Chemical Engineering (ChemE)

MACHINE LEARNING POTENTIALS FOR CATIONIC ZIRCONIUM COMPLEXES GRAFTED ON AMORPHOUS SILICA

Lennard Duynkerke (5360374)

Thesis MSc Applied Physics
Physics for Energy
Delft University of Technologies

Delft, 01-04-2025
Supervisors: prof. dr. E. A. Pidko
dr. K.R. Rossi
dr. A.A. Kolganov

ABSTRACT Accurately modeling heterogeneous catalytic systems while maintaining computational efficiency is a persistent challenge, as conventional methods like Density Functional Theory (DFT) offer high accuracy but are computationally expensive, whereas classical force fields provide efficiency without precision. In recent years, Machine Learning Potentials (MLPs) have emerged as a powerful tool to bridge the gap between the efficiency of classical force fields and the precision of first-principles methods. In this study, I assess the accuracy, efficiency, and limitations of MACE MLP-models when applied to a challenging catalytic system: cationic zirconocene hydride grafted onto an amorphous silica slab model. My results demonstrate that MACE models, even with minimal training data, achieve impressive accuracy with energy RMSE below 0.05 eV/atom and force errors under 0.2 eV/Å, highlighting the efficiency of foundational models. Nonetheless, challenges such as a sub-unity slope in energy predictions and dynamically unstable MD simulations due to catastrophic forgetting remain, even with the application of active learning techniques. A novel multihead replay technique shows promise in enhancing stability, though additional validation is necessary. Furthermore, thermodynamic reweighting proves effective in refining bond length distributions, especially with hybrid functionals like PBE0+D3, but its robustness remains sensitive to model accuracy and bias. Overall, these results demonstrate the potential of MLP-based approaches in accelerating calculations by orders of magnitude while emphasizing the importance of thorough validation for accurate predictions.

Table of Contents

1 Introduction

The development of efficient and selective catalysts is a central challenge in chemical engineering, with profound implications for energy sustainability and industrial efficiency. Nowadays, computational modeling plays a crucial role in accelerating catalyst discovery, with Density Functional Theory (DFT) serving as the predominant tool for characterizing reaction energetics and active site behavior [1–4]. However, despite its success, DFT is inherently constrained by its steep computational cost and limited scalability, as its calculations scale cubically with the number of electrons [5, 6]. In contrast, classical force fields offer a robust and computationally efficient alternative, but they lack the accuracy needed for applications in heterogeneous catalysis [7–9].

In recent years, Machine Learning Potentials (MLPs) have emerged as a powerful tool to bridge the gap between the efficiency of classical force fields and the precision of first-principles methods [6]. Most modern MLPs are based on Graph Neural Networks (GNNs), which represent atomic structures as graphs where nodes correspond to atoms and edges encode bonds [10]. Through message passing, these networks iteratively exchange information between neighboring atoms, allowing for a dynamic representation of local atomic environments. Early MLPs relied primarily on interatomic distances to describe atomic environments, but later developments incorporated higher-order geometric features, such as bond angles [11] and dihedral angles [12], leading to significantly improved accuracy in capturing local chemical interactions. State-of-the-art MLP architectures, such as MACE [13] and NequIP [14], further enforce physical symmetries via equivariant representations, allowing them to capture many-body interactions and long-range correlations with higher fidelity.

Apart from the novel message passing scheme, one of MACE's most significant innovations is the introduction of foundational models, such as MACE-MP-0 and MACE-MPA-0, trained on the MPtrj dataset [15]. Foundational models enable researchers to fine-tune the model for specific chemical systems using a minimal amount of additional data. This approach not only extends the applicability of MLPs to a wide variety of chemistries, but also significantly reduces the data and computational cost required for training compared to developing models from scratch.

Since it was first introduced, MACE has been successfully used to model a wide range of materials and their properties, including applications in heterogeneous catalysis [16–18]. However, applying MLPs to complex catalytic systems remains challenging, requiring careful selection of training data and electronic structure methods, construction of an active learning loop, and validation of their transferability under various conditions [18].

A more out-of-the-box application of MLPs is their integration with thermodynamic reweighting, a technique rooted in statistical mechanics and traditionally used to correct sampling biases in molecular simulations [19]. This technique becomes significantly more powerful when paired with Machine Learning Potentials (MLPs) trained on high-level DFT data. By reweighting large trajectories from inexpensive functionals, MLPs allow observables to be reconstructed as though they were sampled from a more accurate reference potential — all without the computational cost of running high-level DFT simulations directly.

In this work I explore the application of MLPs, active learning loops, and thermodynamic reweighting to model heterogeneous catalytic systems. In doing so, I assess the accuracy, efficiency, and limitations of these methods. First, I provide an in-depth discussion of my methodology, detailing the integration of MLPs with reweighting techniques. Next, I demonstrate my approach, by applying it to a complex catalytic system, namely cationic zirconocene hydride grafted onto an amorphous silica slab model. The organometallic core $(^{Bu}Cp)_2ZrH_2$ is a derivative of the initial Ziegler-Natta catalysts [20–23], and makes the system particularly challenging due to the dynamic interactions between the silica support, the organic cyclopentadienyl ligands, and the transition metal center. Finally, I assess the broader applicability of my method, define its limitations, and summarize key findings.

2

Theory

This chapter describes essential theory needed for my research. It provides a literature overview and is structured in three parts. In the first part, I will present some theory on the chemical systems I simulated, the second part explores machine learning potentials, particularly MACE, and the final part provides some background on thermodynamic reweighting.

2.1. Surface Organometallic Catalysts

Surface organometallic catalysts (SOMCs) are a powerful approach to generate single-site catalysts with known coordination spheres, facilitating the rational design of heterogeneous catalysts. SOMCs consist of two main components: the organometallic core and the oxide support [24]. In my research, the oxide support of interest is amorphous silica, and the organometallic core is $(^{Bu}Cp)_2ZrH_2$, often used for the polymerization of α-olefins [23]. In this section, I first discuss the properties of the oxide support and the various models used to simulate it, then elaborate upon the organometallic core.

2.1.1. Amorphous silica (a-SiO_2)

Commercial catalysts are often amorphous, lacking long-range order, rather than crystalline materials [25]. This is because they are typically inexpensive, have more tunable physical properties such as porosity, and exhibit significantly higher activities and productivities [26]. Silica itself is generally inert; however, the large surface area of mesoporous silica is ideal for maximizing catalyst density while maintaining accessibility to reactants. Catalysts can be grafted onto the surface after preparing the porous material or directly incorporated during synthesis [27, 28].

However, the heterogeneity in composition and structure makes identifying their active sites extremely challenging. Furthermore, a small fraction of surface sites may contribute most of the reactivity [30]. Therefore, computational simulations are essential to understanding the behavior of catalysts on amorphous materials. Computational investigations of a-SiO_2 primarily use two models to approximate the material: cluster models and slab models, as illustrated in Fig. 2.1 [29].

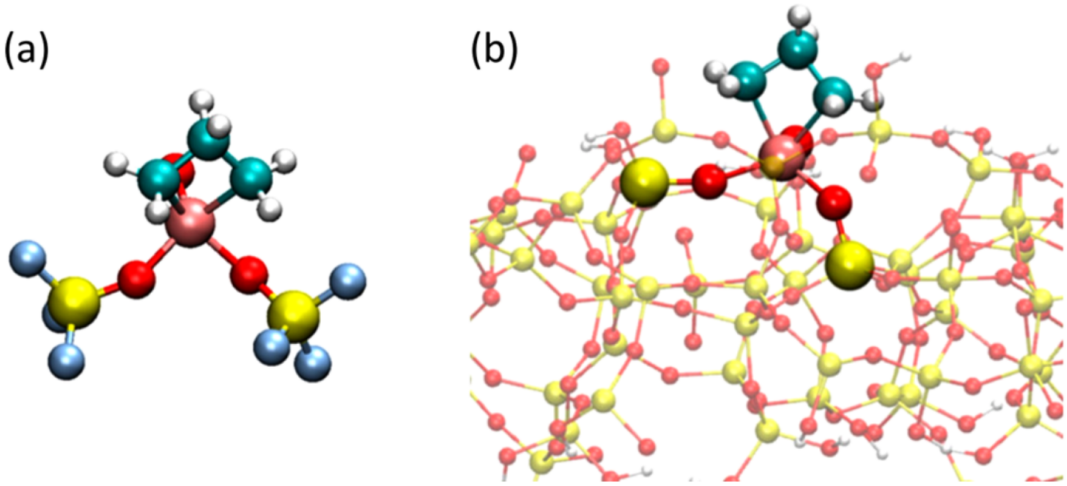

Figure 2.1: (a) A small cluster model, and (b) a slab model, both representing a molybdenacyclobutane attached to amorphous silica during ethene metathesis (O = red; Mo = pink; Si = yellow; H = white; C = teal; F = light blue) [29].

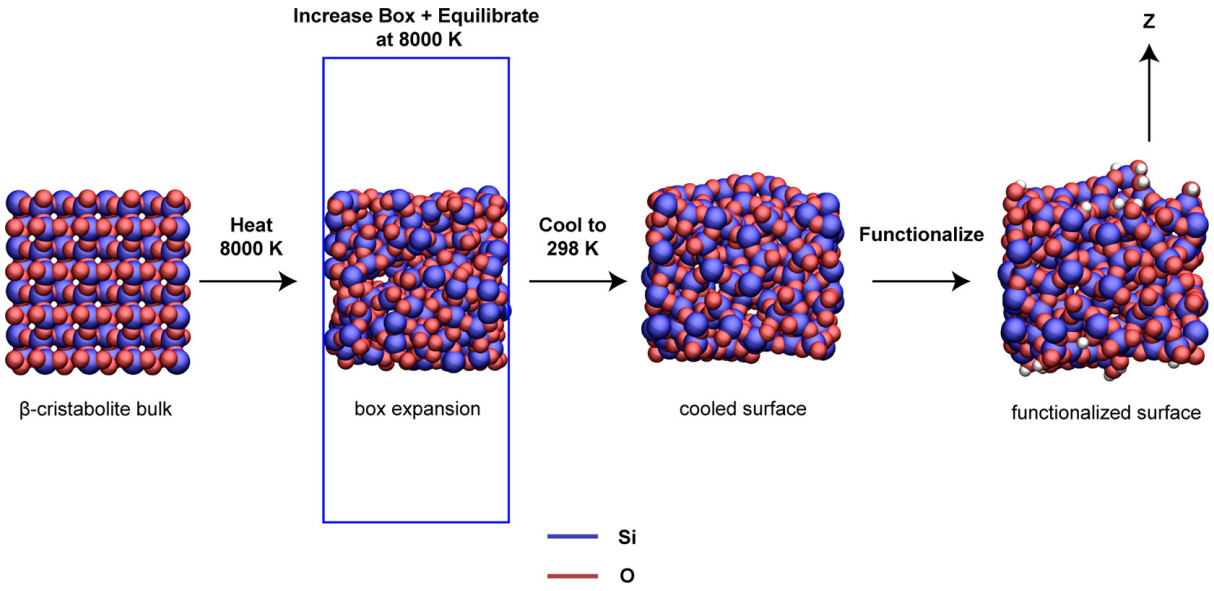

Figure 2.2: Schematic illustration of the melt-cleave-quench-functionalize procedure for the generation of silica slab models [27].

2.1.2. Cluster models

Cluster models are created by isolating a small portion of the bulk material to represent the active site and its surroundings. The primary advantage of these models is that they require minimal structural assumptions, making them effective for identifying key local structural features. However, they often neglect important long-range interactions and may not capture the full diversity of an amorphous material [31, 32]. Utilizing several medium-sized cluster models derived from a larger amorphous model can mitigate this issue. Nevertheless, statistical analysis indicates that thousands of such models may be necessary for reliable reactivity predictions [29, 33].

2.1.3. Slab models

Slab models, unlike cluster models, are technically periodic and avoid artificial boundaries and capping atoms. They naturally incorporate long-range electronic and geometric effects, thereby overcoming some limitations of finite-sized cluster models [32]. However, slab models also have drawbacks. Large supercells are required to minimize artificial interactions between adsorbates and their periodic images, and sophisticated correction schemes are necessary for charged systems to achieve size convergence [29, 34].

Amorphous silica slab models are typically created using a melt-quench-cleave process. This involves melting crystalline silica, quenching it below its melting temperature to form an amorphous solid, and cleaving it to create the desired interface geometry (e.g., planar or cylindrical). This procedure results in a large number of under-coordinated Si and O atoms, which are then saturated by the addition of OH and H groups, respectively. Finally, the surface atoms are identified, creating a functionalized surface [27, 35, 36].

2.1.4. Organometallic core

Ziegler and co-workers discovered that mixtures of triethylaluminum and zirconium acetylacetonate polymerize ethylene to high-density polyethylene under mild conditions in 1953, and two years later Natta reported that $TiCl_4$ and Et_2AlCl mixtures polymerize propylene to stereoregular products [22]. The organometallic core $(^{Bu}Cp)_2ZrH_2$ (where Cp = cyclopentadienykl) is a derivative from these initial Ziegler-Natta catalysts, and is often used for the polymerization of α-olefins. [23].

The combination of the organometallic core and a silica slab model exists in two configurations: "chemically" adsorbed (CA) and "physically" adsorbed (PA). In the CA scenario, the zirconocene reacts with the silica substrate, releasing a hydrogen molecule and creating a covalent Zr-O bond. In contrast, when no chemical reaction occurs, the interaction remains physical, and the core's attachment to the substrate is maintained by electrostatic forces between the electronegative silanol group and the electropositive zirconocene core (see Fig. 2.3).

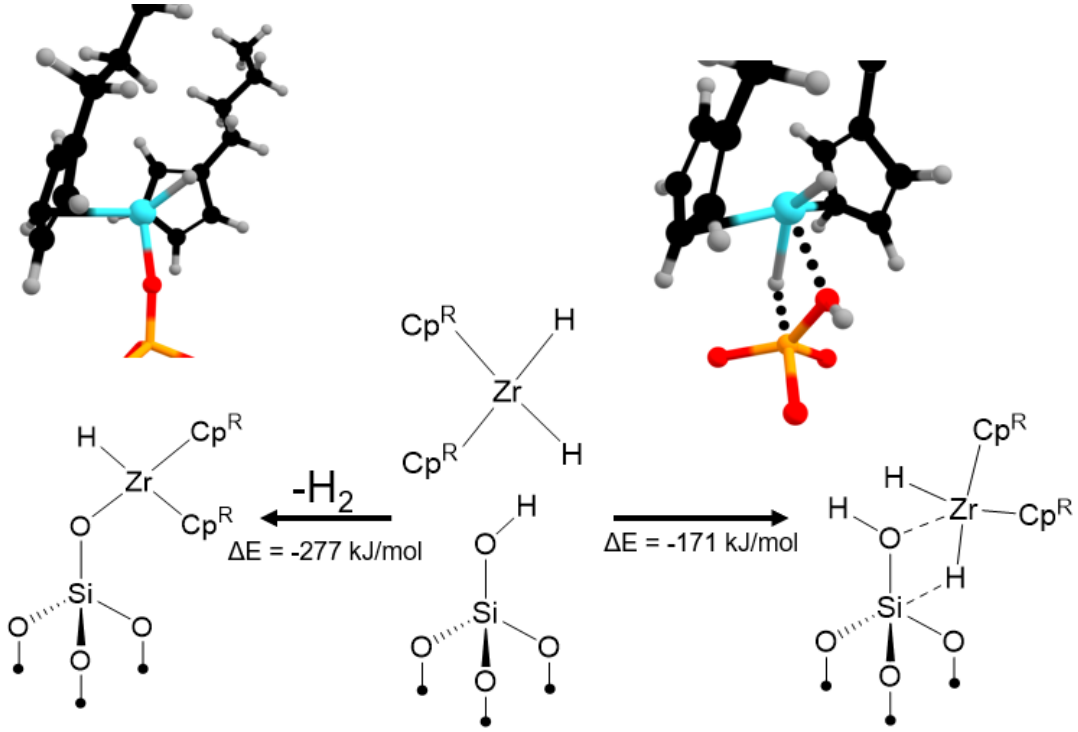

Figure 2.3: Visual representation of the organometallic core $(^{Bu}Cp)_2ZrH_2$ and its two configurations. The left hand side shows the CA configuration, with the core bonded to the silica substrate by a covalent bond, while the right-hand side shows the PA configuration, with the core's attachment to the substrate maintained by electrostatic forces.

2.2. Density Functional Theory

Density Functional Theory (DFT) is an essential building block for modern chemistry, particularly in the study of electronic properties. In DFT, the electron density $\rho(\mathbf{r})$ is used as the fundamental variable instead of the many-body wave function. The use of electron density is much more tractable because it depends on only three coordinate variables, regardless of the number of electrons N, whereas the many-body wave function is a function of $3N$ coordinates. This approach has been justified by the Hohenberg-Kohn theorems [37].

2.2.1. Hohenberg-Kohn Theorems

The first Hohenberg-Kohn theorem proves that a functional of the electron density $F[\rho]$ exists that can be used to solve the Schrödinger equation. The second defines an important property of the functional: the electron density that minimizes the energy of the overall functional is the true electron density corresponding to the full solution of the Schrodinger equation [1].

As such, the Hohenberg-Kohn theorems state a one-to-one correspondence between the ground-state electron density and the external potential. Consequently, once the ground-state electron density is known, the external potential is uniquely determined. This implies all physical properties associated with the ground-state wave function, such as the kinetic energy $E_{kin}[\rho]$ and the electron-electron interaction energy $E_{ee}[\rho]$, can be unambiguously derived from the electron density [37].

The theorem also gives a variational principle. When the energy functional is defined as in Eq. 2.1

$$E_v[\rho] = E_{kin}[\rho] + \int \rho(\mathbf{r})\, v(\mathbf{r})\, d\mathbf{r} + E_{ee}[\rho] \tag{2.1}$$

for some external potential $v(\mathbf{r})$, the functional satisfies the inequality

$$E_v[\rho] \geq E_v[\rho_0] = E_0 \tag{2.2}$$

where ρ_0 and E_0 represent the ground-state electron density and energy, respectively, under the potential $v(\mathbf{r})$. Thus, by searching for the electron density that minimizes the energy functional $E_v[\rho]$, the ground-state electron density can be determined.

2.2.2. Kohn-Sham equations

While the Hohenberg-Kohn theorems state that energy is a functional of electron density, they do not hint how to derive the electron density. To this end, Kohn and Sham propose a self-consisted scheme to approximate functional $E_{kin}[\rho]$ using "orbitals" [37]. The electron density $\rho(r)$ is now defined in terms of the solution to the Kohn-Sham equations $\psi_j(\mathbf{r})$ by

$$\rho(r) = \sum_j \psi_j(\mathbf{r})\psi_j^*(\mathbf{r}) \tag{2.3}$$

The Kohn-Sham equations are

$$\frac{h^2}{2m}\nabla^2\psi_j(\mathbf{r}) + V_{eff}(\mathbf{r})\psi_j(\mathbf{r}) = \epsilon_j\psi_j(\mathbf{r}) \tag{2.4}$$

where $V_{eff}(\mathbf{r})$ is the effective potential.

The fact that $V_{eff}(\mathbf{r})$ itself is a complicated function of $\rho(\mathbf{r})$, makes these equations hard to solve directly. There is a straightforward numerical method for solving this problem though, namely by using an iterative strategy:

1. Estimate the overall electron density $\rho(\mathbf{r})$.
2. Use this trial density to define the effective potential.
3. Now solve Eq. 2.4 numerically, defining a new electron density.
4. Repeat this process until the old and new electron densities match closely enough.

Note that a self-consistent solution is reached much more quickly if a better initial approximation is available. For this reason, it can sometimes be helpful to store the electron density and related information from a large calculation for use in starting a subsequent similar one. In the next sections I describe approaches to solve the Kohn-Sham problem. Operationally, I adopted their implementation in CP2K, an open source electronic structure code [38].

The Kohn-Sham approximation has paved the way for Density Functional Theory (DFT) calculations with sufficient accuracy for practical applications. One such application is ab initio Molecular Dynamics (aiMD), where interatomic forces are computed on-the-fly using DFT. However, the enhanced accuracy and predictive power of aiMD simulations come at a significant computational cost [39].

2.2.3. Exchange-correlation Functionals

DFT only describes a precisely mathematical problem once the exchange-correlation functional (XC functional) has been specified. The exact form of the XC functional is unknown though, and various approximations exist. It is crucial to understand the similarities and differences between the various functionals that are commonly used, before actually choosing a functional to use. One useful classification of functionals is made by John Perdew, called Jacob's ladder of DFT [40]. The lower the step of the ladder, the simpler the functional, and, in general, the higher the step, the higher the accuracy of the functional. The XC functional V^{XC} is the functional derivative of the exchange-correlation energy (XC energy), given by

$$V^{XC}[\rho] = \frac{\delta E^{XC}[\rho]}{\delta\rho} \tag{2.5}$$

The XC energy E^{XC} is divided in two separate terms, an exchange term E^X and a correlation term E^C. The former is normally associated with the interactions between moving electrons of the same spin, while the latter essentially represents those between electrons of opposite spin [41]

$$E^{XC}[\rho] = E^X[\rho] + E^C[\rho] \tag{2.6}$$

2.2.4. LDA

The first step above the Hartree-Fock earth is the local density approximation (LDA). This is the simplest approach to represent the XC functional and assumes that the XC energy at any point in space is a function of the electron density at that point in space only. The local XC potential in the Kohn-Sham equations (Eq. 2.4) is defined as the XC potential for the spatially uniform electron gas with the same density as the local electron density, namely

$$V_{XC}^{LDA}(\mathbf{r}) = V_{XC}^{electrongas}[n(\mathbf{r})] \tag{2.7}$$

Some of the most commonly used LDA functionals are those developed by Slater, Perdew, and Wang (SPWL) [42] [43], Vosko, Wilk, and Nusair (VWN) [44], Perdew and Zunger (PZ81) [45], and the Padé approximation of the latter (PADE) [46]. The LDA is surprisingly accurate for a method this simple, notwithstanding some typical deficiencies, such as the inadequate cancellation of self-interaction contributions. LDA's limitations become particularly evident in systems where precise energy differences are critical, such as in chemical reactions and small molecules with steep density gradients. In such cases, its inherent deficiencies, such as inadequate treatment of self-interaction and poor handling of van der Waals

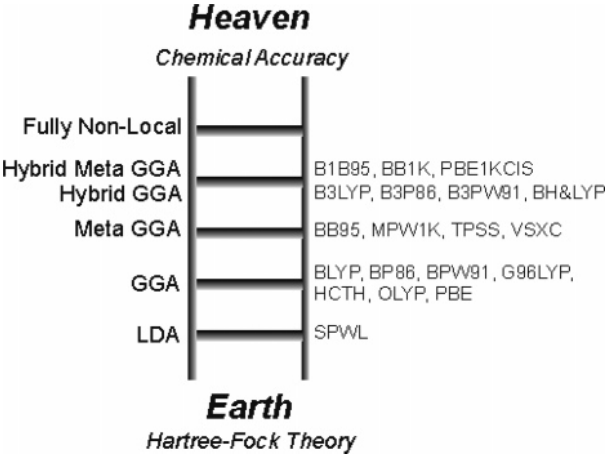

Figure 2.4: Jacob's ladder for the five generation of DFT functionals, according to the vision of J. Perdew [40] with indication of some of the most common DFT functionals within each step [41].

interactions, lead to significant inaccuracies, making it unsuitable for many chemical applications, including heterogeneous catalysis [3].

LDA is suitable for systems having slowly varying densities, but occasionally good results for systems with relatively large density gradients have also been observed. A partial explanation for this success lies in the systematic cancellation of errors. In fact, LDA typically underestimates E^X but overestimates E^C, sometimes resulting in unexpectedly good E^{XC} values [41].

2.2.5. GGA

The next approximation to the Kohn-Sham functional is the General Gradient Approximation (GGA). The physical idea behind the GGA is simple: real electron densities are not uniform, so including information on the spatial variation in the electron density improves the approximation. The equations on which the GGA are based, are valid for slowly varying densities. In the GGA, the XC functional is expressed as

$$V_{XC}^{GGA}(\mathbf{r}) = V_{XC}[n(\mathbf{r}), \nabla n(\mathbf{r})] \tag{2.8}$$

Nonempirical GGA functionals satisfy the uniform density limit as well as several known, exact properties of the exchange–correlation hole. Two widely used nonempirical functionals that satisfy these properties are the Perdew–Wang 91 (PW91) [43] functional and the Perdew–Burke–Ernzerhof (PBE) [47] functional. Other functionals, like BLYP [48] [49], are partially-empirical because some parameters were obtained via empirical fittings. As a result, PBE and PW91 offer reasonable accuracy across a broad range of systems, whereas BLYP is more accurate for systems similar to those used in its parametrization, e.g. in main-group organic molecules [3].

In general, the GGA methods represent a significant improvement over the LDA methods. In fact, GGA methods tend to give better total energies, atomization energies, structural energy differences, and energy barriers. The GGA methods tend to expand and soften bonds, compensating for the LDA tendency to overbind. However, the accuracy of GGA methods is still not enough for a correct description of many chemical aspects of molecules. For example, the GGA typically fails for van der Waals interactions. In the case of the solid state, GGA functionals do not yield significantly better results than LDA, nor in the calculation of ionization potentials and electron affinities. Furthermore, the differences obtained when using different GGAs are often almost as large as those between individual GGAs and LDA functional [41].

2.2.6. meta-GGA

The third step of Jacob's ladder is defined by meta-GGA functionals, which include information from $n(\mathbf{r})$, $\nabla n(\mathbf{r})$, and $\nabla^2 n(\mathbf{r})$. In practice, the kinetic energy density of the Kohn-Sham orbitals,

$$\tau(\mathbf{r}) = \frac{1}{2} \sum_{\text{occupied states}} |\nabla \phi_i(\mathbf{r})|^2 \tag{2.9}$$

contains the same physical information as the Laplacian of the electron density $\nabla^2 n(\mathbf{r})$, so $\tau(\mathbf{r})$ may be used in meta-GGA functionals instead of $\nabla^2 n(\mathbf{r})$. Commonly used meta-GGA functionals are TPSS [50], SCAN [51], and M06-L [52]. The former two are non-empirical functionals, while the latter is classified as an empirical functional [3].

2.2.7. Hybrid GGA

The important fourth step of the ladder in fig. 2.4 is hybrid GGA, whose functionals are most commonly used in quantum chemistry calculations with localized basis sets. A critical feature of the exchange energy E^X in hybrid GGA is that is non-local, i.e. a functional based on E^X cannot be evaluated at one particular spatial location, unless the electron density is known for all spatial locations. The exchange energy can be written as

$$E^X(\mathbf{r}) = \frac{1}{2n(\mathbf{r})} \int d^3r' \frac{\left| \sum_{\text{occupied states}} \phi_i^*(\mathbf{r}')\phi_i(\mathbf{r}) \right|^2}{|\mathbf{r}-\mathbf{r}'|} \tag{2.10}$$

Functionals that include contributions from the exact exchange energy with a GGA functional are classified as a hyper GGA or hybrid GGA functional. A widely used hybrid GGA functional is the B3LYP functional [53],

$$V_{XC}^{B3LYP} = V_{XC}^{LDA} + \alpha_1(E^X - V_X^{LDA}) + \alpha_2(V_X^{GGA} - V_X^{LDA}) + \alpha_3(V_C^{GGA} - V_C^{LDA}) \tag{2.11}$$

Here, V_X^{GGA} is the B88 exchange functional [48], V_C^{GGA} is the Lee-Yang-Parr correlation functional [49], and α_1, α_2, and α_3 are three numerical parameter, hence the name B3LYP. The three parameters were empirically chosen to optimize the performance of the functional for a sizable set of molecular properties (formation energies, bond lengths, etc.) [3].

Another widely used hybrid functional, and one that I will use during my research is PBE0, which mixes the PBE exchange functional with the exact Hartree-Fock exchange in a fixed ratio of 1:3. The PBE0 functional is defined in [54] as:

$$V_{XC}^{PBE0} = V_{XC}^{PBE} + \frac{1}{4}(E^X - V_X^{PBE}) \tag{2.12}$$

Hybrid functionals have allowed a significant improvement over GGAs for many molecular properties. For this, they have become a very popular choice in quantum chemistry and are now widely used. However, in solid-state physics this type of functional was much less successful due to difficulties in computing the exact-exchange part within a plane-wave basis set. It is important to note that B3LYP does not satisfy the uniform density limit. Thus, it would not be expected to perform especially well, as indeed it does not, in predictions for bulk materials, especially metals. Examples of other hybrid density functionals include B3P86 [53] [43], B3PW91 [53] [55], and BH&HLYP [48] [49].

2.2.8. Hybrid Meta GGA

Another step up one finds the hybrid meta GGA functionals. These functionals combine the exact exchange energy with a meta GGA functional. These methods represent an improvement over the previous formalisms, particularly in the determination of barrier heights and atomization energies. Examples of hybrid meta GGA functionals include B1B95 [48] [56], and PBE1KCIS [55] [57] [58].

2.2.9. DFT-D correction

One conceptually simple remedy for the shortcomings of lower level DFT regarding dispersion forces is to simply add a dispersion-like contribution $\propto 1/r^6$ to the total energy between each pair of atoms in a material [59]. This idea is used in the DFT-D2 method, which augments the calculated total energy as given in Eq. 2.13:

$$E^{DFT} = E^{DFT} + S\sum_{i<j} \frac{C_{ij}}{r_{ij}^6} f_{damp}(r_{ij}) \tag{2.13}$$

where r_{ij} is the distance between atoms i and j, C_{ij} is a dispersion coefficient for atoms i and j, which can be calculated directly from tabulated properties of the individual atoms, and $f_{damp}(r_{ij})$ is a damping function to avoid unphysical behavior for small distances. The only empirical parameter S is estimated separately for each DFT functional and applied uniformly to all pairs of atoms [3, 60].

DFT-D3 offers a more precise dispersion correction than DFT-D2 by incorporating three-body interactions, dynamically adjusting $C_{6,ij}/r_{ij}^6$ based on atomic coordination environments, and introducing an additional $C_{8,ij}$ term to enhance long-range accuracy [61]. The expression for the DFT-D3 correction is given in Eq. 2.14:

$$E^{DFT} = E^{DFT} + S_6\sum_{i<j} \frac{C_{6,ij}}{r_{ij}^6} f_{damp}(r_{ij}) + S_8\sum_{i<j} \frac{C_{8,ij}}{r_{ij}^8} f_{damp}(r_{ij}) + S_{ATM}E_{ATM} \tag{2.14}$$

where r_{ij} is the distance between atoms i and j, $C_{6,ij}$ and $C_{8,ij}$ are dispersion coefficients for atoms i and j, E_{ATM} represents the Axilrod-Teller-Muto (ATM) three-body term [62], and S_6, S_8, S_{ATM} are various scaling factors. In my research, I use the DFT-D3(BJ) correction, which employs the Becke-Johnson damping function for $f_{damp}(r_{ij})$ to improve short-range interactions [63, 64].

2.3. Machine learning potentials

Over the past decade, machine learning potentials (MLPs) are being developed and applied to describe energy and forces in chemical systems, effectively bridging the gap between the computational efficiency of classical force fields and the accuracy of quantum mechanics DFT. The development of machine learning potential involves four key components: dataset, representation, training and evaluation, and simulation [6]. Chapter 3 will provide a detailed description of these steps for my specific project. This section, however, focuses on the available models and methods to train MLPs in general. There are two primary classes of models used: neural net based MLPs and kernel based MLPs. Since my project focuses on neural net based MLPs, I will not elaborate upon kernel based MLPs, but instead provide a more in-depth discussion of recent neural net based MLPs, such as SchNet, DimeNet, and MACE.

2.3.1. Neural net based MLPs

The neural network architecture is often chosen for predicting the potential energy surface of a given structure due to its nonlinear character, which enables it to solve complex tasks. A significant advantage of this model is its flexible architecture, allowing adaptability to various data shapes and types. Moreover, compared to kernel-based techniques, its computational costs scales more favorably with an increasing dataset size. However, this high degree of versatility necessitates a more extensive search space for optimal hyperparameters, rendering the tuning process rather intricate [6, 10, 65, 66].

A visual representation of the available neural net based MLPs is shown in Fig. 2.5. Following the first neural networks with linear and convolution layers, SchNet, introduced in 2017, circumvented the limitations of convential convolution layers by applying a continous filter convolutional layer, capable of managing continuous data like atom positions [65]. Next, in 2018, CGCNN introduced a Graph Neural Network (GNN) to treat molecules as a graph in which the nodes are atoms and the edges represent atom connections [10]. In 2020 Dimenet was published, incorporating bond angles via directional message passing. This allowed for the discernment of e.g. molecules with hexagonal and two triangular geometries, indistinguishable by distance only GNNs [11].

In 2021, the SpinConv neural network expressed angular information using latitude and longtitude formats relative to source and target nodes, enabling the study of higher-order effects beyond triplet order [66]. Also in 2021, the geometric message passing neural network (GemNet) incorporated direct edge embedding and two-hop message passing, allowing for consideration of dihedral angles during node embedding [12]. The Spherical Channel Network (SCN), proposed in 2022, and the EquiformerV2, introduced in 2023, both implemented equivariant representations and achieved state-of-the-art performances [67, 68]. In 2023, MACE introduced higher order message passing, addressing limitations like high computational cost and poor scalability of other equivariant message passing neural networks (MPNNs) [13].

A key advantage of neural net based MLPs is the application of foundation models, like MACE-MP-0 [15]. The foundational model is pretrained on the MPtrj dataset, which consists of approximately 1.5M configurations, roughly ten times the approximately 150k unique structures from the Materials Project [69]. Therefore, MACE-MP-0 is not only capable of molecular dynamics simulation across a wide variety of chemistries itself, but also allows users to fine-tune the foundation model with their own specific dataset. As a result, the model substantially reduces the amount of data and/or training time required, compared to training models from scratch. Moreover, foundation models are very promising for modeling competitive multi-component processes, for which training a system-specific, on-the-fly active learning model would be expensive or even prohibitive.

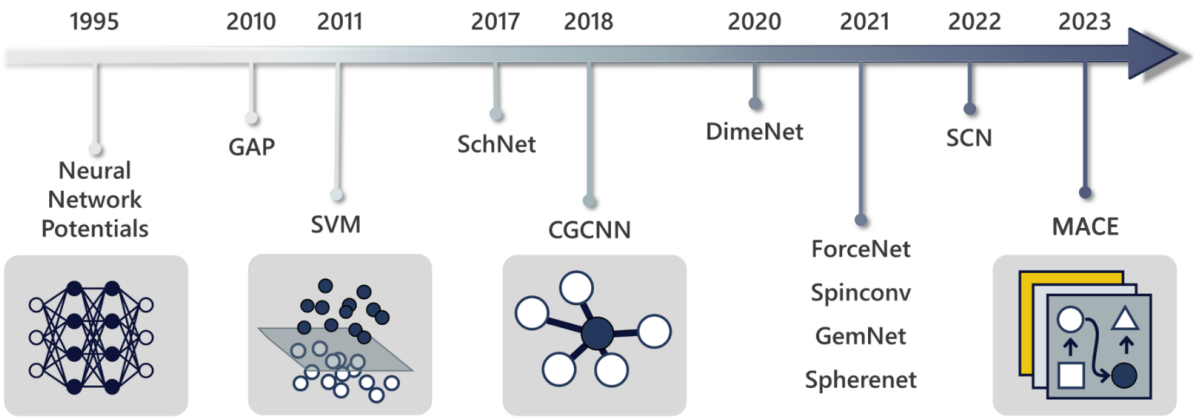

Figure 2.5: Visual representation of the chronological history of machine learning potentials, modified from [6].

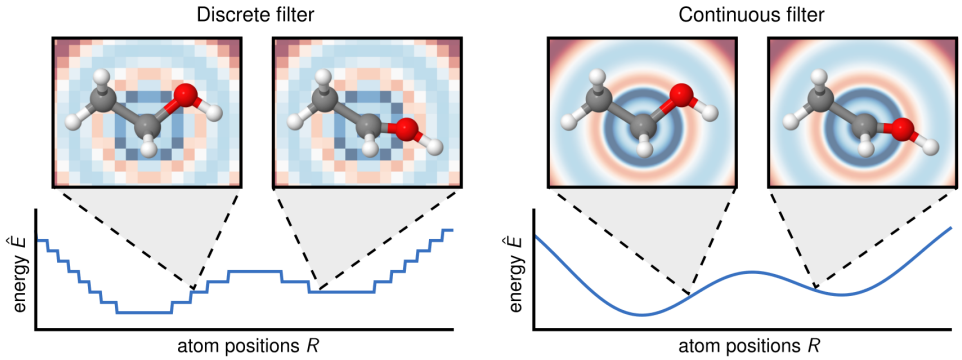

Figure 2.6: The discrete filter (left) is not able to capture the subtle positional changes of the atoms resulting in discontinuous energy predictions $\hat{E}$ (bottom left). The continuous filter captures these changes and yields smooth energy predictions (bottom right) [65]

2.3.2. SchNet

Early neural networks primarily consisted of linear and convolutional layers. In deep learning, convolutional layers typically operate on discretized signals, such as image pixels or video frames. For these signals, it is straightforward to define the filter on the same grid. However, this approach is not feasible for unevenly spaced inputs, such as the atomic positions in a molecule (see Fig. 2.6). To address this, SchNet introduced continuous filters, capable of handling unevenly spaced data, particularly atoms positioned arbitrarily in space [65].

Given n objects $X^l = (x_1^l, ..., x_n^l)$ where $x_i^l \in \mathbb{R}^F$ at locations $R = (\mathbf{r}_1, ..., \mathbf{r}_n)$ where $\mathbf{r}_i \in \mathbb{R}^D$, the continuous-filter convolutional layer l requires a filter-generating function W^l that maps from a position to the corresponding filter values, see Eq. 2.15. In SchNet this filter-generating function is modeled with a neural network [65].

$$W^l : \mathbb{R}^D \to \mathbb{R}^F \tag{2.15}$$

The output $\mathbf{x}_i^{l+1}$ for the convolutional layer at position $\mathbf{r}_i$ is then given by

$$\mathbf{x}_i^{l+1} = (X^l * W^l)_i = \sum_j \mathbf{x}_j^l \circ W^l(\mathbf{r}_i - \mathbf{r}_j) \tag{2.16}$$

where "$\circ$" represents the element-wise multiplication.

SchNet, like CGCNN, uses relative distances $\left\|\mathbf{x}_{ij}\right\|$, a 2-body invariant, to scalarize geometric information [70]

$$\mathbf{s}_i^{(t+1)} := \mathbf{s}_i^{(t)} + \sum_j f_1\left(s_j^{(t)}, \left\|\mathbf{x}_{ij}\right\|\right) \tag{2.17}$$

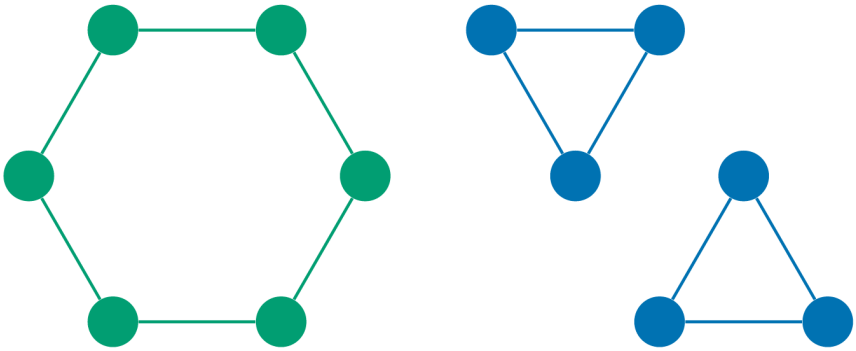

Figure 2.7: A standard non-directional GNN cannot distinguish between a hexagonal (left) and two triangular molecules (right) with the same bond lengths, since the neighborhood of each atom is exactly the same. An example of this would be cyclohexane and two cyclopropane molecules with slightly stretched bonds, when the GNN either uses the molecular graph or a cutoff distance of $c < 2.5$A. Directional message passing solves this problem by considering the direction of each bond [11].

2.3.3. DimeNet

The next generation of neural net based MLPs, like DimeNet and GemNet, goes beyond using only the absolute distance between two atoms, also taking angular information into account. Distances and angles, represented by $\mathbf{x}_{ij} \cdot \mathbf{x}_i k$, among 3-body invariants are utilitzed as shown in Eq. 2.18 [70]

$$\mathbf{s}_i^{(t+1)} := \sum_j f_1\left(\mathbf{s}_i^{(t)}, \mathbf{s}_j^{(t)}, \sum_{k \neq j} f_2\left(\mathbf{s}_j^{(t)}, \mathbf{s}_k^{(t)}, \left\|\mathbf{x}_{ij}\right\|, \mathbf{x}_{ij} \cdot \mathbf{x}_i k\right)\right) \tag{2.18}$$

GNNs represent each atom i via an atom embedding $\mathbf{s}_i \in \mathbb{R}^S$. These embeddings are updated in each layer by passing messages along the molecular graph edges. However, GNNs do not use the full distance matrix, as doing so would mean passing messages globally between all pairs of atoms, which increases computational complexity and can lead to overfitting. Instead, GNNs often apply a cutoff distance c, which means they cannot distinguish between certain molecules [71].

One example of two indistinguishable atoms for GNNs with a low cutoff value, are a hexagonal (e.g. cyclohexane) and two triangular molecules (e.g. cyclopropane) with the same bond lengths, since the neighborhoods of each atom are identical for each atom. DimeNet addresses this limitation by incorporating the directions to neighboring atoms, not just their distances. The directional embedding $\mathbf{m}_{ji}$ associated with the atom pair ji can be thought of as a message being sent from atom j to atom i. Therefore, DimeNet embeds each atom i using a set of incoming messages $\mathbf{m}_{ji}$, i.e. $\mathbf{s}_i = \sum_j \mathbf{m}_{ji}$, and updates the message $\mathbf{m}_{ji}$ based on the incoming messages $\mathbf{m}_{kj}$. Hence, as illustrated in Fig. 2.8 the update function and aggregation scheme for message embeddings in DimeNet is defined as

$$\mathbf{m}_{ji}^{(l+1)} = f_1\left(\mathbf{m}_{ji}^{(l)}, \sum_{k \neq i} f_2\left(\mathbf{m}_{kj}^{(l)}, \mathbf{e}_{RBF}^{(ij)}, \mathbf{a}_{SBF}^{(ki,ij)}\right)\right) \tag{2.19}$$

where $\mathbf{e}_{RBF}^{(ij)}$ represents distance d_{ij}, and $\mathbf{a}_{SBF}^{(ki,ij)}$ is a joint representation of angles $\alpha_{(ki,ij)}$ between message embeddings and interatomic distances d_{kj} [11].

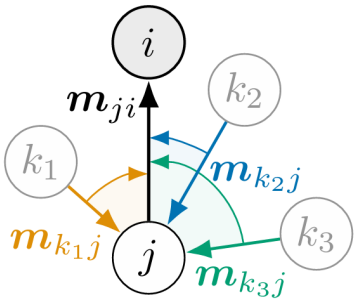

Figure 2.8: Aggregation scheme for message embeddings [11].

2.3.4. MACE

Up to this point I have discussed geometric message passing GNNs, which propagate only local scalar quantities such as distances and angles. In contrast, equivariant GNNs propagate geometric quantities, such as vectors and relative positions or higher order tensors. Higher order spherical tensors, denoted as $\mathbf{h}_{i,l} \in \mathbb{R}^{2l+1 \times f}$, begin at order $l = 0$ for scalar quantities, $l = 1$ for vector quantities, and extend to arbitrary orders up to $l = L$. These higher order tensors are updated via tensor products $\bigotimes$ of neighborhood features $\mathbf{h}_j$ for all $j \in \mathcal{N}_i$, combined with higher order spherical harmonic representations Y of the relative displacement $\hat{\mathbf{x}}_{ij}$

$$\mathbf{h}_i^{(t+1)} := \mathbf{h}_i^{(t)} + \sum_{j \in \mathcal{N}_i} Y(\hat{\mathbf{x}}_{ij}) \bigotimes_{\mathbf{w}} \mathbf{h}_j^{(t)} \tag{2.20}$$

where the weights $\mathbf{w}$ of the tensor product are computed via a learnt radial basis function of the relative distance, i.e. $\mathbf{w} = f(\left\|\mathbf{x}_{ij}\right\|)$ [70].

The key innovation in the MACE network is a new message construction mechanism. The messages $\mathbf{m}_i^{(t)}$ are expanded using a hierarchical body order expansion

$$\mathbf{m}_i^{(t)} = \sum_j \mathbf{u}_1\left(\mathbf{s}_i^{(t)}, \mathbf{s}_j^{(t)}\right) + \sum_{j_1, j_2} \mathbf{u}_2\left(\mathbf{s}_i^{(t)}, \mathbf{s}_{j_1}^{(t)}, \mathbf{s}_{j_2}^{(t)}\right) + \ldots + \sum_{j_1, \ldots, j_\nu} \mathbf{u}_\nu\left(\mathbf{s}_i^{(t)}, \mathbf{s}_{j_1}^{(t)}, \ldots \mathbf{s}_{j_\nu}^{(t)}\right) \tag{2.21}$$

where the $\mathbf{u}$ functions are learnable, the sum runs over the neighborhood of i, and ν corresponds to the maximum correlation order, which is the body order of the message function minus one.

Crucially, by writing the sum $\sum_{j_1,\ldots,j_\nu}$ which includes self-interaction, MACE achieves a computational efficient parameterization of the tensor product structure. This parameterization allows MACE to avoid the exponential scaling of computational cost with the body order ν. The approach eliminates the need to symmetrize or generate all k-tuples, as required in more traditional many-body expansions, using the summation $\sum_{j_1 < \ldots < j_\nu}$, like DimeNet does [11]. An analogy for this method is calculating the product $(a + b + \ldots)^k$, which implicitly includes terms like $a^l b^{k-l}$ instead of calculating each $a^l b^{k-l}$ term individually [13].

2.4. Thermodynamic Reweighting

Thermodynamic reweighting is a technique rooted in statistical mechanics and traditionally used to correct sampling biases in molecular simulations [19]. In order to understand the relevance of thermodynamic reweighting for my case, one has to approach molecular dynamics (MD) from a thermodynamical point of view. MD effectively samples configurations according to a Boltzmann distribution, governed by the potential energy surface (PES), which defines forces and energies and thus shapes the sampled distribution (see Fig. 2.9). The PES, and therefore the sampled distribution, depends directly on the chosen density functional theory (DFT) functional, whose accuracy dictates how closely the PES reflects reality. Thermodynamic reweighting addresses this dependence by allowing trajectories generated under one PES to be reweighted as though sampled from another distribution.

In order to properly reweight an observable or distribution, one uses the expression in Eq. 2.22 for the weights

$$w_i = e^{-\frac{\Delta E}{k_B T}} \quad (i \in structures) \tag{2.22}$$

where w_i is the weight for structure i, ΔE the difference in energy of the structure and the respective structure in the reference distribution, k_B Boltzmann's constant, and T the temperature.

The weights are subsequently used to reweight an observable as expressed in Eq. 2.23 and Eq. 2.24

$$A_{i,new} = A_i \frac{w_i}{\sum_i w_i} \quad (i \in structures) \tag{2.23}$$

$$\langle A \rangle = \frac{\sum_i A_i w_i}{\sum_i w_i} \quad (i \in structures) \tag{2.24}$$

where w_i is the weight for structure i determined by Eq. 2.22, A_i is the observable for structure i, and $\langle A \rangle$ represents the weighted average of the observable across the distribution.

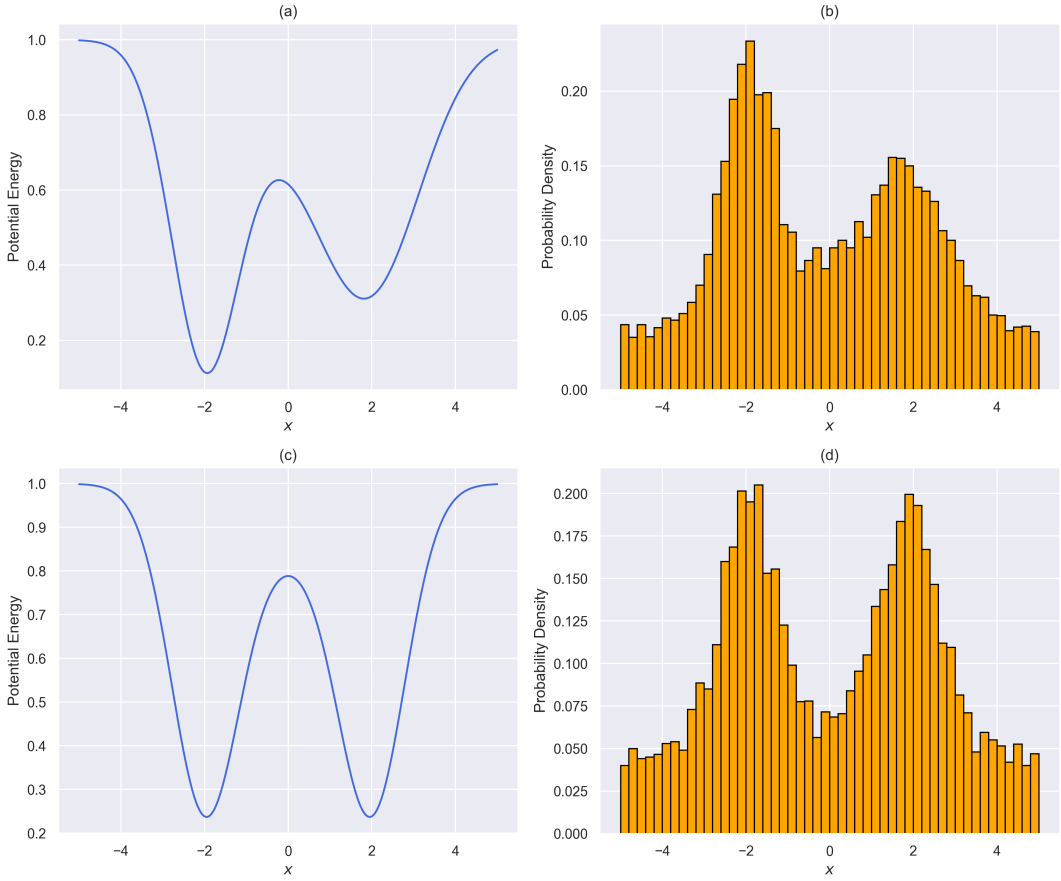

Figure 2.9: Comparison of one-dimensional potential energy surfaces (PES) and their corresponding probability density distributions: (a) an asymmetric PES with its derived distribution (b), and (c) a symmetric PES with the associated distribution (d).

In Fig. 2.9 I report an illustrative example of thermodynamic reweighting. Panel (a) shows a 1D PES with a global minimum at $x = -2$ and a local minimum at $x = 2$, alongside its corresponding probability distribution $P(x) \propto e^{\beta E(x)}$ in panel (b). Panels (c) and (d) depict a similar 1D PES and probability distribution, but with symmetric global minima at $x = -2$ and $x = 2$. Reweighting an observable sampled from (b) to (d) increases the relative occurrence of the configuration at $x = 2$ compared to $x = -2$.

Originally developed for Monte Carlo and MD studies, reweighting enables researchers to extract high-fidelity thermodynamic properties without sampling new and computationally expensive trajectories [19, 72–74]. With the rise of MLPs, trained on high-level DFT data, thermodynamic reweighting has become significantly more powerful. By reprocessing large trajectories from inexpensive functionals, MLPs allow observables to be reconstructed as though they were sampled from a more accurate reference potential — all without the computational cost of running high-level DFT simulations directly.

3

Methods

3.1. Data

The initial dataset consists of 54 PBE+D3 trajectories of the zirconocene core $(^{Bu}Cp)_2ZrH_2$ grafted onto an amorphous silica slab model in two configurations, "chemically" adsorbed (CA) and "physically" adsorbed (PA) (see Section 2.1.4). Trajectories in the dataset were generated under three strain levels and three simulation methods - aiMD at 335K, aiMD at 773K, and Velocity Squared Molecular Dynamics (VSMD) [75–77] - producing 18 unique conditions (2 configurations × 3 strains × 3 methods). For each combination I have three trajectories with different seed numbers, running for 10,000 time steps, 540,000 structures in total. To minimize temporal autocorrelation, only 50th structures were retained, reducing autocorrelation to ~0.5 (see Fig. 3.1). The three trajectories per combination were merged into 18 reduced datasets of 600 structures each.

From these, the CA average strain trajectory at 353K was selected for CP2K single-point energy and force calculations using seven DFT functionals: LDA, PBE, PBE+D3, BLYP+D3, TPSS+D3, PBE0+D3, and B3LYP+D3. Here the suffix '+D3' refers to the DFT-D3 correction with Becke-Johnson dampening function (see Section 2.2.9). Each functional's dataset was randomly divided into six batches of 100 structures. One batch was reserved for model testing, while the remaining five formed training sets of 400 structures each, with a different batch excluded in each set to introduce stochastic variation among the models. These training sets were subsequently used to train models for thermodynamic reweighting.

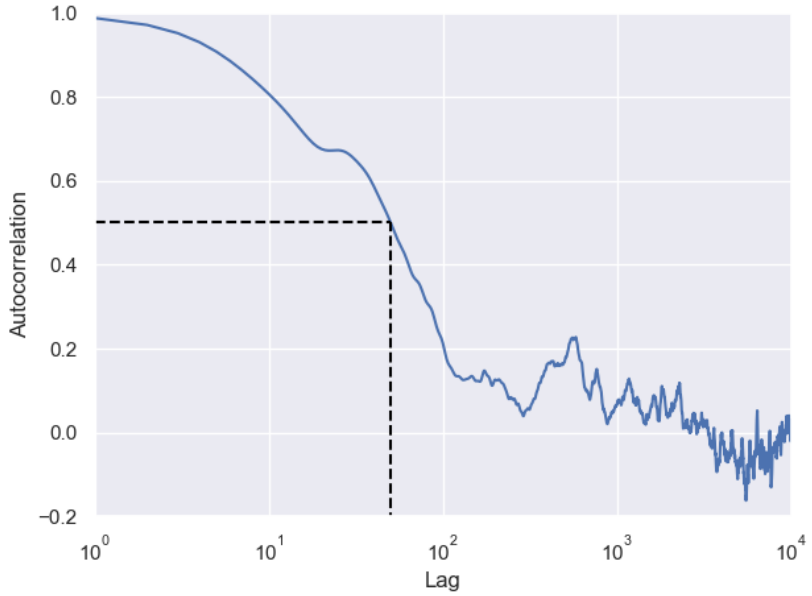

Figure 3.1: Temporal autocorrelation of a full trajectory's coordinates, plotted on a logarithmic x-axis. Dashed lines indicate the autocorrelation at a lag of 50.

3.2. MACE models training strategy

The first model I deployed was a MACE-MP-0 medium foundation model, fine-tuned on the PA, average strain, 353K full trajectory of 10,000 time steps. It was trained for 50 epochs without stochastic weight averaging (SWA), using an 80/20 training-validation split, with batch sizes of 4 and 2. This model was then used to predict energies and forces across all reduced trajectories, with error analysis presented in Section 4.1.

Due to the model's inability to run physically sensible MD simulations, I prepared a more diverse dataset to train more robust models. Five training sets of 1,000 structures were randomly drawn from the 10,200 structures of the reduced dataset. Each set underwent the same 80/20 split for training and validation, with batch sizes unchanged. I then trained five models from these datasets, and implemented an active learning loop to enhance performance even further. The active learning loop consists of the following steps (see Fig. 3.2 for a visual representation):

1. Run MD using trained MACE models and Langevin thermostat in ASE. Evaluate standard deviation between the five models energies after each 0.5 fs time step and exit MD once the standard deviation exceeds a threshold;
2. Extract structures of the last 25 MD time steps before the simulation collapsed;
3. Load coordinates into CP2K and compute forces with PBE+D3 functional;
4. Expand training sets with the calculated CP2K coordinates/forces;
5. Train the five MACE models for 10 more epochs using the updated training set;
6. Repeat process from step 1.

The models were subsequently used as MACE calculators in ASE to perform MD simulations. The simulations were initialized at 300K and regulated at 353K using a Langevin thermostat. The simulation was run for 10,000 time steps with a 0.5 fs integration step, and the results are presented in Section 4.2.

Finally, I trained another ensemble of five fine-tuned MACE-MP-0 medium MLP models for 75 epochs on five 400-structure datasets as defined in Section 3.1. SWA was applied during the final 25 epochs to refine energy predictions. Each dataset was split 80/20 for training and validation, using batch sizes of 4 and 2, respectively. This procedure was repeated for each functional, producing ensemble models capable of energy prediction and error estimation through inter-model variance.

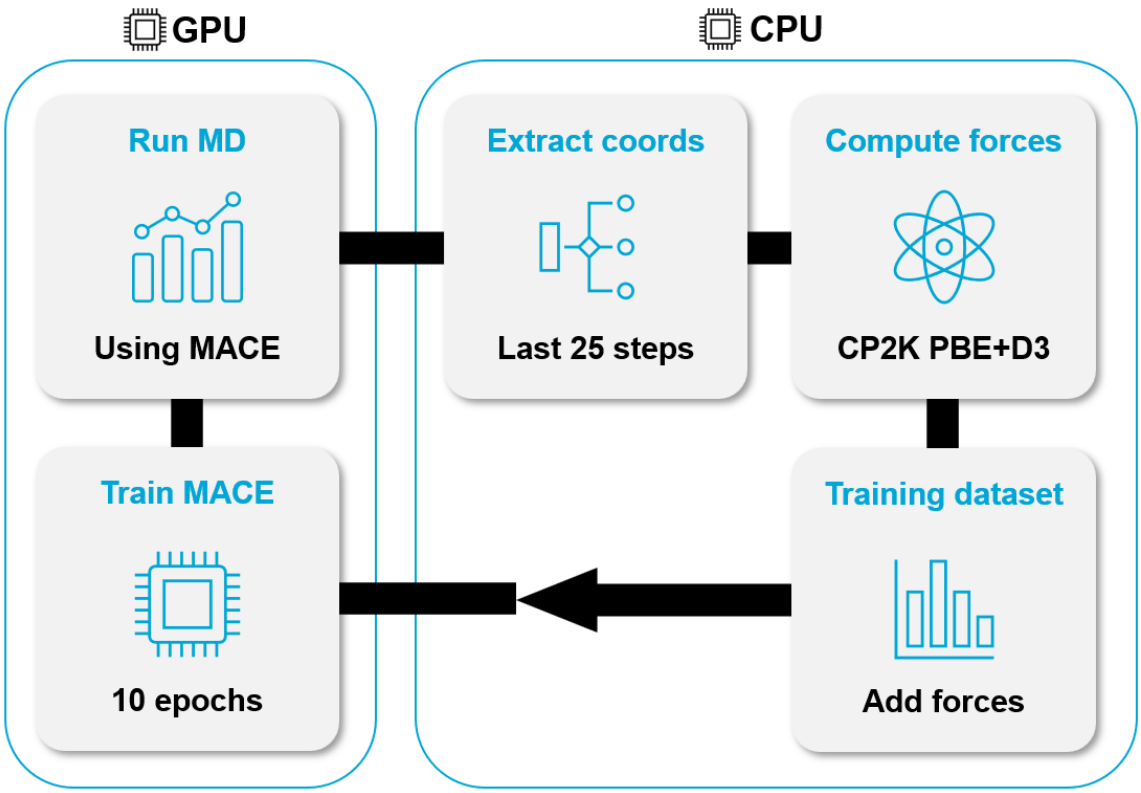

Figure 3.2: Active learning loop used to further improve the ensemble model used to perform MD. The loop consists of (1) Running MD with MACE models until the prediction gets inaccurate; (2) Extracting structures of the last 25 MD steps; (3) Loading these into CP2K to calculate forces; (4) Expanding the training sets by these coordinates/forces/energies; (5) Retraining the MACE models for 10 epochs.

In order to minimize the energy error of the thermodynamic reweighting models, I set up another active learning protocol, initializing from the ensemble models. The loop consists of the following steps:

1. Evaluate energies using MACE across a broader dataset with known energy values.
2. Select the 100 configurations with the highest error, and integrate them into the training set, targeting areas where the model struggles most.
3. Retrain MACE models for 10 more epochs using the revised training set.
4. Repeat process from step 1.

3.3. Thermodynamic Reweighting

To compare the different functionals, I first correct the total energy by the individual atom contributions

$$E^0_{DFT} = E_{DFT} - \sum_j N_j E^j_{DFT} \quad j \in atoms \tag{3.1}$$

Where E^0 is the corrected energy, DFT corresponds to the respective DFT functional, j represents the chemical symbol, and N_j stands for the number of atoms in the structure with chemical symbol j.

A comparison of the corrected energies across different functionals is presented in Fig. 3.3. Among them, B3LYP+D3 shows the greatest similarity to PBE0+D3, with TPSS+D3 following closely. Notably, B3LYP+D3 and BLYP+D3 have a significant offset to PBE0+D3, about 0.2 eV/atom. This is consistent with the way these functionals are constructed: while PBE, PBE+D3, TPSS, and PBE0 share a common foundation in the PBE potential, B3LYP+D3, BLYP+D3, and LDA stem from alternative formulations.

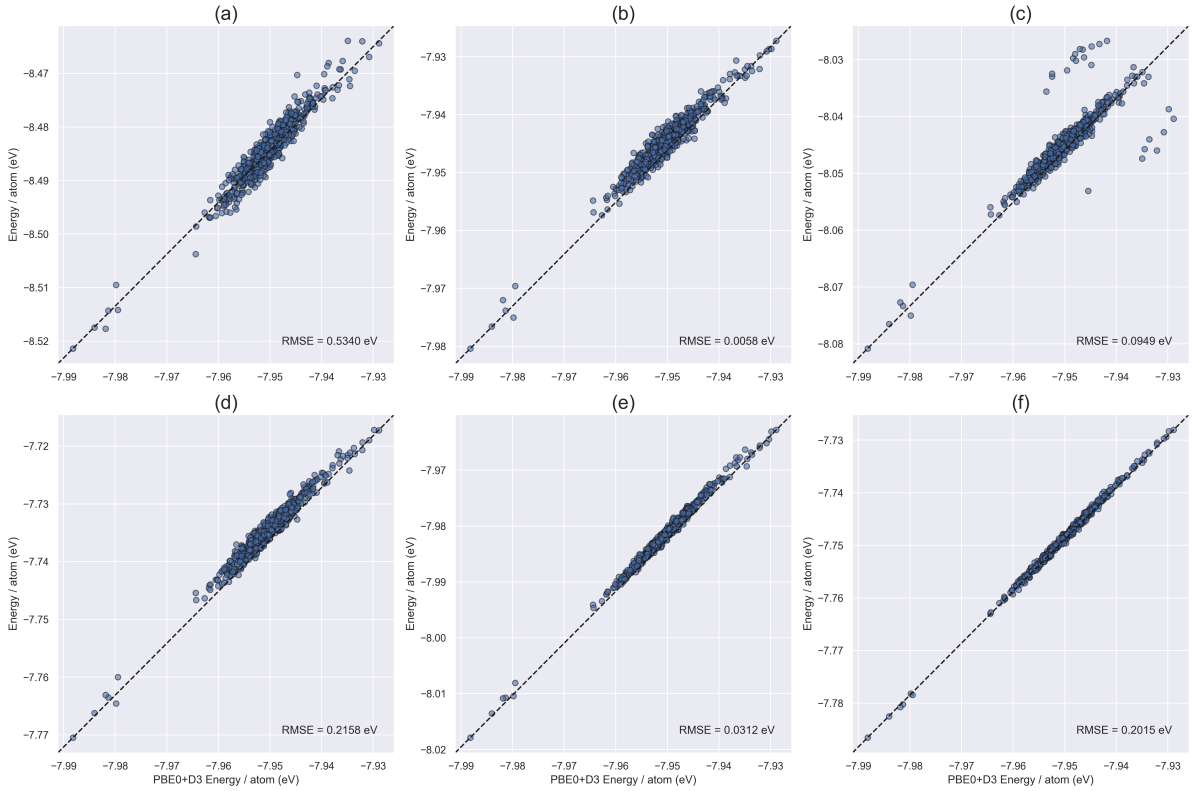

Figure 3.3: Parity plots comparing corrected energies per atom across various DFT functionals, benchmarked against PBE0+D3. The x-axis shows the corrected PBE0+D3 energy per atom, while the y-axis gives the corrected energy per atom for (a) LDA (b) PBE (c) PBE+D3 (d) BLYP+D3 (e) TPSS+D3 and (f) B3LYP+D3.

For each of the 540,000 structures I calculated the Zr-H and Si-H bond lengths, provided such bonds existed within the structure. Each structure was assigned a weight w_i^{MLP} computed according to Eq. 3.2 using energies E^0_{MLP} predicted by each of the trained ensemble models. These weights were then used to reweight the bond lengths distribution accordingly.

$$w_i^{MLP} = e^{-\frac{E^0_{PBE+D3} - E^0_{MLP}}{k_B T}} \qquad (i \in structures) \tag{3.2}$$

Where w_i is the weight for that structure with i ranging from 1 to 30,000, E^0_{PBE+D3} is the corrected DFT energy of the input data, E^0_{MLP} the corrected MLP energy (for the six different ensemble models), k_B Boltzmann's constant, and T the temperature of the system.

4

Results

4.1. Statics

4.1.1. Energies and Forces

The model's performance was first tested on a trajectory structurally similar to the training data to evaluate its predictive accuracy. Specifically, the model, trained on a PA average strain trajectory, was evaluated on a CA maximum strain reduced trajectory with known forces and energies. This setup retained the same organometallic core and substrate but introduced differences in strain and attachment geometry, providing a practical test of the model's generalization capabilities.

The results of the test in Fig. 4.1 show excellent agreement for forces, with an RMSE of 0.098 eV/Å and a near-ideal slope of 0.99, where data points align tightly along the parity line. The error decreases for larger forces and increases for near-zero forces, which is intuitive from both physical and machine learning intuition. From a physical perspective, larger forces arise from steep gradients on the PES, providing clearer learning signals. From a machine learning standpoint, MACE's RMSE-driven training objective gives larger forces more weight in error calculations, ensuring the model prioritizes their accuracy over smaller, less consequential forces.

The energy prediction error of 0.17 eV/atom is also good, but the 0.52 slope deviates significantly from the parity line. This imbalance likely stems from the model's training process, which prioritizes force accuracy a 100 times more than energy

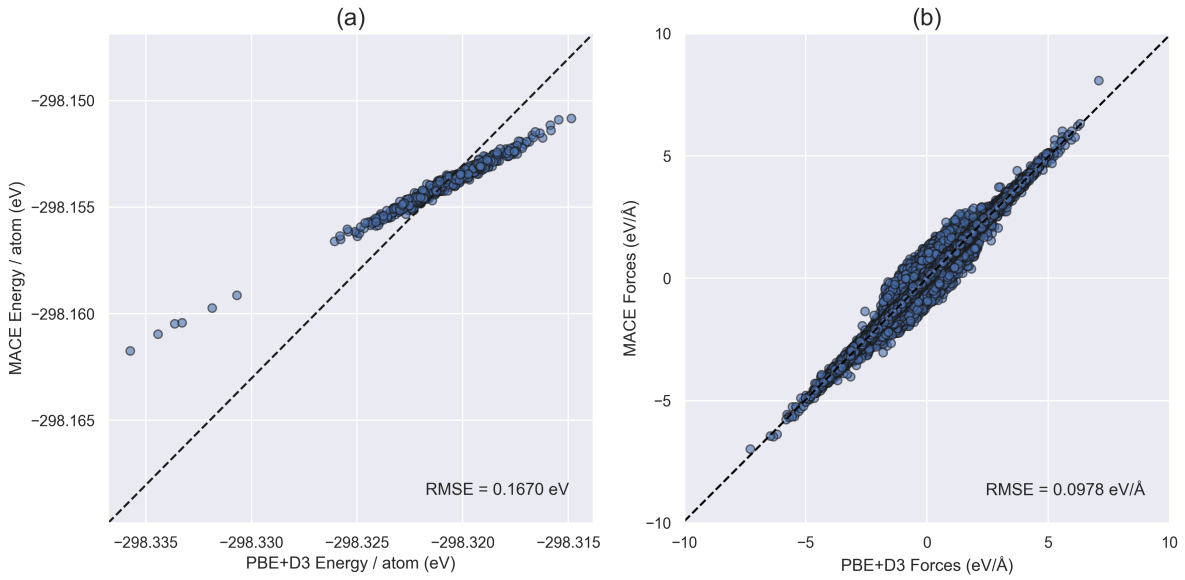

Figure 4.1: Parity plots of MACE-predicted energy (a) and forces (b) versus DFT PBE+D3 references. The model was trained on a PA average strain full trajectory and evaluated on a CA maximum strain reduced trajectory, both derived from aiMD simulations at 353K. The dashed diagonal line represents the ideal parity line with a slope of 1.

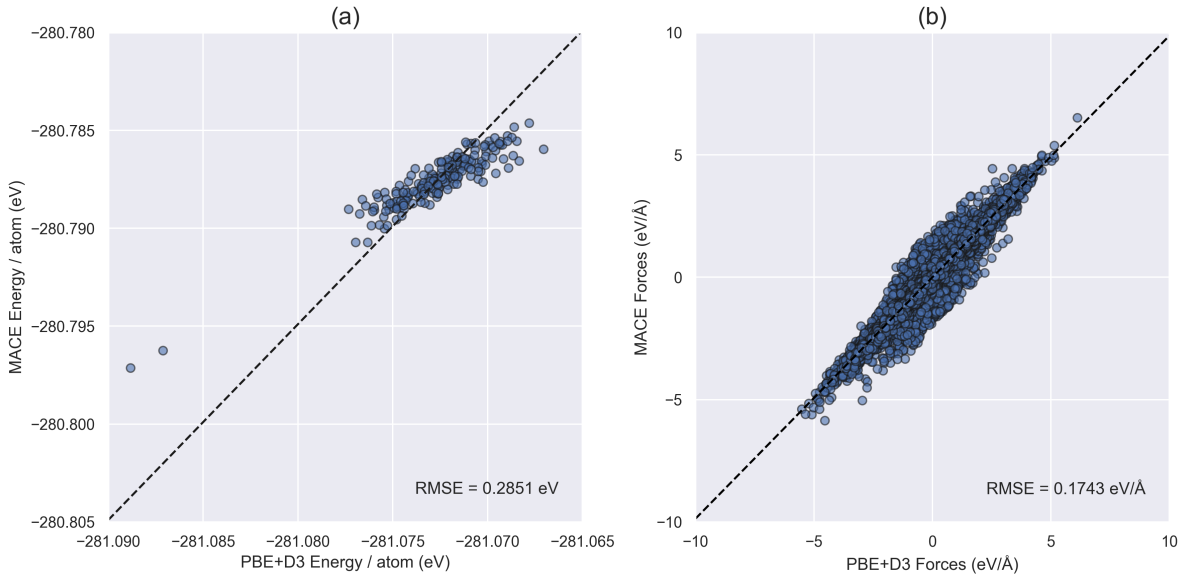

Figure 4.2: Parity plots of MACE-predicted energy (a) and forces (b) versus DFT PBE+D3 references. The model was trained on a PA average strain full trajectory and evaluated on a highly strained silica reduced trajectory, both derived from aiMD simulations at 353K. The dashed diagonal line represents the ideal parity line with a slope of 1.

accuracy. Alternative explanations, such as limited data diversity, insufficient model capacity, or premature training termination, were tested. Yet, models trained with more varied datasets, larger architectures, and longer training durations still displayed the same behavior, ruling out these possibilities.

Testing the model on a system with the same organometallic core but a smaller, more strained silica substrate yielded similar trends to those in Fig. 4.2, though the RMSE for forces and energies increased by roughly 75%. This outcome is consistent with expectations, given the more substantial difference between this system and the training set, compared to the test system used in Fig. 4.1.

4.1.2. PA Trajectories

Next, I evaluated how the model performed across different trajectories to uncover any patterns or structures where it might struggle. Since the model was trained on PA data, I began by predicting energies and forces of other PA trajectories. The energy error distributions in Fig. 4.3 show only slight differences between the trajectories, with an impressive maximum error of just 0.025 eV/atom. The VSMD trajectories, however, demonstrate slightly higher errors than the aiMD trajectories, which aligns with expectations given that the model was trained on aiMD data.

Interestingly, error distributions across strain strengths reveal that structures with minimum strain yield the highest errors, while maximum strain structures exhibit the lowest. Although counterintuitive at first, this pattern is consistent with observations from Fig. 4.1 and Fig. 4.2, where lower energies appear to be predicted less accurately than higher energies. This can be explained by the nature of the PES again: high-strain structures correspond to higher-energy configurations that lie on steeper regions of the PES, making small deviations in geometry lead to larger energy changes, which the model captures more easily. In contrast, low-strain structures sit in flatter regions of the PES, where small deviations cause only minor energy differences, making errors more proportionally significant and harder to learn.

The force error distributions in Fig. 4.4 follow a similar pattern, with higher errors for VSMD trajectories and comparable spreads across PA trajectories. However, the improved accuracy in high-strain trajectories doesn't carry over to force predictions. One hypothesis is that while high-strain structures provide clearer energy gradients, they also introduce more localized, anharmonic distortions, especially near strained bonds, leading to more irregular force patterns that deviate from the smoother forces the model was trained on.

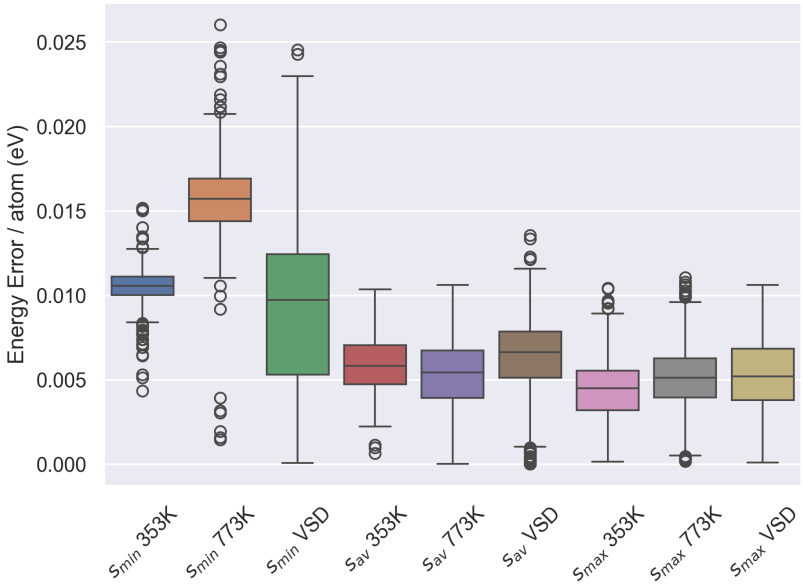

Figure 4.3: Error distribution of MACE energy predictions for various PA trajectories presented as a boxplot. Strain conditions are labeled s_{min}, s_{av}, s_{max} for minimum, average, and maximum strain respectively, while labels 353K, 773K, VSD refer to aiMD at 353K, aiMD at 773K, and Velocity Squared Molecular Dynamics.

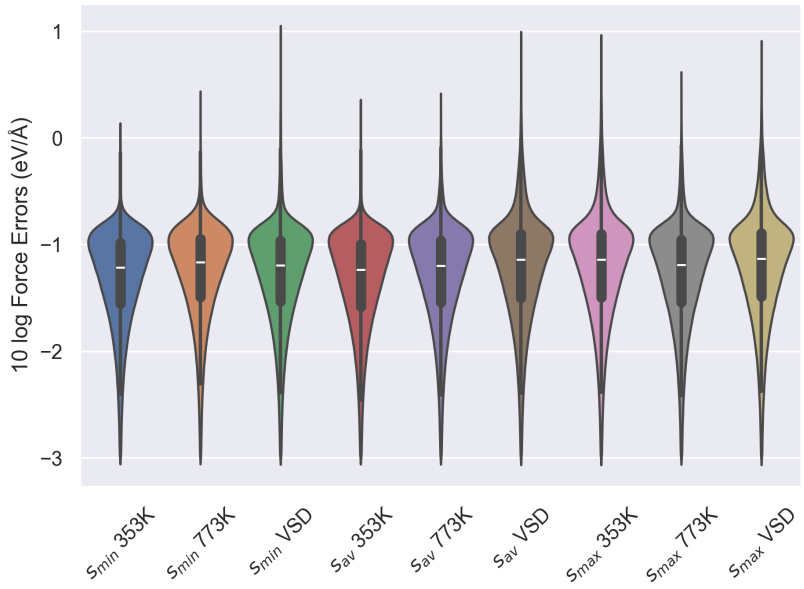

Figure 4.4: Error distribution of MACE force predictions for various PA trajectories presented as a violin plot. Strain conditions are labeled s_{min}, s_{av}, s_{max} for minimum, average, and maximum strain respectively, while labels 353K, 773K, VSD refer to aiMD at 353K, aiMD at 773K, and Velocity Squared Dynamics.

4.1.3. CA Trajectories

When comparing CA and PA trajectories, error distributions across energies and forces show remarkably consistent trends. The force error distributions in Fig. 4.6 closely mirror those from the PA trajectories in Fig. 4.4, including the higher errors in the VSMD trajectories. Energy errors in Fig. 4.5, on the other hand, range between 0.14 and 0.20 eV/atom, which is a sharp increase from the PA errors. This increase, while expected due to the model's training on PA data, remains surprising for its magnitude.

Consistent with prior results, the highest errors remain in trajectories with minimum strain, while maximum strain trajectories show the lowest errors. What stands out is the performance of the 773K aiMD trajectories, which consistently yield lower errors than the 353K aiMD or VSMD trajectories, contrary to the 773k PA trajectories.

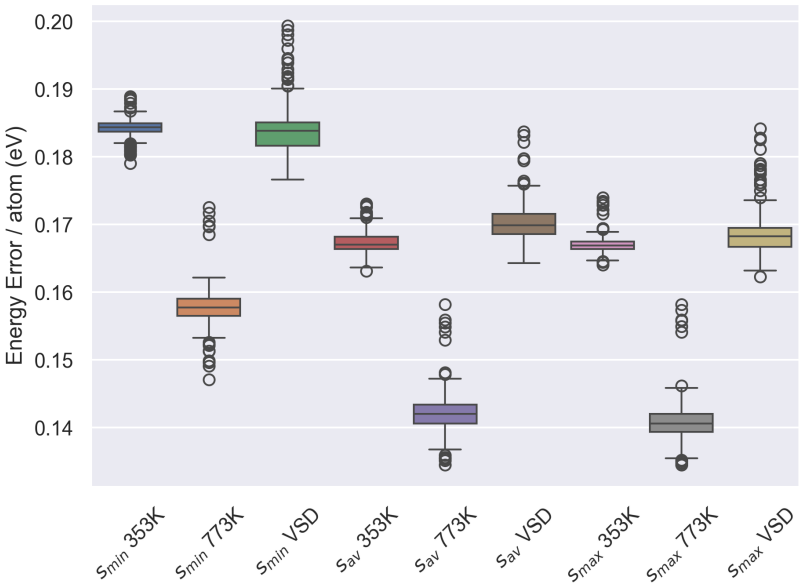

Figure 4.5: Error distribution of MACE energy predictions for various CA trajectories presented as a boxplot. Strain conditions are labeled s_{min}, s_{av}, s_{max} for minimum, average, and maximum strain respectively, while labels 353K, 773K, VSD refer to aiMD at 353K, aiMD at 773K, and Velocity Squared Dynamics.

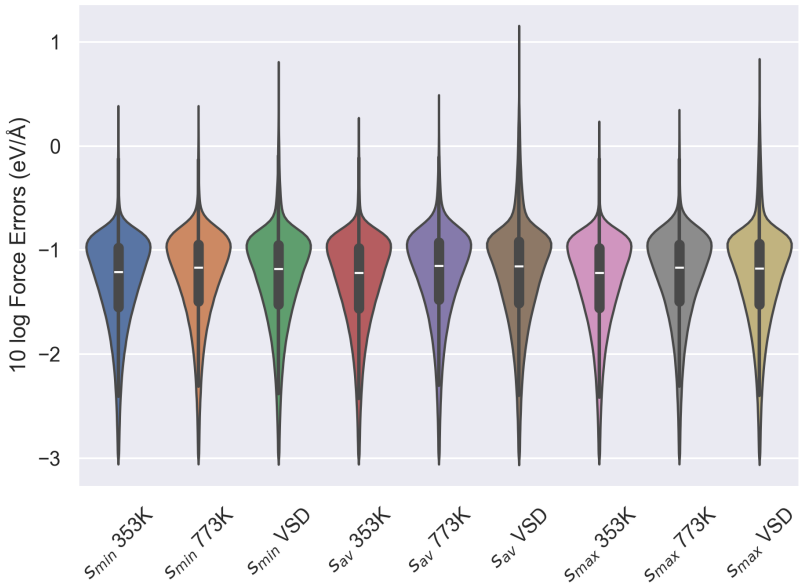

Figure 4.6: Error distribution of MACE force predictions for various CA trajectories presented as a violin plot. Strain conditions are labeled s_{min}, s_{av}, s_{max} for minimum, average, and maximum strain respectively, while labels 353K, 773K, VSD refer to aiMD at 353K, aiMD at 773K, and Velocity Squared Dynamics.

4.1.4. Improved model

To verify whether the differences between trajectories were an artifact of the previous model or a consistent trend, I trained a new model on a 1000 randomly selected data points across all trajectories. The results, displayed in Fig. 4.8 and Fig. 4.7, show that the previously observed order-of-magnitude gap between CA and PA trajectories is no longer present, though the overall patterns remain unchanged. This suggest that the original model was slightly overfitted to the CA trajectories. However, the differences are not purely an artifact of overfitting; certain trajectories, like 773K aiMD, are indeed easier to predict, while low-strain trajectories remain more difficult due to their weaker energy gradients.

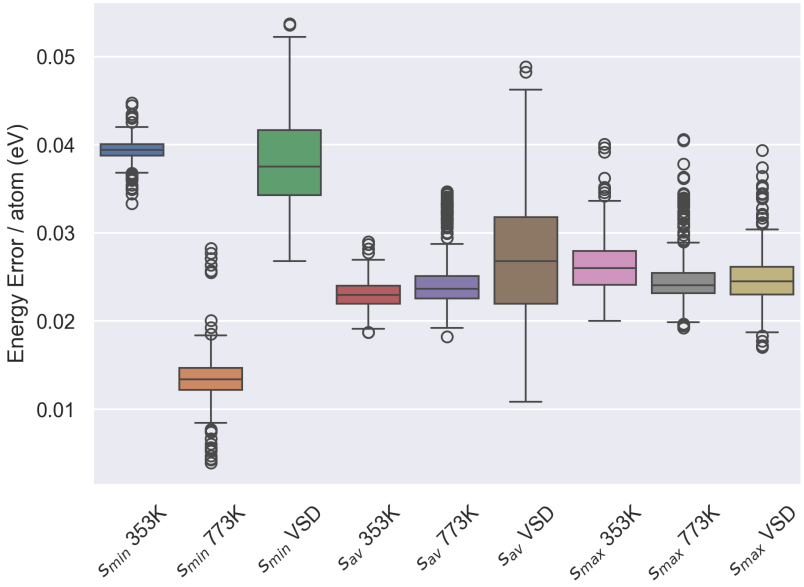

Figure 4.7: Error distribution of MACE energy predictions on various PA trajectories for a model trained on a 1000 randomly selected structures from all trajectories. Strain conditions are labeled s_{min}, s_{av}, s_{max} for minimum, average, and maximum strain respectively, while labels 353K, 773K, VSD refer to aiMD at 353K, aiMD at 773K, and Velocity Squared Dynamics.

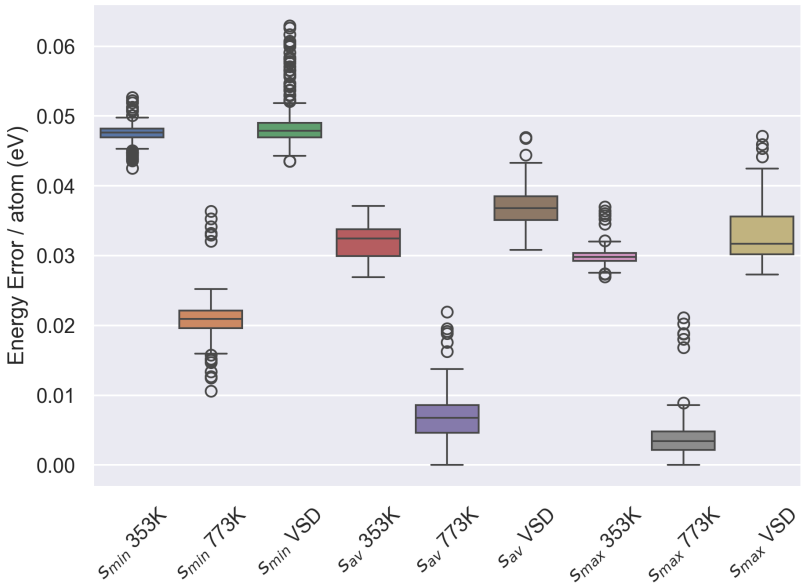

Figure 4.8: Error distribution of MACE energy predictions on various CA trajectories for a model trained on a 1000 randomly selected structures from all trajectories. Strain conditions are labeled s_{min}, s_{av}, s_{max} for minimum, average, and maximum strain respectively, while labels 353K, 773K, VSD refer to aiMD at 353K, aiMD at 773K, and Velocity Squared Dynamics.

4.1.5. Training sets

In this final test before running MD, I adjusted the dataset size and training duration to observe their impact on performance. Both factors improved accuracy initially but plateaued beyond a certain threshold (see Fig. 4.9), where additional data or additional training cycles did not improve the model any further.

Remarkably, even a model trained on a single test structure and single validation structure achieved reasonable accuracy. While models trained for 25 or 50 epochs still showed significant errors, models trained for 200 or 400 epochs achieved decent energy errors below 0.1 eV/atom while training for less than 5 minutes.

Obviously, a model trained on a 12-structure dataset for 50 epochs significantly outperformed the single-structure counterpart, with further improvements seen for 24, 48, and 600 structures. Interestingly, 48 structures and 50 epochs were needed to surpass the performance of a 200-structure model trained for just one epoch. These results emphasize that small datasets can already yield strong performance, and adding more diverse training data or training time does not always translate to better performance.

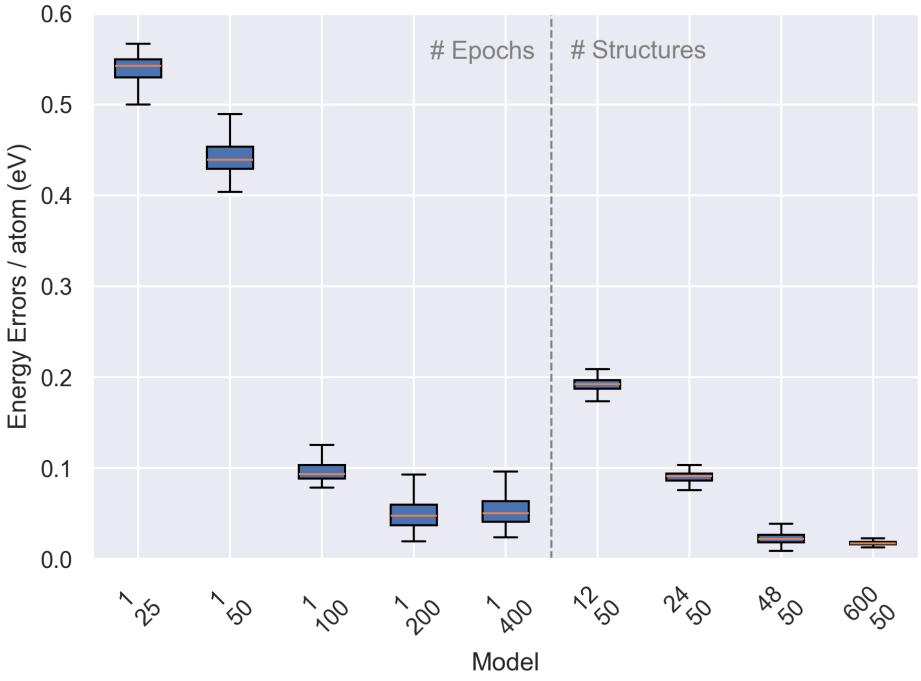

Figure 4.9: Box plot comparing energy error distributions for MACE models trained with varying dataset sizes (left) and different numbers of epochs (right). Each label denotes the number of structures and the number of epochs respectively. All models were trained on PA, maximum strain, 773K aiMD data and tested on a reduced CA, minimum strain, VSMD trajectory.

4.1.6. Compute time

To make a fair comparison between MACE and first-principles methods, it is important to take both accuracy and compute time into account. While MACE occasionally underperforms in accuracy, its speed advantage is undeniably substantial. To illustrate this, I conducted a single-point calculation on a structure taken from a CA, average strain, 353 K aiMD trajectory, using CP2K (48 cores of an AMD EPYC 9654 96-Core CPU) and MACE (NVIDIA V100 GPU), and compared the result in Table 4.1. Due to sequential stages in CP2K, where most cores remain idle, the estimated TFLOPS for LDA and PBE+D3 may be somewhat overestimated. However, as the calculation time increases, more parallel processing is utilized, yielding a more reliable PBE0+D3 estimate.

It is important to recognize that these benchmarks are based on single-point calculations, where CP2K spends considerable time converging the wavefunction. In contrast, MD simulations benefit from wavefunction availability from the previous time step, greatly speeding up DFT calculations. For example, performing PBE+D3 with an LDA wavefunction as initial guess, cuts computation time from 77 minutes to 12 minutes. While this reduces the relative speed advantage in MD scenarios, MACE remains orders of magnitudes faster than DFT.

Table 4.1: MACE compute time for a single point calculation compared to different DFT functionals, namely LDA, PBE+D3, and PBE0+D3.

Computational Method	Compute Time (h)	Compute Time (s)	TFLOPS/s
DFT LDA	0:46:08	2,768	38,752
DFT PBE+D3	1:17:21	4,641	68,454
DFT PBE0+D3	9:12:30	33,150	488,962
MACE	0:00:00.192	0.192	24

4.2. Dynamics

With a deeper understanding of the model from extensive testing on static structures, I employed the trained MACE model as a calculator in ASE to perform MD simulations. To improve model robustness, I trained five new models on different datasets and refined them using the active learning loop shown in Fig. 3.2. I then conducted MD simulations in ASE for 10,000 time steps with a Langevin thermostat and a 0.5 fs time step.

The results in Fig. 4.12 show seemingly contrasting insights. Panel (a) confirms that force predictions are highly accurate, and panel (b) shows reasonable energy predictions, though the slope deviation persists. The temperature profile in panel (c) behaves as expected, stabilizing around 353K after initialization at 300K, suggesting proper thermostat behavior. However, panel (d) reveals a major issue: the system has effectively disintegrated, with the original silica structure completely lost. Further inspection shows that hydrogen atoms begin escaping well before the 250th time step, indicating instability far earlier than expected.

In order to better understand why the model fails, Batatia et al. [13] provide an insightful illustration, which I reproduced in Fig. 4.11. Since the MACE model is trained only on the train and validation sets (red and green regions), it must extrapolate outside this domain. When the PES diverges significantly, the model becomes unreliable, leading to MD failures. Retraining on the last pre-failure datapoints enables the model to refine its understanding of these regions, gradually extending its predictive reach. However, if the model does not effectively generalize or learns only a narrow PES portion, it will fail repeatedly.

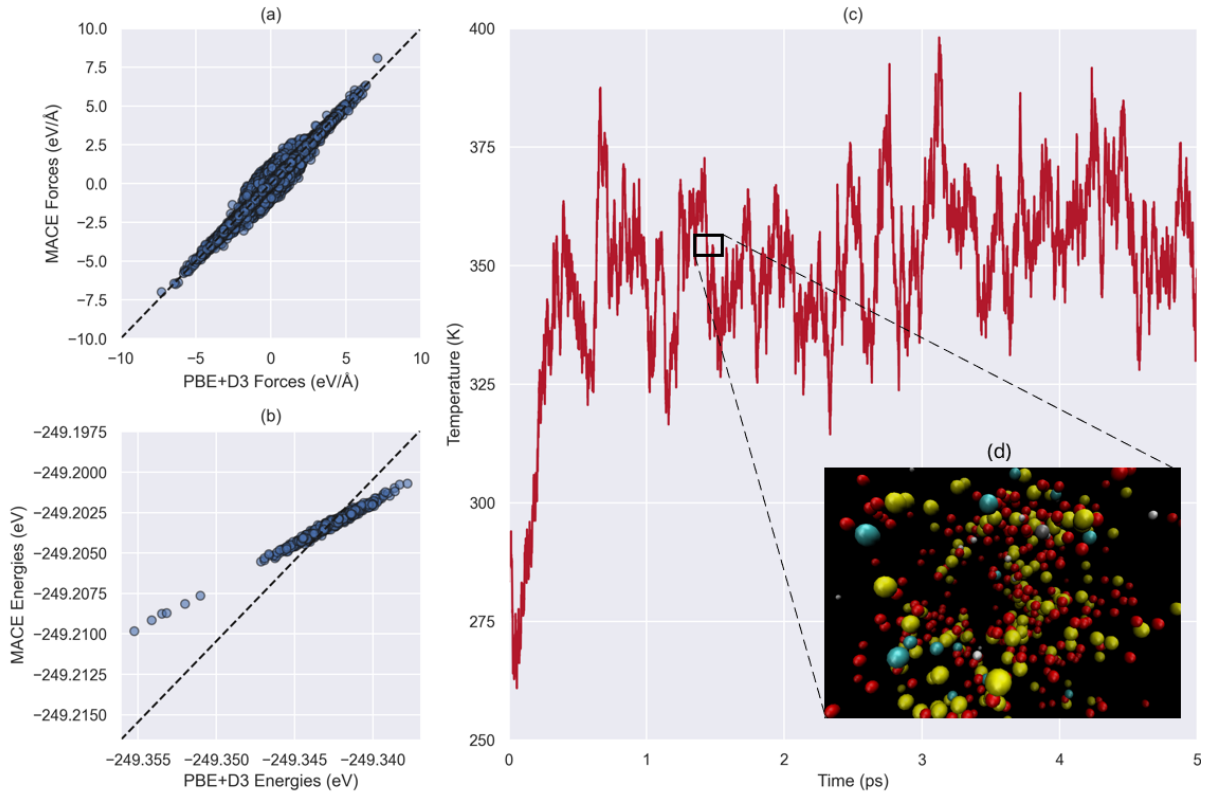

Figure 4.10: (a) Parity plots of MACE energies and (b) forces, (c) temperature evolution over 10,000 simulation steps (0.5 fs timestep), and (d) a visualization of the system at 1.5 ps.

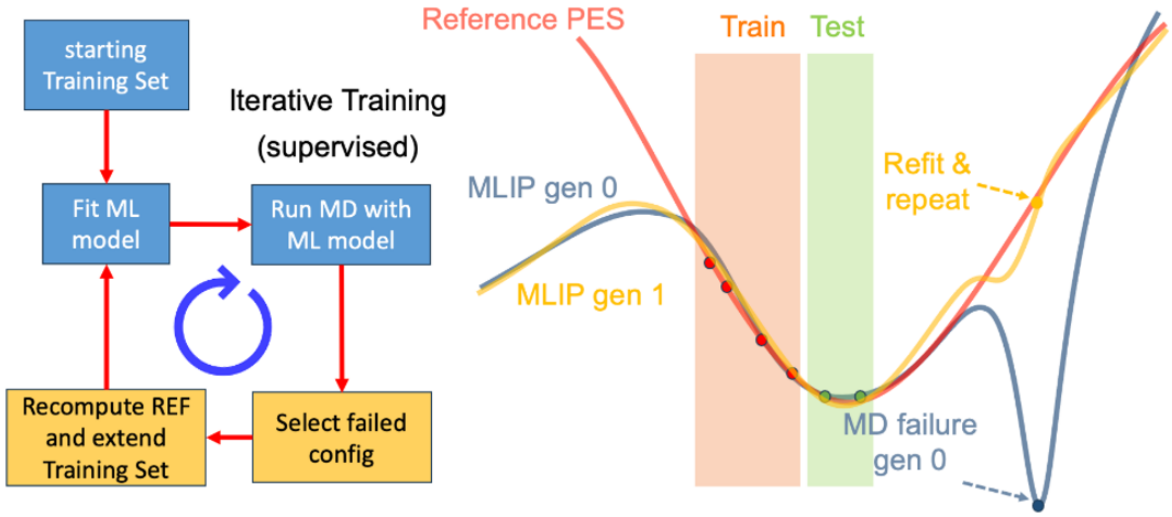

Figure 4.11: Illustration of the active learning approach. The model learns the PES using available data from the train set (red region) and validation set (green region). It interpolates within this range but must extrapolate beyond it, leading to eventual MD failure when encountering an unexplored PES region. By adding the last pre-failure datapoints to the training set and retraining the model, subsequent iterations progressively improve the model's PES knowledge, allowing MD to advance further [13].

To address this, I implemented the active learning loop shown in Fig. 3.2, systematically tuning parameters to enhance model robustness. I altered the number of active learning iterations, stored more or fewer structures before retraining, and experimented with different standard deviation thresholds. Additionally, I tested a fixed MD time step approach, instead of waiting for the standard deviation threshold to trigger retraining. Finally, I expanded the training set with over 250 structures containing random coordinate perturbations to improve PES coverage. Despite these efforts, none of the approaches prevented hydrogen atoms from escaping, even when assigning them the same mass as carbon atoms. Moreover, active learning iterations did not lead to any notable improvements in the models or their MD performance.

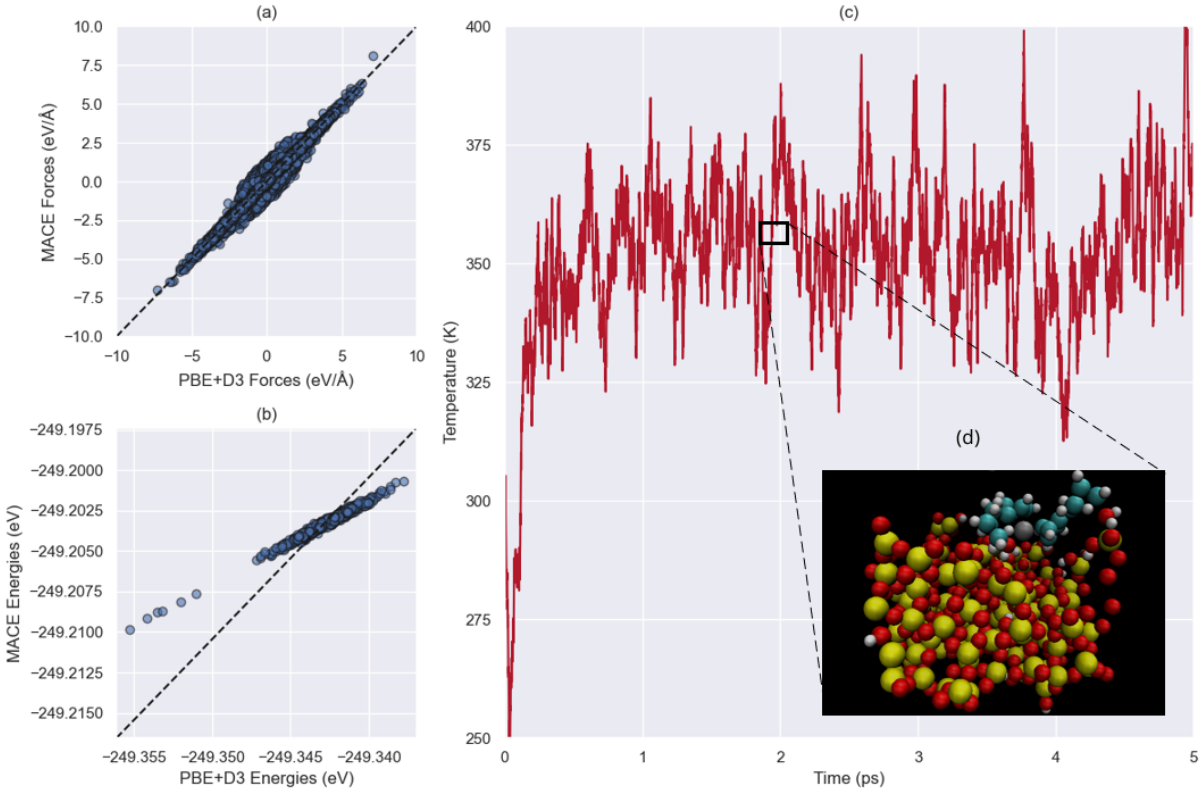

Figure 4.12: (a) Parity plots of MACE energies and (b) forces, (c) temperature evolution over 10,000 simulation steps (0.5 fs timestep), and (d) a visualization of the system at 2 ps using the new MACE "Multihead Replay Fine-Tuning" feature.

It wasn't until I had already moved on to the thermodynamic reweighting section and begun writing, that I discovered the MACE team had released multiple new versions since my initial installation in September. Curious about potential improvements, I trained a new model using their "Multihead Replay Fine-Tuning" feature. This method incorporates not only the user-provided training data but also selected data from the foundational model's original training set. As a result, the trained model ran for 10,000 MD steps without a single hydrogen atom escaping; an outcome unattainable in my previous experiments without this feature. The core issue had been "catastrophic forgetting": standard training enabled the model to learn the PES of the new dataset but caused it to lose key physical information embedded in the foundational model.

Although this section does not feature extensive graphs, its insights remain valuable. I successfully demonstrated that MACE models can be used for MD simulations, but enabling multihead replay fine-tuning is essential. Secondly, I developed automation scripts that streamline MACE training across CPUs and GPUs. This is an important contribution since MACE does not natively support this functionality. My GitHub repository contains three active learning scripts:

- **Standard Deviation Active Learning:** The script `run_gpu.sh` executes MACE training on a SLURM GPU node while `run_cpu.sh` extracts forces and runs CP2K-based DFT calculations on a CPU node. The MD process runs until a predefined standard deviation threshold is met, at which point retraining and model updates are triggered automatically.
- **Fixed-Step Active Learning**: Instead of relying on the standard deviation threshold, this approach increments MD runtime as model accuracy improves. The number of MD steps follows $N = 250\sqrt{i}$, where i is the active learning iteration.
- **Random perturbations**: By applying random perturbations to structures and calculating forces and energies via DFT in CP2k, the script `rattle.py` expands the explored PES space. It can function independently or complement the previous methods.

4.3. Thermodynamic Reweighting

4.3.1. 1D visualization

In the next step of my analysis, I utilized MACE models to perform thermodynamic reweighting on the Zr-H and Si-H bond lengths of the PA, maximum strain, 353K aiMD full trajectory. I chose this specific trajectory because it differs in all key aspects (adsorption mode, strain, temperature) from the one I trained my models on, namely the CA, average strain, 773K aiMD reduced trajectory. I first examined the influence of reweighting on one-dimensional bond length distributions (Fig. 4.13a-f), followed by an analysis of reweighted KDE plots (Fig. 4.13g-i and Fig. 4.15) to detect differences in the two-dimensional bond length relationships.

A closer inspection of the Zr-H bond length distributions in Fig. 4.13a-c reveals that both distributions maintain a similar shape, with a peak centered at approximately 1.82 Å. However, the reweighted trajectory shows noticeably lower kurtosis compared to the original. This effect likely originates from the inclusion of exact exchange in PBE0+D3, which better localizes electrons compared to PBE+D3. As a result, the electron density becomes more concentrated within bonding regions, leading to stronger and shorter bonds and thereby reducing kurtosis. Interestingly, Fig. 4.13c shows that distributions reweighted by BLYP+D3 and TPSS+D3 models are almost identical to the PBE0+D3 reweighted distribution, all of which display sharper peaks compared to the original PBE+D3 distribution. This observation still aligns with the hypothesis, as BLYP+D3, TPSS+D3, and PBE0+D3 all improve exchange effects relative to PBE, leading to superior electron localization.

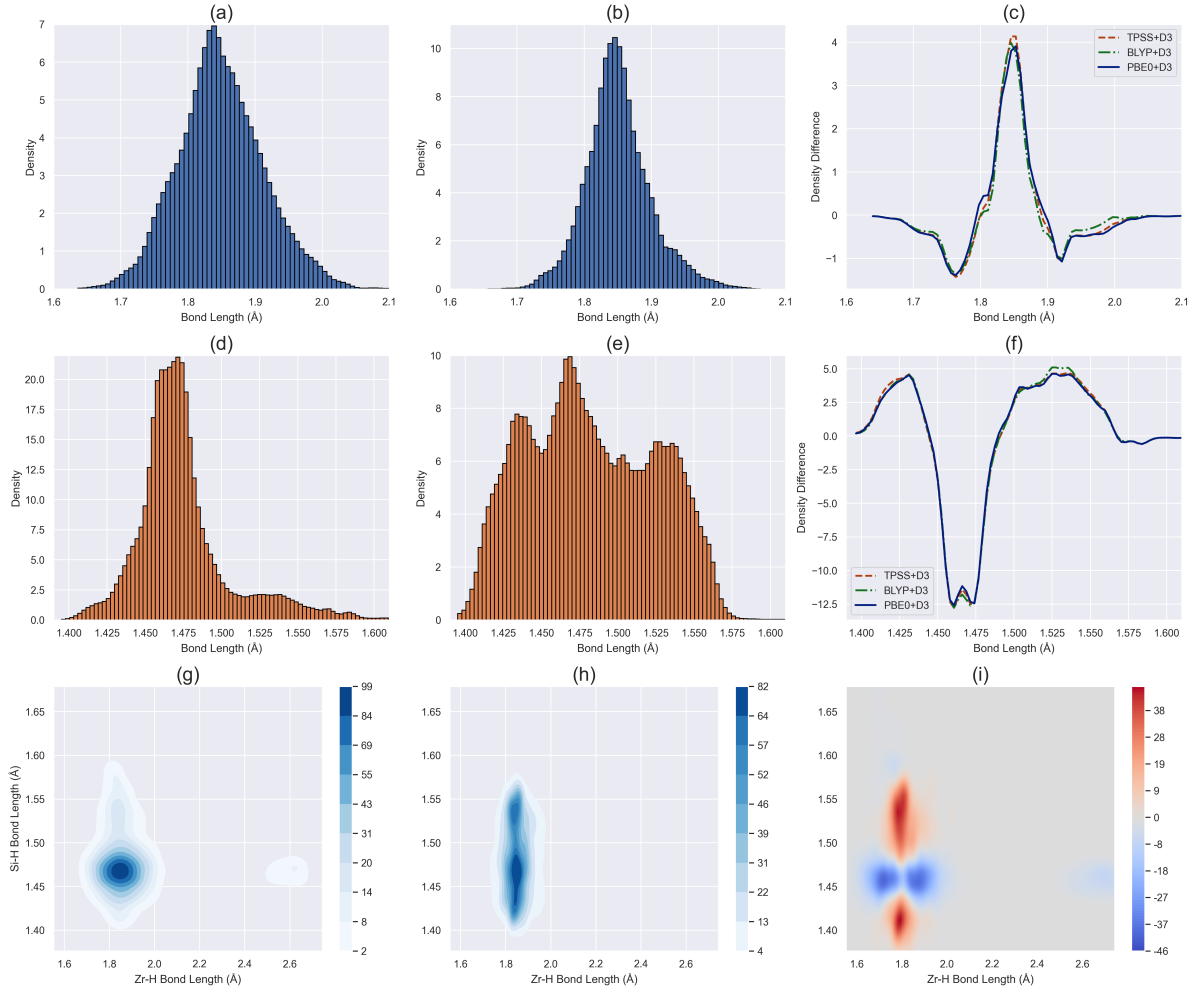

Figure 4.13: Visualization of thermodynamic reweighting applied to the PA, maximum strain, 353K aiMD full trajectory. Panel (a) shows the original Zr-H bond length distribution, panel (b) the reweighted distribution obtained using a MACE model trained on the PBE0+D3 CA, average strain, 773K aiMD reduced trajectory, and panel (c) the difference between the original and reweighted distributions for functionals BLYP+D3, TPSS+D3, PBE0+D3. Panels (d), (e), and (f) repeat this analysis for Si-H bond lengths. The distributions in these plots are slightly smoothed with a Gaussian filter of σ=1. Panel (g) shows a KDE plot of Si-H versus Zr-H bond lengths, panel (h) the reweighted version, and panel (i) the difference between the two KDE plots.

Whereas the Zr-H reweighted distribution retains a similar shape to the original, the Si-H reweighted distribution changes significantly (see Fig. 4.13d-f). The new distribution now shows three distinct peaks at 1.43 Å, 1.47 Å, and 1.53 Å, with the central peak being slightly higher. The original distribution, however, contained a single dominant peak at 1.47 Å and a minor bump at 1.53 Å. These pronounced differences make it difficult to maintain the hypothesis of improved exchange effects. Instead, they point towards a significant energy discrepancy, likely leading to high weights as specified in Eq. 3.2.

There are two potential sources of such high weights: the PBE+D3 DFT energy in the equation or the predicted energy from the MACE model. Although it would be natural to suspect an error in the latter, Fig. 3.3 shows considerable variance in the PBE+D3 energies, often deviating significantly from the PBE0+D3 parity line for lower energies. On the other hand, the MACE models in Fig. 4.14 show a slope of approximately 0.7, indicating a systematic error. However, the model error is not uniform, as TPSS+D3 and BLYP+D3 exhibit greater inaccuracies at high energies, whereas PBE0+D3 tends to perform worse at lower energies. Considering the reweighted distributions for BLYP+D3, TPSS+D3, and PBE0+D3 are nearly superimposed (Fig. 4.13), it is reasonable to conclude that the primary source of energy differences lies not within the MACE models, but rather in the PBE+D3 data variability. The MACE models likely adjust the distribution to correct for these high-variance data points.

Still, the slope of the energy predictions for all MACE models remains below one, indicating a systematic error for the highest and lowest energy predictions. This gives rise to a flatter energy landscape with lower highs and higher lows, resulting in a more homogeneous bond length distribution. Additionally, some functionals, particularly TPSS, show significant uncertainty, as indicated by the standard deviation between models in Fig. 4.14. Accounting for and propagating this uncertainty during reweighting would likely lead to an even more homogeneous distribution. As a result, the final reweighted distribution becomes a convolution of the adjusted high-variance PBE+D3 data and the inherently homogenized predictions from the MACE models.

One question remains, however: *why does the Zr-H distribution narrow slightly while the Si-H distribution broadens significantly when reweighted?* A closer look at the initial bond distributions reveals that the Zr-H distribution is naturally wide, ranging from 1.7 to 2.0 Å, indicating more flexible bonding. Conversely, the Si-H distribution is comparatively sharp, spanning from 1.4 to 1.50/1.55 Å, which points to tighter bonds that are closer to equilibrium. The broadening of the Si-H distribution can be attributed to the adjusted high-variance PBE+D3 data, while MACE's tendency to overpredict low-energy configurations further contributes to this effect. Enhanced localization from PBE0+D3 functionals does not play a role here, as the Si-H distribution is already quite compact. Therefore, the observed broadening likely stems from a combination of the former two factors. In contrast, the initially broader Zr-H distribution narrows slightly, driven not only by improved exchange effects but also by MACE's systematic errors that truncate high-energy tails. Since the Si-H distribution is twice as tight, this truncation effect has a reduced impact, making the narrowing more evident in the Zr-H distribution. While the adjusted high-variance PBE+D3 data and MACE's overprediction of low-energy configurations still influence the Zr-H distribution, their effects are comparatively weaker, resulting in a tighter distribution.

The last row of Fig. 4.13 effectively condenses the information from the upper two rows. Fig. 4.13g shows the original Zr-H and Si-H distributions as a KDE plot, while Fig. 4.13h applies the reweighting to the same data. This plot clearly illustrates the broadening of the Si-H distribution and the slight sharpening of the Zr-H distribution. The difference plot in Fig. 4.13i further emphasizes these effects, with blue areas indicating a sharper Zr-H distribution and red areas pointing to a broader Si-H distribution.

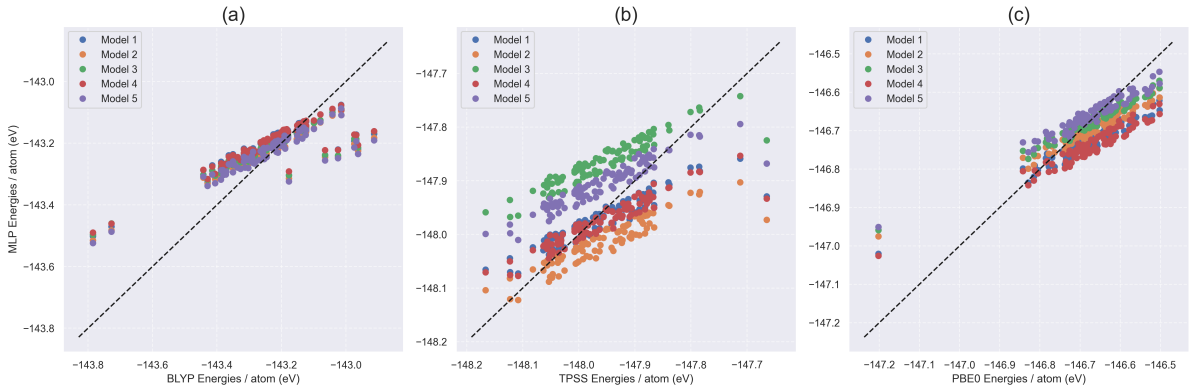

Figure 4.14: Predictions of MACE models trained on DFT data from the CA, average strain, 773K aiMD reduced trajectory for different DFT functionals, evaluated on an out-of-sample test set. The x-axis corresponds to the DFT energy for functionals (a) BLYP+D3, (b) TPSS+D3, and (c) PBE0+D3, while the y-axis shows the MACE predictions.

4.3.2. 2D visualization

While Figs. 4.7 and 4.8 demonstrate satisfactory predictions for most trajectories, I specifically chose the CA, average strain, 773K full trajectory to minimize potential errors arising from different trajectory conditions. This trajectory contains 30,000 data points, 600 of which were used for training my models.

When examining Fig. 4.15, certain characteristics appear consistently across all panels, while others differ among subplots. Each distribution exhibits three regions of higher density, but the spot corresponding to a Si-H bond length of 1.45 Å and a Zr-H bond length of 2.6 Å mostly disappears in the reweighted distributions. Instead, the reweighted plots split the peak at an Si-H bond length of 2.0 Å into two separate spots based on Zr-H bond lengths at 1.75 Å and 2.05 Å. This separation suggests that the original dataset merged multiple equilibrium positions into a single peak, and reweighting helps separate them based on their energetic and configurational differences.

It is interesting to see how all the reweighted distributions prioritize the region corresponding to an Si-H bond length of 1.5 A and a Zr-H bond length of 2.0 A, apart from the PBE0 reweighted trajectory, which prioritizes the spot closest to the original maximum, at Si-H length 2.0 and Zr-H length 1.75 A. This is probably due to PBE0's hybrid nature, which mixes in exact exchange, shifting the balance between different local minima and leading to a different reweighting pattern.

However, what stands out most in this plot is how the size and shape of the spot at a Si-H bond length of 1.5 Å and a Zr-H bond length of 2.0 Å change with the choice of functional. Whereas LDA produces a broad distribution in both bond length directions, PBE and BLYP+D3 appear to constrain the Zr-H dimension, TPSS+D3 further narrows the Si-H spread, and PBE0+D3 provides the clearest, most localized estimate of the true equilibrium bond length, as expected. This indicates that the MACE models are indeed different and they actually learn the features available in the original DFT functional dataset.

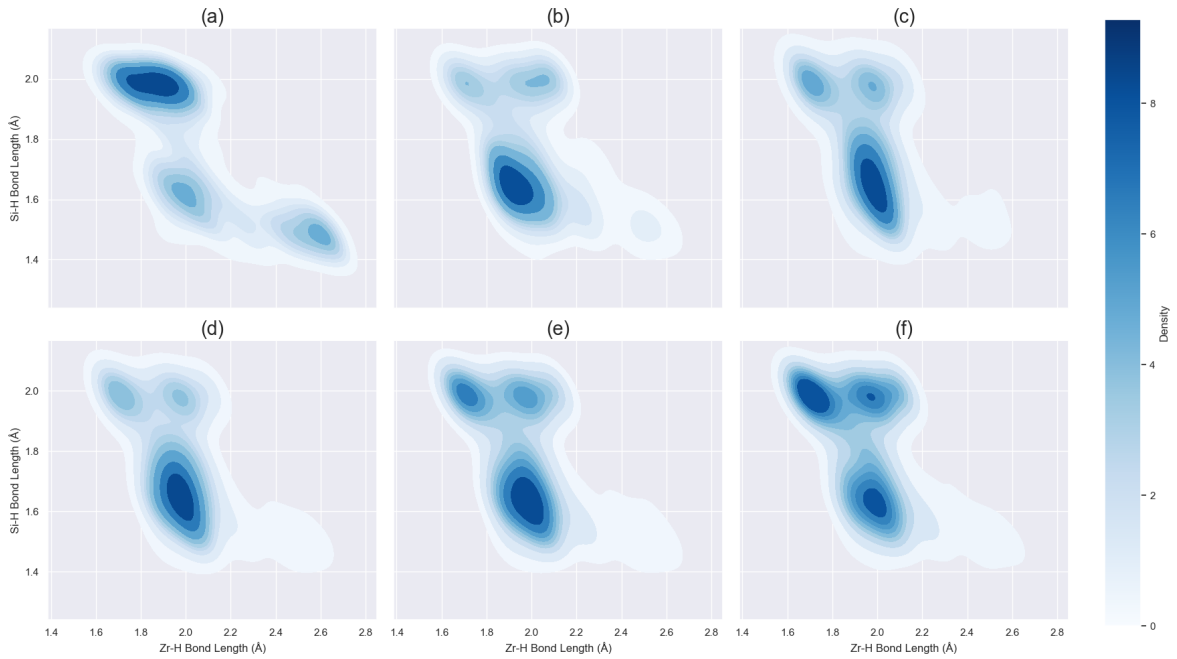

Figure 4.15: KDE visualization of reweighting applied to Si-H and Zr-H bond length distributions from the CA, average strain, 773K aiMD full trajectory. Panel (a) displays the original distribution, while panels (b) through (f) show the distribution reweighted by models trained on the CA, average strain, 773K reduced trajectory using different DFT functionals, respectively (b) LDA, (c) PBE, (d) BLYP+D3, (e) TPSS+D3, and (f) PBE0+D3.

5

Conclusion

The results from this study reveal both strengths and limitations of MACE MLP-models when applied to a challenging catalytic system: zirconocene hydride grafted onto an amorphous silica slab model. The parity plots for energy and force predictions consistently show low RMSE across all trajectories, with force errors remaining below 0.2 eV/Å and data points aligning closely with the parity line. The energy RMSE is also reliably below 0.05 eV/atom, but the sub-unity slope hints at systematic deviations, particularly at the energy range boundaries.

One key conclusion from this study is the minimal data and training requirements needed to achieve accurate predictions. A model trained on just one training structure and one validation structure for 200 epochs (i.e. less than three minutes of training) already attains an RMSE of 0.05 eV/atom on an out-of-sample test dataset. While adding more training structures further enhances the model's performance, this remarkable efficiency highlights the inherent robustness of the foundational model, significantly reducing GPU training time.

Despite accurately predicting forces and carefully set up active learning loops, the models struggled to maintain a physically stable reaction for more than a few hundred iterations, with hydrogen atoms escaping uncontrollably. This instability results from catastrophic forgetting, where the model, when trained on a new dataset, effectively overwrites fundamental physical knowledge embedded in the foundational model. The novel multihead replay fine-tuning technique has shown promising improvements in maintaining stability, but further experiments are needed to validate the physical accuracy of these visually stable outcomes.

Thermodynamic reweighting has demonstrated its potential to refine bond length distributions, though the extent of improvement varies with the functional used. The consistent narrowing of the Zr-H distribution across trajectories suggests that models trained on hybrid functionals data like PBE0+D3 effectively mitigate electron delocalization typically associated with pure GGA methods, although the model's systematic error also plays a role. Additionally, the KDE visualization of bond length relationships shows how reweighting enhances the separation of overlapping equilibrium configurations, resulting in clearer and more interpretable bond dynamics.

Nevertheless, the robustness of the reweighting approach depends heavily on the accuracy of the MACE models themselves. Erroneous predictions can distort the reweighted distributions, emphasizing the need for comprehensive model validation before application. Due to the limited availability of non-PBE+D3 data in this study, a direct comparison between reweighted distributions and actual DFT functional distributions was not feasible. Future work should focus on a more detailed assessment of MACE model accuracy and systematically compare reweighted distributions with those derived from high-accuracy DFT calculations, such as PBE0+D3.

In short, MACE models demonstrate great potential for accurate and efficient energy and force predictions. Their ability to reweight existing lower-level DFT data makes them highly practical in many scenarios. However, caution is warranted when using MACE for dynamic simulations or tasks that require exceptionally high accuracy. These results reinforce the idea that MLPs serve as a robust intermediary between classical force fields and advanced ab initio methods.

References

[1] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Physical Review*, 136(3B):B864–B871, November 1964. Publisher: American Physical Society.

[2] J. K. Nørskov, M. Scheffler, and H. Toulhoat. Density Functional Theory in Surface Science and Heterogeneous Catalysis. *MRS Bulletin*, 31(9):669–674, September 2006. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 9 Publisher: Springer International Publishing.

[3] David S. Sholl and Janice A. Steckel. *Density functional theory: a practical introduction*. Wiley, Hoboken, N.J, 2009.

[4] Valeria Butera. Density functional theory methods applied to homogeneous and heterogeneous catalysis: a short review and a practical user guide. *Physical Chemistry Chemical Physics*, 26(10):7950–7970, 2024. Publisher: Royal Society of Chemistry.

[5] Yunkai Zhou, Yousef Saad, Murilo L. Tiago, and James R. Chelikowsky. Self-consistent-field calculations using Chebyshev-filtered subspace iteration. *Journal of Computational Physics*, 219(1):172–184, November 2006.

[6] Seokhyun Choung, Wongyu Park, Jinuk Moon, and Jeong Woo Han. Rise of machine learning potentials in heterogeneous catalysis: Developments, applications, and prospects. *Chemical Engineering Journal*, 494:152757, August 2024.

[7] S. Lifson and A. Warshel. Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules. *The Journal of Chemical Physics*, 49(11):5116–5129, December 1968.

[8] Jon R. Maple, Uri Dinur, and Arnold T. Hagler. Derivation of force fields for molecular mechanics and dynamics from ab initio energy surfaces. *Proceedings of the National Academy of Sciences*, 85(15):5350–5354, August 1988. Publisher: Proceedings of the National Academy of Sciences.

[9] Luca Monticelli and D. Peter Tieleman. Force Fields for Classical Molecular Dynamics. In *Biomolecular Simulations*, pages 197–213. Humana Press, Totowa, NJ, 2013. ISSN: 1940-6029.

[10] T. Xie and J.C. Grossman. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, 120(14), 2018.

[11] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional Message Passing for Molecular Graphs, April 2022. arXiv:2003.03123 [physics, stat].

[12] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. GemNet: Universal Directional Graph Neural Networks for Molecules. In *Advances in Neural Information Processing Systems*, volume 34, pages 6790–6802. Curran Associates, Inc., 2021.

[13] Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields, January 2023. arXiv:2206.07697 [cond-mat, physics:physics, stat].

[14] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nature Communications*, 13(1):2453, May 2022. arXiv:2101.03164 [physics].

[15] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Samuel M. Blau, Vlad Cărare, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Fabio Falcioni, Edvin Fako, Andrea C. Ferrari, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Christian Holm, Jad Jaafar, Stephan Hofmann, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn, Christopher Sutton, Thomas D. Swinburne, Jules Tilly,

Cas van der Oord, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry, March 2024. arXiv:2401.00096 [cond-mat, physics:physics].

[16] X. Cheng, C. Wu, J. Xu, Y. Han, W. Xie, and P. Hu. Leveraging Machine Learning Potentials for In-Situ Searching of Active sites in Heterogeneous Catalysis. *Precision Chemistry*, 2(11):570–586, 2024.

[17] Brook Wander, Joseph Musielewicz, Raffaele Cheula, and John R. Kitchin. Accessing Numerical Energy Hessians with Graph Neural Network Potentials and Their Application in Heterogeneous Catalysis. *The Journal of Physical Chemistry C*, 129(7):3510–3521, February 2025. Publisher: American Chemical Society.

[18] Amir Omranpour, Jan Elsner, K. Nikolas Lausch, and Jörg Behler. Machine Learning Potentials for Heterogeneous Catalysis. *ACS Catalysis*, January 2025. Publisher: American Chemical Society.

[19] Alan M. Ferrenberg and Robert H. Swendsen. New Monte Carlo technique for studying phase transitions. *Physical Review Letters*, 61(23):2635–2638, December 1988.

[20] Karl Ziegler. Aluminium-organische Synthese im Bereich olefinischer Kohlenwasserstoffe. *Angewandte Chemie*, 64(12):323–329, 1952. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ange.19520641202.

[21] G. Natta, P. Pino, P. Corradini, F. Danusso, G. Moraglio, E. Mantica, and G. Mazzanti. Crystalline high polymers of -olefins. *Journal of the American Chemical Society*, 77(6):1708–1710, 1955.

[22] Vincenzo Busico. Giulio Natta and the Development of Stereoselective Propene Polymerization. In Walter Kaminsky, editor, *Polyolefins: 50 years after Ziegler and Natta I: Polyethylene and Polypropylene*, pages 37–57. Springer, Berlin, Heidelberg, 2013.

[23] Damien B. Culver, Rick W. Dorn, Amrit Venkatesh, Jittima Meeprasert, Aaron J. Rossini, Evgeny A. Pidko, Andrew S. Lipton, Graham R. Lief, and Matthew P. Conley. Active Sites in a Heterogeneous Organometallic Catalyst for the Polymerization of Ethylene. *ACS Central Science*, 7(7):1225–1231, July 2021. Publisher: American Chemical Society.

[24] Christophe Copéret, Aleix Comas-Vives, Matthew P. Conley, Deven P. Estes, Alexey Fedorov, Victor Mougel, Haruki Nagae, Francisco Núñez-Zarur, and Pavel A. Zhizhko. Surface Organometallic and Coordination Chemistry toward Single-Site Heterogeneous Catalysts: Strategies, Methods, Structures, and Activities. *Chemical Reviews*, 116(2):323–421, January 2016. Publisher: American Chemical Society.

[25] Richard Zallen. *The Physics of Amorphous Solids*. Wiley, Mörlenbach, 2008.

[26] John G. Howell, Yi-Pei Li, and Alexis T. Bell. Propene Metathesis over Supported Tungsten Oxide Catalysts: A Study of Active Site Formation. *ACS Catalysis*, 6(11):7728–7738, November 2016. Publisher: American Chemical Society.

[27] Pubudu N. Wimalasiri, Nuong P. Nguyen, Hasini S. Senanayake, Brian B. Laird, and Ward H. Thompson. Amorphous Silica Slab Models with Variable Surface Roughness and Silanol Density for Use in Simulations of Dynamics and Catalysis. *The Journal of Physical Chemistry C*, 125(42):23418–23434, October 2021. Publisher: American Chemical Society.

[28] Craig Vandervelden, Amy Jystad, Baron Peters, and Marco Caricato. Predicted Properties of Active Catalyst Sites on Amorphous Silica: Impact of Silica Preoptimization Protocol. *Industrial & Engineering Chemistry Research*, 60(35):12834–12846, September 2021. Publisher: American Chemical Society.

[29] Bryan R. Goldsmith, Baron Peters, J. Karl Johnson, Bruce C. Gates, and Susannah L. Scott. Beyond Ordered Materials: Understanding Catalytic Sites on Amorphous Solids. *ACS Catalysis*, 7(11):7543–7557, November 2017. Publisher: American Chemical Society.

[30] Baron Peters and Susannah L. Scott. Single atom catalysts on amorphous supports: A quenched disorder perspective. *The Journal of Chemical Physics*, 142(10):104708, March 2015.

[31] Jarosław Handzlik, Robert Grybos, and Frederik Tielens. Structure of Monomeric Chromium(VI) Oxide Species Supported on Silica: Periodic and Cluster DFT Studies. *The Journal of Physical Chemistry C*, 117(16):8138–8149, April 2013. Publisher: American Chemical Society.

[32] Christopher S. Ewing, Abhishek Bagusetty, Evan G. Patriarca, Daniel S. Lambrecht, Götz Veser, and J. Karl Johnson. Impact of Support Interactions for Single-Atom Molybdenum Catalysts on Amorphous Silica. *Industrial & Engineering Chemistry Research*, 55(48):12350–12357, December 2016. Publisher: American Chemical Society.

[33] Maciej Gierada and Jarosław Handzlik. Active sites formation and their transformations during ethylene polymerization by the Phillips CrOx/SiO2 catalyst. *Journal of Catalysis*, 352:314–328, August 2017.

[34] Christopher S. Ewing, Saurabh Bhavsar, Götz Veser, Joseph J. McCarthy, and J. Karl Johnson. Accurate amorphous silica surface models from first-principles thermodynamics of surface dehydroxylation. *Langmuir: the ACS journal of surfaces and colloids*, 30(18):5133–5141, May 2014.

[35] Nuong P. Nguyen and Brian B. Laird. Generation of Amorphous Silica Surfaces with Controlled Roughness. *The Journal of Physical Chemistry A*, 127(46):9831–9841, November 2023. Publisher: American Chemical Society.

[36] Gil M. Repa and Lisa A. Fredin. Predicting Electronic Structure of Realistic Amorphous Surfaces. *Advanced Theory and Simulations*, 6(11):2300292, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adts.202300292.

[37] Yusuke Nomura and Ryosuke Akashi. Density functional theory, November 2022. arXiv:2210.07647 [cond-mat].

[38] Thomas D. Kühne, Marcella Iannuzzi, Mauro Del Ben, Vladimir V. Rybkin, Patrick Seewald, Frederick Stein, Teodoro Laino, Rustam Z. Khaliullin, Ole Schütt, Florian Schiffmann, Dorothea Golze, Jan Wilhelm, Sergey Chulkov, Mohammad Hossein Bani-Hashemian, Valéry Weber, Urban Borštnik, Mathieu Taillefumier, Alice Shoshana Jakobovits, Alfio Lazzaro, Hans Pabst, Tiziano Müller, Robert Schade, Manuel Guidon, Samuel Andermatt, Nico Holmberg, Gregory K. Schenter, Anna Hehn, Augustin Bussy, Fabian Belleflamme, Gloria Tabacchi, Andreas Glöß, Michael Lass, Iain Bethune, Christopher J. Mundy, Christian Plessl, Matt Watkins, Joost VandeVondele, Matthias Krack, and Jürg Hutter. CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics*, 152(19):194103, May 2020.

[39] Thomas D. Kühne. Ab-Initio Molecular Dynamics. *WIREs Computational Molecular Science*, 4(4):391–406, July 2014. arXiv:1201.5945 [cond-mat, physics:physics].

[40] John P. Perdew. Jacob's ladder of density functional approximations for the exchange-correlation energy. In *AIP Conference Proceedings*, volume 577, pages 1–20, Antwerp (Belgium), 2001. AIP. ISSN: 0094243X.

[41] Sérgio Sousa, Pedro Fernandes, and Maria Ramos. General Performance of Density Functionals. *The journal of physical chemistry. A*, 111:10439–52, November 2007.

[42] J. C. Slater and James C. Phillips. *Quantum Theory of Molecules and Solids Vol. 4: The Self-Consistent Field for Molecules and Solids. Physics Today*, 27(12):49–50, December 1974.

[43] John P. Perdew and Wang Yue. Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation. *Physical Review B*, 33(12):8800–8802, June 1986.

[44] S. H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of Physics*, 58(8):1200–1211, August 1980.

[45] J. P. Perdew and Alex Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Physical Review B*, 23(10):5048–5079, May 1981. Publisher: American Physical Society.

[46] S. Goedecker, M. Teter, and J. Hutter. Separable dual-space Gaussian pseudopotentials. *Physical Review B*, 54(3):1703–1710, July 1996.

[47] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77(18):3865–3868, October 1996.

[48] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38(6):3098–3100, September 1988. Publisher: American Physical Society.

[49] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, January 1988.

[50] Jianmin Tao, John P. Perdew, Viktor N. Staroverov, and Gustavo E. Scuseria. Climbing the Density Functional Ladder: Nonempirical Meta–Generalized Gradient Approximation Designed for Molecules and Solids. *Physical Review Letters*, 91(14):146401, September 2003. Publisher: American Physical Society.

[51] Jianwei Sun, Adrienn Ruzsinszky, and John P. Perdew. Strongly Constrained and Appropriately Normed Semilocal Density Functional, June 2015. arXiv:1504.03028 [cond-mat].

[52] Yan Zhao and Donald G. Truhlar. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *The Journal of Chemical Physics*, 125(19):194101, November 2006.

[53] Axel D. Becke. Density-functional thermochemistry. IV. A new dynamical correlation functional and implications for exact-exchange mixing. *The Journal of Chemical Physics*, 104(3):1040–1046, January 1996.

[54] John P. Perdew, Matthias Ernzerhof, and Kieron Burke. Rationale for mixing exact exchange with density functional approximations. *The Journal of Chemical Physics*, 105(22):9982–9985, December 1996.

[55] John P. Perdew and Yue Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Physical Review B*, 45(23):13244–13249, June 1992. Publisher: American Physical Society.

[56] E. J. Baerends and O. V. Gritsenko. A Quantum Chemical View of Density Functional Theory. *The Journal of Physical Chemistry A*, 101(30):5383–5403, July 1997. Publisher: American Chemical Society.

[57] Joseph B. Krieger, Jiqiang Chen, Gerald J. Iafrate, and Andreas Savin. Construction of An Accurate Self-interaction-corrected Correlation Energy Functional Based on An Electron Gas with A Gap. In A. Gonis, N. Kioussis, and M. Ciftan, editors, *Electron Correlations and Materials Properties*, pages 463–477. Springer US, Boston, MA, 1999.

[58] Yan Zhao, Nathan E. Schultz, and Donald G. Truhlar. Design of Density Functionals by Combining the Method of Constraint Satisfaction with Parametrization for Thermochemistry, Thermochemical Kinetics, and Noncovalent Interactions. *Journal of Chemical Theory and Computation*, 2(2):364–382, March 2006. Publisher: American Chemical Society.

[59] Marcus A. Neumann and Marc-Antoine Perrin. Energy ranking of molecular crystals using density functional theory calculations and an empirical van der waals correction. *The Journal of Physical Chemistry. B*, 109(32):15531–15541, August 2005.

[60] Stefan Grimme. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry*, 27(15):1787–1799, 2006. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20495.

[61] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. 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):154104, April 2010.

[62] B. M. Axilrod and E. Teller. Interaction of the van der Waals Type Between Three Atoms. *The Journal of Chemical Physics*, 11(6):299–300, June 1943.

[63] Axel D. Becke and Erin R. Johnson. A simple effective potential for exchange. *The Journal of Chemical Physics*, 124(22):221101, June 2006.

[64] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry*, 32(7):1456–1465, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21759.

[65] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[66] Muhammed Shuaibi, Adeesh Kolluru, Abhishek Das, Aditya Grover, Anuroop Sriram, Zachary Ulissi, and C. Lawrence Zitnick. Rotation Invariant Graph Neural Networks using Spin Convolutions, June 2021. arXiv:2106.09575 [cs].

[67] C Lawrence Zitnick, Abhishek Das, Adeesh Kolluru, Janice Lan, Muhammed Shuaibi, Anuroop Sriram, Zachary Ulissi, and Brandon Wood. Spherical Channels for Modeling Atomic Interactions.

[68] Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations, March 2024. arXiv:2306.12059 [physics].

[69] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, July 2013.

[70] Chaitanya K. Joshi, Cristian Bodnar, Simon V. Mathis, Taco Cohen, and Pietro Liò. On the Expressive Power of Geometric Graph Neural Networks, March 2024. arXiv:2301.09308 [cs, math, stat].

[71] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks?, February 2019. arXiv:1810.00826 [cs, stat].

[72] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, February 1977.

[73] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.540130812.

[74] Andrew L. Ferguson. BayesWHAM: A Bayesian approach for free energy estimation, reweighting, and uncertainty quantification in the weighted histogram analysis method. *Journal of Computational Chemistry*, 38(18):1583–1605, July 2017.

[75] Graeme Henkelman and Hannes Jónsson. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *The Journal of Chemical Physics*, 111(15):7010–7022, October 1999.

[76] Shantanu Roy, Waldemar Hellmann, and Stefan Goedecker. A Bell-Evans-Polanyi principle for molecular dynamics trajectories and its implications for global optimization. *Physical Review E*, 77(5):056707, May 2008. arXiv:0705.0838 [physics].

[77] Sandro E. Schönborn, Stefan Goedecker, Shantanu Roy, and Artem R. Oganov. The performance of minima hopping and evolutionary algorithms for cluster structure prediction. *The Journal of Chemical Physics*, 130(14):144108, April 2009.