and computation*, 11(8):3696–3713.

[139] Maillard, O. and Munos, R. (2009). Compressed least-squares regression. *Advances in neural information processing systems*, 22.

[140] Maple, J. R., Hwang, M.-J., Stockfisch, T. P., Dinur, U., Waldman, M., Ewig, C. S., and Hagler, A. T. (1994). Derivation of class II force fields. i. methodology and quantum force field for the alkyl functional group and alkane molecules. *Journal of Computational Chemistry*, 15(2):162–182.

[141] Marsalek, O. and Markland, T. E. (2017). Quantum dynamics and spectroscopy of ab initio liquid water: The interplay of nuclear and electronic quantum effects. *The journal of physical chemistry letters*, 8(7):1545–1551.

[142] Melchionna, S. (2000). Constrained systems and statistical distribution. *Physical Review E*, 61(6):6165.

[143] Melchionna, S., Ciccotti, G., and Lee Holian, B. (1993). Hoover npt dynamics for systems varying in shape and size. *Molecular Physics*, 78(3):533–544.

[144] Monserrat, B., Brandenburg, J. G., Engel, E. A., and Cheng, B. (2020). Liquid water contains the building blocks of diverse ice phases. *Nature communications*, 11(1):5757.

[145] Munoz, J. M., Batatia, I., and Ortner, C. (2022). Boost invariant polynomials for efficient jet tagging. *Machine Learning: Science and Technology*, 3(4):04LT05.

[146] Musaelian, A., Batzner, S., Johansson, A., Sun, L., Owen, C. J., Kornbluth, M., and Kozinsky, B. (2023). Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579.

[147] Musil, F., Grisafi, A., Bartók, A. P., Ortner, C., Csányi, G., and Ceriotti, M. (2021). Physics-inspired structural representations for molecules and materials. *Chemical Reviews*, 121(16):9759–9815.

[148] Musil, F., Zaporozhets, I., Noé, F., Clementi, C., and Kapil, V. (2022). Quantum dynamics using path integral coarse-graining. *The Journal of Chemical Physics*, 157(18).

[149] Najibi, A. and Goerigk, L. (2018). The nonlocal kernel in van der waals density functionals as an additive correction: An extensive analysis with special emphasis on the b97m-v and ωb97m-v approaches. *Journal of Chemical Theory and Computation*, 14(11):5725–5738.

[150] Nandi, A., Conte, R., Qu, C., Houston, P. L., Yu, Q., and Bowman, J. M. (2022). Quantum calculations on a new ccsd (t) machine-learned potential energy surface reveal the leaky nature of gas-phase trans and gauche ethanol conformers. *Journal of Chemical Theory and Computation*, 18(9):5527–5538.

[151] Neese, F., Wennmohs, F., Becker, U., and Riplinger, C. (2020). The orca quantum chemistry program package. *The Journal of chemical physics*, 152(22).

[152] Nigam, J., Pozdnyakov, S., and Ceriotti, M. (2020). Recursive evaluation and iterative contraction of n-body equivariant features. *The Journal of Chemical Physics*, 153(12):121101.

[153] Nigam, J., Pozdnyakov, S., Fraux, G., and Ceriotti, M. (2022). Unified theory of atom-centered representations and message-passing machine-learning schemes. *The Journal of Chemical Physics*, 156(20):204115.

[154] Ohto, T., Dodia, M., Xu, J., Imoto, S., Tang, F., Zysk, F., Kühne, T. D., Shigeta, Y., Bonn, M., Wu, X., et al. (2019). Accessing the accuracy of density functional theory through structure and dynamics of the water–air interface. *The journal of physical chemistry letters*, 10(17):4914–4919.

[155] Paige, C. C. and Saunders, M. A. (1982). LSQR: An Algorithm for Sparse Linear Equations and Sparse Least Squares. *ACM Transactions on Mathematical Software (TOMS)*, 8(1):43–71.

[156] Páll, S., Zhmurov, A., Bauer, P., Abraham, M., Lundborg, M., Gray, A., Hess, B., and Lindahl, E. (2020). Heterogeneous parallelization and acceleration of molecular dynamics simulations in gromacs. *The Journal of Chemical Physics*, 153(13):134110.

[157] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems*.

[158] Pfau, D., Spencer, J., de G. Matthews, A., and Foulkes, W. (2020). Ab-initio solution of the many-electron schrödinger equation with deep neural networks. *Phys. Rev. Research*, 2:033429.

[159] Pickard, C. J. (2022). Ephemeral data derived potentials for random structure search. *Physical Review B*, 106(1):014102.

[160] Pickard, C. J. and Needs, R. (2011). Ab initio random structure searching. *Journal of Physics: Condensed Matter*, 23(5):053201.

[161] Plé , T., Lagardère, L., and Piquemal, J.-P. (2023). Force-field-enhanced neural network interactions: from local equivariant embedding to atom-in-molecule properties and long-range effects. *Chemical Science*.

[162] Pozdnyakov, S. N. and Ceriotti, M. (2022). Incompleteness of graph neural networks for points clouds in three dimensions. *Machine Learning: Science and Technology*, 3(4):045020.

[163] Pozdnyakov, S. N., Willatt, M. J., Bartók, A. P., Ortner, C., Csányi, G., and Ceriotti, M. (2020). Incompleteness of atomic structure representations. *Physical Review Letters*, 125(16):166001.

[164] Purvis III, G. D. and Bartlett, R. J. (1982). A full coupled-cluster singles and doubles model: The inclusion of disconnected triples. *The Journal of Chemical Physics*, 76(4):1910–1918.

[165] Qu, C. and Bowman, J. M. (2019). A fragmented, permutationally invariant polynomial approach for potential energy surfaces of large molecules: Application to n-methyl acetamide. *The Journal of Chemical Physics*, 150(14):141101.

[166] Quaranta, V., Behler, J., and Hellström, M. (2018). Structure and dynamics of the liquid–water/zinc-oxide interface from machine learning potential simulations. *The Journal of Physical Chemistry C*, 123(2):1293–1304.

[167] Rai, B. K., Sresht, V., Yang, Q., Unwalla, R., Tu, M., Mathiowetz, A. M., and Bakken, G. A. (2022). Torsionnet: A deep neural network to rapidly predict small-molecule torsional energy profiles with the accuracy of quantum mechanics. *Journal of Chemical Information and Modeling*, 62(4):785–800.

[168] Raimbault, N., Athavale, V., and Rossi, M. (2019a). Anharmonic effects in the low-frequency vibrational modes of aspirin and paracetamol crystals. *Physical Review Materials*, 3(5).

[169] Raimbault, N., Grisafi, A., Ceriotti, M., and Rossi, M. (2019b). Using gaussian process regression to simulate the vibrational raman spectra of molecular crystals. *New Journal of Physics*, 21(10):105001.

[170] Raimbault, N., Grisafi, A., Ceriotti, M., and Rossi, M. (2019c). Using gaussian process regression to simulate the vibrational raman spectra of molecular crystals. *New Journal of Physics*, 21(10):105001.

[171] Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. (2014). Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7.

[172] Rappoport, D. and Furche, F. (2010). Property-optimized gaussian basis sets for molecular response calculations. *The Journal of chemical physics*, 133(13).

[173] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

[174] Reilly, A. M. and Tkatchenko, A. (2013). Understanding the role of vibrations, exact exchange, and many-body van der waals interactions in the cohesive properties of molecular crystals. *The Journal of chemical physics*, 139(2):024705.

[175] Reinhardt, A. and Cheng, B. (2021). Quantum-mechanical exploration of the phase diagram of water. *Nature communications*, 12(1):588.

[176] Riniker, S. (2018). Fixed-charge atomistic force fields for molecular dynamics simulations in the condensed phase: an overview. *Journal of chemical information and modeling*, 58(3):565–578.

[177] Roos, B. O., Taylor, P. R., and Sigbahn, P. E. (1980). A complete active space scf method (casscf) using a density matrix formulated super-ci approach. *Chemical Physics*, 48(2):157–173.

[178] Rosenbrock, C. W., Gubaev, K., Shapeev, A. V., Pártay, L. B., Bernstein, N., Csányi, G., and Hart, G. L. (2021). Machine-learned interatomic potentials for alloys and alloy phase diagrams. *npj Computational Materials*, 7(1):24.

[179] Rowe, P., Deringer, V. L., Gasparotto, P., Csányi, G., and Michaelides, A. (2020). An accurate and transferable machine learning potential for carbon. *The Journal of Chemical Physics*, 153(3):034702.

[180] Salomon-Ferrer, R., Case, D. A., and Walker, R. C. (2013). An overview of the amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(2):198–210.

[181] Satorras, V. G., Hoogeboom, E., and Welling, M. (2021). E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR.

[182] Schaaf, L. L., Fako, E., De, S., Schäfer, A., and Csányi, G. (2023). Accurate energy barriers for catalytic reaction pathways: an automatic training protocol for machine learning force fields. *npj Computational Materials*, 9(1):180.

[183] Schoenholz, S. S. and Cubuk, E. D. (2020). Jax m.d. a framework for differentiable physics. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc.

[184] Schran, C., Thiemann, F. L., Rowe, P., Müller, E. A., Marsalek, O., and Michaelides, A. (2021a). Machine learning potentials for complex aqueous systems made simple. *Proceedings of the National Academy of Sciences*, 118(38):e2110077118.

[185] Schran, C., Thiemann, F. L., Rowe, P., Müller, E. A., Marsalek, O., and Michaelides, A. (2021b). Machine learning potentials for complex aqueous systems made simple. *Proceedings of the National Academy of Sciences*, 118(38):e2110077118.

[186] Schütt, K., Unke, O., and Gastegger, M. (2021). Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR.

[187] Schütt, K., Kindermans, P.-J., Felix, H. E. S., Chmiela, S., Tkatchenko, A., and Müller, K.-R. (2017). Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *NIPS*, pages 991–1001.

[188] Scott, W. R., Hünenberger, P. H., Tironi, I. G., Mark, A. E., Billeter, S. R., Fennen, J., Torda, A. E., Huber, T., Krüger, P., and Van Gunsteren, W. F. (1999). The gromos biomolecular simulation program package. *The Journal of Physical Chemistry A*, 103(19):3596–3607.

[189] Shapeev, A. V. (2016). Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Modeling & Simulation*, 14(3):1153–1173.

[190] Shaw, D. E., Adams, P. J., Azaria, A., Bank, J. A., Batson, B., Bell, A., Bergdorf, M., Bhatt, J., Butts, J. A., Correia, T., et al. (2021). Anton 3: twenty microseconds of molecular dynamics simulation before lunch. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–11.

[191] Shipley, A. M., Hutcheon, M. J., Needs, R. J., and Pickard, C. J. (2021). High-throughput discovery of high-temperature conventional superconductors. *Physical Review B*, 104(5):054501.

[192] Sidiropoulos, N. D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E. E., and Faloutsos, C. (2017). Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582.

[193] Simeon, G. and De Fabritiis, G. (2023). Tensornet: Cartesian tensor representations for efficient learning of molecular potentials. *arXiv preprint arXiv:2306.06482*.

[194] Slater, J. C. and Koster, G. F. (1954). Simplified lcao method for the periodic potential problem. *Phys. Rev.*, 94:1498–1524.

[195] Smith, D. G., Burns, L. A., Simmonett, A. C., Parrish, R. M., Schieber, M. C., Galvelis, R., Kraus, P., Kruse, H., Di Remigio, R., Alenaizan, A., et al. (2020a). Psi4 1.4: Open-source software for high-throughput quantum chemistry. *The Journal of chemical physics*, 152(18).

[196] Smith, J. S., Isayev, O., and Roitberg, A. E. (2017a). Ani-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific data*, 4(1):1–8.

[197] Smith, J. S., Isayev, O., and Roitberg, A. E. (2017b). ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science*, 8(4):3192–3203.

[198] Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O., and Roitberg, A. E. (2018). Less is more: Sampling chemical space with active learning. *The Journal of chemical physics*, 148(24):241733.

[199] Smith, J. S., Nebgen, B. T., Zubatyuk, R., Lubbers, N., Devereux, C., Barros, K., Tretiak, S., Isayev, O., and Roitberg, A. E. (2019). Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature communications*, 10(1):1–8.

[200] Smith, J. S., Zubatyuk, R., Nebgen, B., Lubbers, N., Barros, K., Roitberg, A. E., Isayev, O., and Tretiak, S. (2020b). The ani-1ccx and ani-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific data*, 7(1):134.

[201] Smith, J. S., Zubatyuk, R., Nebgen, B., Lubbers, N., Barros, K., Roitberg, A. E., Isayev, O., and Tretiak, S. (2020c). The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific Data*, 7(1):1–10.

[202] Sun, H., Jin, Z., Yang, C., Akkermans, R. L. C., Robertson, S. H., Spenley, N. A., Miller, S., and Todd, S. M. (2016). COMPASS II: extended coverage for polymer and drug-like molecule databases. *Journal of Molecular Modeling*, 22(2).

[203] Thölke, P. and De Fabritiis, G. (2022). Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations*.

[204] Thölke, P. and Fabritiis, G. D. (2022). Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations*.

[205] Thomas, J., Chen, H., and Ortner, C. (2022). Body-ordered approximations of atomic properties. *Archive for Rational Mechanics and Analysis*, 246(1):1–60.

[206] Thomas, N., Smidt, T., Kearnes, S., Yang, L., Li, L., Kohlhoff, K., and Riley, P. (2018). Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*.

[207] Thürlemann, M. and Riniker, S. (2023). Hybrid classical/machine-learning force fields for the accurate description of molecular condensed-phase systems. *Chemical Science*, 14(44):12661–12675.

[208] Timmermann, J., Kraushofer, F., Resch, N., Li, P., Wang, Y., Mao, Z., Riva, M., Lee, Y., Staacke, C., Schmid, M., et al. (2020). Iro 2 surface complexions identified through machine learning and surface investigations. *Physical review letters*, 125(20):206101.

[209] Tu, N. T. P., Rezajooei, N., Johnson, E. R., and Rowley, C. (2023). A neural network potential with rigorous treatment of long-range dispersion. *Digital Discovery*.

[210] Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

[211] Unke, O. T., Chmiela, S., Gastegger, M., Schütt, K. T., Sauceda, H. E., and Müller, K.-R. (2021). Spookynet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nature communications*, 12(1):7273.

[212] Unke, O. T. and Meuwly, M. (2019). Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of chemical theory and computation*, 15(6):3678–3693.

[213] Unke, O. T., Stöhr, M., Ganscha, S., Unterthiner, T., Maennel, H., Kashubin, S., Ahlin, D., Gastegger, M., Sandonas, L. M., Tkatchenko, A., and Müller, K.-R. (2022). Accurate machine learned quantum-mechanical force fields for biomolecular simulations.

[214] van Der Oord, C., Dusson, G., Csányi, G., and Ortner, C. (2020). Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials. *Machine Learning: Science and Technology*, 1(1):015004.

[215] Vanommeslaeghe, K. and MacKerell Jr, A. (2015). Charmm additive and polarizable force fields for biophysics and computer-aided drug design. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(5):861–871.

[216] Vassilev-Galindo, V., Fonseca, G., Poltavsky, I., and Tkatchenko, A. (2021). Challenges for machine learning force fields in reproducing potential energy surfaces of flexible molecules. *The Journal of Chemical Physics*, 154(9).

[217] Wales, D. (2003). *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge Molecular Science. Cambridge University Press.

[218] Wales, D. J. and Doye, J. P. (1997). Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116.

[219] Wales, D. J. and Scheraga, H. A. (1999). Global optimization of clusters, crystals, and biomolecules. *Science*, 285(5432):1368–1372.

[220] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9):1157–1174.

[221] Wang, L. P., Chen, J., and Van Voorhis, T. (2013). Systematic parametrization of polarizable force fields from quantum chemistry data. *Journal of Chemical Theory and Computation*, 9(1):452–460.

[222] Wang, L. P., Martinez, T. J., and Pande, V. S. (2014). Building force fields: An automatic, systematic, and reproducible approach. *Journal of Physical Chemistry Letters*, 5(11):1885–1891.

[223] Wang, Y., Fass, J., Kaminow, B., Herr, J. E., Rufa, D., Zhang, I., Pulido, I., Henry, M., Bruce Macdonald, H. E., Takaba, K., and Chodera, J. D. (2022). End-to-end differentiable construction of molecular mechanics force fields. *Chem. Sci.*, 13:12016–12033.

[224] Weigend, F. and Ahlrichs, R. (2005). Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305.

[225] Weiler, M., Geiger, M., Welling, M., Boomsma, W., and Cohen, T. S. (2018). 3d steerable cnns: Learning rotationally equivariant features in volumetric data. *Advances in Neural Information Processing Systems*, 31.

[226] Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784.

[227] Wigner, E. (2012). *Group theory: and its application to the quantum mechanics of atomic spectra*, volume 5. Elsevier.

[228] Willatt, M. J., Musil, F., and Ceriotti, M. (2018). Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Physical Chemistry Chemical Physics*, 20(47):29661–29668.

[229] Woodruff, D. P. et al. (2014). Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157.

[230] Yang, M., Bonati, L., Polino, D., and Parrinello, M. (2022). Using metadynamics to build neural network potentials for reactive events: the case of urea decomposition in water. *Catalysis Today*, 387:143–149.

[231] Yang, M., Raucci, U., and Parrinello, M. (2023). Reactant-induced dynamics of lithium imide surfaces during the ammonia decomposition process. *Nature Catalysis*, pages 1–8.

[232] Young, T. A., Johnston-Wood, T., Deringer, V. L., and Duarte, F. (2021). A transferable active-learning strategy for reactive molecular force fields. *Chemical science*, 12(32):10944–10955.

[233] Zaidi, S., Schaarschmidt, M., Martens, J., Kim, H., Teh, Y. W., Sanchez-Gonzalez, A., Battaglia, P., Pascanu, R., and Godwin, J. (2023). Pre-training via denoising for molecular property prediction. In *The Eleventh International Conference on Learning Representations*.

[234] Zaverkin, V., Holzmüller, D., Bonfirraro, L., and Kästner, J. (2023). Transfer learning for chemically accurate interatomic neural network potentials. *Physical Chemistry Chemical Physics*, 25(7):5383–5396.

[235] Zaverkin, V., Holzmüller, D., Steinwart, I., and Kästner, J. (2021). Fast and sample-efficient interatomic neural network potentials for molecules and materials based on gaussian moments. *Journal of Chemical Theory and Computation*, 17(10):6658–6670.

[236] Zaverkin, V. and Kastner, J. (2020). Gaussian moments as physically inspired molecular descriptors for accurate and scalable machine learning potentials. *Journal of Chemical Theory and Computation*, 16(8):5410–5421.

[237] Zen, A., Brandenburg, J. G., Klimeš, J., Tkatchenko, A., Alfè, D., and Michaelides, A. (2018). Fast and accurate quantum monte carlo for molecular crystals. *Proceedings of the National Academy of Sciences*, 115(8):1724–1729.

[238] Zeni, C., Rossi, K., Glielmo, A., and De Gironcoli, S. (2021). Compact atomic descriptors enable accurate predictions via linear models. *The Journal of Chemical Physics*, 154(22):224112.

[239] Zhang, L., Han, J., Wang, H., Saidi, W., Car, R., et al. (2018). End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. *Advances in neural information processing systems*, 31.

[240] Zhang, L., Onat, B., Dusson, G., McSloy, A., Anand, G., Maurer, R. J., Ortner, C., and Kermode, J. R. (2022). Equivariant analytical mapping of first principles hamiltonians to accurate and transferable materials models. *Npj Computational Materials*, 8(1):158.

[241] Zhang, Y., Xia, J., and Jiang, B. (2021). Physically motivated recursively embedded atom neural networks: incorporating local completeness and nonlocality. *Physical Review Letters*, 127(15):156002.

[242] Zhou, Y., Zhang, W., Ma, E., and Deringer, V. L. (2023). Device-scale atomistic modelling of phase-change memory materials. *Nature Electronics*, pages 1–9.

[243] Zubatyuk, R., Smith, J. S., Leszczynski, J., and Isayev, O. (2019). Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Science advances*, 5(8):eaav6490.

[244] Zubatyuk, R., Smith, J. S., Nebgen, B. T., Tretiak, S., and Isayev, O. (2021). Teaching a neural network to attach and detach electrons from molecules. *Nature Communications*, 12(1):4870.