EPFL

Thèse n° 9710

Modelling of metal alloys in realistic conditions with machine learning

Présentée le 15 juin 2023

Faculté des sciences et techniques de l'ingénieur
Laboratoire de science computationnelle et modélisation
Programme doctoral en science et génie des matériaux

pour l'obtention du grade de Docteur ès Sciences

par

Nataliya LOPANITSYNA

Acceptée sur proposition du jury

Prof. J. Brugger, président du jury
Prof. M. Ceriotti, directeur de thèse
Prof. J. Kitchin, rapporteur
Prof. B. Grabowski, rapporteur
Prof. A. R. Natarajan, rapporteur

École polytechnique fédérale de Lausanne

2023

Acknowledgements

I want to express my sincere gratitude to all those who have supported me throughout my PhD journey. First and foremost, I extend my appreciation to my supervisor, Michele Ceriotti, whose invaluable guidance and encouragement have been instrumental in shaping my research approach. His exceptional expertise and unyielding scientific enthusiasm have left an indelible mark on my research pursuits.

I am deeply grateful to my colleagues, collaborators, and friends, Dr. Guillaume Fraux and Dr. Ben Mahmoud, whose contributions to our shared research goals were indispensable. Collaborating with them has been a true privilege.

I would like to convey my heartfelt thanks to all the members of the Laboratory of Computational Science and Modelling (COSMO), who have not only been my professional colleagues and my closest friends. I am particularly delighted to have shared my office with Chiheb and Edoardo, who brightened my days and borne with my plants and singing. Former COSMO members Piero, Daniele, Venkat, Felix and Andrea G made my move to a new country smooth and comfortable. I express my gratitude to Giulio for introducing me to the fascinating world of symmetry functions. I want to thank Kevin R for sharing his experience and being available to my questions and doubts. I appreciate Rocio for maintaining our connections beyond the lab and her invaluable advice and support. My appreciation extends to Alex, Philip, Raymond, Chiheb, Matthias, Joe, Divya and Victor for enhancing my last year's PhD experience with their refined taste in coffee culture, food, and music. I thank Filippo, whose dedication to running became a constant source of inspiration for me. I will never forget the fun-filled days with my fellow countrymen, Arslan and Sergey. I am grateful to Federico, Davide, and Lorenzo for providing constant support, enriching my understanding of physics and Italian culture. To all the current members of COSMO, Kevin, Wei Bin, Jigyasa, Hanna, and Yaolong, thank you for being an integral part of this journey. I sincerely thank Anne Roy and Isabel Nzazi for ensuring a seamless and enjoyable lab experience.

My heartfelt gratitude goes to my friends outside the lab, Linnea, Anya, Polina, Anton, Dima,

and Denis, for the precious time spent together and their camaraderie. Also, Liza, Borya, Maks, Tanya, and Danya, your enduring friendship and the memories we created together have meant the world to me.
In this journey, the presence of Andrea by my side has been invaluable. I am deeply grateful for his support, for providing strength and stability throughout life's challenges, and for enriching this experience with his love and encouragement.
Finally, I am eternally grateful to my family—my mom, Oksana, Nastya, Natasha, and Marina—for their steadfast faith and unconditional support. Their encouragement and unwavering love have been the bedrock upon which this milestone has been achieved.

Lausanne, 09 May 2023 N.L.

Abstract

Computer simulations based on statistical methods have emerged as a powerful tool for studying structure-property relationships at the atomistic level. However, to provide reliable insights into materials in realistic conditions, it is essential to accurately describe their behaviour at finite temperatures. While ab initio calculations offer the flexibility to study any stoichiometry and chemical complexity, their scalability and computer resource requirements limit their application to large systems and timescales. Machine learning interatomic potentials (MLIP) overcome this limitation by approximating quantum mechanical (QM) potential energy surfaces at a fraction of the cost. Despite their advantages, ML methods based on atom-centred density have been constrained to systems with 4-5 chemical elements.

This thesis aims to address these challenges by focusing on two aspects: 1) accurately describing finite temperature effects, and 2) enabling ML models based on atom densities representations to describe systems with a large number of chemical elements. To address finite temperature effects, we employ a combination of machine learning and statistical sampling methods, using elemental nickel as a prototypical material with a wide application temperature range. Our framework covers bulk, interfacial, and defect properties from 100 to 2500 K and models nuclear quantum fluctuations and electronic entropy when necessary. The presented framework is versatile and, when paired with an appropriate potential, can be readily applied to complex alloys and various material classes.

Another problem we tackle in this thesis is how to describe different chemistries with ML. We generate a dataset covering a wide range of concentrations of 25 d-block transition metals and apply a scheme to compress chemical information in lower-dimensional space. The resulting model demonstrates semi-quantitative accuracy for prototypical alloys and is stable for extrapolation. We use this model to study element segregation in an equimolar 25-element alloy, reproducing in a computational setting Cantor et al.'s experiments. Our observations are used to define data-driven Hume-Rothery rules for alloy design guidance. Furthermore, we investigate three prototypical alloys (CoCrFeMnNi, CoCrFeMoNi, and IrPdPtRhRu), determining their stability and short-range order behaviour of their constituents.

Key words: machine learning potentials, finite temperature effects, alchemical compression, high-entropy alloys, exploration of chemical space

Sommario

Le simulazioni al computer basate su metodi statistici sono diventate uno strumento potente per lo studio delle relazioni tra struttura e proprietà a livello atomistico. Tuttavia, per fornire informazioni affidabili sui materiali in condizioni realistiche, è essenziale descrivere accuratamente il loro comportamento a temperature finite. Sebbene i calcoli ab initio offrano la flessibilità di studiare qualsiasi stechiometria e complessità chimica, la loro scalabilità e il costo computazionale ne limitano l'applicabilità per sistemi di grande taglia e tempi lunghi. I potenziali interatomici machine learning (ML) superano questa limitazione approssimando, a una frazione del costo, la superficie di energia potenziale derivante da calcoli ab initio. Nonostante questi vantaggi, i metodi ML basati sulla densità atomica sono stati finora limitati a sistemi di al piú 4 o 5 elementi chimici.

Questa tesi si propone di affrontare i problemi sopraelencati concentrandosi su due aspetti: 1) descrivere accuratamente gli effetti di temperatura finita e 2) consentire ai modelli di ML basati sulle rappresentazioni delle densità degli atomi di descrivere sistemi con un gran numero di elementi chimici. Per affrontare gli effetti di temperatura finita, abbiamo utilizzato una combinazione di metodi ML e di campionamento statistico, concentrandoci sul nichel elementare quale tipico materiale caratterizzato da un ampio intervallo di temperature di utilizzo. Il nostro aproccio è in grado di trattare le proprietà del bulk, delle interfacce e dei difetti da 100 a 2500 K, includendo, ove necessario, le fluttuazioni quantistiche nucleari e l'entropia elettronica. Il metodo presentato è versatile e, se abbinato a un potenziale appropriato, può essere facilmente applicato a leghe complesse e a varie classi di materiali.

Un altro problema trattato è quello della descrizione di diverse specie chimiche in un formalismo ML. A tal fine, abbiamo generato un dataset che copre un'ampia gamma di concentrazioni per 25 metalli di transizione del blocco *d* e abbiamo applicato uno schema per comprimere le informazioni chimiche in uno spazio a bassa dimensionalità. Il modello risultante dimostra un'accuratezza semi-quantitativa ed è stabile in regime estrapolativo. Abbiamo utilizzato questo modello nello studio della segregazione in una lega equimolare di 25 elementi, riproducendo, in ambito computazionale, gli esperimenti di Cantor et al. Sfruttando

tali osservazioni, abbiamo definito regole di Hume-Rothery basate sui dati, in grado di indirizzare la produzione di nuove leghe. Inoltre, abbiamo analizzato tre leghe prototipiche (CoCrFeMnNi, CoCrFeMoNi e IrPdPtRhRu), determinandone la stabilità e l'ordine atomico a corto raggio.

Parole chiave: potenziali machine learning, effetti di temperatura finita, compressione alchemica, leghe ad alta entropia, esplorazione dello spazio chimico

Contents

List of Figures

List of Tables

1 Introduction

1.1 Computational modelling of materials at realistic conditions

Computational modelling has been used for several decades to gain a qualitative, mechanistic understanding of the atomic-scale phenomena that underlie the structure-property relations in materials[1]. Recent developments in this field have made it possible to achieve predictive accuracy for several structural, mechanical and functional properties, assisting the design and optimization of materials for both fundamental and technical applications [2–8].

To date, one of the most common approaches in computational modelling is static-lattice techniques, which are employed to determine the atomic arrangement of minimum energy for the system simulated and the corresponding elastic, energetic or functional properties. The procedure involves the evaluation of the appropriate energy (for example, the lattice energy in the simulation of an ionic crystals), which is then minimized with respect to all relevant degrees of freedom, e.g. cell dimensions and atomic coordinates in condensed phase simulations. Static-lattice methods are essentially 'zero Kelvin' calculations with no representation of thermal effects. The most primitive way of including the latter is via the 'harmonic' or 'quasi-harmonic' approximations[9, 10]. This approach has been used to compute thermodynamic and elastic properties of zirconium hydrides[11] and chalcogenides[12], to study melting curves of some elemental metals[13] etc. While it was previously believed that taking into consideration anharmonicity did not considerably increase the accuracy of the results[14], Jörg Neugebauer and co-authors have demonstrated based on first principles methods that anharmonic corrections play a crucial role in the accurate description of materials, notably metals and metal alloys [15–19].

In-depth investigation of material's properties at finite temperature becomes available with the use of molecular dynamics (MD) simulations. By resolving Newton's equations of motion for a

group of particles that represent the simulated system, MD explicitly takes into consideration the kinetic energy of the atoms. Thus, the time averages along a molecular dynamics trajectory sampling a representative statistical ensemble can be used to derive various thermodynamic observables under realistic conditions, such as tensile properties [20, 21], thermal transport properties [22–24], interface properties [25, 26]. Furthermore, MD simulations can be used in combination with the concept of thermodynamic integration [27, 28] to accurately quantify the anharmonic contribution to free energy, which is fundamental for a fair comparison to experiment [15–19].

The classical behavior of the atoms is implicitly assumed by the use of Newton's equations of motion. However, as can be observed from the thermodynamic characteristics of light metals [29], nuclear quantum effects (NQE) become more apparent at the Debye temperature and below, which, in the case of materials like lithium niobate (LiNbO3) and lithium tantalate (LiTaO3), can be significantly higher than room temperature [30]. There is a range of methods to accurately account for NQE by projecting the quantum problem onto a system with multiple classical degrees of freedom, but doing so increases substantially the computational cost. In path integral molecular dynamics (PIMD)[31–33], for instance, atoms are replaced by P beads connected by harmonic potentials, which raises the computational cost to run a simulation using the PIMD formalism P times as compared to the cost of a single trajectory and makes it challenging to combine with ab initio methods.

All of these simulation methods rest upon the underlying description of atomic interactions, representing the physics and chemistry of the system, and require means to compute accurately and efficiently the potential energy for a given configuration of atoms – the potential energy surface (PES) of the system. The PES can be defined from first principles by solving the Schrödinger equation for every nuclear configuration,for example, using an approximate form such as density-functional theory (DFT) [34, 35]. Although this approach is accurate and transferable across different chemistries, it is computationally demanding, limiting the size of simulations to ~1000 atoms and time scales to hundreds of picoseconds. As a result, large-scale and long-time simulations have historically relied on interatomic potentials (IPs), which are, in most cases, empirical parametrizations of the PES based on physically-motivated functional forms [36–42]. IPs acquire linear scaling with respect to the number of atoms at the expense of accuracy and transferability.

The advent of machine learning (ML) has advanced the field of IP's development and permitted to strike a balance between computational efficiency ab initio accuracy and transferability, introducing machine learning interatomic potentials (MLIP) based on the regression of energy and forces from reference electronic structure calculations. MLIPs have reduced consider-

ably the effort needed to thoroughly investigate the structural and mechanical properties of materials [43–45], and to evaluate the finite-temperature thermodynamics of materials with first-principles accuracy, which has made it possible, for instance, to investigate the finite-temperature mechanical properties of iron [46] to determine the subtle difference in free energy between different phases of water [47], or to study the phase diagram of hybrid perovskite materials [48].

In chapter 4, we demonstrate the combination of machine-learning potentials with thermodynamic integration and finite-temperature sampling to compute bulk and interfacial properties of materials from cryogenic temperatures up to above the melting point. We also use a recently-developed scheme to predict the electronic density of states [49] to take into account the impact of electronic excitations, without the need to perform additional electronic-structure calculations.

1.2 Interpolation of interatomic interaction using machine learning

Data-driven approaches have been widely adopted in atomistic modelling for diverse applications spanning from advanced data analytics [50] to the generative design of materials with optimal properties [51]. In this thesis, our primary focus will be on the application of machine learning techniques to accelerate electronic structure calculations: predicting and analyzing the relationship between a specific atomic configuration and property computed from first principles [52–58]. In particular, a lot of progress has been made in building MLIPs, in which machine learning models are trained on a few quantum mechanical calculations to reconstruct the PES of materials. Fitting of MLIP has been done successfully for various systems including metallic alloys [59, 60], amorphous materials[58, 61], phase change materials[62], proving it's able to capture diverse chemistries and bonding and enable modelling of systems with large degrees of freedom which were not accessible before with DFT [63]. Continuous improvements are being made to MLIP's performance and efficiency, as well as internal process optimization, feature reduction[64, 65], and data utilization[66–68]. Additional acceleration has been achieved by exploiting GPU-optimized libraries like PyTorch [69] and TensorFlow[70].

The construction of an MLIP requires three ingredients: i) a descriptive dataset, which samples the phase space of interest, ii) a representation (descriptors or fingerprints) of atomistic structures, which is communicated to the algorithm, iii) the regression algorithm itself. Each element impacts the ultimate accuracy and has been thoroughly researched, yielding a wide range of potentials and approaches to their creation. Below, we provide a brief review of the most common practices in the field. Among various algorithms used for MLIPs are linear

regression (LR), kernel ridge regression (KRR), neural networks (NN), support vector machines (SVM). Commonly used examples of the MLIPs are the high-dimensional neural network potential (NNP) [53, 71], the Gaussian approximation potential (GAP)[72–74], the spectral neighbor analysis potential (SNAP)[75–78] and moment tensor potentials (MTP)[79–81] among others[82–95].

In the realm of MLIPs, the choice of descriptors is often (but not necessarily) tied to the method employed, as developers frequently provide a comprehensive package for fitting an MLIP that includes the internally generated and optimized descriptors for a specific workflow. Despite distinct research efforts, a principle that is common to the majority of MLIP descriptors is the representation of atomic properties and positions through the transformation of Cartesian coordinates, in such a way that physically-motivated requirements like smoothness and invariance to translations, rotations, and permutations of atoms of the same type are fulfilled. These atom-density-based descriptors, such as atom-centred symmetry functions [96] and the smooth overlap of atomic positions (SOAP)[72], have a wide range of applicability, from gas-phase molecules to bulk solids, and are not limited to a specific application. These representations can be tuned to reflect other physical and chemical principles. Recent research has shown that MLIPs that use such representations, which closely reflect principles such as locality, the multiscale nature of interactions, and the similarities in the behavior of elements from the same group in the periodic table, tend to be more robust, transferable, and data-efficient[97].

Creating MLIPs requires reference data for training, typically computed from first principles. While most ML models focus on computing energies and forces, recent developments in the field have expanded the range of evaluated properties to include molecular dipoles[98, 99], polarizabilities[100, 101], electron density[102], Hamiltonians[103, 104] and others. Additionally, since the community aims for open, reproducible, and FAIR [105] research, a growing number of publicly accessible datasets based on experiments and electronic structure calculations are created, encompassing a variety of materials from molecules to inorganic materials [106–111]. However, most of these databases lack the structural variety essential for developing robust and accurate MLIP since they comprise perfect crystal structures or a single representative of configurations encountered experimentally. Since there is no universal approach for creating descriptive datasets, and it typically demands manual adjustments tailored to a particular scientific problem, numerous techniques have been proposed to enhance dataset efficiency, such as de novo structure generation [112] and high-throughput random structure searching [113]. Active learning schemes are one such approach, where the training of an MLIP is combined with a QM workflow and the data points are generated on the fly to improve the predictions in the areas of low confidence [67, 114, 115].

Despite the progress made in MLIP development, these models still fall short in their transferability and ability to generalize to different systems, which is crucial for studying and exploring the properties of materials across varying elemental compositions. One of the main challenges in this respect is the unfavourable scaling of MLIP complexity with the number of chemical components in a system. In this thesis, we use the chemical embedding optimization of descriptors introduced in a prior study [97] to solve the scaling issue with the number of chemical components. A more in-depth discussion of this issue can be found in Section 2.4.4.

1.3 Modelling of multi-component systems with machine learning interatomic potentials

Almost 20 years have passed since independent work from the groups of Yeh[116] and Cantor[117] showed that mixing up to 20 metallic elements in roughly equal parts leads to a smaller-than-expected number of distinct phases, with some corresponding to disordered solid solutions of 4-6 elements. These so-called high-entropy alloys (HEAs) have since become the subject of intense study.[118] On a fundamental level, the observation of the existence of an extended single-phase stability region for alloys with multiple principal components was surprising, and from a technological standpoint it opened up the possibility of designing new materials that defy the limitations of conventional metallurgy and alloy engineering.[119, 120]

Besides their metallurgical and mechanical applications, HEAs have been found to be promising catalysts[121, 122], especially in electrocatalysis[123–125]. They can efficiently reduce overpotentials and boost activities for, e.g., water splitting[126–136], the oxygen reduction reaction [132, 135, 137–139], or the methanol oxidation reaction[137, 140–143] while exhibiting very good stability under reaction conditions. These unusual properties are linked to their multi-elemental character, which gives rise to four core effects[144, 145]: the entropy, 'sluggish diffusion' (not observed in some alloys[146]), lattice distortion and 'cocktail effect'. While the former two enhance the stability, the latter two can explain the high activity in catalysis. First, lattice distortions occur due to atoms being surrounded by atoms of many different atomic radii leading to stress and strain. This alters the electronic structure of the alloy. For example, the water splitting activity of a family of AlNiCoIrX (X = Mo, Cr, Cu, Nb, V) is superior to IrO2 because the lattice distortion leads to shorter Ir-O bonds[131]. Second, the 'cocktail effect' describes unexpected, synergistic effects of the chosen composition. For instance, the non-noble metal HEA CoCrFeMoNi shows activity for the oxygen reduction reaction similar to that of Pt.

From the computational perspective, modelling HEAs poses a number of distinct challenges.

The presence of multiple components requires relatively large simulation cells to unveil microstructures or order-disorder behaviour, while sluggish diffusion requires long time scales and accelerated sampling techniques to overcome free-energy barriers to atom diffusion. Due to the associated computational costs of running finite temperature simulations, most studies using ab initio methods to compute properties of high-entropy alloys (HEAs), such as elastic constants and phase stabilities, are limited to 0K calculations [147, 148]. To address this limitation, machine learning interatomic potentials (MLIPs) have been employed to study phenomena in HEAs like phase transitions, melting, and dislocation dynamics through Monte Carlo (MC) or molecular dynamics (MD) simulations, requiring time and size scales beyond the capabilities of ab initio methods[149–151].

However, significant obstacles remain in applying MLIPs to computational studies of multicomponent alloys. One is that the complexity of modern machine learning models grows steeply with the number of different elements due to the unfavourable scaling of their associated feature space sizes. As a result, the computational and memory requirements to evaluate full feature vectors limit the chemistry of the system explored thus far with MLIPs to a specific combination of 4-5 components. Recently this issue has been addressed by developing "alchemical" contractions of the Smooth Overlap of Atomic Position (SOAP)[72] features[97], and by constructing iteratively contracted version of the high-order features [152], as well as by introducing tensor-reduced representation [153].

Another issue relates to the importance of sampling, as large and descriptive datasets are needed to provide a comprehensive description of the energy landscape of such diverse systems. Most of the proposed datasets available are based on known structures, and thus models trained using such datasets are only applicable to a limited configurational space. For example, the Open Catalyst Project [108] have clearly stated that previous datasets are inappropriate for their adsorption task.

In Chapter 5, we describe our strategy for tackling these challenges, detailing the construction of an MLIP for 25 d-block elements and its application to model the finite-temperature thermodynamics of HEAs.

List of publications

The list of papers resulting from the original work discussed in this thesis is shown below in chronological order of publication:

1. **Lopanitsyna, N.**, Ben Mahmoud, C., Ceriotti, M., 2021. Finite-temperature materials modelling from the quantum nuclei to the hot electrons regime. Phys. Rev. Materials 5, 043802

2. **Lopanitsyna, N.**, Fraux, G., Springer, M.A., De, S. and Ceriotti, M., 2023. Modeling high-entropy transition-metal alloys with alchemical compression. Phys. Rev. Materials 7, 045802

2 Machine learning methods in atomistic modelling

2.1 Introduction

Machine learning (ML) has transformed materials science through data-driven approaches, enabling advances such as the development of better Li-ion battery cathode materials using ML algorithms[154], the prediction of the structure of unseen proteins[155], and the improvement of synthesis conditions for novel zeolites[156], to name a few notable examples. Historically, the material design depends on experimental trials and errors, which slows down the research and discovery process at the expense of energy and source material. Data-driven computational chemistry has made it possible to speed up the sampling and analysis of complex structure-properties landscapes. This progress has pushed the limits of previously accessible time/length scales and chemical spaces, providing researchers with recommendations and actionable insights for experiments.

ML algorithms can capture complex relationships between inputs and outputs by learning from provided data samples without relying on a rigidly defined functional form. For example, in materials science, ML is commonly used to predict various material properties, such as the glass transition temperature[157], thermal conductivity[158], bulk and shear moduli[159], and band gap[160], based on chemistry and structural information. Additionally, much progress has been made in developing machine learning interatomic potentials (MLIPs) [59, 161, 162] to accelerate predicting the potential energy surfaces for a given configuration of atomic positions. While many of these applications provide deep insights in their fields of interest, they all share a similar abstract infrastructure. The fundamental workflow for ML involves four main steps: data collection (or generation), input characterization (feature engineering), model selection and model validation. This chapter will cover the aspects related to feature engineering and model selection in the context of building for atomistic machine learning

applications. The data generation, training procedures, and model evaluation will be discussed in the specific sections dedicated to each model.

ML models can be categorized based on the available input information, which falls into two main categories: supervised and unsupervised learning. In supervised learning, the training data is labelled and consists of input values (such as the structures of different materials, e.g. the unit cell definition of a perfect crystal) and their associated output values (such as materials property values e.g. melting temperature). The goal of the ML model is to derive an optimized function that can accurately predict the output values from a given specific set of input values. MLIPs are an example of a supervised learning task, where the energies (and forces) are predicted from structural inputs. To build such a model, one could use any supervised learning algorithm, such as linear regression, feed-forward neural networks (NN) [163, 164], support vector regressors (SVR)[165], and kernel ridge regression (KRR)[166], to name a few.

If the available dataset only includes input values, unsupervised learning can perform dimensionality reduction, identify patterns, and cluster similar data points to improve data or feature selection. Notable unsupervised learning models include principal component analysis (PCA)[167], CUR decomposition[65] and farthest point sampling (FPS)[168].

Designing an accurate and efficient ML model hinges on the choice of input representation communicated to the algorithm. Features represent input data and map it to properties of interest. In this chapter, we focus on atom density-based representations that map Cartesian coordinates to a feature space characterizing the atomic environments contained in them[169, 170]. In addition, we will discuss the limitations of this representation, which includes unfavourable scaling with the number of chemical elements and explore possible solutions to address these issues.

2.2 Supervised methods

A supervised learning task aims to find a function f that can explain the relationship between inputs $\mathbf{X}$ and outputs $\mathbf{y}$ using a set of examples. This function $f_{\boldsymbol{\omega}}$ belongs to some class of parametric functions $\mathscr{F}$ and it is defined by its parameters $\boldsymbol{\omega}$, which are also addressed as model parameters. The performance of the model is measured with a loss function $\mathscr{L}$. For example, for a set of input-output $(\boldsymbol{X}_i, y_i)$, the prediction of $f_{\boldsymbol{\omega}}(\boldsymbol{X}_i) = \hat{y}_i$ and the loss is $\mathscr{L}_{\boldsymbol{\omega}}(y_i, \hat{y}_i)$. Then the best model, defined by the best $\boldsymbol{\omega}^*$ is chosen by minimising the loss function $\mathscr{L}_{\omega}$ for a given dataset:

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \mathscr{L}_{\omega}(\mathbf{y}, \hat{\mathbf{y}})$$

The functional form of f can be very different depending on the employed algorithm. Below, we provide a brief overview of the approaches used in the thesis.

2.2.1 Linear regression

First, we start with a linear formulation of the problem for a scalar variable y from a vector of observations $\boldsymbol{X} \in \mathbb{R}^N$. Specifically, we want to determine the linear least-squares estimator for a given input-output pair of variables $(\boldsymbol{X}, y)$, and find coefficients $\boldsymbol{\omega^1}, \omega^0$ that yield an estimator for y in the form:

$$\widehat{y} = \boldsymbol{X}^\top \boldsymbol{\omega^1} + \omega^0 \quad \text{(estimator for } y\text{)} \tag{2.1}$$

It is customary to refer to this model as a linear regression model since the observations are being combined linearly (or a linear model is being fitted to the observations in $\boldsymbol{X}$ to estimate y). This simply corresponds to:

$$\widehat{y} = \sum_j^N \omega_j^1 X_j + \omega^0$$

in terms of the individual entries of ω and X. Therefore, this construction defines the function $f(\boldsymbol{X})$ to the choice

$$f(\boldsymbol{X}) = \boldsymbol{X}^\top \boldsymbol{\omega^1} + \omega^0$$

where the coefficients $\boldsymbol{\omega^1}, \omega^0$ are now the parameters of the model. Once the optimal coefficients $\boldsymbol{\omega}^{1*}$ and ω^{0*} are determined, evaluating $f(\boldsymbol{X})$ for a specific $\boldsymbol{X}$, will result in an estimate for y, i.e.,

$$\widehat{y} = \boldsymbol{X}^\top \boldsymbol{\omega}^{1*} + \omega^{0*} \quad \text{(estimate for } y\text{)}$$

To determine the function f and the coefficients $\boldsymbol{\omega^1}, \omega^0$, different criteria could be used as the loss function. For example, in the case of the mean-square-error criterion, the parameters of the model can be found by minimizing the mean square error (MSE) over $\boldsymbol{\omega^1}$ and ω^0. Now the loss function is $\mathscr{L}_{\boldsymbol{\omega^1}, \omega^0}(\mathrm{y}, \hat{\mathrm{y}}) = \mathbb{E}\left[(y - \hat{y})^2\right]$ and the optimal coefficients are defined by:

$$\boldsymbol{\omega}^{\mathbf{1}*}, \omega^{0*} = \underset{\boldsymbol{\omega}^{\mathbf{1}}, \omega^{0}}{\operatorname{argmin}} \mathbb{E}\left[(y - \hat{y})^2\right]$$

The solution formulated by equation 2.1 faces certain challenges when applied in practice due to ill-conditioning, redundancy, and overfitting. To address these challenges, it is common practice to introduce a so-called regularization term. This approach involves incorporating an element that penalizes the loss function values by discouraging large valued model parameters, thereby encouraging solutions that have specific desirable properties. Depending on the choice of the regularization approach, linear regression models take different names. One commonly used form of regularization is the L^2 regularization, and its adoption in a linear model is often referred to as ridge regression. In ridge regression, the cost function $\mathscr{L}_{\boldsymbol{\omega}^{\mathbf{1}}, \omega^{0}}$ is replaced by a regularized version $\mathscr{L}_{\text{ridge}}$ that includes a penalty term based on the squared norm of the model's weights vector:

$$\mathscr{L}_{\text{ridge}} = \lambda |\boldsymbol{\omega}|^2 + \frac{1}{N} \sum_{n}^{N} \left(y_n - \boldsymbol{X}_n^{\top} \boldsymbol{\omega}\right)^2 \tag{2.2}$$

Here, $\lambda > 0$ represents the regularization factor. The term $\lambda |\boldsymbol{\omega}|^2$ in equation 2.2 is referred to as a penalty term because it penalizes large values of $\boldsymbol{\omega}$. By promoting solutions with smaller norms, ridge regression helps reduce the risk of overfitting the training data. The presence of the penalty discourages solutions which allow the model parameters to follow too closely the training points. This introduces a controllable amount of error over the training data, in exchange for a more generalizable performance across unseen test points. The aforementioned balancing exercise is commonly called the bias-variance trade-off and is key to containing the overfitting tendency in presence of arbitrarily flexible (e.g. overparametrized) models. Intuitively, the addition of penalty terms ensures that small variations in the observed data do not result in significant changes in inference decisions. Another practical advantage of introducing a penalty term is that it enters the least squares (LS) solution as an eigenvalue lifting component. This aspect stabilizes the inversion problem in the LS formulation.

Linear models are widely used in machine learning due to their simplicity and interpretability. However, the linear relationship they assume between input features and output variables obviously imposes limitations on their ability to accurately model more complex relationships. This limitation has led to the increasing importance of non-linear models in recent years.

Non-linear models are essential for modelling complex relationships and achieving high accuracy in various machine-learning tasks. They can learn to project input data in a high-dimensional space and capture intricate patterns and relationships between input features

and output variables. This enables them to achieve state-of-the-art performance in various machine learning tasks such as image and speech recognition, natural language processing, and recommendation systems.

In the next section, we will delve into the workings of non-linear models and explore how they can be used to model complex relationships in materials science.

2.2.2 Neural networks

Neural networks (NN) have revolutionized the field of machine learning by enabling the modelling of complex and nonlinear relationships between inputs and outputs. At their core, neural networks are composed of simple computational units called neurons, which are organized into layers. These layers can learn increasingly complex representations of the input data, enabling neural networks to excel at tasks such as image and speech recognition, natural language processing, and time series prediction [171–173]. In this section, we will explore a widely used architecture called a feed-forward neural network.

We start again with a collection of inputs $\boldsymbol{X}$ and outputs $\boldsymbol{y}$. First, the input layer is fed with a vector of observations $\boldsymbol{X_i}$. Then it passes through several hidden layers followed by a single node output layer. The value y_i^j of a node i in a hidden layer j is given by:

$$y_i^j = f_i^j\left(b_i^j + \sum_{k=1}^{N_{j-1}} a_{k,j}^{j-1,j} \cdot y_k^{j-1}\right) \tag{2.3}$$

where N_{j-1} stands for the number of nodes in the previous layer $j-1$. The notation $a_{n_1 n_2}^{l_1 l_2}$ is used for the weights connecting node n_1 in layer l_1 with node n_2 in layer l_2 where $l_2 = l_1 + 1$ and the superscript 0 is assigned to the input layer. Additionally, there is a bias node connected to all nodes in the hidden layers and to the output node by a bias weight b_i^j, where j refers to the layer of the node and i is its number within this layer. Basically, the expression in the brackets is a linear combination of the values of the nodes in the previous layer. To add a non-linearity, activation function f_i^j is introduced, which maps the input values to a new range of output values when passing through a node. Different non-linear functions could be used as an "activation function", provided they are differentiable, a necessary condition to allow for gradient-based optimization of the network's weights. The complete analytic expression of the output estimator could be written as a set of nested activation functions acting on linear combinations of the values in the previous layer, e.g. the NN has 2 hidden layers 5 nodes each and gets three features at the input layer:

$$\hat{y_n} = f_1^3\Big(b_1^3 + \sum_{k=1}^{5} a_{k1}^{23} \cdot f_k^2\Big(b_k^2 + \sum_{j=1}^{5} a_{jk}^{12} \cdot f_j^1\Big(b_j^1 + \sum_{i=1}^{3} a_{ij}^{01} \cdot \boldsymbol{X}_n\Big)\Big)\Big) \tag{2.4}$$

To fit a feed-forward neural network to a series of training data, it is customary to calculate the derivatives of the loss with respect to the nodes' weights by using chain rule throughout its architecture. This approach takes the name of back-propagation [174], and it's a key component of training a feed-forward neural network.

2.3 Unsupervised methods

Computational materials design often involves the analysis of complex data sets that often contain a large number of variables and relationships that are difficult to extract using conventional techniques. Unsupervised methods offer a powerful approach to uncovering patterns and structures in these data sets, without requiring prior knowledge or assumptions. In this section, we will introduce three specific unsupervised methods that are frequently used in materials science: CUR decomposition, farthest-point sampling (FPS), and principal component analysis (PCA). By understanding the principles behind these techniques and their specific applications in materials science, researchers can improve their ability to analyze and interpret data and gain deeper insights into the underlying properties of materials.

2.3.1 CUR decomposition

CUR decomposition [175, 176] is a matrix factorization method used for dimensionality reduction of feature or data space, where the elements of the initial matrix are used to determine the most relevant features or samples. For example, a given matrix $\mathbf{X}$ consisting of vectors $\boldsymbol{X}_n$ can be approximated with a lower-rank matrix $\tilde{\mathbf{X}}$ constructed from the selected k columns C and rows R of $\mathbf{X}$ and $\mathbf{U}$ a $k \times k$ matrix:

$$\mathbf{X} \approx \tilde{\mathbf{X}} = \mathbf{CUR} \tag{2.5}$$

CUR decomposition is advantageous for tasks like selecting optimal feature vectors or subsets because it uses actual elements of the matrix. We will discuss how it can be used for feature selection (along columns) in this overview, but the same procedure can be applied to select the samples (along rows). First, to select columns, we compute an "importance score" for every column of the initial matrix:

$$\pi_c = \sum_{i}^{k} (v_c^i)^2, \tag{2.6}$$

where v_c is the c-th component of the right singular vector and k is the number of columns to be selected. Many CUR methods use a probabilistic approach to select features, which ensures that if there are multiple similar features, they have approximately the same probability of being selected. However, to achieve a deterministic selection process, the column with the highest score is chosen at each step, and an orthogonalization procedure is used to avoid selecting multiple similar features. Once the column with the highest score is selected, all remaining columns in X are orthogonalized with respect to it.

To reduce the number of fingerprints, the CUR decomposition method is iteratively applied by selecting the column with the highest score at each step, orthogonalizing the remaining columns to avoid selecting nearly-identical features, and re-computing the singular value decomposition (SVD) solution based on the orthogonalized matrix until all desired features are chosen to build the reduced feature matrix. The accuracy of the approximation can be computed as:

$$\epsilon = \|\mathbf{X} - \mathbf{CUR}\|_F / \|\mathbf{X}\|_F$$

The number of features to be selected can be fixed or increased until a desired threshold is met.

2.3.2 Farthest-point sampling

Farthest-point sampling (FPS) is a deterministic algorithm that selects the point that is farthest away from the previously selected points in each iteration. In other words, FPS chooses points that are as diverse as possible for the given set of inputs - allowing for uniform sampling. The algorithm is initialized by picking any first sample. To keep its deterministic nature, it is customary to start from the first sample present in the dataset. Then, after calculating a distance metric between each point in the dataset, it finds the next selected point as:

$$k = \text{argmax}\left(\min_j \left| X - X_j \right|\right)$$

where j refers to all of the vectors that have already been selected. In practice, at every iteration, one computes the distance matrix of each available point to all the samples selected thus far. By keeping, for each point, the minimum distance separating it from the selected samples, we have an indication of the degree of uniformity of the sampling at each step. To find the next selection, one simply selects the point that is at the maximum of the minimum distance vector we just calculated. It is often called min-max selection due to the operative way to obtain the next selected indices. The process continues until all the desired points have been selected.

This technique can be used for an effective feature selection or generating a diverse dataset.

2.3.3 Principal component analysis

Principal Component Analysis (PCA) is a widely used unsupervised technique that serves various purposes, such as dimensionality reduction, feature extraction, and data visualization. The primary objective of PCA is to create a new basis (usually of a lower dimensionality) for representing data in a more insightful way that can reveal underlying low-dimensional patterns, as the observations often include intercorrelated and noisy data. The new basis, called principal components (PC), comprises a set of orthogonal variables that are linear combinations of the original input variables and constructed such that the first principal component captures the highest variance of the data when projected onto a scalar, the second principal component captures the second highest variance, and so on.

Even though there are various methods to obtain principal components from a set of data, in practice, it often involves solving an eigenvalue-eigenvector problem for a symmetric matrix that is positive-semidefinite. Let us consider a mean-centered collection of inputs $\boldsymbol{X} = [\boldsymbol{X_1}, \boldsymbol{X_2}...\boldsymbol{X_N}]$, where $X_n \in \mathbb{R}^L$, then the covariance matrix is given by $C = X^\top X$. By calculating the eigenvectors $\mathbf{E} = \mathbf{e}_k$ of the covariance matrix, we can remap the input dataset into an identically sized space, where the dimensions are uncorrelated to each other:

$$\mathbf{P} = \mathbf{XE}$$

The eigenvectors of C are represented by **E**, and are L-dimensional orthonormal vectors. Now to reduce the dimensionality, the eigenvectors can be sorted according to the size of their corresponding eigenvalues and the desired number l of principal components can be chosen on the basis of the cumulative variance expressed. Once the number of components to keep is decided, it is sufficient to truncate the sum of the reconstruction to the desired l features by using a subset of the eigenvector matrix $\mathbf{E}^l = \{\mathbf{e}_i\}_{i<l}$. The initial input can thus be projected in terms of principal components:

$$\mathbf{P} = \mathbf{XE}^l$$

The resulting embedding will have dimension $N \times l$, and will contain the first principal components of the dataset. Usually such a projection is used to represent complex, high dimensional data in few dimensions to visually inspect the presence of clusters or obvious linear correla-

tions with available properties.

2.4 Representation of atomic structures

The previous section discussed ML methods commonly used to solve materials science problems. However, every ML workflow relies heavily on the choice of descriptors and their ability to capture relevant information from raw data. The selection of descriptors is task-specific, and choosing a suitable descriptor can significantly impact the ML model's performance. This section focuses on descriptors that are used to learn potential energy surfaces (PES). The problem of describing a PES with a functional form has been of scientific interest for a long time. An early example is the Lennard-Jones potential, where pairwise distances are used as descriptors of the atomic configurations. However, as forcefields were being developed to handle systems with increasingly complex potential energy surfaces, it became clear that the descriptor must respect the internal symmetries of a considered structure, including translational, rotational, and permutational symmetries, as well as encode information beyond that of a two-body problem. Finally, the descriptor should be complete, ensuring that it is unique for every structure, and two different arrangements of atoms must yield different fingerprints.

A successful approach to address these challenges is to derive features from the structure's atom density, where the structure is represented as a set of "local environments" and a local decomposition of the total potential energy is mapped to each of them by a training procedure. One can then obtain the objective PES by summing all the contributions coming from local environments. Several density-based descriptors have been developed, including Behler Parinello Symmetry functions [53] and Smooth Overlap of Atomic Positions (SOAP)[72], which have been effectively applied to various systems.

One drawback of density-based descriptors is that they scale quadratically with the number of chemical species, posing computational constraints on the number of components evaluated within a single force field. However, the scaling issue can be circumvented by creating an embedding that maps features built for real chemical elements into a continuous lower-dimensional space of pseudo-elements. This approach enables the development of models that can generalize across different elements and structures and has shown promising results in predicting the properties of materials[97]. It is built upon the atom-centred density correlation framework[177], which encompasses most of the widespread descriptors for atomic-scale ML, and that is essentially equivalent to the moment tensor potentials[178] and the atomic cluster expansion[179].

2.4.1 Atom density-based representations

Before discussing specific examples of atom density-based representations, we want to give a brief overview of the formalism, generalising the description of density-based representations. We will be using Dirac notation for the atom density descriptors to follow the footprints laid out by the authors in Ref.[177]. For a more detailed explanation, we invite the reader to consult the references [170, 177].

Within this notation, one can introduce a ket vector $|A\rangle$, which represents complete structural and chemical information about structure A. For this structure A, we want to represent the atom density field ρ in the position space x as a sum of local contributions. For this purpose, we place smooth localised functions (i.e. a gaussian g) on top of every atom i with coordinates r_i. In bra-ket notation it can be written it as:

$$\langle \mathbf{x} \mid A;\rho\rangle = \sum_{i\in A} \langle \mathbf{x} \mid \mathbf{r}_i; g\rangle \equiv \sum_{i\in A} g\left(\mathbf{x}-\mathbf{r}_i\right).$$

Similarly, we can rewrite the expression for every local contribution associated with an atomic environment:

$$\langle a\mathbf{x} \mid A;\rho_i\rangle = \sum_{j\in A_i} \delta_{aa_j} \langle \mathbf{x} \mid \mathbf{r}_{ij}; g\rangle f_{\text{cut}}\left(r_{ij}\right) \tag{2.7}$$

We added a channel a distinguishing different chemical elements and the cut-off function f_{cut}, which limits the number of considered atomic environments to those that fall within a given cut-off. The r_{ij} indicates the distance between the central atom i and its neighbour atom j. Now, the representation is translationally invariant, as it is localised and centred on an atom i.

To achieve rotational invariance, one can perform a Haar integration by averaging over the symmetry group $SO(3)$. However, if the ket vector is directly averaged, all information about angular, and more broadly, higher body-order correlations, would be lost. In ref [180], it is demonstrated that in order to include high-order correlations between atomic environments, tensor products of the ket can be taken before applying the Haar integral. Omitting for simplicity the indication of A, this reads as follows:

$$\left\langle a_1\mathbf{x}_1;\dots a_\nu\mathbf{x}_\nu \mid \overline{\rho_i^{\otimes\nu}}\right\rangle = \sum_{k=0,1}\int_{SO^3} d\hat{R}\left\langle a_1\mathbf{x}_1\left|\hat{R}\hat{i}^k\right|\rho_i\right\rangle\dots\left\langle a_\nu\mathbf{x}_\nu\left|\hat{R}\hat{i}^k\right|\rho_i\right\rangle \tag{2.8}$$

where $\overline{\rho_i^{\otimes\nu}}$ is a tensor product of ν atom-centred fields averaged over all possible improper rotations. The sum over k indicates the inversion symmetry and the operator $\hat{i}$ represents

inversion, while $\hat{R}$ represents rotation.

2.4.2 Symmetry functions

Behler-Parinello symmetry functions (SFs) are another representative of atom density-based descriptors. The SFs can be derived from 2.8 using the bra-ket notation by projecting the $SO(3)$ invariant ket onto an appropriate test function G. Below, we will consider the functional form of two types of SFs reflecting radial (G^2) and angular (G^3) correlations, which have been used in this thesis. In general, the functional form of Behler-Parrinello SF represents a product of Gaussians and the cutoff function $f_c(r_{ij})$, where r_{ij} refers to the distance between the atom i and its neighbour j. The cutoff function is a smooth function, which takes the value of a monotonically decreasing function up to the cutoff radius r_c and zero beyond the cutoff radius, which reflects the decaying strength of the interatomic interactions. r_c determines the boundaries of the atomic environments and should be chosen considering energy convergence with respect to its value.

We start the discussion with the "radial" SFs, describing the radial distribution of neighbours inside the cutoff sphere:

$$G_i^2 = \sum_{j=1}^{N_{\text{atom}}} e^{-\eta(r_{ij}-r_s)^2} \cdot f_c(r_{ij}) \tag{2.9}$$

Here η is a parameter which controls the width of the Gaussians and r_s – a shifting radius – displaces the center of the Gaussians improving the sensitivity of the symmetry functions at specific radii. A family of "radial" SFs could be generated by varying these two parameters.

Fingerprints based on the radial distribution alone could not provide satisfactory accuracy of the atomic environment description for systems with complex, directional bonding; however, they provide essential robustness and stability to an MLIP.

To describe angular dependencies, an "angular" type of SFs is introduced, which can be expressed as follows:

$$G_i^3 = 2^{1-\zeta} \sum_{j \neq i} \sum_{k \neq i,j} [(1 + \lambda \cdot \cos\theta_{ijk})^{\zeta} \cdot e^{-\eta(r_{ij}^2 + r_{ik}^2 + r_{jk}^2)} \cdot f_c(r_{ij}) \cdot f_c(r_{ik}) \cdot f_c(r_{jk})]$$

Additional angular functions depending on the angle θ_{ijk} centred at the atom i are used. Multiplication by three cutoff functions guarantees that G_i^3 becomes zero if any of the pair

distances are greater than r_c. Similarly to the "radial" SFs, here η also corresponds to the width of the Gaussians. The distribution of angles could be adjusted by varying the ζ parameter. The λ parameter could take the values of +1 or -1, shifting the maxima of the cosine term and providing a better description for different values of θ_{ijk}.

To sum up, the parameters r_c, θ, r_s, ζ and λ determine the shape of the SFs. A set of SFs is generated by spanning these parameters over a meaningful range of values. Then, one of the unsupervised methods, such as CUR [175, 176], is usually used to customize the set of SFs for a specific dataset. This step allows for unbiased coverage of the configuration space while reducing the number of SFs without compromising the precision of the PES description.

2.4.3 Smooth overlap of atomic positions

Another representation, Smooth Overlap of Atomic Positions (SOAP) [72], can be derived from Eq. 2.8 by projecting Eq. 2.7 onto an orthonormal basis of radial functions $R_n(x) \equiv \langle x \mid n \rangle$ and a basis of spherical harmonics $Y_m^l(\hat{\mathbf{x}}) \equiv \langle \hat{\mathbf{x}} \mid lm \rangle$. Then, the expansion coefficients of localised atom density will be expressed as:

$$\left\langle anlm \mid \rho_i \right\rangle = \sum_{j \in A_i} \delta_{aa_j} \int d\mathbf{x} \langle nl \mid x \rangle \langle lm \mid \hat{\mathbf{x}} \rangle \left\langle \mathbf{x} \mid \mathbf{r}_{ji}; g \right\rangle$$

The 3-body-order representation (the power spectrum) can be obtained from the Eq. 2.8 by fixing the body order expansion term to $\nu = 2$:

$$\left\langle a_1 n_1; a_2 n_2; l \mid \overline{\rho_i^{\otimes 2}} \right\rangle = \frac{1}{\sqrt{2l+1}} \sum_m (-1)^m \left\langle a_1 n_1 lm \mid \rho_i \right\rangle \left\langle a_2 n_2 l(-m) \mid \rho_i \right\rangle, \tag{2.10}$$

highlighting that $\mid \overline{\rho_i^{\otimes 2}} \rangle$ represents a symmetrized, 3-body correlation of the atom density centred on the i-th atom. If the expansion includes $n_{\max}$ radial functions and a maximum angular momentum channel of $l_{\max}$, the power spectrum will consist of $n_{\max}^2 l_{\max}$ elements. For a system containing multiple species, this scaling results in a significant computational cost associated with both the size of feature vectors and the amount of data points required to train such a model.

2.4.4 Alchemical compression of representations

As shown in Eq. 2.10, the number of components grows quadratically with the number of species because each element is considered independently in the neighbour density. The

generalization to higher-ν correlations leads to an even steeper increase, but for most of the multi-component problems, the computational cost is prohibitive even for two-neighbours correlations. Here we give a brief overview of the approach introduced in Ref. [180]. Similarities in the behaviour of elements have inspired the construction in the periodic table[181], and are routinely used to inform materials design and optimization. Instead, elements should be mapped to a continuous n_{alch} -dimensional space, where each chemical species is mapped to n_{alch} *pseudo-species* with a set of coupling coefficients $\mathbf{u}_{\text{alch}}$. Then, the density coefficients can be contracted as

$$\left\langle bnlm \mid \overline{\tilde{\rho}_i^{\otimes 1}} \right\rangle \equiv \sum_a u_{ba} \left\langle anlm \mid \overline{\rho_i^{\otimes 1}} \right\rangle, \tag{2.11}$$

where we use $\tilde{\rho}$ to indicate the alchemically-compressed neighbor density (Fig. 2.1). We note that similar ideas were applied – without optimizing the contraction coefficients – in the context of atom-centred symmetry functions[182, 183], and that a systematic, rather than data-driven, compression has also been recently applied to an 8-element alloy system in the context of atomic cluster expansion potentials[184]. Moreover, there is a large design space of variations on a theme: separate coupling coefficients could be used depending on angular (l) and/or radial (n) channel, and it would be possible to jointly contract over chemical and radial components – which was shown to be effective in reducing the number of features with minimal information loss[185].

To conclude this overview, we note that the alchemical coefficients $\mathbf{u}_{\text{alch}}$ enter the expression for the $\nu = 2$ features in a quadratic fashion, so they cannot be directly determined using linear algebra, even if one uses a linear model based on the contracted features. In ref. [97], this issue was tackled with an iterative strategy, alternating a solution of the linear problem with fixed $\mathbf{u}_{\text{alch}}$ and a gradient descent on the coupling coefficients. In the thesis, instead, we implemented the model using the PyTorch framework[69], allowing us to use automatic differentiation and gradient descent to optimize $\mathbf{u}_{\text{alch}}$ and the model weights simultaneously.

2.5 Fitting a machine learning interatomic potential

The accurate determination of potential energy surfaces (PES) is a fundamental requirement for molecular dynamics simulations, enabling the prediction of thermodynamic properties, reaction mechanisms, and the calculation of molecular properties. Ab initio calculations can provide highly accurate potential energy surfaces but are computationally expensive, and their scaling limits make their application to larger systems challenging. ML has emerged as a powerful tool for accelerating the sampling of *ab initio* potential energy surfaces, enabling the development of more efficient interatomic potentials. While it is necessary to be able to predict atomic energies and forces as a function of atomic coordinates in order to define a potential, a

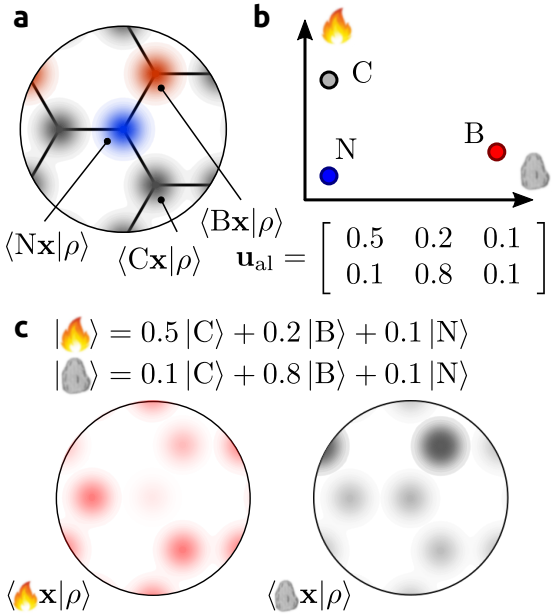


Figure 2.1: Different interpretations of the alchemical compression scheme. (a) In a conventional density-correlation ML scheme, each type of atoms is associated with a separate density. (b) The entries in the alchemical compression matrix $\mathbf{u}_{\mathrm{alch}}$ can be interpreted as describing the "character" of each physical element in terms of n_{alch} pseudoelements - a concept that is not dissimilar from the notion of "classical elements". (c) The structure can be also seen as described in terms of a density of pseudo-elements, for which each site contains a contribuition from each of the compressed channels.

subtle but basic requirement is to do so while keeping the obtained potential conservative. This is possible, provided that forces are learnt through the positional derivatives of the descriptors, an approach that is commonly adopted in modern machine learning potential architectures. In this section, we discuss the use of ML for fitting interatomic potentials, including the representation of the potential energy as a sum of local energy contributions and the learning of forces as derivatives of the potential energy surface. We will also provide examples of machine learning potentials and their applications.

2.5.1 Linear potentials

Since potential fitting is a straightforward supervised learning exercise, it is not surprising that a very simple and effective way to obtain a potential is to associate a linear function to the atomic systems' descriptors. Linear potentials have limited fitting power and strongly depend on the resolution of the underlying descriptors but provide, in turn, an easy-to-train and interpret framework for forces and potential fitting.

We call the matrix of targets (containing energies $E^{(i)}$ and forces $f^{(i)}$ on each atom i) as $\mathbf{Y}^{(i)}$. Its rows will contain the information available for each structure, i.e:

$$\mathbf{Y}^{(i)} = \left[E^{(i)}, f_1^{(i)x}, f_1^{(i)y} f_1^{(i)z}, ..., f_N^{(i)x} f_N^{(i)y} f_N^{(i)z}\right]$$

where frame i contains N atoms and has a total $E^{(i)}$ energy. To construct a conservative potential, we use as a representation matrix the concatenation of the structural descriptor (e.g. a sum of local descriptors in case of an atom density representation like SOAP) $\mathbf{x}$ and its negative derivatives over the atomic coordinates. The entry for the i structure of such a matrix will have the following form:

$$\mathbf{X}^{(i)} = \left[\mathbf{x}^{(i)}, -\frac{\partial \mathbf{x}^{(i)}}{\partial x_1}, -\frac{\partial \mathbf{x}^{(i)}}{\partial y_1}, -\frac{\partial \mathbf{x}^{(i)}}{\partial z_1}, \ldots, -\frac{\partial \mathbf{x}^{(i)}}{\partial x_N}, -\frac{\partial \mathbf{x}^{(i)}}{\partial y_N}, -\frac{\partial \mathbf{x}^{(i)}}{\partial z_N}\right]$$

The ingredients needed to construct the linear problem are all defined, and thus one can proceed to build a potential with the standard form:

$$\mathbf{Y} = \mathbf{X}\mathbf{W}$$

In this formula, the $\mathbf{W}$ matrix contains the weight vectors relating the structural descriptors and their derivatives to the output forces and energies. Given the high-dimensional nature of the problem, it is customary to add a regularization term in the fitting cost function. To account for the nature of the different numerical length scales between forces and energies, it

is also possible to simply substitute the regularization scalar with a vector of regularizers $\boldsymbol{\lambda}$, depending on the nature of the feature (e.g, whether it pertains to a force element or energy). The final fitting formula is simple:

$$\mathbf{W} = \left(\mathbf{X}^T\mathbf{X} + \boldsymbol{\lambda}\mathbb{I}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

The convenience of linear potentials lies in their simplicity and speed of training. The clear limitation of this approach lies in its impossibility of capturing non-linear correlations between input features and target properties. For this reason, the most successful linear potentials tend to be based on high body order representation descriptors, such as MTPs[186] and ACE[179].

2.5.2 High-dimensional neural network potentials

We continue the discussion on different approaches fitting the potential energy surface with a neural network (NN) potential as proposed in Ref. [187]. Similarly to the previous section, we start with the matrix of targets $\mathbf{Y}$, which contains structures' energies $E^{(i)}$ and atomic forces $\boldsymbol{f}_n^{(i)}$. However, in this approach, a single structure energy with N atoms is represented as a sum of the energy contributions $E_n^{(i)}$ determined by local atomic environments $\boldsymbol{x}_n$:

$$E^{(i)} = \sum_n^N E_n^{(i)}(\boldsymbol{x}_n) \tag{2.12}$$

Each contribution $E_n^{(i)}$ is learnt (and predicted) by a separate atomic NN and then the total energy is recovered as a sum of the predictions, where each prediction is obtained following Eq. 2.4.

Since the functional form of a NN is well-defined, the descriptors are differentiable with respect to atomic positions, and the cut-off function is smooth, the analytical derivative for one force component (here for F_x) can be expressed by applying the chain rule as follows:

$$F_x^{(i)} = -\frac{\partial E^{(i)}}{\partial x} = -\sum_{n=1}^{N}\frac{\partial E_n^{(i)}}{\partial x} = -\sum_{n=1}^{N}\sum_{\mu=1}^{N_{x,n}}\frac{\partial E_n^{(i)}}{\partial \boldsymbol{x}_{n\mu}}\cdot\frac{\partial \boldsymbol{x}_{n\mu}}{\partial x},$$

where N_x, n is a number of descriptors of atom n. The term $\frac{\partial E_n^{(i)}}{\partial \boldsymbol{x}_{n\mu}}$ is given by the architecture of the model and includes the weights of the NN, while $\frac{\partial \boldsymbol{x}_{n\mu}}{\partial x}$ depends on the construction of the descriptor.

The weights of a NN are optimized through a process called training or learning, which usually

involves using gradient backpropagation to iteratively adjust the weights in the direction that minimizes the loss function. One reason why an iterative approach is necessary is that the loss function of a NN is generally non-convex, which means that it has multiple local minima and can be challenging to optimize. The optimization process is repeated iteratively until the weights converge to a point where the loss is minimized.

3 Sampling methods

3.1 Introduction

Atomistic computational methods have become an essential tool in understanding materials' behaviour by offering insights into the structure, dynamics, and mechanisms of processes occurring at the atomic scale. Such insights obtained from computational modelling have helped to identify novel materials[188, 189] for a wide range of applications, including renewable energy[190], catalytic energy conversion[191], and energy storage[2].

More recently, the advent of MLIPs increased the computational efficiency of atomistic simulations, enabling simulations to reach large system sizes and long timescales. In the previous chapter, we focused on the set of tools needed to construct robust potentials. In this chapter, we provide insight on how to put them to best use to obtain realistic simulations.

Firstly, we will provide a brief overview of molecular dynamics (MD), which can be used to calculate the dynamical and statistical properties of many-body systems by sampling all possible states of the system classically. Additionally, we will introduce replica exchange MD, which speeds up the sampling procedure. Secondly, we will emphasize the significance of considering quantum nuclear effects and introduce the path integral formalism. Lastly, we will discuss various techniques for free energy estimation, such as thermodynamic integration (TI), interface pinning (IP), and metadynamics.

3.2 Molecular dynamics

Molecular dynamics (MD) is a computational method used to determine the equilibrium and dynamic properties of classical many-body systems. MD simulations can be seen as computer experiments: the material is represented as a system consisting of N interacting particles, and

the temporal evolution of the system is described by Newtonian dynamics. The positions and momenta of each particle can be calculated by solving Hamilton's equations:

$$\dot{\mathbf{r}}_i = \frac{\partial H(\mathbf{p},\mathbf{r})}{\partial \mathbf{p}_i} = \frac{\mathbf{p}_i}{m_i},$$
$$\dot{\mathbf{p}}_i = -\frac{\partial H(\mathbf{p},\mathbf{r})}{\partial \mathbf{r}_i} = -\frac{\partial V(\mathbf{r})}{\partial \mathbf{r}_i} = \mathbf{F}_i,$$

where r_i, p_i, and F_i correspond to the position, momentum, and force of the i-th particle, while V represents the interatomic potential and H is the classical Hamiltonian of the N-particle system. It is worth noting that problems can arise if the interatomic potential energy function V is not a smooth function of the particles' positions. This is due to the explicit appearance of $\frac{\partial H}{\partial \mathbf{r}}$ in Hamilton's equations, which requires that at least the first derivative of $V(\mathbf{r})$ be continuous.

The system's state at any given time is fully determined by the positions $(r_1, ..., r_N)$ and momenta $(p_1, ..., p_N)$ of each particle. However, analytical solutions to the equations of motion for complex many-body problems are not feasible, and thus MD achieves time evolution iteratively using a numerical integration scheme with a discrete time step Δt. One should choose the time step which is long enough to avoid excessively expensive simulations while still allowing for the study of the system's evolution on the time scale of interest. The typical value of Δt usually varies from 0.5 to 2 fs. It is important that the chosen numerical method conserves total energy and is time-reversible, and is easy to implement in computer code. However, no algorithm can provide a precise solution indefinitely, as errors accumulate with each iteration. Instead, the aim is to obtain a trajectory that is representative of the statistical and time-dependent behaviour of the process being simulated. The velocity Verlet method[192] satisfies all the aforementioned requirements and is commonly used in integration schemes for MD:

$$v\left(t+\frac{1}{2}\Delta t\right) = \boldsymbol{v}(t) + \frac{1}{2}\Delta t\boldsymbol{a}(t)$$
$$\boldsymbol{r}(t+\Delta t) = \boldsymbol{r}(t) + \Delta t\boldsymbol{v}\left(t+\frac{1}{2}\Delta t\right)$$
$$v(t+\Delta t) = v\left(t+\frac{1}{2}\Delta t\right) + \frac{1}{2}\Delta t a(t+\Delta t)$$

The algorithm works in three steps: first, forces are evaluated to obtain accelerations $\boldsymbol{a}$ at time t. Then, velocities v are calculated at time $t+\Delta t/2$. Next, positions $\boldsymbol{r}$ are updated up

to time $t+\Delta t$. Finally, a second evaluation of forces updates velocities at time $t+\Delta t$. The Verlet algorithm is known for accurately conserving energy with a root-mean-square error proportional to Δt^2.

If we run an MD simulation using the above equations, we will effectively be sampling a microcanonical ensemble (NVE), which describes a system with a fixed number of particles, volume, and energy. By exploiting the principle of ergodicity, we can compute the ensemble average of any observable A by taking a time average along the simulation for a long enough trajectory.

$$\langle A \rangle_{\text{ens}} = \lim_{\mathscr{T}\to\infty} \frac{1}{\mathscr{T}} \int_0^{\mathscr{T}} A\big(\mathbf{p}(t), \mathbf{r}(t)\big)\, dt.$$

The idea behind this is that if the system does not exchange particles or energy with the environment, it will eventually sample all possible microstates of the phase space. This is particularly useful in MD because it allows us to compute the averages of quantities that are experimentally measurable based on simulations of finite-size systems. However, the NVE ensemble is typically not the best representative of thermodynamic conditions set in experiments, which is why MD has been extended to other ensembles, such as NVT, where a thermostatting algorithm ensures the sampling of an ensemble at constant temperature, or NPT, where the volume is allowed to freely fluctuate, at the constrained pressure and temperature conditions. [193–196].

3.2.1 Replica-exchange molecular dynamics

The exploration of the potential energy landscape can be accelerated by using replica-exchange molecular dynamics (REMD). In this approach, several system replicas are set up in parallel and run at different temperatures (and/or pressures) independently. Each state of the system could be described by a state vector $\boldsymbol{s}(\boldsymbol{r}_1^{(1)}, \boldsymbol{p}_1^{(1)}, ..., \boldsymbol{r}_i^{(i)}, \boldsymbol{p}_i^{(i)}, ...)$, where $\boldsymbol{r}$ and $\boldsymbol{p}$ refer to the atomic positions and momenta, subscript i indicates the index of the replica and the superscript (i) indicates the index of the state (T_i, P_i). The distribution function of the system can be expressed as a product of Boltzmann factors of all the replicas[197]:

$$P[\boldsymbol{s}(..., \boldsymbol{r}_i^{(i)}, \boldsymbol{p}_i^{(i)}, ...)] = \frac{1}{Z} exp\Big(-\sum_i \frac{H(\boldsymbol{r}_i^{(i)}, \boldsymbol{p}_i^{(i)})}{k_B T^{(i)}}\Big)$$

where $H(\boldsymbol{r}_i^{(i)}, \boldsymbol{p}_i^{(i)})$ is the Hamiltonian of the replica i at the temperature $T^{(i)}$, k_B is the Boltzmann constant and a normalization factor Z. All the replicas are allowed to exchange the entire configuration at every N step in accordance with the Metropolis-Teller algorithm [198].

Basically, the system tries to perform the transfer from the state $\boldsymbol{s}(...,\boldsymbol{r}_i^{(i)},\boldsymbol{p}_i^{(i)},...,\boldsymbol{r}_j^{(j)},\boldsymbol{p}_j^{(j)},...)$ to $\boldsymbol{s}(...,\boldsymbol{r}_i^{(j)},\boldsymbol{p}_i^{(j)},...,\boldsymbol{r}_j^{(i)},\boldsymbol{p}_j^{(i)},...)$. The probability of this transfer could be defined as:

$$P(i,j) = min\Big\{1, \frac{exp(-V(r_j^{(i)})/k_B T^{(i)})exp(-V(r_i^{(j)})/k_B T^{(j)})}{exp(-V(r_i^{(i)})/k_B T^{(i)})exp(-V(r_j^{(j)})/k_B T^{(j)})}\Big\}$$

where $V(\boldsymbol{r}_i^{(i)})$ is the potential energy of the i-th replica.

The probability of swapping between the trajectories increases as the temperature of the replicas gets closer. To ensure the effective exploration of the state space, it is important to have a range of temperatures, with the highest temperature being sufficiently high to allow for escape from free-energy minima and exploration of low-probability regions, while the lower temperature replicas probe the various stable states corresponding to free-energy minima. The number of replicas should also be large enough to ensure proper swapping among adjacent replicas. Common practice is to use geometric spacing to define the temperatures of the replicas. In the thesis, we used the REMD technique to improve the convergence of ensemble averages and also to generate less correlated states while creating datasets for MLIPs.

3.2.2 Path integral molecular dynamics

While classical MD simulations have been widely used to study the behaviour of materials and molecules, they do not take into account nuclear quantum effects (NQE). In many cases, this is a reasonable approximation, as the thermal energy of the system is much greater than the energy differences between quantum states. However, NQE can have a significant impact on certain materials and systems. For example, NQE can affect the structure and dynamics of protons[199]. They can also have a significant impact on thermal reaction rates at metal surfaces: neglecting the wave nature of adsorbed hydrogen atoms and their electronic spin degeneracy can lead to a 10× to 1000× overestimation of the rate constant for temperatures relevant to heterogeneous catalysis[200]. To accurately capture these effects, more sophisticated simulation methods such as path integral molecular dynamics (PIMD) have been developed, which treat both the electrons and nuclei as quantum mechanical objects by using the imaginary time path integral formalism[32].

It is widely known that the static equilibrium properties of a quantum mechanical system can be computed relatively easily by using the isomorphism between the path integral representation of the quantum mechanical partition function and the classical partition function of a fictitious ring polymer[201]. The ring-polymer is constructed of replicas (or “beads”) of the physical system, with corresponding atoms connected by harmonic springs. Using path

integral formalism, the quantum partition function is equivalent to the classical partition function of an extended classical system composed of several replicas of the physical system at an elevated temperature. This is achieved by expressing the canonical Boltzmann distribution at $\beta = 1/(k_B T)$ as a product of P high-temperature distributions at $\beta_P = \frac{\beta}{P}$:

$$Z(N,V,\beta) = \mathrm{Tr}[\exp(-\beta\hat{H})] = \mathrm{Tr}\left[\left(\exp\left(-\beta_P\hat{H}\right)\right)^P\right] \tag{3.1}$$

By executing the Trotter expansion, we can write the partition function as :

$$Z_P(\beta) = \frac{1}{(2\pi\hbar)^P}\int d^P\boldsymbol{p}\int d^P\boldsymbol{r}\, e^{-\beta_p H_P(\mathbf{p},\mathbf{r})},$$

where P represents the number of beads. The Hamiltonian of an extended ring polymer is denoted by $H_P(\mathbf{p},\mathbf{r})$ and expressed as:

$$H_P(\mathbf{p},\mathbf{r}) = \sum_{j=0}^{P-1}\left[\sum_{i=1}^{N}\frac{\boldsymbol{p}_i^{(j)2}}{2m_i} + V\left(\boldsymbol{r}^{(j)}\right) + \sum_{i=1}^{N}\frac{1}{2}m_i\omega_P^2\left(\boldsymbol{r}_i^{(j)} - \boldsymbol{r}_i^{(j+1)}\right)^2\right],$$

where $r^P = r^0$ and $\omega_P = (\beta_P\hbar)^{-1}$.

With this, one can run P MD trajectories, where every quantum particle is described by P classical particles connected with the springs. The computational cost of such a simulation is now P times greater than the cost of a classical MD trajectory, where P is the number of beads. The use of Trotter expansion implies that the partition function approaches the exact description of the quantum system as the number of beads tends to infinity. In practice, an optimal number of beads, required for reasonable accuracy of the results, strongly depends on the investigated system. It is also interesting to note, that if the spring constant becomes infinitely large, all the replicas will collapse into a single entity and behave according to classical MD principles.

3.3 Free energy estimation methods

The free energy is an essential concept in thermodynamics, as it is used to determine phase stability of materials[202], study phase transitions[203, 204] and predict the direction and extent of reactions[205]. To accurately calculate free energy, various methods have been developed, that use statistical mechanical principles within computer simulations, such as molecular dynamics. Despite recent advances, the computation of free energy remains a challenging task.

Directly calculating the absolute free energy of a material is difficult, except for relatively simple systems such as harmonic solids. For example, the Helmholtz free energy is related to the canonical partition function Z as $F = -k_B T \ln Z$, where k_B is Boltzmann's constant and T is the absolute temperature, but it involves integrations over all degrees of freedom in the system, making direct calculation problematic. Nevertheless, the relative free energy can be calculated by determining the work done to transform the system from one state to another, i.e. by using thermodynamic integration methods.

In addition to thermodynamic integration methods, enhanced sampling techniques are also commonly used to estimate relative free energies in atomic systems. These techniques modify the probability distribution of a system to sample rarer events more efficiently, allowing for more accurate sampling and reconstruction of the true free energy landscape. These techniques include metadynamics, replica exchange molecular dynamics, and adaptive biasing force.

In this section, we provide a brief introduction to the various free energy estimation methods used within the scope of this thesis.

3.3.1 Thermodynamic integration

Thermodynamic integration (TI) is a method used to define the free energy difference between two states of a system. As the free energy is a function of the Boltzmann-weighted integral over the entire phase space and not just a function of the system's state, the free energy differences cannot be computed directly using the potential energies. However, one can calculate the free energy difference by integrating the ensemble-averaged enthalpy changes along a chosen path connecting two states. In this section, we provide a brief overview of the TI in application to solid states.

For the case of a solid system, it is sensible to start from the free energy of the harmonic crystal as the reference state, which can be straightforwardly computed as:

$$F_{\mathrm{h}}(V, T_0) = k_B T_0 \sum_{i=1}^{3N-3} \ln \frac{\hbar \omega_i}{k_B T_0} \tag{3.2}$$

where ω_i are phonon frequencies of the crystal with N atoms, and T_0 a low temperature chosen so that the system is close to a local minimum of the potential energy. Note that we use the classical expression because we are ultimately interested in high-temperature values of the free energy. If one wanted to estimate the anharmonic free energy at low temperature, it is possible to do so by a further thermodynamic integration step [206–208]. Starting from the

harmonic reference, one then performs the actual TI step, which involves parameterising a Hamiltonian $\mathscr{H}(\lambda)$ in such a way that $\mathscr{H}(\lambda = 0)$ corresponds to the harmonic potential and $\mathscr{H}(\lambda = 1)$ to the real system. One then evaluates numerically the integral

$$\Delta F = F(\lambda = 1) - F(\lambda = 0) = \int_0^1 \mathrm{d}\lambda \left\langle \frac{\partial \mathscr{H}}{\partial \lambda} \right\rangle_\lambda \tag{3.3}$$

to give the free energy difference between the systems, which is the anharmonic correction to the free energy.

By choosing a sufficiently low T_0, the system is very close to being harmonic, and this term is small and can be computed easily, possibly even just by free energy perturbation. In order to convert between constant-volume and constant-pressure boundary conditions, we perform a constant pressure simulation in conditions that give a mean volume close to that used to compute F_{anh}, and evaluate the distribution of volumes $\rho(V|p,T)$. The Gibbs free energy is then given by

$$G(N,p,T) = pV + F_{anh}(N,V,T) + k_b T \ln\left[\rho(V|p,T)\frac{V}{N}\right], \tag{3.4}$$

which is based on the definition of the isobaric partition function $\mathscr{Z} = \int \mathrm{d}V e^{-\beta p V} e^{-\beta F_{\mathrm{anh}}(N,V,T)} N V^{-1}$ discussed in Ref. [209].

To evaluate the Gibbs free energy at higher temperature, one can then perform a series of NpT simulations at different values of T – possibly using replica exchange to enhance statistical convergence – and evaluate a TI estimate of

$$\frac{G(p,T_1)}{k_B T_1} = \frac{G(p,T_0)}{k_B T_0} - \int_{T_0}^{T_1} \frac{\langle \mathscr{H} + pV \rangle}{k_B T^2} dT \tag{3.5}$$

where $\mathscr{H}$ denotes the total energy.

3.3.2 Interface pinning

The interface pinning (IP) method[210] is a special case of the umbrella sampling technique [211], used to quantify the Gibbs free energy difference between two coexisting phases separated by a flat surface. In the IP method, a harmonic bias potential coupled to an order parameter Φ that discriminates between the two phases of interest is used to analyze a two-phase system and to force the interface to stay in an intermediate state. The Gibbs free energy difference between the phases is determined by the average force that the pinning potential exerts on the system. Let us consider a system where the solid and liquid phases coexist. If the mean value of the order parameter in bulk solid and liquid at a given temperature is $\bar{\phi}_s$ and $\bar{\phi}_l$,

and the sum over all atoms of the order parameter for a given configuration is Φ, the number of solid atoms can be estimated as $N_s = (\Phi - N\phi_l)/(\bar{\phi}_s - \bar{\phi}_l)$ – with the underlying assumption of choosing the dividing surface between the solid and the liquid phase that corresponds to zero excess for the chosen order parameter. With this definition, the Gibbs free energy associated with a two-phase configuration is given by $G(N_s) = \mu_s N_s + (N - N_s)\mu_l + 2\gamma_{\mathrm{sl}} A_{xy}$, where γ_{sl} is the solid-liquid interfacial free energy and A_{xy} the cross-section of the simulation box. When performing a simulation of the interface applying the pinning potential of the form $(\Phi - \Phi_{ref})^2\kappa/2$, the overall free energy reads:

$$\tilde{G}(N_s) = \mu_s N_s + (N - N_s)\mu_l + 2\gamma_{\mathrm{sl}} A_{xy} + (\Phi - \Phi_{ref})^2\kappa/2.$$

Hence, in conditions above or below the melting point, the difference in chemical potential between the solid and the liquid phases leads to the interface fluctuating around an equilibrium position for which $\Phi \neq \Phi_{ref}$, and one can extract:

$$\Delta\mu_{\mathrm{sl}} = \mu_s - \mu_l = -\kappa(\Phi - \Phi_{ref})(\bar{\phi}_s - \bar{\phi}_l) \tag{3.6}$$

By performing multiple simulations at different temperatures, one can identify the dependence of $\Delta\mu_{\mathrm{sl}}$ on T. The temperature at which $\Delta\mu_{\mathrm{sl}} = 0$ identifies the melting point T_m, and the slope is equal to the entropy of melting. To put this method into practice, one should choose a collective variable, which distinguishes between two phases. In this thesis, we used the IP method to compute the melting point of elemental nickel. The results and details of the simulations are provided in the corresponding section.

3.3.3 Metadynamics

Metadynamics[212] is another technique used to enhance MD simulations by adding to the system's Hamiltonian a time-dependent bias linked to some function of the internal coordinates, also called collective variables (CVs). A bias potential helps to overcome energy barriers and sample rare events more efficiently, and moreover, it can be used to reconstruct the free energy surface along the explored direction in CV space. The bias potential is gradually built up over time by depositing small Gaussian-shaped hills, which prevents the system from getting trapped in local minima. The height and width of the hills are controlled by the deposition frequency and the Gaussian width parameter, respectively. If the studied phenomenon can be described with a CV $\Phi(\boldsymbol{r})$, where $\boldsymbol{r}$ defines the state of the system, then the bias potential V_{G} at the time t can be defined as:

$$V_{\mathrm{G}}(\Phi, t) = \int_0^t dt' \omega e^{-\frac{(\Phi_i(R) - \Phi_i(R(t')))^2}{2\sigma^2}}$$

In this context, ω represents the bias deposition, and σ represents the width of the Gaussian distribution for the CV Φ. The energy rate is typically expressed in terms of a Gaussian height W and a deposition stride τ_{G}, where:

$$\omega = \frac{W}{\tau_{\mathrm{G}}}$$

Assuming the simulation is of sufficient length, it is possible to calculate the free energy F using the following equation:

$$\mathrm{V_G}(\Phi, \mathrm{t} \rightarrow \infty) = -\mathrm{F}(\Phi) + \mathrm{C},$$

where C is an additive constant. Metadynamics in the formulation described above has certain limitations, such as oscillatory behaviour around true free energy surfaces and the challenge of determining when to stop the simulation. To address these issues, well-tempered metadynamics [213] was developed as a modification of the standard metadynamics method. By using an adaptive bias factor $\gamma = 1 + \frac{\Delta T}{T}$ to control the height of the Gaussian-shaped potentials deposited during the simulation, well-tempered metadynamics allows for faster convergence to the true free energy surface. Additionally, the adaptive bias factor provides a means of determining when the simulation has reached convergence. The bias factor is set to decrease over time, allowing the system to escape from the shallow basins and converge to the true free energy landscape. In practice, the well-tempered metadynamics is implemented by rescaling the Gaussian heights by a factor as shown below:

$$W = \omega_0 \tau_{\mathrm{G}} e^{-\frac{V_{\mathrm{G}}(\Phi, t)}{k_{\mathrm{B}} \Delta T}}.$$

In this context, ω_0 represents the initial rate, while ΔT is a parameter in temperature units that controls the extent to which the free energy is explored. Unlike standard metadynamics, the bias potential in well-tempered metadynamics does not converge to the negative of the free energy, but rather to a fraction of it, as shown in the following equation:

$$\mathrm{V_b}(\Phi, \mathrm{t} \rightarrow \infty) = -\frac{\Delta \mathrm{T}}{\mathrm{T} + \Delta \mathrm{T}} \mathrm{F}(\Phi) + \mathrm{C}$$

The expression $\frac{\Delta T}{T + \Delta T}$ is commonly known as the bias factor. This results in an improved sampling of the CV space, which corresponds to an effective temperature of $T + \Delta T$. As

ΔT approaches infinity, the method approaches standard metadynamics, while for ΔT = 0, it reduces to regular molecular dynamics (MD). Additionally, it is possible to re-weight well-tempered metadynamics simulations to obtain accurate statistics for any observable of interest.

4 Finite temperature modelling of nickel[1]

4.1 Introduction

In the previous chapters, we have discussed various machine learning methods that are instrumental in building a machine learning interatomic potential (MLIP) and sampling techniques that can be combined with an MLIP to investigate properties using a surrogate quantum mechanical potential energy surface. This chapter outlines the process of constructing an MLIP and its practical integration with a wide range of statistical mechanics techniques, including thermodynamic integration and finite-temperature sampling, using elemental nickel as an example. Moreover, we adopt a recently-developed scheme to predict the electronic density of states, accounting for electronic excitations without the need for additional calculations.

Nickel was selected as the reference system for two primary reasons. First, it has applications across a wide temperature range and holds significant industrial importance, as it serves as a key component in numerous alloys, including steel, Inconel, and Hastelloy. Second, it serves as an ideal benchmarking system, given its extensive experimental studies and the availability of a reasonably accurate empirical interatomic potential.

The MLIP created in this thesis is benchmarked against experiments and density functional theory (DFT) where possible, and the results are compared to an accurate embedded atom model (EAM) potential. We commence with Sec. 4.2, where we summarize reference calculations, neural-network potential construction, and the machine-learning model for the electronic density of states. Subsequently, in Sec. 4.3.1, we demonstrate the accuracy of the machine-learning potential. Lastly, we present the computation of challenging finite-temperature properties of nickel in Sec. 4.3.2.

[1] This chapter is an adaptation of my contribution to Ref. [214], where I was responsible for all aspects except for the construction of the ML model for DOS.

4.2 Constructing a machine learning interatomic potential for nickel

We begin by providing a brief summary of the methods we use to obtain reference properties and train machine-learning models, together with the details that are necessary to reproduce the underlying electronic-structure calculations, the construction of the training set, and the structure of the machine-learning model.

4.2.1 Electronic-structure details

We compute all the energies and forces using density-functional theory (DFT), as implemented in QUANTUM ESPRESSO[215]. We use the PBE exchange-correlation functional[216], together with an ultrasoft pseudopotential[217] with 10 valence electrons for Ni, from the standard solid-state pseudopotential library [218]. The wave function is expanded in plane waves with a cutoff energy of 40Ry. The Brillouin zone sampling uses the Monkhorst-Pack scheme[219] with a k-point density of 0.07 $Å^{-1}$. To improve the convergence of the integral over the k-points mesh, we use the Methfessel-Paxton first-order spreading[220] with a broadening parameter equal to 0.0441 Ry. All the parameters are kept fixed for the whole data set and are converged in terms of energy differences.

All the reference calculations are non-magentic, even though Nickel exhibits ferromagnetic ordering below its Curie temperature (628K). This choice is driven by the fact that treatment of magnetism implies associating additional degrees of freedom describing the magnetic configuration of the system (e.g. spin polarization of atoms for a collinear treatment), that is not compatible with the typical infrastructure of machine-learning potentials, that use only nuclear coordinates as inputs. As an approximate alternative, one would need to perform a separate set of reference calculations below and above the Curie temperature, for instance, through a collinear spin-polarized approximation below the Curie temperature and more complex approximations when the magnetic disorder occurs. However, this approach would inadvertently introduce an undesirable temperature dependence of the potential.

Given that our main goal is to describe high-temperature conditions, where anharmonic contributions to the free energy become important, we prioritize the description of the paramagnetic phase. Furthermore, Ni is a weak ferromagnet, and many of its properties (such as phonon dispersion curves, vacancy formation energies, thermal expansion[221–223]) are only weakly affected by magnetism. Indeed, as shown in Figure 4.1, the equation of state computed with non spin-polarized DFT (blue curve) and collinear spin-polarized DFT (spDFT) (yellow curve, ferromagnetic ordering) exhibits very small differences. The lattice constant changes by less than 1% (3.517Å for DFT vs 3.526 Å for spDFT) and bulk moduli differ by 5% (195GPa for

DFT and 185GPa for spDFT). Nevertheless, there are other properties such as the heat capacity curve shown in Figure 4.8 for which magnetism plays an important role, and incorporating magnetic excitations in a similar way as what we do for electronic excitations is a promising research direction.

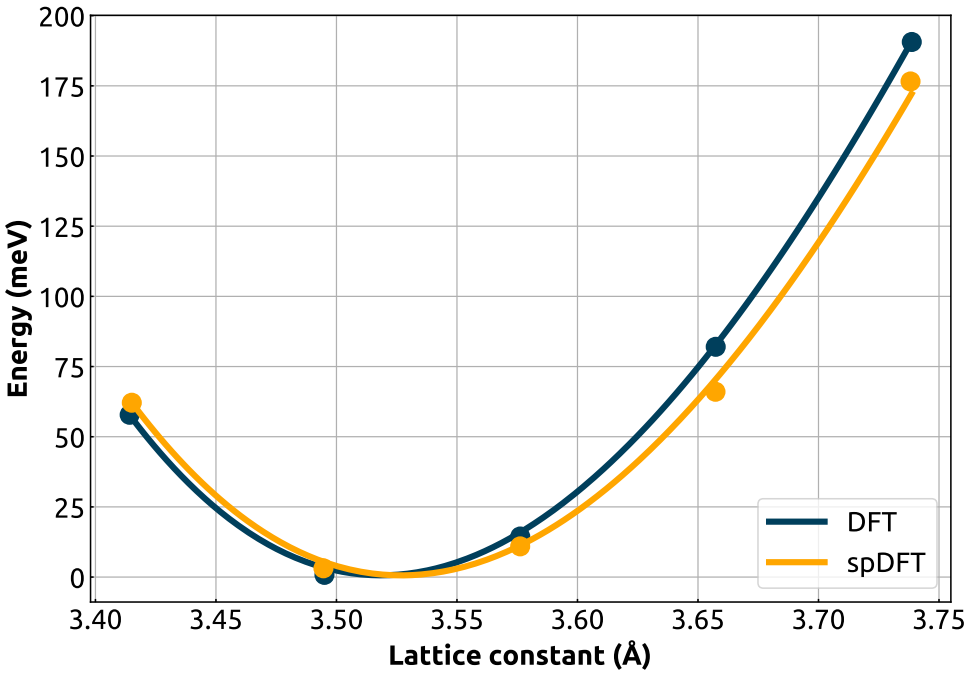


Figure 4.1: Equation of state of FCC nickel, (referenced to the minimum energy): blue dots represent DFT calculations, yellow dots spin-polarized DFT (spDFT) calculations. Solid lines indicate the corresponding fits to a Birch-Murnaghan equation of state.

4.2.2 Training set construction

The structures included into the dataset are selected by an iterative procedure, building on the insights of existing literature in similar systems[46, 224, 225]. Each training structure provides total energy and forces computed by DFT, with the computational settings indicated above. In Table 4.1 we summarize the content of the final version of the dataset. Given the availability of a reliable EAM potential[226], we used it to generate a diverse set of configurations, for which we then recomputed energies and forces using DFT. We first performed a long replica exchange molecular dynamics simulation[227], using the i-PI implementation[228, 229], and including 82 NpT trajectories spanning a broad range of temperatures [100 K, 3200 K] and pressures [−5 GPa, 5 GPa]. From these trajectories, we selected 1000 structures using farthest

Structure type	No. structures	No. atoms
Selected from REMD	988	108
FCC Bulk		
isotropic stress	4	108
shear stress	4	108
uniaxal stress	4	108
displacement of one atom	10	108
Single vacancy	6	107
Single intestitial	18	109
HCP Bulk	22	54
BCC Bulk	10	54
Stacking fault	299	24
Solid-Vacuum Interface		
(100)AA	154	9
(100)AB	110	8
(110)AA	88	13
(110)AB	252	12
(111)AA	110	8
(111)AB	105	9
Solid-liquid interface	17	96
Liquid-vacuum interface	10	108
Other	24	7
Total	2235	–

Table 4.1: Overview of the composition of the training dataset used to fit the neural network potential. The first column shows the number of structures included in each group, and the second column shows the number of atoms included in each supercell.

point sampling (FPS)[168].

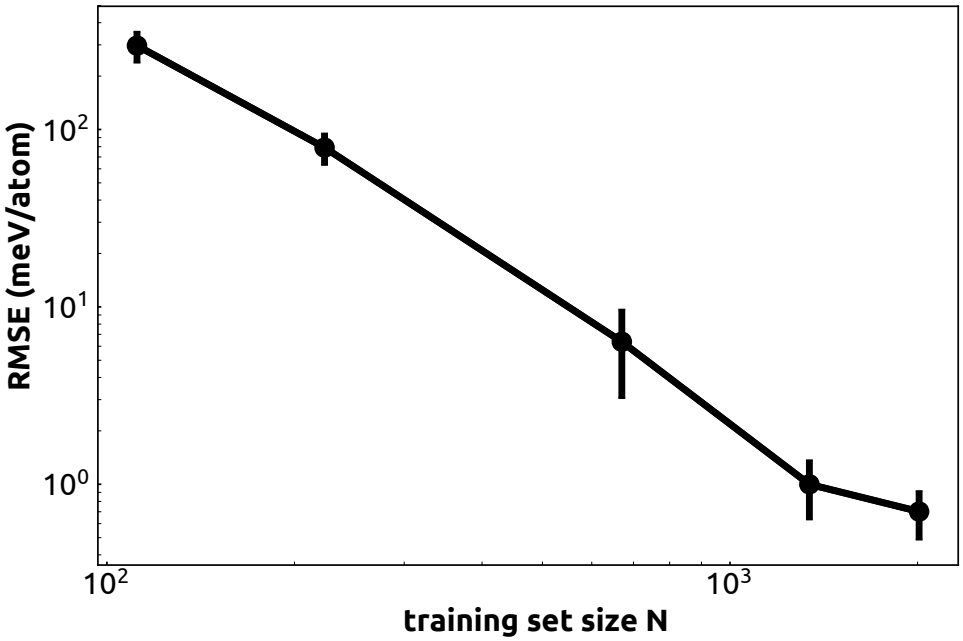
On top of these baseline structures, which provide a diverse set of configurations across the phase diagram of Ni, we incorporate targeted single-point calculations that are used to ensure that configurations that are relevant for important structural and mechanical properties are well represented. In particular, we include 1x1x1 FCC structures stretched and compressed by less than 3% of the equilibrium lattice parameter – that report directly on bulk modulus and elastic constants.

To reproduce accurately defects formation energies, we perform geometry optimization of a single vacancy and interstitial in 3x3x3 FCC cells at 0K, and of a 3x3x1 HCP and 3x3x3 BCC cells, describing metastable phases of Ni. We also include 1x1x6 FCC structures with the x-, y- and z-axes oriented along $[11\bar{0}]$, $[11\bar{2}]$, and [111] directions, distorted to incorporate information on the generalized stacking fault surface, as well as (111), (001), and (110) surfaces created by rigid cleavage of the bulk, leaving a slab which is more than 12Å thick. All the steps of the geometry optimization have been added to the reference set.

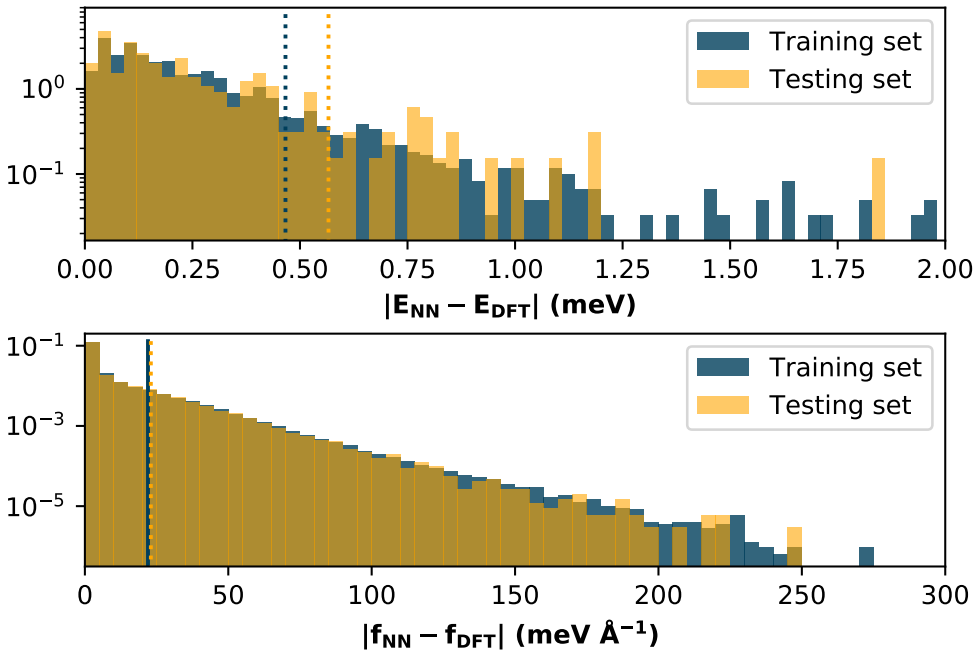
Finally, to ensure that the neural network potential (NNP) samples the repulsive part of the interatomic potential, we add 3x3x3 FCC structures where a position of one atom has been randomized by up to [0.5, 1.2]Å. In total, the training set contains approximately 2200 configurations[230].

4.2.3 Neural network potential

To describe the interatomic potential we train a high-dimensional neural network (NN) following the approach proposed by Behler and Parinello[231]. The framework of the Behler-Parinello NNPs has been discussed extensively in the literature[231–234]. In Sec. 2.5.2, we offer a concise introduction, wherein we show that the atomic energy contributions are represented as a series of nested activation functions acting on linear combinations of the values in the previous layer. The input layer, which describes the geometry of each atom-centred environment, entails a vector of atom-centred symmetry functions, that describe two and three-body correlations between neigbours [96, 177]. The architecture of the NN and the functional form of the symmetry function are analogous to those used in Ref. [176]. The values of the parameters defining the set of symmetry functions were determined by first generating a large set of possible features, combining cutoff distances of 8, 12, 16 and 20 Bohr, and selecting the 50 most informative ones based on a deterministic CUR algorithm, as discussed in Ref. [176]. The parameters of the network are optimized using the N2P2 package[235, 236] to agree with the reference DFT data. The NN architecture includes 2 hidden layers with 25 nodes each. 90% of



(a) Learning curve for the energy RMSE of the NN potential, as a function of the number of structures included in the training set. The points and error bars indicate the mean and standard deviation of five potentials, computed with different random choices of the training points.



(b) The histograms of the differences between predicted with NNP energies and forces and DFT reference for training and testing subsets. The dashed lines show the reported RMSE's.

Figure 4.2: Performance evaluation of the NNP.

the dataset set is used for training, with a random selection including 10% of structures being held out for validation. The RMSEs on the training and testing subsets are 0.45meV/atom and 0.55meV/atom for energies and 22meV/Å and 23meV/Å for forces respectively. In Figure 4.2b we show the errors distribution. These errors – as well as the errors on selected target properties, discussed in Sections 4.3.1 and 4.3.2– are in line with state-of-the-art potentials, and comparable with the typical error of density functional theory. As shown in Figure 4.2a, the model accuracy is limited by the amount of training data, and not by the complexity of the model, so it would be easy, if needed, to further reduce the error by just increasing the train set size. The neural network weights of the model we used in the rest of this work can be downloaded from a public repository[230].

4.2.4 Machine-learning model of the electronic density of states

A NN potential allows to sample phase space in a way that is consistent with ab initio quality energetics. However, it does not give direct access to electronic-structure properties. Recently, ML models have been proposed that give direct access to properties that are related to the electronic degrees of freedom, such as the ground-state charge density [102, 237, 238] and the density of single-particle energy levels (density of states, DOS) [49]. As a first step towards a fully integrated, universal ML scheme that provides a complete surrogate model of quantum mechanical calculations, we train a model relying on a fixed DOS approximation and we use it to predict properties that depend on electronic excitations. We use an atom-centered model for the DOS, where we expand the total DOS of a structure A over a sum of local DOS contributions (LDOS) associated with its atomic environments A_i:

$$\mathrm{DOS}(A,E) = \sum_{i \in A} \mathrm{LDOS}(A_i, E).$$

The reference DFT DOS is constructed with a Gaussian broadening $g_b = 0.1$eV, which ensures that the curves are well-detailed. We use the Fermi energy ε_F of each structure as the energy reference.

We follow the approach introduced in Ref.[239] to determine the mapping between the atomic environment A_i and its contribution to the total DOS. In a nutshell, we introduce a positive-definite scalar kernel $k(A_i, A'_{i'})$ that describes the similarity between two atomic environments. We use in practice the SOAP kernel [72], as implemented in librascal [240]. We then determine the *active set* containing the M most diverse environments found in the training set, and write a Projected Process (PP) approximation of the Gaussian Process (GP) algorithm to express the LDOS as a function of the basis set formed by the kernel between each target environment

and the active set

$$\mathrm{LDOS}(A_i, E) = \sum_{j \in M} x_j(E)\, k(A_i, M_j).$$

The expansion coefficients $\mathbf{x}_M(E)$ are determined separately for each energy channel. We use the pointwise representation of the DOS from Ref.[239], where we discretize the energy axis over a finite range and take the DOS at every energy point as a target of the ML model. Once the model is trained, the DOS of a new structure A_* can be easily obtained from the dot product between the kernel matrix of its atomic environments and the active set, and the energy-dependent expansion coefficients $\mathbf{x}_M$. To monitor the reliability of the predictions, we also implement uncertainty estimation based on a calibrated committee model [241]. We use the DOS model to compute the electronic contributions to several thermodynamic properties, such as the Helmholtz Free energy at finite temperature

$$F^{\mathrm{el}}(T) = U^{\mathrm{el}}(T) - TS^{\mathrm{el}}(T) \tag{4.1}$$

which is decomposed in a contribution from the hot electrons to the band energy

$$U^{\mathrm{el}}(T) = \int_{-\infty}^{\infty} \varepsilon \mathrm{DOS}(\varepsilon) f(\varepsilon - \varepsilon_F, T) \mathrm{d}\varepsilon - \int_{-\infty}^{\varepsilon_{\mathrm{F}}} \varepsilon \mathrm{DOS}(\varepsilon) \mathrm{d}\varepsilon \tag{4.2}$$

and an entropy term

$$S^{\mathrm{el}}(T) = \int_{-\infty}^{\infty} \mathrm{DOS}(\varepsilon) \Big[f(\varepsilon - \varepsilon_F, T) \log\big(f(\varepsilon - \varepsilon_F, T)\big) - (1 - f(\varepsilon - \varepsilon_F, T)) \log\big(1 - f(\varepsilon - \varepsilon_F, T)\big) \Big] \mathrm{d}\varepsilon, \tag{4.3}$$

and the electronic contribution to the high-temperature heat capacity

$$C_v^{\mathrm{el}}(T) = \frac{\partial U^{\mathrm{el}}(T)}{\partial T}. \tag{4.4}$$

These expressions are written in a "non-self-consistent" approximation added *a posteriori*, where we consider the density of states to be fixed to that computed from the Kohn-Sham eigenvalues obtained self-consistently at $T = 0$. The temperature dependence is due to the occupation of the energy levels, which is given by a Fermi function $f(\varepsilon - \varepsilon_F, T)$, and by the Fermi energy ε_F which is computed for the DOS at each temperature by enforcing charge neutrality. To achieve a consistent sampling of the dynamics, where the electronic excitations are accounted for in the ions' interatomic forces, one can follow the approach proposed in Ref. [242].

To train a model of the DOS we use a subset containing 1069 structures of the data set in Tab. 4.1,

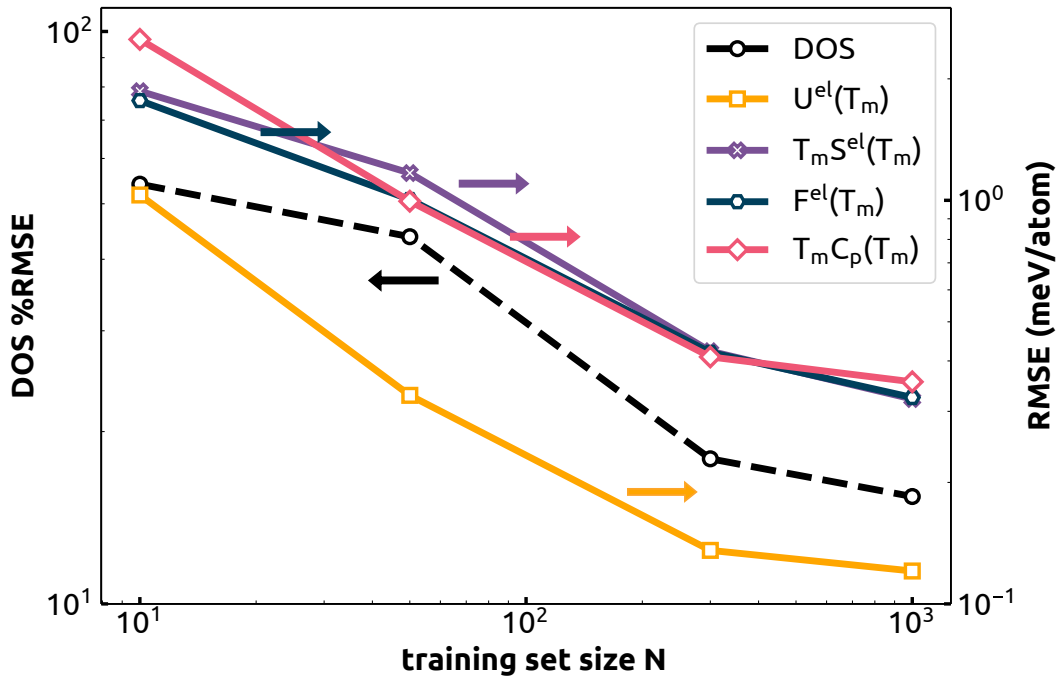


Figure 4.3: Evolution of the prediction errors in the validation set as a function of the training set size for the pointwise representation of the ML DOS (in black), as well as for quantities derived from the DOS prediction for thermal excitations computed at T_m = 1700K (namely, the band energy $U^{\mathrm{el}}(T_m)$, the electronic entropy term $T_m S^{\mathrm{el}}(T_m)$, the free energy $F^{\mathrm{el}}(T_m)$, and the heat capacity, written in energy units, $T_m C_p(T_m)$).The reference DOS is generated with a Gaussian broadening of 0.1eV. The arrows point to the axis on which the errors can be read.

excluding those pertaining to liquid-vacuum and solid-vacuum interfaces due to their need for more careful band alignment. We complement it with 123 independent structures extracted from liquid and solid trajectories at the melting temperature. We use the radial cutoff $r_0 = 6$Å and an atomic density smoothing $\sigma_{at} = 0.45$ for the SOAP features. The active set contains 15000 environments selected by FPS out of the ≈ 127000 that are present in the training set. We determine the regression weights $\mathbf{x}_M$ using a regularization parameter that is optimized by a 10-fold cross-validation scheme, in order to ensure the model is not in the over-fitting regime.

The learning curves, computed by reporting errors on a fixed test set of the predictions of models trained on an increasing fraction of the remaining 1000 structures, are shown in Figure 4.3. The figure shows both the error on the DOS, computed as the integrated root mean square error (RMSE) of the ML DOS and DFT DOS normalized by the integrated standard deviation of the reference DFT DOS, as well as errors for the quantities in Eqs. (4.1-4.4), computed on the predicted DOS and checked against those obtained from the reference DFT curve. Learning curves are not saturating, indicating that a more accurate model could be obtained, if needed, by increasing further the train set size. In practice, this model is sufficiently accurate: even though the normalized error on the DOS is large (%RMSE=14.71% for the largest train set size), this translates into sub-meV errors for the key properties at the melting temperature T_m = 1700K. For the band energy $U^{\mathrm{el}}(T_m)$: %RMSE=3.30% and RMSE=0.12meV/atom; for the entropy $T_m S^{\mathrm{el}}(T_m)$: %RMSE=5.81% and RMSE=0.32meV/atom; for the free energy $F^{\mathrm{el}}(T_m)$: %RMSE=9.04% and RMSE=0.32meV/atom and for the heat capacity $T_m C_p^{\mathrm{el}}(T_m)$: %RMSE=4.25% and RMSE=0.36 meV/atom.

4.2.5 Sampling and thermodynamic integration

To compute finite temperature properties we perform different kinds of standard and accelerated molecular dynamics simulations. Unless otherwise specified, all simulations use a timestep of 2 fs, with a BAOAB integrator[243]. Efficient constant-temperature sampling is achieved by combining stochastic velocity rescaling[244] and a colored-noise Langevin thermostat[245], as implemented in i-PI [229]. Energies and forces are computed using the n2p2 [246] package interfaced with LAMMPS [247]. In constant-pressure simulations, the pressure is controlled with the Bussi-Zykova-Parrinello barostat[196, 248]. The time constants parameters of the barostat and its thermostat are set to 225 fs and 100 fs respectively. To compute self-diffusion coefficients and viscosity we applied weak global velocity rescaling thermostat [244] with a 1 ps time constant, which improves statistical sampling without affecting dynamical properties. To shrink the statistical error on computing the bulk modulus, the heat capacity and the stability of defects, we run replica exchange molecular dynam-

ics (REMD)[227, 249, 250] with a exchange time of 40 fs. Examples of simulations, and the complete set of parameters chosen for interface pinning and metadynamics simulations is provided as commented input files in [230].

4.3 Applications

After having discussed the construction of the machine-learning models we use, and the details of the reference calculations, we now present results that can be obtained when applying them to the prediction of the atomic-scale properties of elemental Ni. We first validate the model by comparing its predictions with explicit density-functional calculations, and then proceed to compute a large number of finite-temperature properties, for which we compare with experimental data and/or previous literature results. We also use an EAM potential[226] to gauge the typical accuracy of a well-established empirical model, and to contrast it with that of a DFT-trained ML scheme. Whenever we compare two computational schemes, we use exactly the same simulation protocol, to ensure that any discrepancy is due to the potential energy surface, and not to finite size effects or other simulation details.

4.3.1 Validation of the potential

To provide a first benchmark of the accuracy of the NNP we predict a few simple, static-lattice properties that can be readily recomputed by DFT. We present bulk properties, defects and interfacial energetics. Most of these quantities are explicitly associated with structures that are included in the training set. For this reason, these tests serve more to demonstrate how the training error is reflected on the properties of interest, rather than to assess the transferability of the NN.

	$\Delta E\|_{fcc}$/(meV/at.)			a_0/Å			
	NNP	DFT	EAM	NNP	DFT	EAM	Exp.
fcc	-	-	-	3.5168	3.5175	3.5200	3.524
hcp	20.8	21.3	22.2	2.4873	2.4801	2.4819	
(c_0)				4.0829	4.0971	4.1048	
bcc	98.3	98.0	67.4	2.7968	2.7962	2.7687	

Table 4.2: The atomic bulk energies of hcp and bcc ideal crystalline structures with respect to the fcc bulk equilibrated at 0K, as well as the equilibrium lattice parameters. Experiments are taken from[251] where the measurements were carried out at 20°C.

Structure and stability of fcc, hcp and bcc phases

The stable structure for crystalline nickel at room temperature and pressure is *fcc*. Higher-energy, meta-stable phases, however, can play a role in different portions of the phase diagram, in the presence of defects, or just to increase the transferability of the NNP. Table 4.2 shows the 0K lattice energy of *bcc* and *hcp* configurations relative to the *fcc* ground state, as well as the relaxed lattice parameters. The sub-meV accuracy of the NN is consistent with the overall

	NNP	EAM	DFT	Exp.
B/GPa	204	180	205	183
B'	4.3	4.6	4.7	–
C_{11}/GPa	275	236	277	243
C_{12}/GPa	167	154	169	153
C_{44}/GPa	130	127	133	128

Table 4.3: Bulk modulus, bulk modulus derivative B' and elastic constants for the NNP, EAM potential, DFT compared with experimental results from Ref. [253].

test and train set errors; the large discrepancy observed for the EAM model for the *bcc* phase is unsurprising, given that the empirical potential is optimized for the stable phases of Ni. Lattice parameters are in excellent agreement with the DFT reference values.

Elastic constants and bulk modulus

The bulk modulus and the elastic constants characterise the response of a material to isotropic and anisotropic deformations. Together with structural properties such as the zero-temperature lattice constants they can be easily measured experimentally and do not require substantial computational resources to obtain from electronic structure calculations, making them good references for benchmarking. We compute the bulk modulus of *fcc* nickel and its derivative by evaluating the change in potential energy when introducing finite isotropic deformations (up to 5% of the equilibrium lattice parameter), and fitting the resulting energy-volume curve to a Birch-Murnaghan equation[252]:

$$\begin{aligned} E(V) = E_0 + \frac{9V_0B_0}{16}\Bigg\{\left[\left(\frac{V_0}{V}\right)^{\frac{2}{3}}-1\right]^3 B_0' + \\ + \left[\left(\frac{V_0}{V}\right)^{\frac{2}{3}}-1\right]^2\left[6-4\left(\frac{V_0}{V}\right)^{\frac{2}{3}}\right]\Bigg\} \end{aligned} \tag{4.5}$$

where E_0 is the minimum lattice energy, V_0 is the reference volume, B_0 is the bulk modulus, and B_0' is the derivative of the bulk modulus with respect to pressure.

For a cubic material the bulk modulus is also linked to the second order elastic constants by the expression:

$$B = \frac{1}{3}(C_{11} + 2C_{12}) \tag{4.6}$$

where the standard Voigt notation is being used for the indices. We estimate the elastic constants by examining the strain energy density for orthorhombic and monoclinic deformations

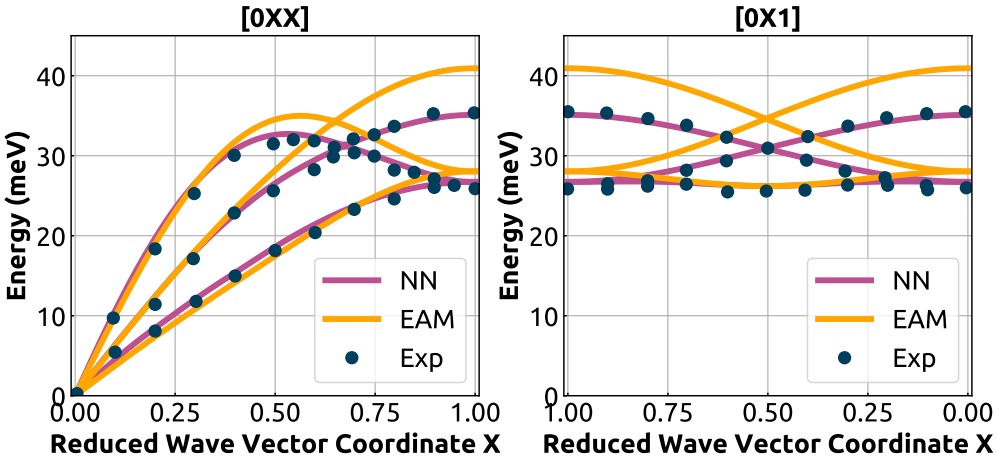


Figure 4.4: Phonon dispersion curves for the EAM potential (yellow), the NNP potential (purple) and experiment(blue dots). DFT results from Ref. [254] are indistinguishable from the experimental values on the scale of the figure.

which corresponds to strain tensors of the form:

$$E_{orth}(\delta)/V = \left\{ \begin{array}{ccc} \delta & 0 & 0 \\ 0 & -\delta & 0 \\ 0 & 0 & \frac{\delta^2}{1-\delta^2} \end{array} \right\} \tag{4.7a}$$

$$E_{mon}(\delta)/V = \left\{ \begin{array}{ccc} 0 & \frac{1}{2}\delta & 0 \\ \frac{1}{2}\delta & 0 & 0 \\ 0 & 0 & \frac{\delta^2}{4-\delta^2} \end{array} \right\} \tag{4.7b}$$

Both matrices define deformations which preserve the volume V of the examined system. The corresponding strain energy densities $\Delta E_{orth}(\delta)/V$ and $\Delta E_{mon}(\delta)/V$ are given by:

$$\Delta E_{orth}(\delta)/V = (E_{tot}(\delta) - E_0)/V = (C_{11} - C_{12})\delta^2 + \mathcal{O}(\delta^3) \tag{4.8a}$$

$$\Delta E_{mon}(\delta)/V = (E_{tot}(\delta) - E_0)/V = \frac{1}{2}C_{44}\delta^2 + \mathcal{O}(\delta^3), \tag{4.8b}$$

where $E_{tot}(\delta)$ denotes the total energy of the deformed system, E_0 is the ideal bulk energy or $E_{tot}(\delta = 0)$. We compute energies for values of $\delta \leq 10\%$, and estimate the elastic constants by fitting the resulting curves to Eq. (4.8). Results, shown in Table 4.3, indicate that the NN reproduces the DFT elastic constants with high accuracy (an error around 2%), and is consistent with previous results for single element bulk metals [43, 46, 224] which also report an error smaller than 4% between DFT and machine-learning potentials.

Phonons

Phonon dispersion curves describe the elastic response of the interatomic potential to a plane wave deformation of wavevector **q**, and can be measured by inelastic neutron or X-ray scattering. DFT has been shown to reproduce closely experimental phonon curves for pure Ni[254]. For this reason, we compare the NNP and the EAM potential with the experimental results. The phonon dispersion curves have been obtained with the small displacement method as implemented in the PHON package[255–257]. In the frame of this method, the position of each atom in the primitive cell is slightly distorted. The force constant matrix is constructed by computing forces acting on all the other atoms in the crystal, using the DFT equilibrium volume. This force constant matrix is used to compute the dynamical matrix at any chosen **q**-vector in the Brillouin zone, which is then diagonalized to yield the squares of the phonon frequencies. The resulting dispersion curves are shown in Fig. 4.4. NNP results are in excellent agreement with experiments and previous DFT calculations [254], while those obtained with the EAM show a deviation up to 20% for the longitudinal mode at the brillouin-zone edge.

Formation energies of point defects

At finite temperature any crystalline system contains an equilibrium concentration of point defects, such as vacancies and interstitial atoms. For a static lattice, and in the case in which bulk Ni is used as a reference state, the ab initio calculation of the single point defect formation energies can be achieved with low effort from the expression:

$$E^f_{def} = E_{def}(N_{def}) - \left[N_{def}/N_0\right] E_0 \tag{4.9}$$

where E_{def} is the final energy of the system with a defect after full ionic relaxation, N_{def} – number of atoms in the system with a defect, while N_0 and E_0 indicate the number of atoms and the energy of a reference supercell corresponding to ideal crystal.

	NNP	EAM	DFT	Experiment
E^f_{vac}, eV	1.52	1.57	1.51	1.4(900-1400K)
E^f_{int}, eV	4.17	4.01	4.2	

Table 4.4: Formation energies of single vacancy and interstitial in bulk Ni for NNP, EAM, DFT and experiment[258] .

We use a relatively large cell size ($3 \times 3 \times 3$ conventional unit cells, corresponding to 108 atoms) which ensures that the interaction of defects through periodic boundaries is negligible. Ionic positions have been fully relaxed using the BFGS algorithm [259–262]. As shown in Table 4.4,

the NNP is in excellent agreement with reference DFT calculations, and in semi-quantitative agreement with experimental data[258], which is however collected at finite temperature, the effect of which is discussed in Section 4.3.2.

Generalised stacking fault

The Generalised stacking fault (GSF) energy is an important property that is related to the response of a material to plastic deformation and fracture. The GSF reports on the energy cost associated with the slip of the crystal along a plane of atoms, with the geometric nature of the deformation being determined by the crystal lattice and symmetries. The only point along a GSF curve that can be probed experimentally is the one corresponding to an intrinsic stacking fault geometry. However it is possible to compute the full curve in simulations, by tilting the repeat vector of an ideal crystalline lattice in a slip plane while keeping all the atoms fixed[263]. The shift of PBCs creates a stacking fault. The deformed system is then relaxed along the direction orthogonal to the slip plane. The full GSF curve can be sampled by introducing larger and larger tilt angles. The GSF energy is defined as:

$$\gamma^{SF}(x,y) = \frac{E[N](x,y) - [N/N_0]E_0}{A_{xy}}, \tag{4.10}$$

where A_{xy} is the cross-section of the supercell. For reference DFT calculations we used an elongated supercell, with a 1x1 dimension along the fixed in-plane lattice vectors, and a 4-fold replication along the [111] direction to minimize interactions between the periodic images of the SF. Both the EAM and the NNP reproduce to excellent accuracy the curve computed with DFT (Fig. 4.5), with a slightly more pronounced overestimation of stable and unstable stacking fault energies by the EAM.

Rigid surface separation

The surface energy of solids controls many technologically-relevant phenomena such as fracture, morphological surface properties etc. Experimentally this property is affected by the presence of defects and impurities, and by surface reconstruction. Computationally, a rigid cleaving of the ideal bulk makes it possible to easily determine whether a potential provides a satisfactory description of the formation of a free surface.

The cleaving potential is computed by evaluating the energy of a bulk solid configuration, in which the lattice spacing between two planes is artificially increased by a separation d. Given the energy $E(N,d)$ of a supercell with N atoms and cross-section A_{xy}, the rigid-surface

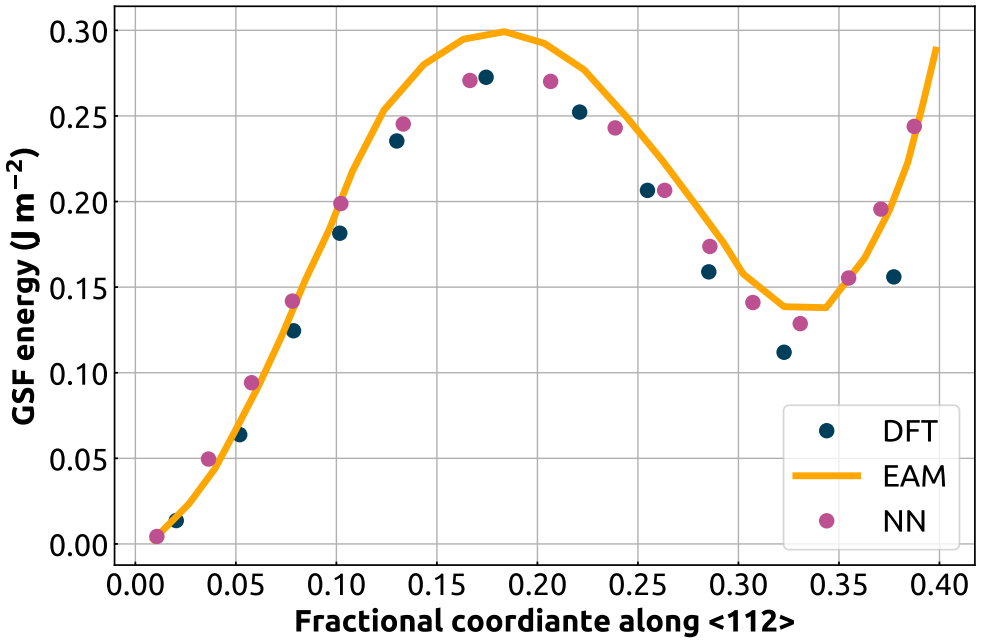


Figure 4.5: Generalized stacking fault curve for bulk Ni along the [112] direction computed using DFT (blue dots), EAM potential (yellow curve), and the present NN potential (purple dots).

Surfaces, mJ/m^2	NNP	EAM	DFT	Experiment
(110)	2468	2087	2440	2280
(001)	2351	1936	2337	2280
(111)	2004	1759	1995	2280

Table 4.5: The surface energy of different surface orientations for NNP, EAM, DFT and experiment. The experimental value is averaged over orientations[264].

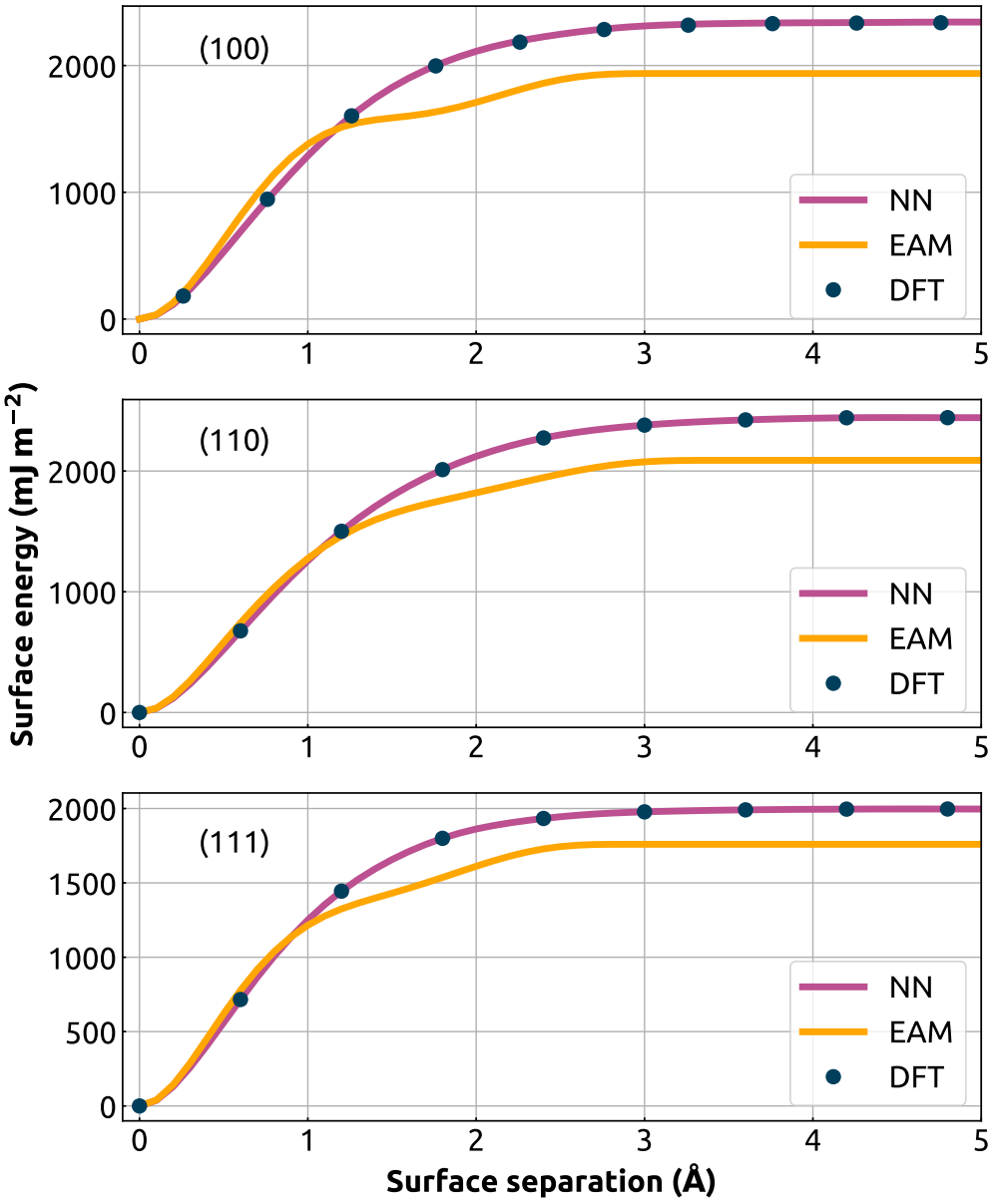


Figure 4.6: The energy vs rigid separation across different surface orientations for DFT (blue dots), EAM (yellow curves) and NNP (purple curves).

cleaving potential is defined as

$$\gamma^{surf}(d) = \frac{E(N,d) - [N/N_0]E_0}{2A_{xy}} \tag{4.11}$$

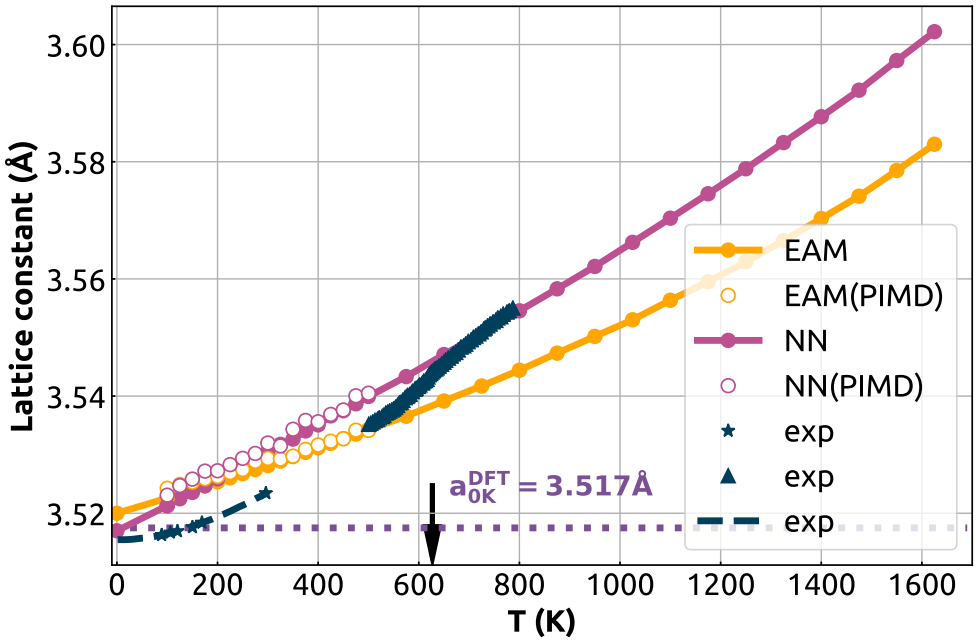
where E_0 is the energy of a reference bulk configuration with N_0 atoms. For our reference calculations, we consider supercells elongated along the (111), (001), and (110) directions, with 8 atomic layers in the direction orthogonal to the surface. The EAM potential captures correctly the order of surfaces stability (table 4.5), although with poor quantitative agreement with DFT, which matches well the experimental estimate[264] (which is an average over multiple orientations). Similar to what was observed for Al in Ref. [265], the EAM cleaving potential displays an unphysical step-like behavior.

4.3.2 Finite temperature properties

Benchmarks on static lattice calculations, such as those discussed in the previous Section, give confidence on the accuracy of the MLP, as they can be compared with little effort with reference DFT calculations. This Section, instead, focuses on properties that require the evaluation of thermodynamic averages at finite temperature. In the low-T regime, quantum fluctuations of the nuclei are also important, while at high temperature magnetic and electronic excitations also play a role in determining the thermophysical properties of Ni. Given that most of the simulations we report in this Section would be impractical when coupled to explicit quantum calculations, we cannot directly compare our results to the DFT reference. We do however compare with existing force fields and with experiments, even though we cannot disentangle the errors associated with the underlying electronic-structure approximations, and those stemming from the NN fit.

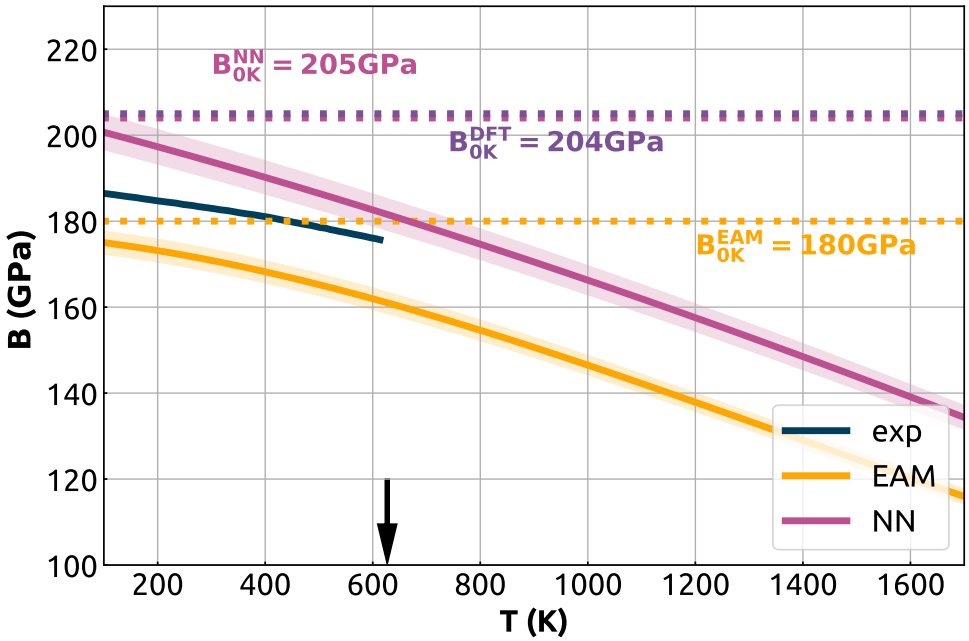
Structural and elastic properties at finite temperatures

We begin by revisiting the bulk properties of Ni incorporating the effect of fluctuations. The Debye temperature of Nickel is around 400 K, and so one can expect a significant effect associated with quantum fluctuations of the nuclei up to and above room temperature. For this reason, we perform simulations using both classical molecular dynamics (that are valid in the high-temperature limit) and with path integral molecular dynamics (PIMD)[31–33] (that incorporate nuclear quantum effects in the low temperature limit). To accelerate convergence of PIMD simulations, we use a finite-difference integrator [269] for the fourth-order Suzuki-Chin factorization of the path integral partition function, [270, 271] as implemented in i-PI[229], that yields converged observables down to about 100K with only four replicas.



(a) Lattice parameter of Ni as a function of temperature for EAM and NN computed classically (filled circles) and with PIMD (empty circles). Experimental data is presented by triangles[266], starts and dashed curve[267].

Black arrow points at the Curie point for Ni.



(b) Bulk modulus of Ni as a function of temperature for EAM and NN. Shaded areas represent corresponding statistical errors. Blue curve indicates experimental data from Ref. [268]. Black arrow points at the Curie point for Ni.

Figure 4.7: Finite-temperature structural and elastic properties of pure Ni, comparing simulations and experiments.

The top panel of Fig. 4.7a shows the behavior of the lattice parameter with temperature, as obtained from REMD simulations of a box of 108 atoms, run for approximately 150ps at each temperature with a possibility to swap between replicas every 40fs. The thermal expansion is similar between the NN and EAM simulations, and both are in good agreement with experiments [266, 267]. Both the EAM and the NN cannot capture the effects of the ferromagnetic transition: the EAM is fitted to low-temperature structural parameters and underestimates the lattice parameter in the high-T regime, while the NN, that is fitted to a non-polarized DFT reference, shows a better agreement above the Curie temperature, and overestimates the lattice parameter in the ferromagnetic phase. Quantum effects on the lattice parameters are small even below the Debye temperature, which justifies using a classical expression to estimate the bulk modulus in this temperature range by considering the volume fluctuations at constant pressure:

$$B(T) = \frac{\langle V \rangle k_B T}{\left\langle V^2 \right\rangle - \langle V \rangle^2}. \tag{4.12}$$

As shown in Fig. 4.7b, the bulk modulus shows a substantial dependency on temperature, with EAM and NN bracketing experimental observations, and exhibiting a similar trend up to the melting point.

Heat capacity

The constant-pressure heat capacity C_p of a ferromagnetic metal such as Nickel is a very challenging quantity for modelling, because it contains features that are associated with excitations on different degrees of freedom and energy scales [46]. As shown in Fig. 4.8, the experimental curve shows a low-temperature limit which is dominated by quantum nuclear effects, tending to zero at low temperature, a peak around the Curie temperature, associated with the ferromagnetic phase transition, and a pronounced increase above the Dulong-Petit limit at high temperature, that is linked to thermal expansion, anharmonic fluctuations, but also to electronic excitations, that make up for half of the deviation at the melting point. Thus, a very accurate interatomic potential is not sufficient to accurately predict the full C_p curve. Within the adiabatic approximation, ionic, electronic and magnetic contributions to the heat capacity could be described separately, provided one can treat them explicitly, as one would do in ab initio molecular dynamics. Here we present a first application of an integrated ML model that incorporates properties beyond the interatomic potential, to have access to contributions beyond those controlled by ionic fluctuations. We focus in particular on the electronic effects, that can be estimated, within a rigid band approximation, from the knowledge of the electron density of states (DOS). The contribution to the internal energy associated with electronic

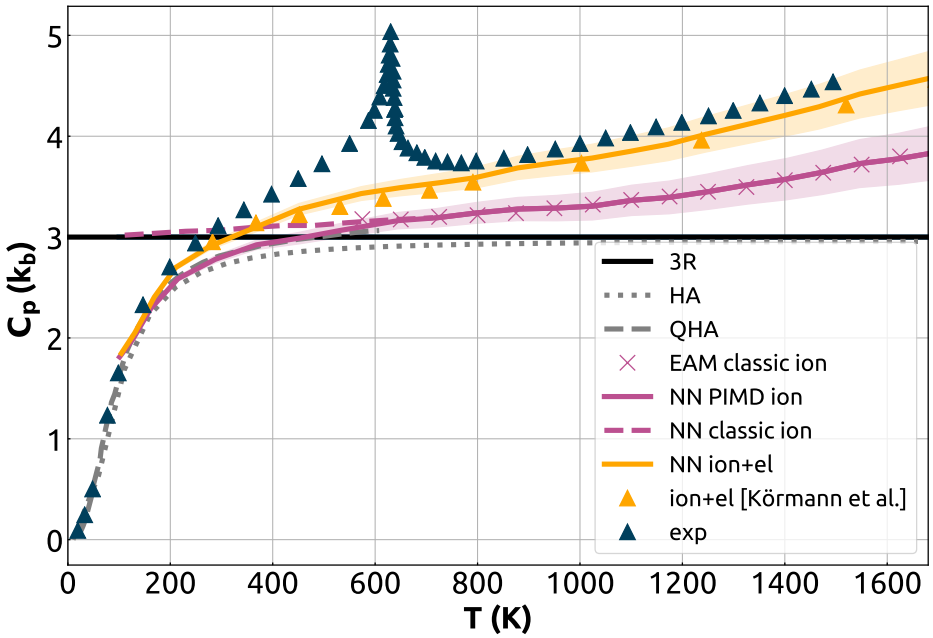


Figure 4.8: Constant pressure heat capacity C_p as a function of temperature. Triangles indicate experimental observations, as well as electronic and vibrational contributions computed by first-principles calculations in Ref. [272], combining quasiharmonic approximation results at low-temperature, and classical ab initio MD at high temperature. Gray lines indicate the heat capacity computed within harmonic and quasiharmonic approximations using the NNP. Solid lines represent the heat capacity computed in this article with PIMD, and including electronic corrections based on a ML model of the DOS. Crosses show the heat capacity computed using an EAM, and without including electronic corrections.

excitations can be computed as in Eq. (4.2). We used a recently-introduced machine learning model of the DOS [239], trained as discussed in Section 4.2.4, to predict the electronic density of states (DOS) for every frame of the REMD simulation, which was then used to estimate the electronic energy U_{DOS} and, by finite differences, the electronic contribution to C_p.

In Figure 4.8 we show the heat capacity as a function of temperature computed from classical molecular dynamics (purple dashed line) using the fluctuation formula

$$C_p = \frac{\langle H^2 \rangle - \langle H \rangle^2}{k_B T^2} \tag{4.13}$$

that deviates dramatically from the experimental curve at low temperature. Results from PIMD, that are evaluated with a fourth order double virial operator heat capacity estimator[273], (purple solid line) display the correct low-temperature behavior, but underestimate by ≈20% the experimental observations at high temperature. The discrepancy (which is also observed in explicit first-principles molecular dynamics[272] and in simulations that use the EAM) is due to electronic contributions, and indeed the curve that incorporates these using the ML model of the DOS (yellow solid line) are in almost perfect agreement with high-temperature measurements, and with previous results obtained, with heroic efforts, using density functional theory and quasi-harmonic simulations in the low-temperature regime [272]. Incorporating quantum nuclei and electronic fluctuations lead to remarkably good agreement with experiments, except for the region around the Curie temperature, where magnetic excitations become important. Even though we do not include them in this model, adding a description of magnetism constitutes an interesting direction for future studies.

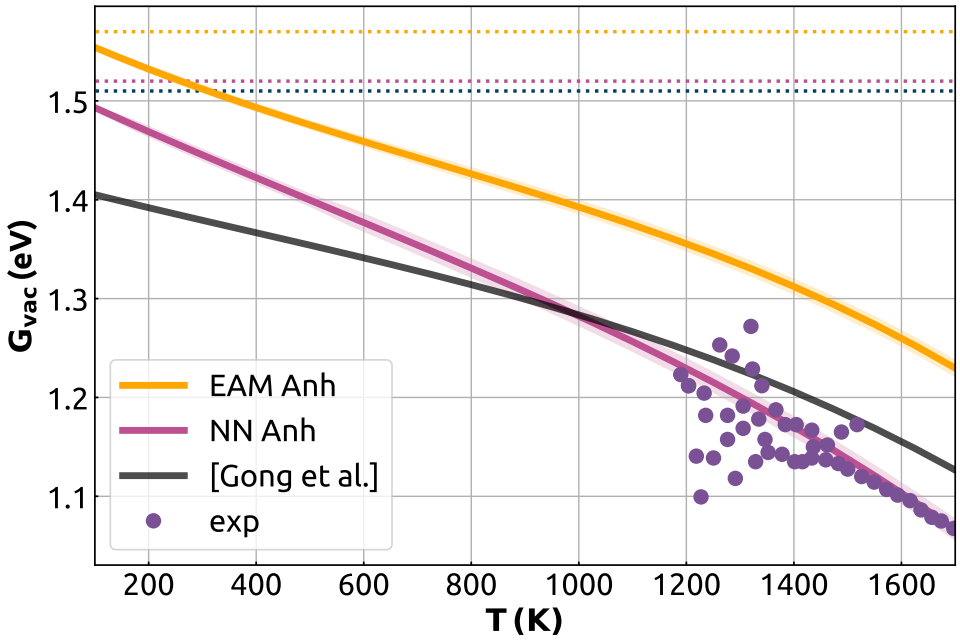


Figure 4.9: The fully anharmonic Gibbs free energy $G(p,T)$ of a single vacancy in fcc nickel obtained with thermodynamic integration. Curves are shown together with the potential energy difference at 0 K. Results for EAM and NN are compared with the DFT curve reported in Ref. [223] based on *ab initio* calculations (we reproduce the curve that does not include electronic or magnetic excitations), and available experimental data[274–277]. Dashed lines indicate the level of 0K energy of formation for NN, EAM and DFT calculated in this work.

Stability of defects

Finite-temperature and quantum fluctuations also affect the stability of defects. We estimate their contribution using thermodynamic integration (TI) [33, 278, 279] that makes it possible to estimate the absolute free energy of a thermodynamic state by a sequence of transformations, and use the values for two different states to estimate their relative stability. For instance, the Gibbs free energy of a single point defect can be easily found with an expression analogous to Eq. (4.9):

$$G_{\mathrm{d}} = G_{\mathrm{defect}} - \frac{N_{\mathrm{defect}}}{N_{\mathrm{perfect}}} G_{\mathrm{perfect}}, \tag{4.14}$$

where G_{defect} and G_{perfect} refer to the absolute free energies of two supercells, one of which includes the defect.

In Section 3.3.1 we describe a thermodynamic path along which the Gibbs free energy of defects can be computed. Since this topic is already covered, we omit the details of thermodynamic integration here.

As shown in Fig. 4.9, at high temperature the contribution from finite-temperature free-energy terms is sizable on the scale of the static defect formation energy (which is around 1.5 eV for the vacancy, see Table 4.4). Even though TI makes it possible to compute this correction with ab initio molecular dynamics [28, 207, 280], the use of a NN potential reduces the cost dramatically, making it feasible to estimate defect formation free energies for more complex defects and for materials with more diverse chemistry and crystallography.

Structure of the melt

One of the simplest and most direct diagnostics of the accuracy of an interatomic potential in the high-temperature limit involves computing the pair correlation function, $g(r) = \langle \delta(r - \mathbf{r}_{ij}) \rangle / (4\pi^2 r^2 \rho)$. As shown in Fig. 4.10, there is an excellent agreement between the NN, the EAM and the experimental results from neutron scattering data [281]. Although the pair correlation function provides only partial information on the structure, the near-perfect agreement indicates that both the EAM and the NN provide an excellent description of the liquid phase of Ni.

Self-diffusion coefficients and viscosity

The self-diffusion coefficient and the viscosity underlie mass transport and convection in the melt. They can be computed rather easily from constant-energy (or weakly-thermostatted)

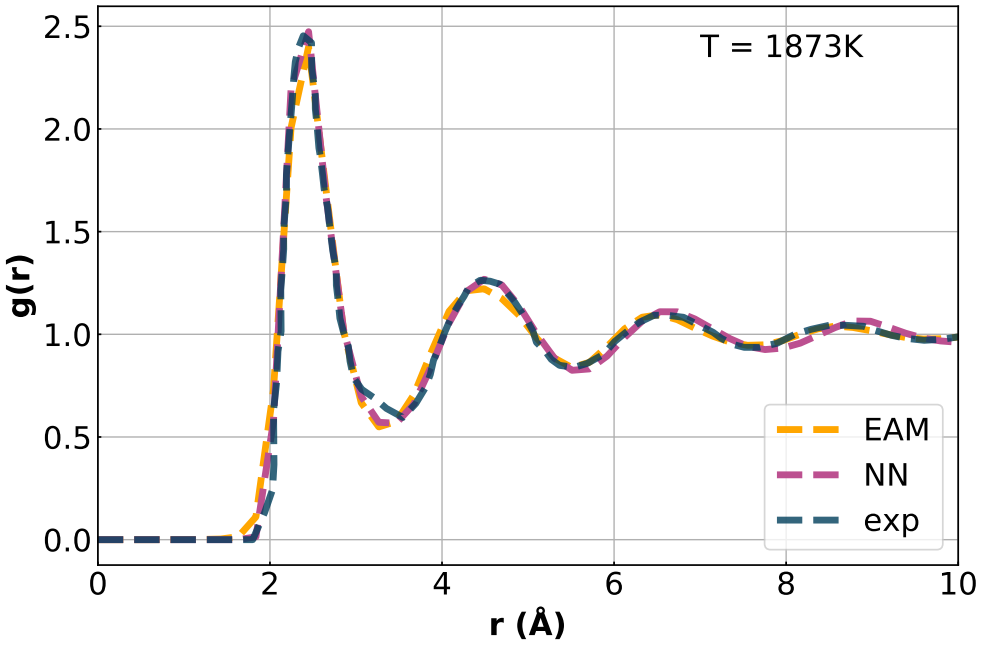


Figure 4.10: Radial distribution function $g(r)$ of liquid nickel. The experimental curve corresponds to Fourier transform of a structure factor obtained from neutron scattering for liquid nickel at $T = 1873K$ [281]. G(r)'s for EAM and NN models are computed from NVT trajectories at $V|_{P=0GPa}$ and $T = 1873K$.

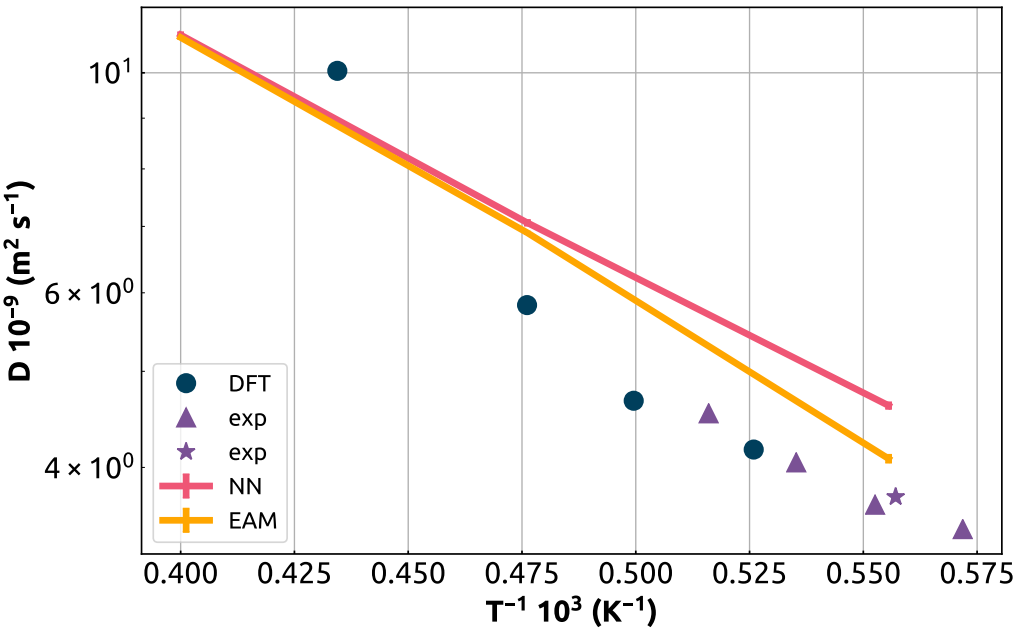


Figure 4.11: Self-diffusion coefficient of liquid nickel as a function of temperature. The triangles [282] and the star [283] indicate experimental measurements, dots indicate the result of AIMD simulations reported in Ref. [284]. NNP and EAM results are shown with statistical errorbars.

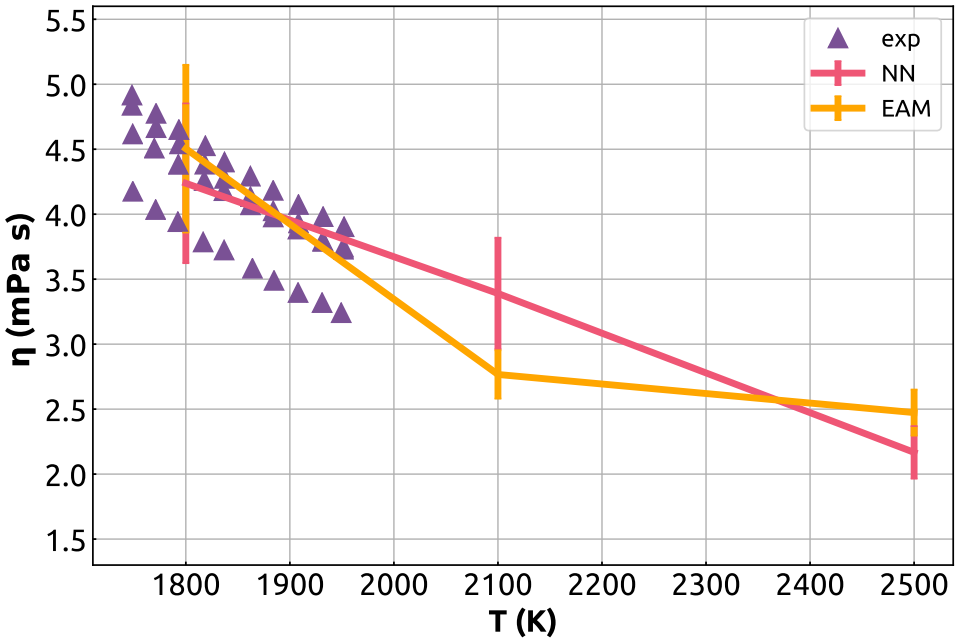


Figure 4.12: Shear viscosity of molten Ni as a function of temperature. The triangles indicate different sets of experiments collected in Ref.[285], while lines with errorbars correspond to NNP and EAM predictions.

molecular dynamics, evaluating the slope of the mean square displacement,

$$D_{\text{sim}} = \lim_{t\to\infty} \frac{\partial}{\partial t} \frac{\left\langle |\boldsymbol{r}(t) - \boldsymbol{r}(0)|^2 \right\rangle}{6N} \tag{4.15}$$

that we compute averaging 10 trajectories of 100-100-50(500-500-50)ps each for NN(EAM) simulations involving 108-256-2048 atoms respectively. The self-diffusion coefficient has a pronounced dependency on the system size which originates from hydrodynamic self-interaction through the periodic boundary conditions. Thus, comparing the results for a cubic simulation box of length L, the diffusion coefficient should be corrected for finite size effects[286, 287]:

$$D_0 = D_{\text{sim}} + 2.837297 k_B T/(6\pi\eta L) \tag{4.16}$$

where D_{sim} is the diffusion coefficient calculated in the simulation, k_B the Boltzmann constant, T the absolute temperature, and η the shear viscosity of the liquid. Thus, performing simulations at different system size at each temperature makes it possible to extract the viscosity as a fitting parameter of the equation (4.16) together with D_0. The diffusion coefficient and viscosity as a function of temperature are shown in Fig. 4.11 and Fig. 4.12, respectively. The predicted values for EAM and the NN potential agree with each other, and are in semi-quantitative agreement with experimental measurements.

Surface tension

The liquid-vapor surface energy γ_{lv} plays an important role in determining wetting and capillary forces, that are relevant e.g. for additive manufacturing. Contrary to solid-vapor surface energies – that can be reasonably estimated by single-point calculations – the liquid-vapor surface tension requires averaging over liquid configurations, and simulations size and time scale that are prohibitive for first-principles molecular dynamics. A practical simulation protocol involves simulating a planar liquid slab, with two free planar surfaces parallel to xy plane, and computing the integral across the slab of the normal and tangential components of the stress σ_n and σ_t [289, 290]

$$\gamma_{\text{lv}} = \frac{1}{2}\int_0^{L_z} [\sigma_{\text{n}}(z) - \sigma_{\text{t}}(z)]\,\mathrm{d}z \tag{4.17}$$

where L_z is the length of the simulation box. Given the slab geometry, this is equivalent to computing the mean value of the stress of the entire simulation box, using $\sigma_n = \langle\sigma_{zz}\rangle$ and $\sigma_t = \langle(\sigma_{xx} + \sigma_{yy})/2\rangle$. To evaluate γ_{lv}, we use a slab containing 927 atoms, with a square cross-section of $\sim 1000\text{Å}^2$ and ~ 10Å spacing between the surfaces, averaging over 400ps of molecular dynamics simulations. As shown in Fig. 4.13, there is a rather large discrepancy

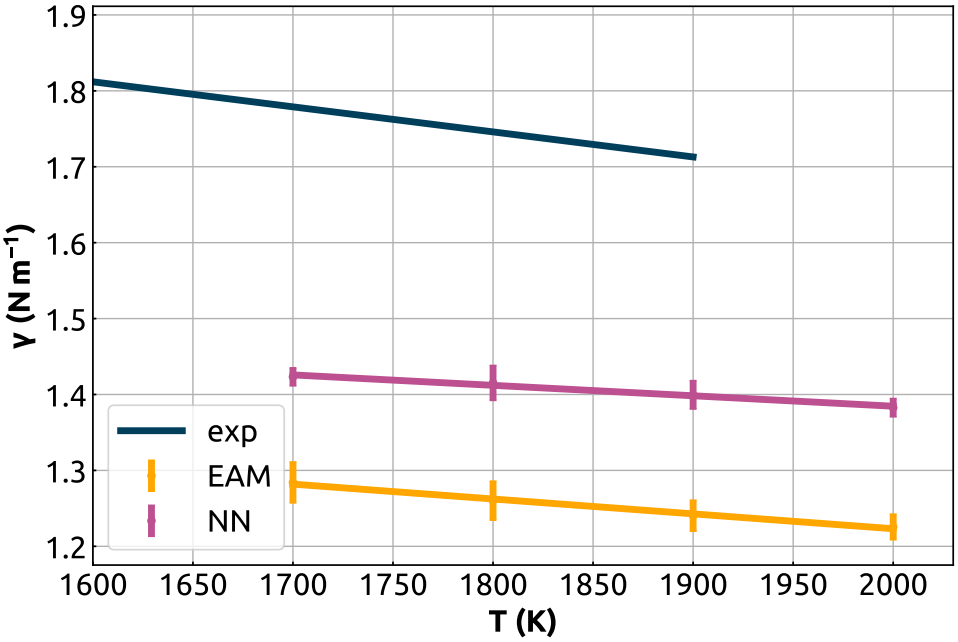


Figure 4.13: Surface tension of a planar interface as a function of temperature, as computed with NN and an EAM, compared with experimental data from Ref. [288].

between theoretical and experimental results for the surface tension, with experimental values being much closer to the solid-liquid interface energy. The NN potential reduces a discrepancy by a third, relative to the EAM, but is still 20% below the measured value at T_m. The observed underestimation of surface tension at temperature T_m aligns with the corresponding underestimation of T_m itself, a deviation that potentially stems from reference calculations using the PBE functional.

Melting point and solid-liquid entropy

Having separately characterized the properties of liquid and solid Ni at finite temperature, we can now turn to the determination of the relationship between the two phases and their interaction. We begin by characterizing the relative stability of the two phases and identifying their coexistence temperature. In order to achieve this, we employ the interface pinning (IP) method[210] (refer to Section 3.3.2). This method operates by applying a harmonic bias potential to a two-phase system, which is coupled to an order-parameter Φ that distinguishes between the two phases of interest. The Gibbs free energy difference between the phases is determined by the average force that the pinning potential exerts on the system. As an

order parameter to differentiate between solid and liquid, we use the same collective variable discussed in Ref. [292], that uses cubic harmonics to identify environments that are *fcc*-like and distinguish them from those that are liquid-like.

We performed several simulations at various temperatures to identify the temperature dependence of the chemical potential $\Delta\mu_{\mathrm{sl}}$ defined by Eq. 3.6. The temperature at which $\Delta\mu_{\mathrm{sl}} = 0$ identifies the melting point T_m, and the slope is equal to the entropy of melting.

As shown in Figure 4.14, the computed melting points for EAM and NNP are 1700K and 1695K respectively – only 2% off the experimental value which is equal to 1728K. The slope of the two curves is also in good agreement with that of the experimental curve, corresponding to ΔS_{sl} = -9.48 (EAM) and -11.5 (NN) mJ/K, to be compared with the experimental value of -10.11[293]/-10.22[294] mJ/K. The excellent agreement with experiments might be somewhat fortuitous, given that a recent DFT-based determination of the melting point of Ni reports T_m = 1570 K for calculations ignoring the magnetic contributions and using a GGA functional[295]. The discrepancy might be due to the use of pseudopotentials in our calculations, or in the accumulation of errors associated with the calculation of absolute free energies by a series of thermodynamic integrations in Ref. [295]: earlier results obtained by coexistence simulations yield a value of 1637 K,[296] which is closer to the one we find here.

Solid liquid interface free energy

The solid-liquid interface free energy plays a crucial role in determining the solidification behavior of materials, both in terms of controlling homogeneous nucleation, and in driving the formation of microstructure that, in turn, influences greatly the final materials properties. Measuring γ_{sl} is however notoriously difficult, which triggered the development of several different methods to estimate it from atomistic modeling[297–299]. Here we use an approach that was first introduced in Ref. [292], that relies on a bias potential to enable the reversible melting of a portion of an elongated simulation box (we use a box that is equivalent to $6 \times 6 \times 18$ *fcc* unit cells, with the interface aligned along the (100) direction), and determine the constant γ_{sl} term in Eq. (3.3.2) based on the free energy difference between a perfect solid and the configurations with two separate solid-liquid interfaces

$$\gamma_{\mathrm{sl}} = \frac{G_{s|l} - G_{s(l)}}{2A_{xy}}. \tag{4.18}$$

This expression is valid at $T = T_{\mathrm{m}}$, and for a planar interface – whereas in out-of-equilibrium conditions [300] or for a finite-size nucleus [301] further subtleties arise including the dependency of the surface excess on the precise location of the solid-liquid dividing surface.

meV/at.	$U^{el}(T_m)$	$T_m S^{el}(T_m)$	$\Delta F^{el}(T_m)$
solid	66.59 ± 0.07	155.37 ± 0.11	-88.78 ± 0.06
liquid	69.55 ± 0.08	157.76 ± 0.27	-88.21 ± 0.25
$\Delta_{\mathrm{liq-sol}}$	2.96 ± 0.15	2.39 ± 0.36	0.57 ± 0.29

Table 4.6: Average band energy, entropy contribution and free energy of solid and liquid phases at the melting temperature of Nickel $T_m = 1700K$, together with their difference. The values are computed from the ML DOS estimated for $\approx$ 15000 snapshots extracted from an NNP simulation of the liquid and solid phase at T_m. The uncertainties are derived by separately computing each quantity using a separate prediction of the calibrated DOS model, and computing the standard deviation of the end results.

We build the bias that compensates for the interface free energy in an adaptive, history-dependent way, using the well-tempered metadynamics [212, 213] technique as implemented in PLUMED[302, 303]. Bias is built from repulsive Gaussians that are 0.007 eV high, have a width equal to 5 CV units (the same Φ order parameter used for the pinning potential) and that are added every 0.5ps. The well-tempered metadynamics bias factor ($\gamma = 1 + \frac{\Delta T}{T}$) is chosen to be 90. Given that, at the melting point, the depth of the well associated with the fully solid and the fully liquid states are equal, a restraint is also applied to restrict sampling and prevent complete melting. A sample PLUMED input is provided with the data record that accompany this publication [230]. As shown in Fig. 4.15 the free energy shows a minimum at large Φ, corresponding to the fully-solid cell, and a plateau close to the restraining potential, corresponding to the presence of a solid/liquid interface. The free energy of this plateau makes it possible to estimate $\gamma_{\mathrm{sl}} = 0.272$ and $0.253\ \mathrm{Jm}^{-2}$ for the EAM and the NN potentials. The results are in a good agreement with previous calculations based on the capillary fluctuation method: $0.234\ \mathrm{Jm}^{-2}$ and $0.325\ \mathrm{Jm}^{-2}$ (calculated for the (100) surface in Refs. [304] and [305], respectively); $0.287\ \mathrm{Jm}^{-2}$ (averaged over different orientations, Ref. [306]).

Finite-electron-temperature effects

Electronic and magnetic fluctuations contribute substantially to the high-temperature thermophysics of nickel, as evidenced for instance by the heat capacity curve in Fig. 4.8. We can use the ML model of the DOS to compute the contributions to the free energy associated with electronic excitations, Eq. (4.1), averaged over trajectories of the bulk solid and liquid phases at temperatures around T_m. The difference $\Delta F^{\mathrm{el}}(T) = F_{\mathrm{l}}^{\mathrm{el}}(T) - F_{\mathrm{s}}^{\mathrm{el}}(T)$ could shift the chemical potential curve in Fig. 4.14, leading to a change in the predicted T_m. As shown in Table 4.6, even though the electronic excitations give a very substantial contribution to the free energy of Ni around T_m, the contributions from the solid and the molten phases cancel out almost perfectly, so that the impact on the melting temperature is less than 10K – in agreement with

the observations made in Ref. [295]. It should also be noted that converging these quantities to the level required to resolve the small difference between solid and liquid phases is far from trivial – both in terms of the statistical error over a MD trajectory, and in terms of the ML error computed following Ref. [241], by first generating a committee of predictions for the DOS, and then using each curve to obtain a separate estimation of $\Delta F^{\mathrm{el}}(T)$.

The averaged DOS over the solid and liquid phases at T_m, shown in Fig. 4.16, demonstrate that the cancellation between $F_{\mathrm{l}}^{\mathrm{el}}(T)$ and $F_{\mathrm{s}}^{\mathrm{el}}(T)$ is to be expected, given the small differences observed in the density of states, particularly in the vicinity of the Fermi level. Larger effects could appear in systems that, upon melting, undergo a substantial change in electronic properties, e.g. from semiconducting to metallic.

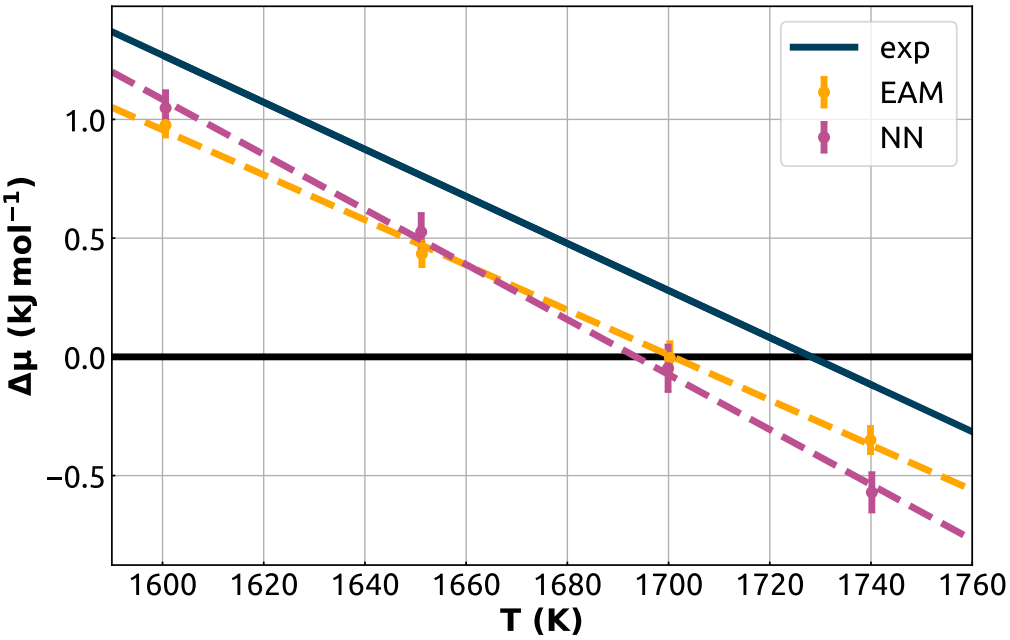


Figure 4.14: Chemical potential difference between solid and liquid phases of pure Ni as a function of temperature for EAM, NN and experiment[291]. The intersection of the yellow and purple lines with the black abscissa identifies the melting point for the corresponding potential.

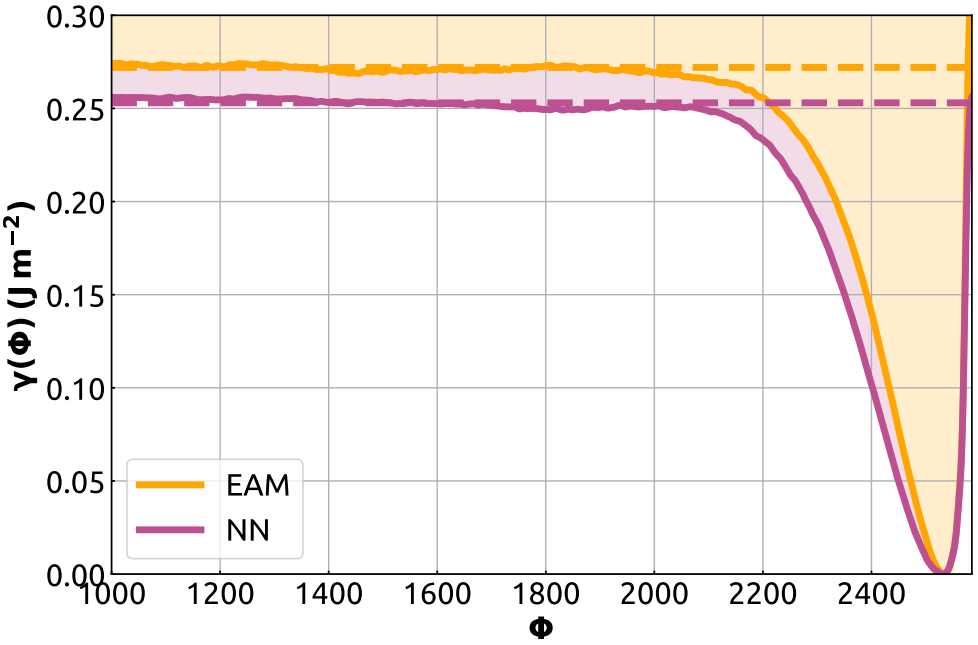


Figure 4.15: The curves show the converged free-energy profiles obtained by performing metadynamics simulations of a two-phase Ni system, and using a one-dimensional CV that measures the number of solid-like atoms. The curves are aligned with respect to the free-energy of the bulk solid state, and scaled by the surface area so that the depth of the well corresponds to the interfacial free energy.

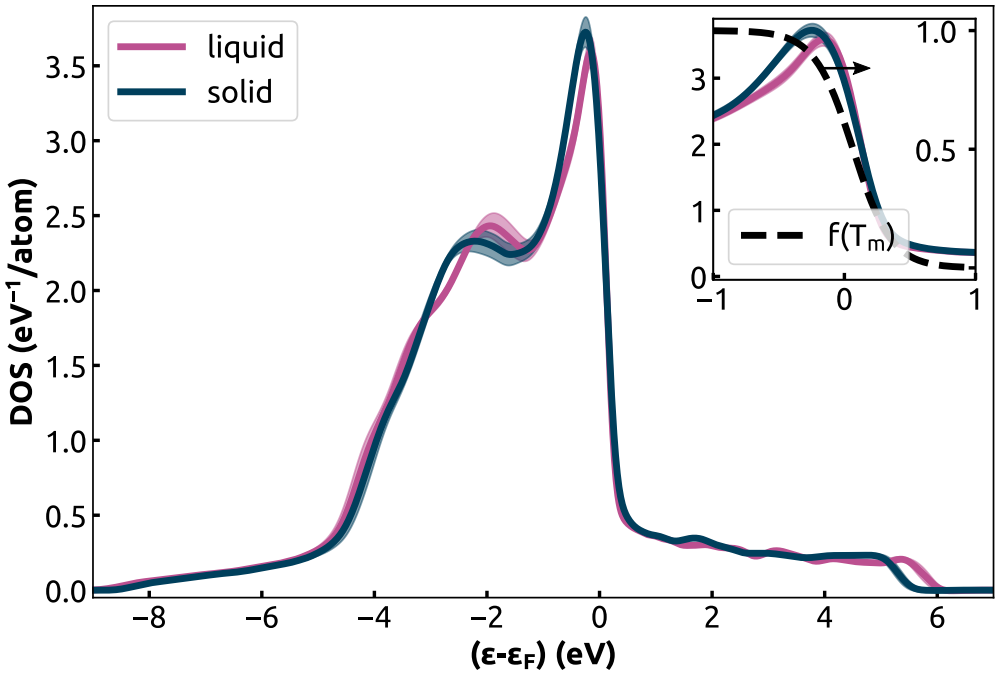


Figure 4.16: Average predicted DOS curve for the solid and liquid trajectories at the melting temperature $T_m = 1700K$. The shaded area represent the standard deviation of DOS(E) over the considered trajectories, and the inset shows a close-up of the region around the Fermi energy. The dashed curve represents the Fermi-Dirac function $f(\epsilon - \epsilon_F, T_m)$

5 High-entropy alloys[1]

5.1 Introduction

In the previous chapter, we demonstrated how machine learning models can be employed to account for thermal excitations of electrons and ions at high temperatures, as well as for quantum effects at low temperatures. The workflow presented does not assume a specific system of choice and can be applied to different systems, provided an interatomic potential is available.

However, constructing a potential for an arbitrary system with an arbitrary number of elements presents a significant challenge: most ML frameworks are limited to 4-5 species due to poor scaling and the inability to learn chemistry regularities, treating chemical elements as uncorrelated entities. To overcome these limitations, we adopt the approach first introduced in Ref. [97].

Using this scheme, we developed a general-purpose ML model for studying bulk high-entropy alloys (HEAs). The first step was to generate a dataset containing 25 transition metals, covering a wide range of stoichiometries. To ensure the dataset's diversity and insightfulness, we established a protocol similar to quasi-random structures.

The model trained on this dataset offers an accurate and transferable ML potential for 25 transition metals, along with an intuitive interpretation of their relationships. The model's validation indicates accuracy comparable to electronic-structure methods across a vast chemical space.

[1] This chapter is an adaptation of my contribution to Ref. [307], where I was responsible for dataset generation and curation, contributed to the model implementation, and carried out the tuning, training, and validation of the model.

We use this potential to reproduce computationally the seminal Cantor experiments on the decomposition of multi-element mixtures, and find a qualitative behavior in the affinity between different species that is consistent with well-known HEAs, allowing us to introduce a data-driven version of the Hume-Rothery rules to guide alloy design. We then study three alloy compositions: the prototypical Cantor alloy CoCrFeMnNi, the Mn→Mo counterpart with enhanced catalytic performance, and PdPtIrRuRh, another promising catalytic composition. In all cases, we observe a tendency for phase separation at low temperatures and short-range order indicative of the thermodynamic drive to de-mix in high-temperature conditions.

5.2 Computational details

We start with a concise summary of the details of the calculations we perform in this work, covering the reference electronic-structure calculations, the construction of the training set, the architecture of the ML model, as well as the details of the sampling protocol that we use for simulations in Sections 5.5 and 5.6. In the Appendix of Ref. [307] we provide representative examples of the typical simulation setup, and additional convergence tests.

5.2.1 Electronic-structure details

All the reference energies and forces are computed using density-functional theory (DFT), as implemented in the VASP code [308], with the PBESol exchange-correlation functional [309]. The core electrons are treated implicitly using projector augmented wave (PAW) potentials [310]. We choose conservative values for the convergence parameters of the electronic structure calculation (see the Appendix of Ref. [307] for details): the wave function is expanded in plane waves with a cutoff energy of 550 eV, and the Brillouin zone sampling uses a Γ centered Monkhorst-Pack scheme [311] with an interval between k-points along reciprocal lattice vector $0.04\,\pi$ Å^{-1}. Even though transition metals often exhibit magnetism, either in the pure phases or in alloys, we perform all our calculations without spin polarization. Even disregarding the fact that ML models that can deal with magnetism are still at a very early stage[312], one should consider that we aim to cover a broad chemical range, that includes materials which require different types of approaches to describe accurately their magnetic behavior - band magnetism within the local spin density approximation,[313] non-colinear magnetism,[314] Hubbard-U calculations[315], etc. This makes non-polarized calculations a reasonable approximation within the scope of the present work (see also the Appendix of Ref. [307]), even though this limits the accuracy of our reference and our model for magnetic systems - which for example would not be able to predict the stabilization of *bcc* iron over the close-packed polymorphs.

5.2.2 Training set construction

We generated an original dataset including 25 *d*-block elements, i.e. all transition metals excluding those that are not listed in Ref. [316] as relevant for HEAs (Tc, Cd, Re, Os, Hg). We generate a total of 25 thousand structures, following a protocol that ensures quasi-random sampling of this high dimensional phase space. We created four subsets of structures based on *bcc* and *fcc* lattices containing 36 or 48 atoms, respectively. All lattice parameters are defined by the average atomic volume of the elements in a structure and scaled up or down by up to 10% at random to simulate compression and expansion. The structures in the first three classes include from 3 to 8 randomly selected elements, and in the fourth – from 3 to 25. In the first class, we included only perfect crystal structures, with random compositions. For the three remaining classes, we shuffled atomic positions around their ideal lattice sites (using a Gaussian distribution of atomic displacement with a standard deviation of 0.2 Å in the second and fourth classes, and 0.5 Å in the third), to incorporate the information about finite positional deviations in crystals.

For every class of structures, we generated 100'000 random configurations and selected around 7'000 of the most diverse from every subset using Farthest Point Sampling (FPS)[176] in radial spectrum feature space.

5.2.3 Machine-learning model

We build ML models based on density-correlation representations, combining an atomic-energy baseline, ridge regression based on pair and 3-body correlation features, and a multi-layer perceptron[317] based on the 3-body features. Here we discuss briefly the functional form of the different terms, and outline the training strategy we followed. The atomic-energy baseline is simply a linear model that depends exclusively on the nature of the atom at the centre of each environment, a_i

$$V^{(\mathrm{aeb})}(A_i) = w^{(\mathrm{aeb})}_{a_i}. \tag{5.1}$$

Even though we train on atomization energies (and so the large dependency of the atomic energies on the details of the pseudopotentials is not an issue) we still find that $V^{(\mathrm{aeb})}$ captures a large fraction of the target variance, and facilitates learning. The second term we consider is a set of pair energies. We use 12 GTO basis functions, with a Gaussian width of 0.25Å, a cutoff of 6Å and radial scaling following Ref. [97]; we expand the density in spherical harmonics and in 12 radial function, enumerated by the n index, and obtained by orthogonalizing Gaussian-type orbitals that cover the range of distances up to the cutoff radius (see e.g. Ref. [170] for a precise definition). We use different weights depending on the nature of the two atoms, so that in

practice the contribution to the potential reads

$$V^{(2\text{ B})}(A_i) = \sum_{an} w^{(2\text{ B})}_{a_i an} \left\langle an \mid \overline{\rho_i^{\otimes 1}} \right\rangle. \tag{5.2}$$

The third term involves 3-body correlations (SOAP features), computed on top of alchemically-contracted density coefficients, with a linear model

$$V^{(3\text{ B})}(A_i) = \sum_{bnb'n'l} w^{(3\text{ B})}_{bnb'n'l} \left\langle bnb'n'l \mid \overline{\tilde{\rho}_i^{\otimes 2}} \right\rangle. \tag{5.3}$$

We use the same set of weights irrespective of the atom type, because in a 3-body descriptor the nature of the central atom is encoded in the density associated with the Gaussian at $r = 0$, so that the compression of the dependency of potentials on the central atom type is achieved implicitly and with the same contraction coefficients used for the neighbor density.

Finally, we include a non-linear term that takes the compressed power-spectrum as input, and feeds it into a Behler-Parrinello-style[231] multi-layer perceptron[317]. First, a linear filter projects the power-spectrum features into 80 input neurons, $\xi^{(0)}$, to which hyperbolic tangent activation functions are applied. A second linear layer combines the outputs of the neurons, feeding them to one hidden layer of the same size. Finally, the outputs are linearly combined to yield the atomic energy

$$\begin{aligned} \xi^{(0)}_q(A_i) &= \sum_{bnb'n'l} w^{(NN,0)}_{qbnb'n'l} \langle bnb'n'l | \overline{\tilde{\rho}_i^{\otimes 2}} \rangle, \\ V^{(\text{NN})}(A_i) &= F(\xi^{(0)}(A_i)) \end{aligned} \tag{5.4}$$

We use this simple neural network — built on top of the compressed power-spectrum features — because we want a simple and well-understood term that can incorporate non-linearity without exploding the design space, and because we want to show that our alchemical compression scheme can be readily applied to several well-established ML schemes. It is possible (and likely) that alternative frameworks, e.g. increasing further the body order, may allow for a better-performing model, but as we shall see this approach is sufficient to achieve state-of-the-art accuracy together with a stable and interpretable model.

The parameters of $V^{(3\text{B})}$ and $V^{(\text{NN})}$ implicitly include the alchemical coupling matrix $\mathbf{u}_{\text{alch}}$; for this reason, we optimize all models with gradient descent, relying on backpropagation as implemented in PyTorch[69]. A ridge penalty term is included on all weights, to reduce the risk of overfitting. We find that (possibly due to the presence of large linear components that contribute a quadratic term to the L^2 loss) a deterministic L-BFGS optimizer[318] performs much better than stochastic gradient descent.

5.2.4 Sampling details

Molecular dynamics (MD) is well-suited to describe structural relaxation of the atomic coordinates. However, long-range diffusion in the solid phase occurs through vacancies, and is too slow to be simulated explicitly by MD. To overcome this time scale problem, we use a combination of techniques to facilitate thorough sampling of atomic ordering. Our base protocol involves performing molecular-dynamics simulations in the constant-temperature/constant-pressure *NpT* ensemble[319]. We use a conservative time step of 2 fs, an isotropic barostat [196] with a time constant of 200 fs coupled to an optimal-sampling colored-noise thermostat[245], and an aggressive thermostat for the ions, alternating an optimal-sampling Langevin equation with a stochastic velocity rescaling[244] with a time constant of 10 fs. We accelerate sampling of the compositional (dis)order by performing Monte Carlo steps in which the nature of two atoms in the system is exchanged, with a Metropolis acceptance criterion[320]. We perform on average one exchange attempt per MD time step. Both the MD and the MC step conserve the Boltzmann distribution (except for a negligible finite time-step error), and so the combined MD/MC protocol is consistent with canonical sampling. In order to further accelerate sampling, we also use replica exchange molecular dynamics (REMD)[321] – a technique in which multiple trajectories at different temperatures are performed in parallel. Periodically, structures are exchanged between temperatures, using a Monte Carlo procedure that preserves the Boltzmann distribution for each thermodynamic state. The fact that each trajectory is brought through cycles of heating and annealing accelerates conformational sampling and reduces the correlation time of observables that are associated with activated events at low temperature. Unless otherwise specified, we use temperature replicas distributed according to a geometric progression between two extremal values T_{min} and T_{max}. For all MD/MC simulations we use the i-PI universal force engine[229], that includes an implementation of element exchange moves[322] and a flexible implementation of replica exchange[323].

5.3 Alchemical learning

As discussed in the methods section 2.4.4, the compression scheme in Eq. (2.11) is just one of the many approaches one could take to reduce the dimensionality of the density expansion coefficients. One of the appealing features of this specific implementation is that it can be interpreted relatively easily, and that it allows us to extract physical-chemical insights through an introspection of the model parameters and performance.

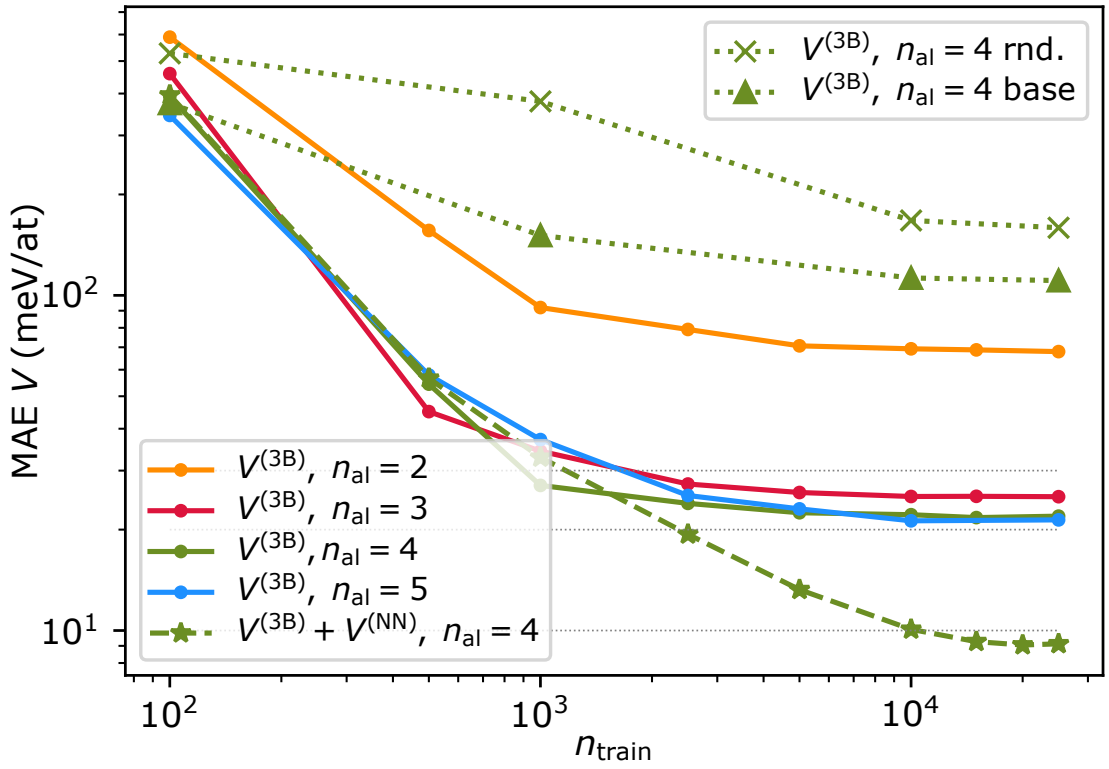


Figure 5.1: Learning curves for different models. Full lines correspond to models built using only $V^{(aeb)}$ and $V^{(3B)}$, with n_{alch} pseudo-elements (all optimized iteratively). The dotted green curves are obtained with a $\mathbf{u}_{alch}$ filled with uniform random numbers (rnd.) and with the weights we use as an initial guess for the optimized models (base), that are built based on physical priors following the scheme discussed in Ref. [97]. The dashed green line corresponds to a model that includes $V^{(aeb)}$ and $V^{(3B)}$, as well as the full set of pair potentials and a non-linear term built on top of the contracted power spectrum features $V^{(NN)}$.

5.3.1 Learning curve analysis

We begin by considering linear models based on contracted power-spectrum features, supplemented by an atomic energy baseline term, $V^{(\mathrm{aeb})} + V^{(\mathrm{3B})}$. We perform separate training exercises, using only energy as targets, and restricting the alchemical contraction to 2, 3, 4, 5 pseudoelements. For each model we compute learning curves by converging the loss at a given number of training structures n_{train}, then increase the train size and continue the optimization restarting from the previous weights. Given that the optimization procedure is rather demanding, we do not perform multiple train/test split, but use consistently the same shuffle with up to 25'000 structure used for training and a hold-out set containing 500 configurations used for testing. Even though the accuracy does depend slightly on the shuffle, and on the initialization of the weights, we find that the qualitative observations we present here are robust.

Figure 5.1 shows a behavior similar to that observed in Ref. [97] for an analogous exercise on the elpasolites data set[324]: at the smaller train set sizes a very aggressive compression is effective at obtaining a robust model, but with more training data the learning curves saturate. Increasing the number of pseudo-elements n_{alch} delays saturation, but the improvement going from $n_{\mathrm{alch}} = 3$ to $n_{\mathrm{alch}} = 4$ is negligible, and the learning curves for $n_{\mathrm{alch}} = 5$ sits almost exactly at the same value. This indicates that, from the point of view of 3-body interactions, 3-4 pseudo-elements are sufficient to saturate the descriptive power of a linear model. Note that the optimization of $\mathbf{u}_{\mathrm{alch}}$ is critical to achieve such efficient compression: a model that uses fixed, random values for the contraction weights, as well as one that uses a fixed, physically-inspired initialization of $\mathbf{u}_{\mathrm{alch}}$, lead to an order of magnitude increase in the saturation error, even with $n_{\mathrm{alch}} = 4$ (Fig. 5.1).

Given the saturation of $V^{(\mathrm{3B})}$, we proceed to increase the effective body-order of the potential adding a non-linear NN layer on top of the contracted power spectrum, $V^{(\mathrm{NN})}$, which introduces about 160′000 additional model parameters, mostly associated with the contraction of the $|\overline{\tilde{\rho}_i^{\otimes 2}}\rangle$ features to the 80 input features of the NN. Furthermore, we also include a non-compressed two-body potential $V^{(\mathrm{2B})}$, for which we also consider a slightly larger cutoff distance. This 2-body term, on its own, does not improve significantly the limiting accuracy of the model (reinforcing the notion that the alchemical contraction is converged) but we include it because it is inexpensive to compute, and has been shown in the past to lead to more stable models, whose performance degrade more gently in the extrapolative regime[325, 326]. Incorporating a non-linear term in the model allows to overcome the saturation of the learning curve (Fig. 5.1, dashed green line). The non-linear $n_{\mathrm{alch}} = 4$ model reaches a validation-set mean absolute error (MAE) below 10 meV/atom. We discuss further the accuracy of this model

(that we will refer to as the HEA25-4-NN) in Section 5.4.

5.3.2 A 3D periodic table for the transition metals

The alchemical coupling matrix associates to each of the physical elements a vector of size n_{alch}, that can be regarded as the "composition" of that element in terms of a set of pseudo-elements (Fig. 2.1b). Thus, different atomic species can be seen as points in a continuum space, and can be visualized as such to gain insights into the data-driven similarities that arise from the optimization of $\mathbf{u}_{\text{alch}}$ to achieve the most accurate regression of the target. To make the visualization independent on unitary transformations of the weight matrix, we perform a principal component analysis.

The eigenvalues of the covariance matrix indicate the magnitude of the various components (their *explained variance*), and provide another indication of the importance of successive increases in the dimensionality of the alchemical space. We observe a quick decrease of the explained variance, with the fourth component typically amounting to less than 2% of the variance (Fig. 5.2, inset). This confirms that the first three components provide sufficient descriptive power to capture the difference in behavior between transition metals. We can then look at how the d-block elements appear when projected along the top three principal components of $\mathbf{u}_{\text{alch}}$ (Fig. 5.2). We focus on the weights from the HEA25-4-NN model, but the qualitative features of the alchemical projections are similar also for other models in Fig. 5.1 (see the Appendix of Ref. [307]). The elements are arranged in a way that is strongly reminiscent of their placement in the d block: the third principal direction corresponds to the period, while the first two dimensions are associated with a semicircular arrangement, with the elements appearing in the same order as the columns in the conventional periodic table. Interestingly, this arrangement is reminiscent of that used for the d block in some of the alternative representations of the periodic table, such as the Benfey spiral[327]. It indicates that, from the point of view of the construction of an interatomic potential, zinc is closer to scandium then it is to the atoms in the middle of the transition metals block.

5.3.3 Alchemical interpolation

The elements we have not considered leave a clear gap in the arrangement of the alchemical coupling weights, and it is interesting to see how accurate a model that places rhenium and osmium between tungsten and iridium fares in predicting their properties without additional fitting.

We pick 60 structures from the hold-out set, containing distorted configurations with random

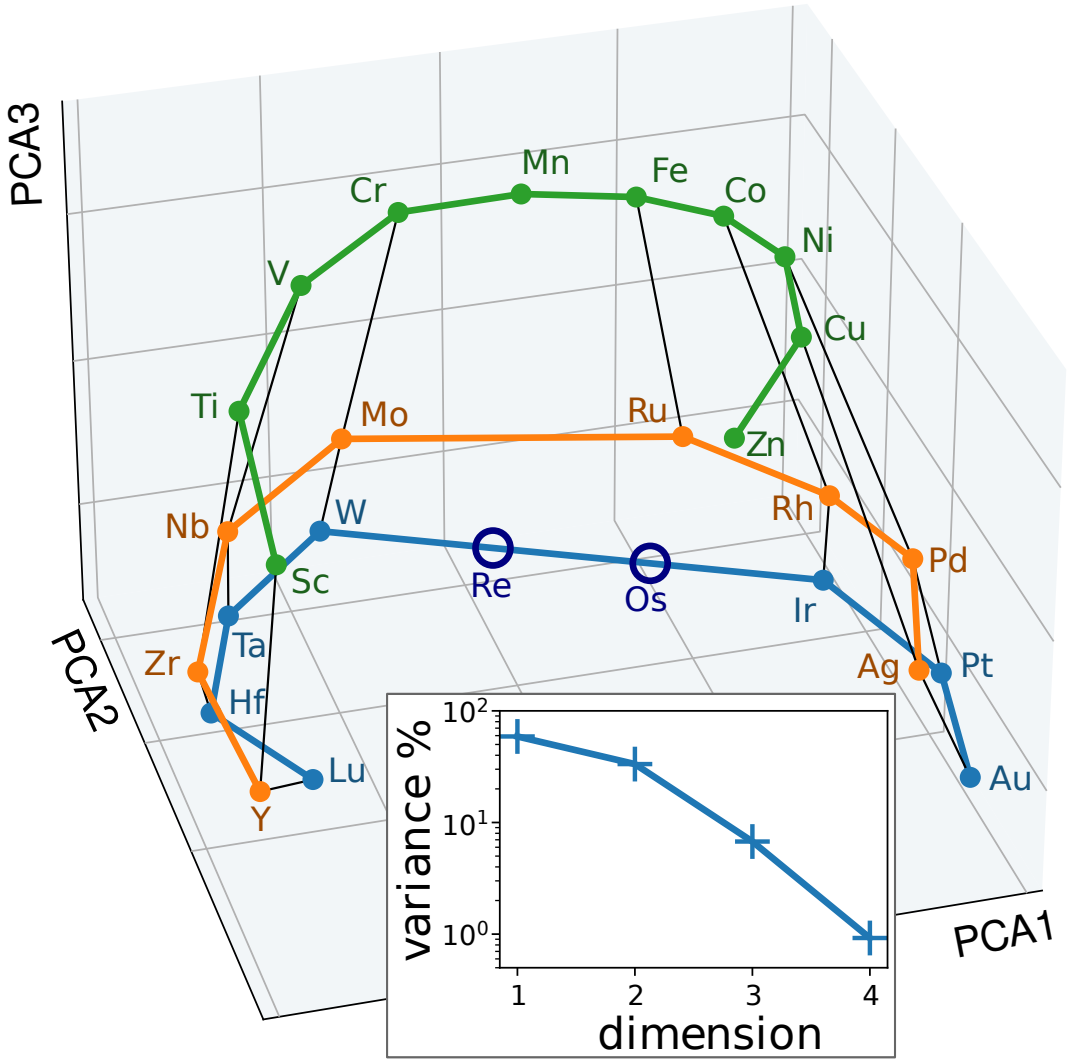


Figure 5.2: Top-3 principal components of the alchemical coupling matrix $\mathbf{u}_{\text{alch}}$ for the HEA25-4-NN model. The periods are highlighted with orange, blue and green lines, and the columns are indicated by black thin lines. Interpolated positions for Re and Os are indicated with empty circles. The inset shows the decay of the explained variance for the four principal components.

composition. The MAE for these structures when using the $n_{\text{alch}} = 4$ model using only $V^{(\text{aeb})}$ and $V^{(3B)}$ is 13 meV/atom. We then substitute some random atoms with Re and Os, without changing the positions, and re-compute their energies with analogous DFT settings.

We then build a model in which we simply take the parameters optimized for the 25-elements dataset, and complete them by adding atomic-energy baselines for Re and Os (obtained by training on the residual a two-parameter model that depends exclusively on the Re and Os content) and by adding pseudoelement weights that interpolate linearly between W and Ir (see Fig. 5.2):

$$u_{b\text{Re}} = \frac{2}{3}u_{b\text{W}} + \frac{1}{3}u_{b\text{Ir}}, \quad u_{b\text{Os}} = \frac{1}{3}u_{b\text{W}} + \frac{2}{3}u_{b\text{Ir}}. \tag{5.5}$$

The powerspectrum model weights are unchanged: we are effectively interpolating in pseudoelement space. The resulting model yields exactly the same predictions for structures that do not contain Os and Re, and has a MAE of only 24 meV/atom for the test structures that include the two species (see also the Appendix of Ref. [307]). The model is also sufficiently stable to run molecular dynamics simulations for Re and Os containing structures.

This example underscores the advantages of the interpretable functional form we use to implement alchemical dimensionality reduction. It also opens up the possibility of designing simulation protocols that include smooth "alchemical transformations", in a similar spirit as the framework pioneered by von Lilienfeld et al.[328]. For example, one could use thermodynamic integration to compute the change in chemical potential associated with an element substitution by running simulations with a mixed potential, in which the alchemical coupling weights are gradually transformed between the values associated with two elements.

5.4 Validation of the potential

We now assess the accuracy and stability of the model we use in the rest of this work, which combines a 4-pseudoelement contraction of the powerspectrum with a multi-layer perceptron. We aim to provide benchmarks that are easy to reproduce, but that reflect the performance of the model in relevant simulation tasks, and we envisage that any comparative study would include most of these and not only cross-validation statistics. To contextualize and provide a reference scale for our results, we report in the Appendix of Ref. [307] similar validation results for the general-purpose, universal graph neural network M3GNet[329]. In all cases HEA25-4-NN, which admittedly has a narrower scope of applicability, outperforms M3GNet by a large margin.

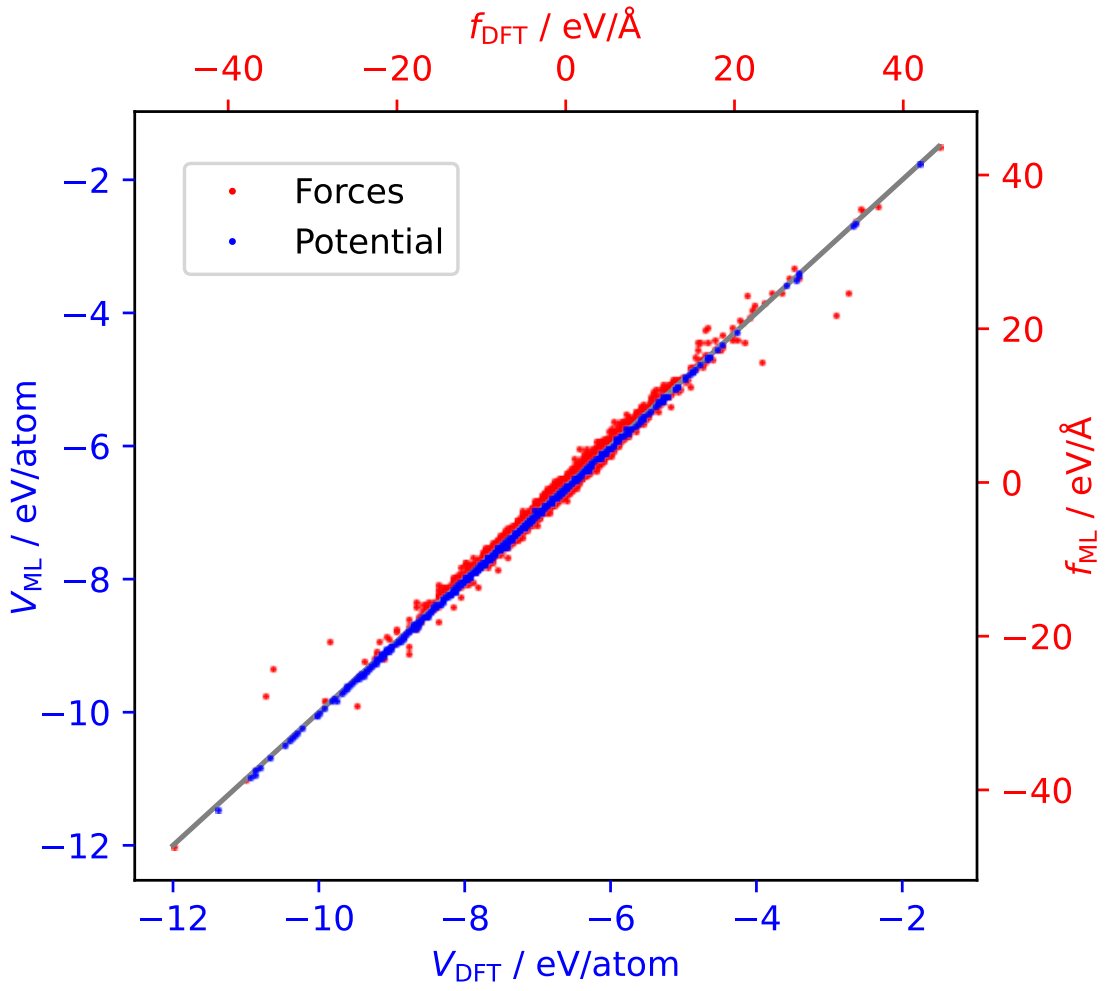


Figure 5.3: Parity plot between reference energy and forces and the values computed with the HEA25-4-NN model, for a hold-out set of 500 structures, randomly selected from the training set. Energy error: 10 meV/atom mean absolute errror (MAE), 14 meV/atom root mean square error (RMSE), Force error: 190 meV/Å MAE, 280 meV/Å RMSE.

5.4.1 Hold-out validation of the HEA25-4-NN model

We train the HEA25-4-NN potential by progressively increasing the train set size, until we run the final optimization on 25'000 structures, including forces for 2'000 of them. We hold out 500 structures and use them for validation. The parity plot between targets and predictions demonstrates the accuracy of the model (Fig. 5.3), which is remarkable given the diversity of the dataset, that contains random combinations of up to 25 elements, and highly-distorted structures.

5.4.2 Binary convex hulls

Even though the HEA25-4-NN is clearly geared towards multi-component simulations, it is important that it also provides reasonable results for simpler compositions, as these may appear spontaneously when complex alloys de-mix and form precipitates. We collect 1438 binary intermetallic structures out of more than 146k crystal structures from the Materials Project database[330], and re-compute their energies with single-point calculations using our DFT setup, as well as with the HEA25-4-NN model. We discard 23 structures for which our DFT calculations did not converge and 10 that correspond to configurations that are too dissimilar from the bulk structures we consider here (see Appendix of Ref. [307]). For the remaining structures, the MAE error for the cohesive energy is 62 meV/at. and for the formation energies is 63 meV/at, which is higher than the cross-validation error, but still remarkably accurate for extrapolative predictions. It is worth noting that the MAE discrepancy between our DFT calculations and those saved in the MP records is 65 meV/at.; this is due to the significant difference in the details of the electronic structure calculations, e.g. the use of Hubbard U corrections for some structures in the MP protocol, and neglect of spin polarization in ours. This observation underscores that the details of the electronic structure calculations can have an impact comparable to the accuracy of our ML model. We then use this data to compute binary convex-hull diagrams for all element pairs. In Fig. 5.4 we show a representative example for the Ti–Pt system. The overall shape of the hull is usually well-reproduced, but often HEA25-4-NN predicts different stable polymorphs than DFT, and/or mis-predicts the stability of certain compositions (as it is the case for $TiPt_2$ in the figure). However, these qualitative errors are usually associated with situations in which a small energy shift can bring a composition above the hull boundary, and even in a fully ab initio study it would not be possible to determine conclusively its thermodynamic stability. The full list of hulls is included in the Appendix of Ref. [307]. Fig. 5.4 also shows an overview of the accuracy of the prediction of formation energies for all phases (stable and unstable) as a function of composition. Errors are not uniform: some elements such as Mn, that have the tendency of

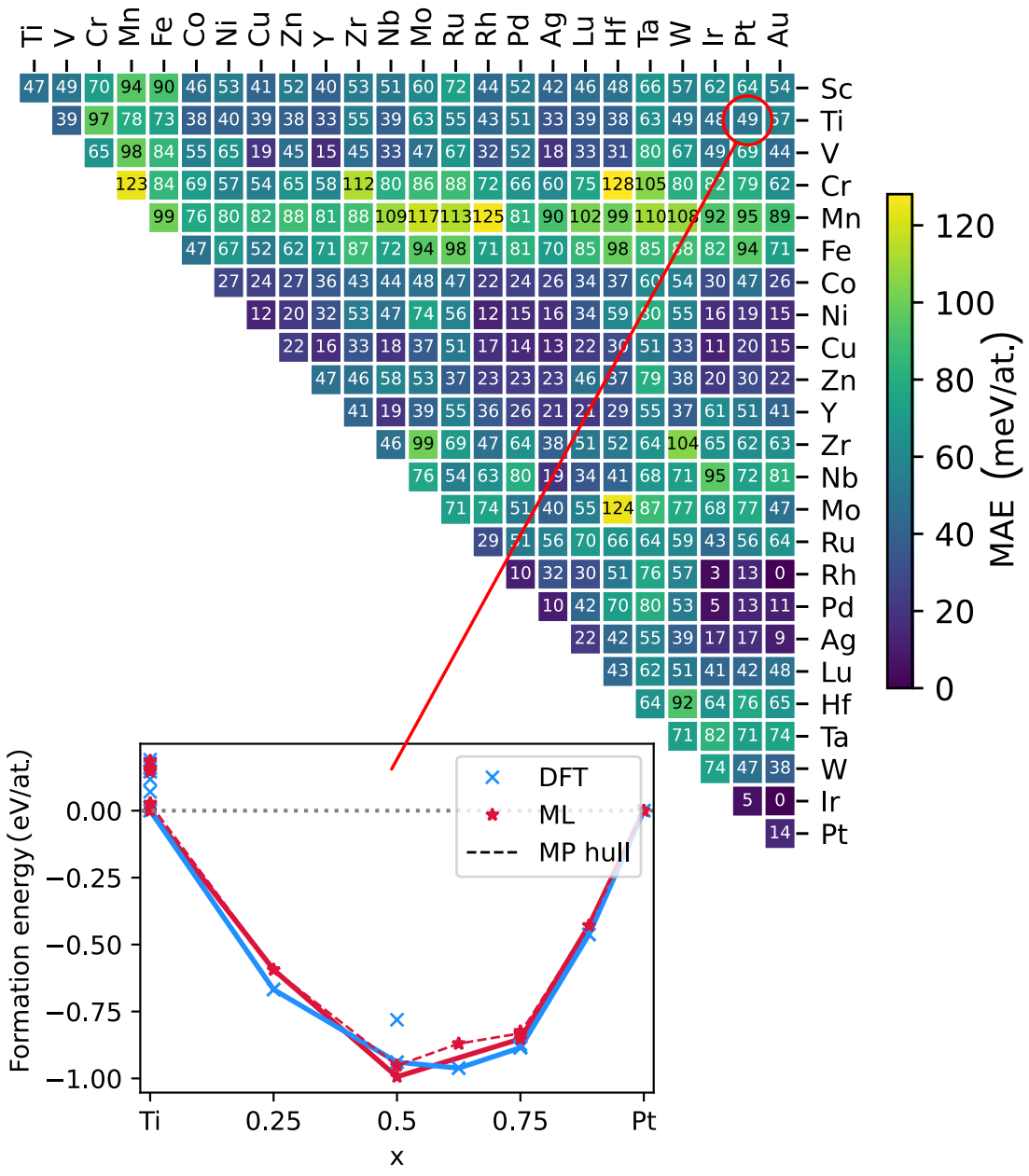
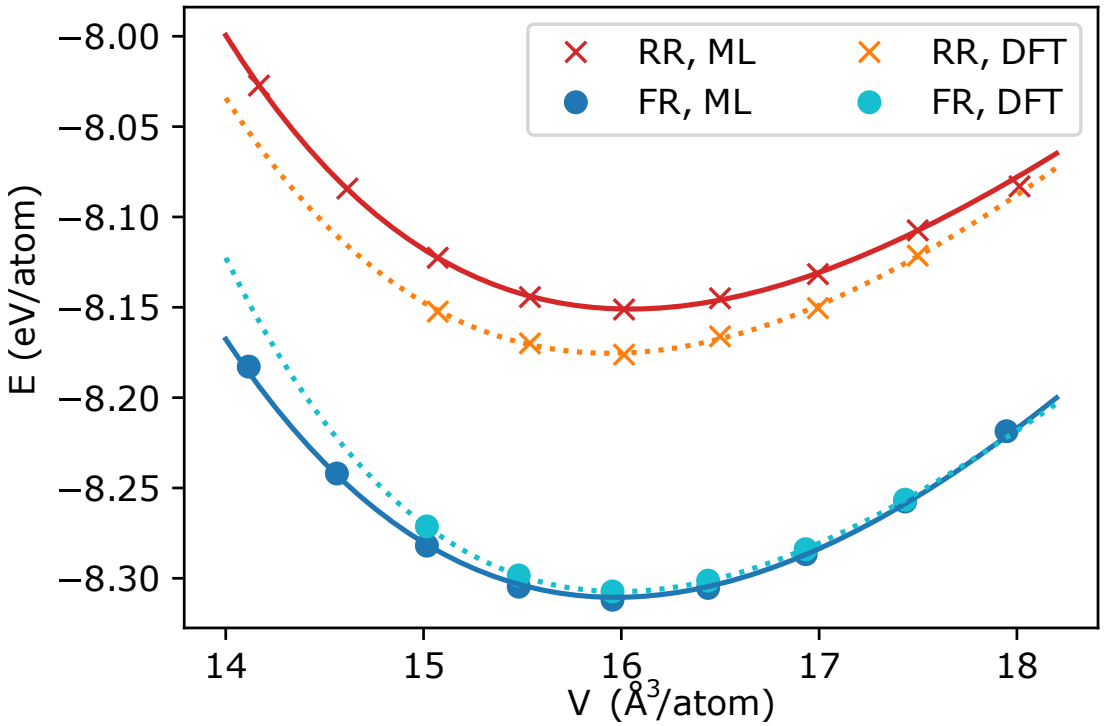


Figure 5.4: MAE for the formation energy of binary compounds from the Materials Project database. The inset shows a representative hull plot for the Ti–Pt system, highlighting the hulls obtained from the single-point DFT calculations and the ML predictions. The dashed line identifies the structures that are stable based on the energies available in the Materials Project database.

forming complex crystal structures, yield larger errors, while others such as Cu or Ni usually yield errors comparable to the validation set. It would be trivial to improve the accuracy of the model for binary structures and pure element polymorphs by including this small number of additional structures in the training set. We chose not to do that to avoid introducing biases in the accuracy depending on the different abundance of structures in the MP database. In the future, we plan to extend systematically our training set to incorporate disordered and liquid structures.



	RR, ML	RR, DFT	FR, ML	FR, DFT
E_0 / eV/at.	-8.151	-8.175	-8.310	-8.307
V_0 / Å^3/at.	16.04	15.94	15.97	16.00
B_0 / GPa	132	141	146	162
B_0'	6.7	6.1	4.8	8.0

Figure 5.5: Equation of state for the random relaxed (RR) and fully relaxed (FR) structures (see text for the full definition), computed with the HEA25-4-NN potential and with the reference DFT. Birch-Murnaghan parameters for cohesive energy (E_0), equilibrium volume (V_0), bulk modulus (B_0), bulk modulus derivative (B_0') are given in the table.

5.4.3 Energy and equation of state

We prepare a $5 \times 5 \times 5$ *fcc* supercell, containing 5 atoms of each of the 25 elements, arranged randomly on the lattice. We relax the geometry of the structure, and the volume of the supercell, using the HEA25-4-NN potential. We refer to this structure as the random relaxed (RR) structure. Starting from the same configuration, we also perform a slow annealing trajectory, combining molecular dynamics and atom exchange moves, to obtain a structure in which the arrangement of elements is not random, but more energetically favourable. We refer to this structure as the fully-relaxed (FR) structure. In both cases, the atoms relax away from *fcc* lattice positions, and the resulting structure within the supercell is rather disordered. We then introduce an isotropic compression or expansion of the two structures, relaxing the coordinates of the atoms within the cell, and fit a Birch-Murnaghan equation of state to the resulting energy-volume curves. We repeat the fixed-cell relaxation with the reference DFT, and compare the resulting equations of state (Fig. 5.5). The error on the cohesive energy E_0 is comparable to the test error (24meV for $E_0^{(\mathrm{RR})}$, 3meV for $E_0^{(\mathrm{FR})}$), and much smaller than the energy gain associated with the annealing of the lattice occupations ($E_0^{(\mathrm{RR})} - E_0^{(\mathrm{FR})}$ is about 150 meV/atom), indicating that HEA25-4-NN is reliable for assessing the energetics of ordering in a random alloy. The equilibrium volume and bulk modulus for the two structures are also in good agreement, with errors below 1% and 10 %, respectively – comparable with the typical discrepancy between different DFT approximations or between DFT and experiments.

5.4.4 Molecular dynamics

As a further demonstration of the accuracy and the stability of this potential, we perform two constant-pressure MD/MC trajectories, one at $T = 300$K and one at $T = 5000$K, each starting from a random arrangement of 5 atoms for each of the 25 elements (a total of 125 atoms) arranged on an *fcc* lattice. The trajectories are 10ps long, with on average one attempt at exchanging a pair of atoms every 2fs. We save a configuration every 100fs, and perform DFT calculations to compare energy and forces with those obtained from the ML potential. Fig. 5.6 shows that the low-temperature trajectory, where major rearrangements of the atoms occur but the structure remains approximately *fcc*, has an accuracy comparable to that measured on the validation set. The high-temperature run exhibits a higher error. However, the main component of the error is a rigid shift of the energies, and the trajectory remains stable – which is remarkable given that we observe complete melting, and the potential is trained exclusively on distorted solid structures.

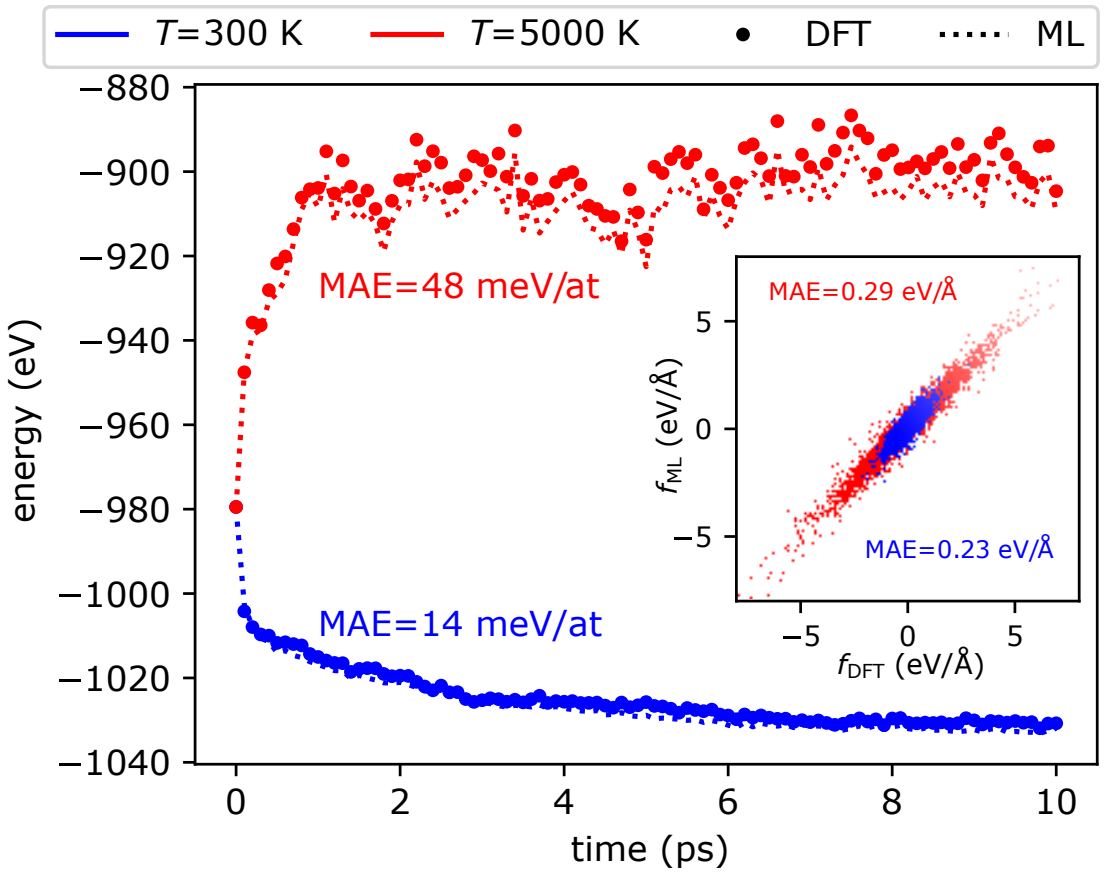


Figure 5.6: Comparison between the potential energy evaluated along two 10ps MD/MC trajectories, and that recomputed by DFT for 100 snapshots. The inset shows the parity plot for the force components computed for those structures. Energies have a MAE of 14 (48) meV/atom and forces a component MAE of 0.23 (0.29) eV/Å for the 300 (5000) K trajectory.

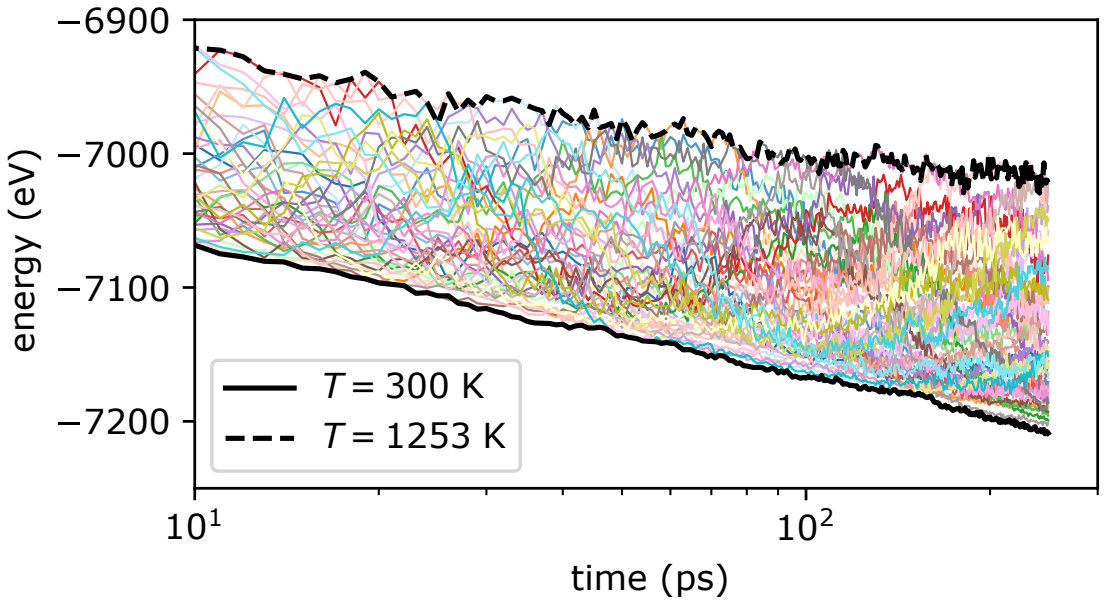


Figure 5.7: Trajectories of the potential energy for the 40 replicas used in one of the REMD simulations of a 864-atoms box of the HEA_{all}. Each color corresponds to a different initial configuration, that goes through cycles of heating and cooling due to REMD exchanges, accelerating the equilibration of the simulation at each temperature. The collection of trajectory segments corresponding to the extremal temperatures $T = 300$ K and $T = 1253$ K are highlighted with thicker, black lines. The logarithmic time scale refers to the MD integration time, but should not be interpreted as physical time given the presence of MC steps and replica exchange moves.

5.5 Temperature-dependent segregation in a Cantor-style alloy

In a seminal experiment, Cantor et al.[117] investigated the development of microstructure during the solidification of equimolar mixtures of 16 and 20 elements. We aim to perform a similar experiment in a computational setting, assessing the propensity of different elements to pair together or segregate, while covering the full component palette allowed by our model. This poses considerable challenges beyond the chemical complexity: kinetic trapping plays an important role in the physics of HEAs, and simulating vacancy-assisted atom diffusion requires time scales that are unattainable in brute-force atomistic modelling. In order to accelerate sampling and achieve (partial) equilibration, we run replica exchange simulations combining molecular dynamics and atom swap moves (REMD/MC), as described in Section 5.2.4.

Fig. 5.7 shows a representative trajectory for a 864-atoms cell, starting from *fcc* configurations, and including equimolar composition of all 25 elements (a composition we will refer to as $\mathrm{HEA_{all}}$). The slow, logarithmic relaxation of the low-temperature replica is indicative of the glassy dynamics of the system, which does not equilibrate completely even after millions of MD/MC steps (see the Appendix of Ref. [307]). For this reason, we perform multiple independent (and longer) simulations with a smaller box size (see the Appendix of Ref. [307]). The qualitative observations on the local ordering are robust, even though the precise arrangement of atoms in the low-temperature regime, as measured by the element-resolved pair correlation functions, differ noticeably between trajectories.

5.5.1 Relative pair probabilities for the $\mathrm{HEA_{all}}$ alloy

The pair correlation functions (Fig. 5.8) display broad, liquid-like peaks at both the highest and the lowest temperature we considered. In fact, simulations show little diffusion (except for some occasional bursts of activity at the high end of the temperature range) and the system can be characterized as an amorphous (or nano-crystalline) solid. The broadening of the peaks can be at least in part attributed to the diversity of pair distances between atomic species: some, like Cr-Cr, peak at distances as short as 2Å, others, such as Y-Y, peak at about 3.7Å. Note that typical distances in same-element pairs do not always match those found in the pure solid, underscoring the fact that the HEA25-4-NN can capture the effects arising from the heterogeneous chemical environments found in this alloy. For this reason, and given the disordered structure that develops in the supercell, we analyze structural correlations using a coarse-grained definition in which the first coordination shell extends up to a distance $r = 3.75$ Å, the second up to $r = 6.25$ Å and the third up to $r = 8$ Å, which is the largest distance we consider given the size of the box. We then define a variation on a theme of the short-range

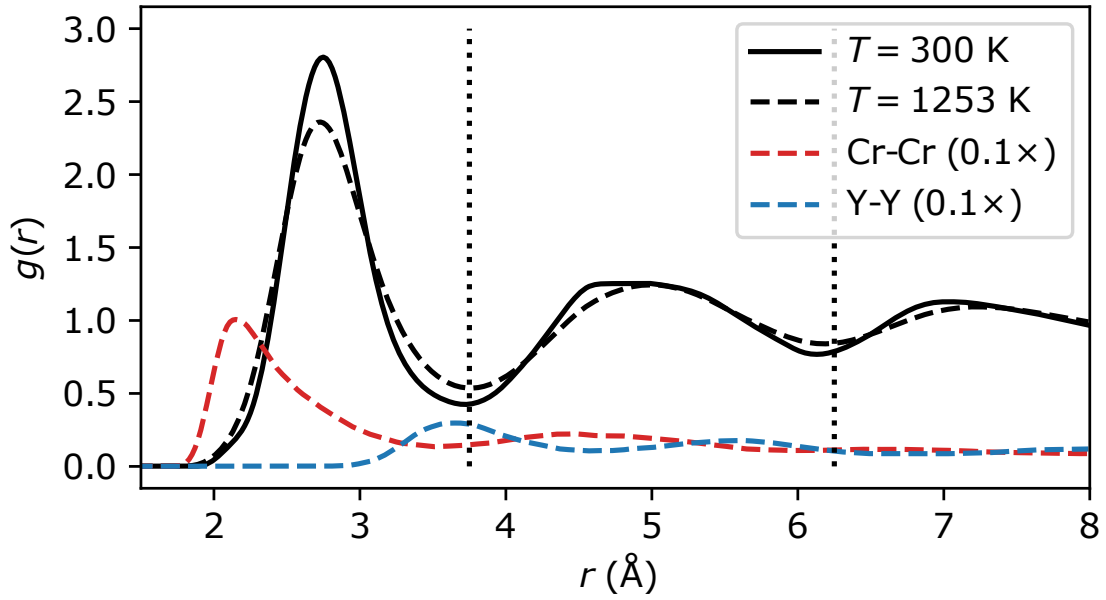


Figure 5.8: Pair correlation functions computed on a the $T = 300$ K (full) and $T = 1253$ K (dashed lines) replicas of a $\mathrm{HEA_{all}}$ box. Black lines correspond to the unresolved pair correlation, while red (Cr-Cr) and blue (Y-Y) lines provide representative examples of pair correlations resolved by species. The vertical dotted lines indicate the regions used in the definition of the pair ordering.

order parameter [331], which we dub the relative pair probability (RPP)

$$\mathrm{RPP}_{\Delta r}(\mathrm{A},\mathrm{B}) = \frac{p_{\Delta r}(\mathrm{A},\mathrm{B})}{p_{\Delta r}(\star,\star)} \frac{\rho^2}{\rho_\mathrm{A}\rho_\mathrm{B}} \tag{5.6}$$

which computes the number of pairs between species A and B that occur within a range Δr of distances, divided by the number of all pairs found in that same region, and normalized by the number density of the two species, $\rho_{\mathrm{A,B}}$ and the overall number density ρ. RPP = 1 indicates that the two species are as likely to be found within a given separation range than any atom pair. RPP > 1 (< 1) indicate that they are more (less) likely to be found in that distance range.

Qualitatively, the value of the RPP in the first coordination shell is indicative of the propensity of two elements to cluster together or to separate from each other. However, the values cannot be interpreted in isolation, without considering the overall setup of the simulation: the finite size of the supercell, the imperfect equilibration, and the many-body interactions between all 25 species mean that the strong affinity between Y and Au, or the poor compatibility of Mn and Pd, do not necessarily imply the same quantitative effect when considered as part of a different overall composition. Fig. 5.9 shows a heat-map representation of $\mathrm{RPP}_{\Delta r}(\mathrm{A},\mathrm{B})$ for the $\mathrm{HEA_{all}}$ at 300 K and 1253 K, and for the three regions indicated in Fig. 5.8. A few qualitative

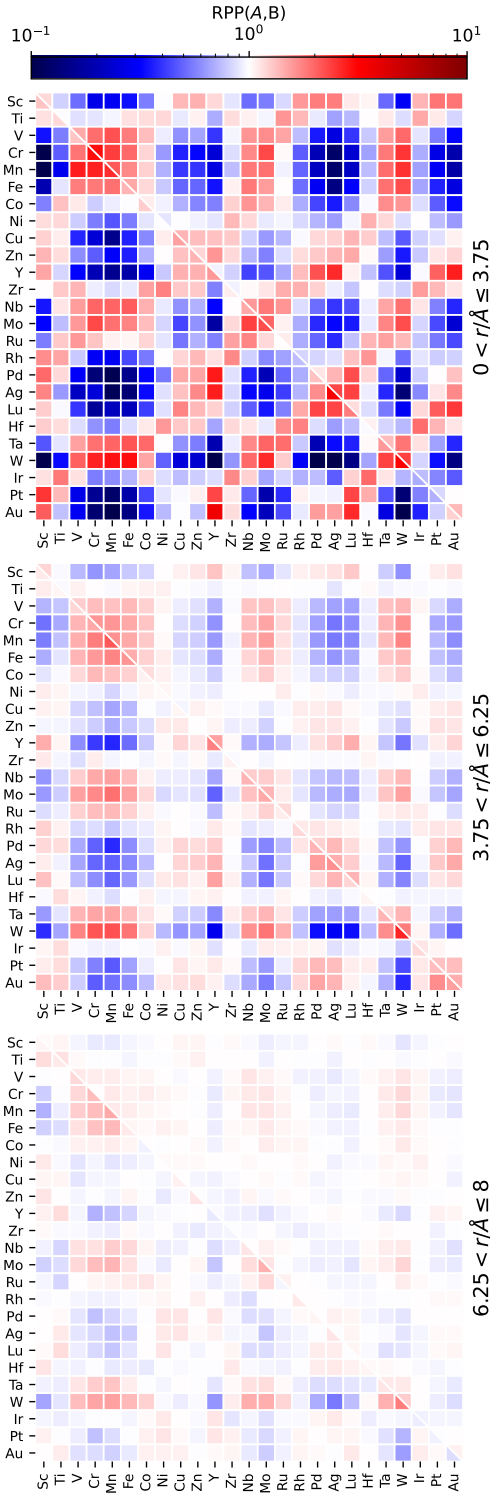


Figure 5.9: A plot of the relative pair probability for all atom pairs and the three regions corresponding to the first, second, and third peaks in the total pair correlation function (Fig. 5.8). Each plot shows results for simulations of HEA_{all} at both 300 K (lower-left corner) and 1253 K (top-right corner), averaged over the trajectories and discarding the first 100 ps (50'000 combined MD/MC steps).

observations can be made. First, in our simulations HEA_{all} evolves to be far from random. Certain atom pairs have a strong tendency to associate or separate at low temperature, and the high-temperature samples (which are well equilibrated) show similar, even though less pronounced, trends. This correspondence is interesting, as it suggests one may use high-temperature trajectories, that are easier to converge, to extract insights on the propensity of different species for association. The trends observed in the second and third region are very similar to those in the first-extended-neighbor shell, although progressively less pronounced: given the finite size of the simulation, and incomplete equilibration, the simulation does not generate clear-cut phase-separated regions.

Considering the RPP along the elements, one can observe a clear periodicity in behavior. Sc, Y, Hf, as well as the noble metals, Cu and Zn, tend to separate from V, Cr, Mn, Fe, which on the other hand have a tendency to cluster together, and also have positive associations to their heavier counterparts Nb, Mo, Ta, W. On the other hand, Sc, Y and (to a lesser degree) Hf associate strongly with noble metals, Cu, and Zn. The noble metals, Cu and Zn also tend to cluster together. Ti, Co, Ni, Zr, Ru, Ir have less clear-cut associations, and are closer to having a random distribution throughout the box. Another way of looking at the association plots in Fig. 5.9 is to check for consistency with known high-entropy alloys. The Cr-Mn-Fe-Co-Ni system is one of the prototypical sets of HEA formers, and indeed we observe strong mutual association tendency between Cr-Mn-Fe in the first shell, and also with Co and Ni in the second extended shell. Second-shell mutual association is also observed for noble-metal based compositions such as Ni-Cu-Pd-Pt-Au. Let us reiterate that strong mutual association for a group of elements in the HEA_{all} runs is a necessary, but not sufficient, conditions for that group of elements to be good HEA-forming candidates. For instance, some elements may have a strong tendency to form ordered intermetallics and might separate out of the mixture.

5.5.2 Data-driven Hume-Rothery rules

This analysis allows us to substantiate and quantify some of the empirical principles that are used in the design of HEAs, such as Hume-Rothery rules[332] that stipulate what elements can be substituted for each other with little effect on the HEA-forming propensity. We use the first-neighbor affinity of each species to all the other elements in the alloy to define a measure of dissimilarity as

$$d_{\mathrm{RPP}}(\mathrm{A},\mathrm{B})^2 = \sum_{\mathrm{X}} \left[\log_{10} \frac{\mathrm{RPP}_1(\mathrm{A},\mathrm{X})}{\mathrm{RPP}_1(\mathrm{B},\mathrm{X})} \right]^2, \tag{5.7}$$

that, roughly speaking, measures the relative strength of interactions between the two species and the other components. Two elements with a small distance are predicted to behave similarly, and vice versa. Fig. 5.10 paints a picture that is consistent with the observations

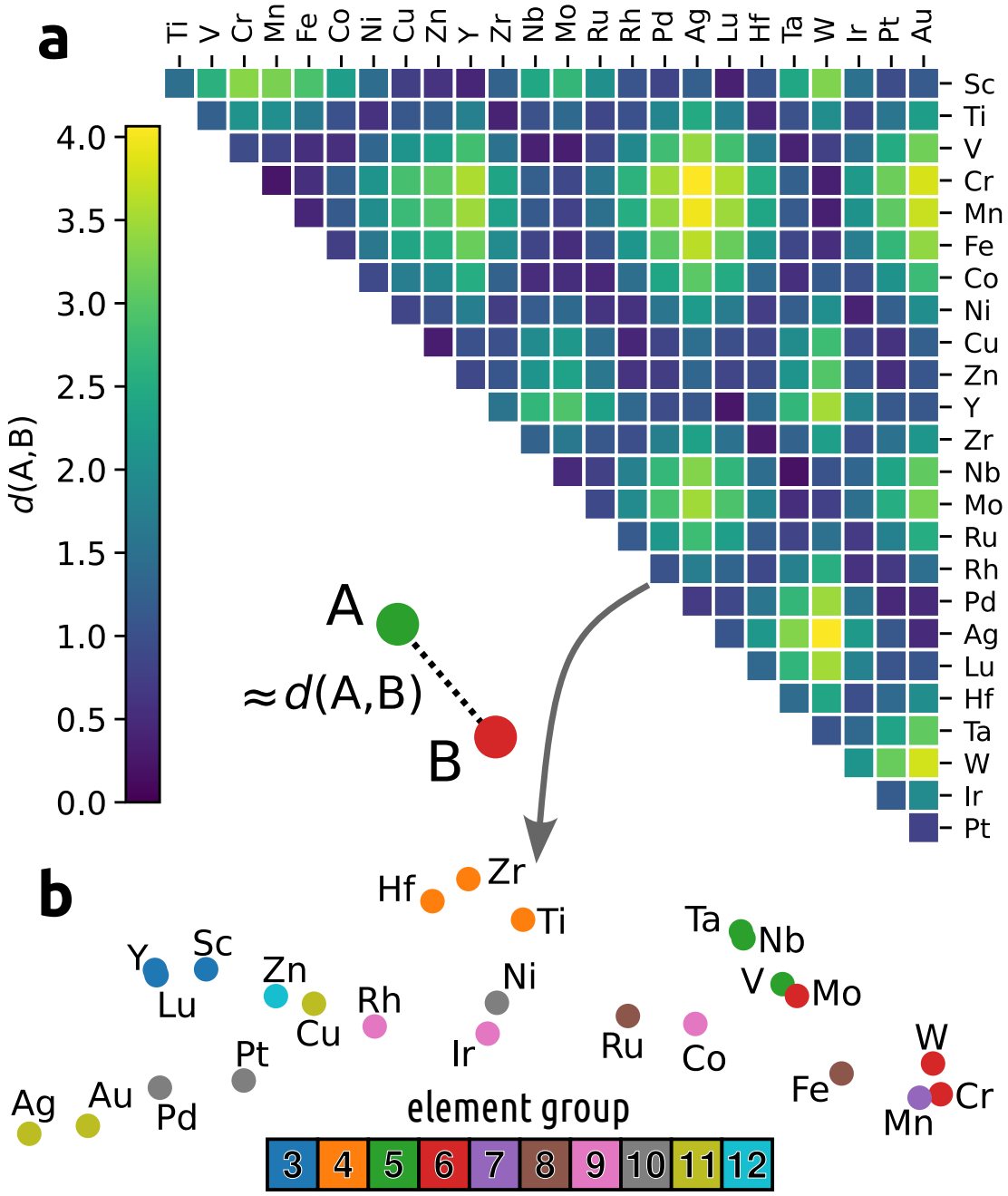


Figure 5.10: (a) Element similarity matrix based on the RPP distance (5.7) for the nearest-neighbor shell, in the $\mathrm{HEA}_{\mathrm{all}}$ simulation at $T = 1253$ K. (b) The element similarity map (color-coded based on the group of the various transition metals) is built by applying metric multi-dimensional scaling to the distance matrix, and provides a visual aid to recognize groups of elements that have similar affinity patterns to the other d-block metals.

we made on short and mid-range order between the elements in the HEA_{all}, and with much of the common wisdom in HEA research.We base this analysis on the high-temperature simulations to obtain a statistically-converged, and somewhat more nuanced, definition, but the qualitative features of the map are similar to those one would obtain from the RPP computed at T = 300 K. Elements in the same group usually show strong similarity, but this is not always the case: for example, Cu is more similar to Zn than to Ag. The similarity matrix can also be converted to a 2D map, in which the Euclidean distance between elements approximates their RPP-based similarity (also shown in Fig. 5.10), which provides an easy-to interpret visual representation of a set of data-driven rules to design HEAs. The element similarity that can be inferred from the RPP-based map differ – both quantitatively and conceptually – from that associated with the alchemical coupling matrix in Fig. 5.2. Whereas the weights are associated with the similarity in terms of the interatomic potential, the RPP similarity is a result of the collective behavior of the HEA_{all}at the prescribed thermodynamic conditions, not unlike the relation between a pair potential and the potential of mean force. This means, for example, that one could compute d_{RPP} for a different alloy composition (extending or refining the assessment of alloying behavior), from a different type of interatomic potential, or even from experimental data on partial structure factors.

5.6 Bulk structure of high-entropy alloys for catalysis

Having demonstrated the accuracy of the HEA25-4-NN model, and used it to investigate the mutual affinity of the full set of 25 transition metals we considered in a Cantor-type computational experiment, we now turn our attention to a more focused study of three specific equimolar compositions. The first is the prototypical CoCrFeMnNi alloy, which was reported by Cantor et al.[117] in their seminal paper. This alloy is also known to be effective as a catalyst[333–335]. Furthermore, we investigate CoCrFeMoNi[126, 336, 337],as an example of an alloy obtained by element substitution that has been broadly studied for its improved mechanical and tribological properties [338, 339], as well as a catalyst of oxygen evolution reactions. We then consider IrPdPtRhRu[128, 340–343] as an example of an alloy based on sixth period elements that has recently received much attention as a catalyst for hydrogen evolution, and is often synthesized in the form of nanoparticles.

To model the alloys, we used *fcc* lattices with 500 atoms per cell ($5 \times 5 \times 5$ super cell). We ran two independent REMD/MC runs according to Section 5.2.4 with a timestep of 2 fs and 32 temperature replicas, logarithmically spaced between 300 K and 1253 K. We discard the first 100ps for equilibration. Given that all these alloys maintain a regular *fcc* structure throughout the simulation, we analyze their structure in terms of Cowley's short-range order[331] (SRO),

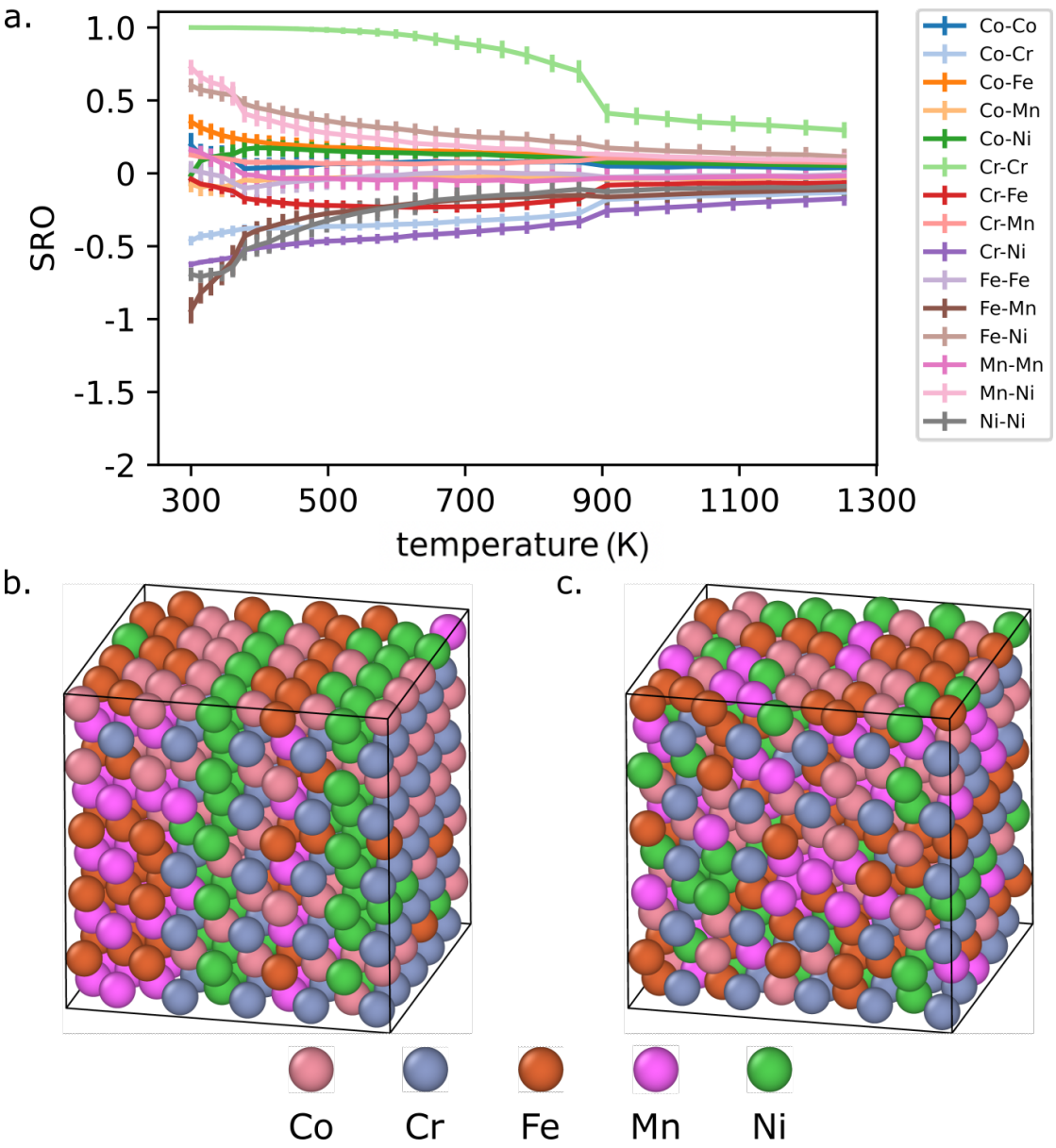


Figure 5.11: a. Cowley's short-range (SRO) parameters for the first shell in CoCrFeMnNi HEA, shown for the 10 replicas between 300 and 1253 K, averaged over the last 1000 steps and two independent runs. At low temperatures, a tendency of Fe-Mn segregation can be seen. In contrast, Cr is very well mixed. There are two phase transformations around 400 K and 900 K. The y-axis is adjusted to the example shown in Fig. 5.13 to facilitate comparison. b,c. snapshot from MC/MD simulations at $T = 300\,\mathrm{K}$ and at $T = 720\,\mathrm{K}$, respectively. In the 300 K snapshot, two planes of Ni can be seen, while in the higher temperature snapshot, Cr order is evident (see the Appendix of Ref. [307]).

which is commonly used in the study of HEAs and takes a value of zero when atoms are distributed fully randomly, becomes negative for pairs of atoms that tend to cluster together, and tends to one when two atom types never appear as first neighbours. In the Appendix we also report an analysis in terms of the RPP that incorporates second-neighbor and long-range correlations. In interpreting these results, one should consider similar considerations to those we discussed for the HEA_{all} simulations: (1) the SRO (and the RPP) are only meaningful for homogeneous phases, and in case of phase separation the values computed for the whole cell serve only to signal the occurrence of a phase transition; (2) a combination of finite-size effects and glassy behavior can hinder reaching full equilibrium in simulations; (3) since they allow for atom exchanges, our simulations cannot give quantitative indications on whether different phases are only metastable, nor on the kinetics of diffusion processes that are required for precipitation.

We start by analyzing the Cantor alloy CoCrFeMnNi. The SRO computed at different temperatures (Fig. 5.11a, plotted for all element combinations) indicate the presence of at least two phase transitions. The high-temperature phase is homogeneous and disordered, but shows substantial ordering, particularly for the Cr-Cr pair. At approximately 900 K we observe a first transition, that is associated with the ordering of Cr atoms. The SRO for the Cr-Cr pair tends to one (as there are almost no first-neighbor chromium atoms) but the RPP show a clear increase of second-neighbor Cr-Cr pairs, consistent with the formation of a simple cubic sublattice. The other elements remain relatively disordered, and no discontinuous behavior is observed in the SRO. As the temperature is reduced further, a second transition occurs around 400 K. The most prominent structural transformation is the formation of (100) Ni planes, separated by (Co,Fe,Mn)-rich regions forming a layered superstructure. Fig. 5.11b,c show snapshots of the simulations at 300 K and 720 K, that give an idea of the partially-ordered structure of the two phases.

Substituting Mn with Mo changes the segregation behavior significantly (Fig. 5.12a): the SRO parameters are generally smaller, with the largest segregation tendency found for the Mo-Ni atom pair. The tendency of Cr to form a cubic sublattice is less pronounced than CoCrFeMnNi, and one only sees the increase of SRO parameters at around 500 K. At low temperature, (100) planes of Ni form that are very similar to those observed in the Mn-based counterpart (Fig. 5.12b,c), that are separated by (Co,Fe,Mo)-rich regions. Given the sizable energy errors of the ML models, as well as those of the underlying DFT reference, one should not overinterpret the details of the structures we observe. Even if *fcc* CoCrFeMnNi is paramagnetic, neglect of magnetism in the presence of several elements which form ferromagnetic phases is worrisome (see e.g. Ref. [344] for a thorough discussion of magnetism in CoCrFeMnNi and CoCrFeMoNi). That said, our observations provide strong indications of the tendency to form partly ordered

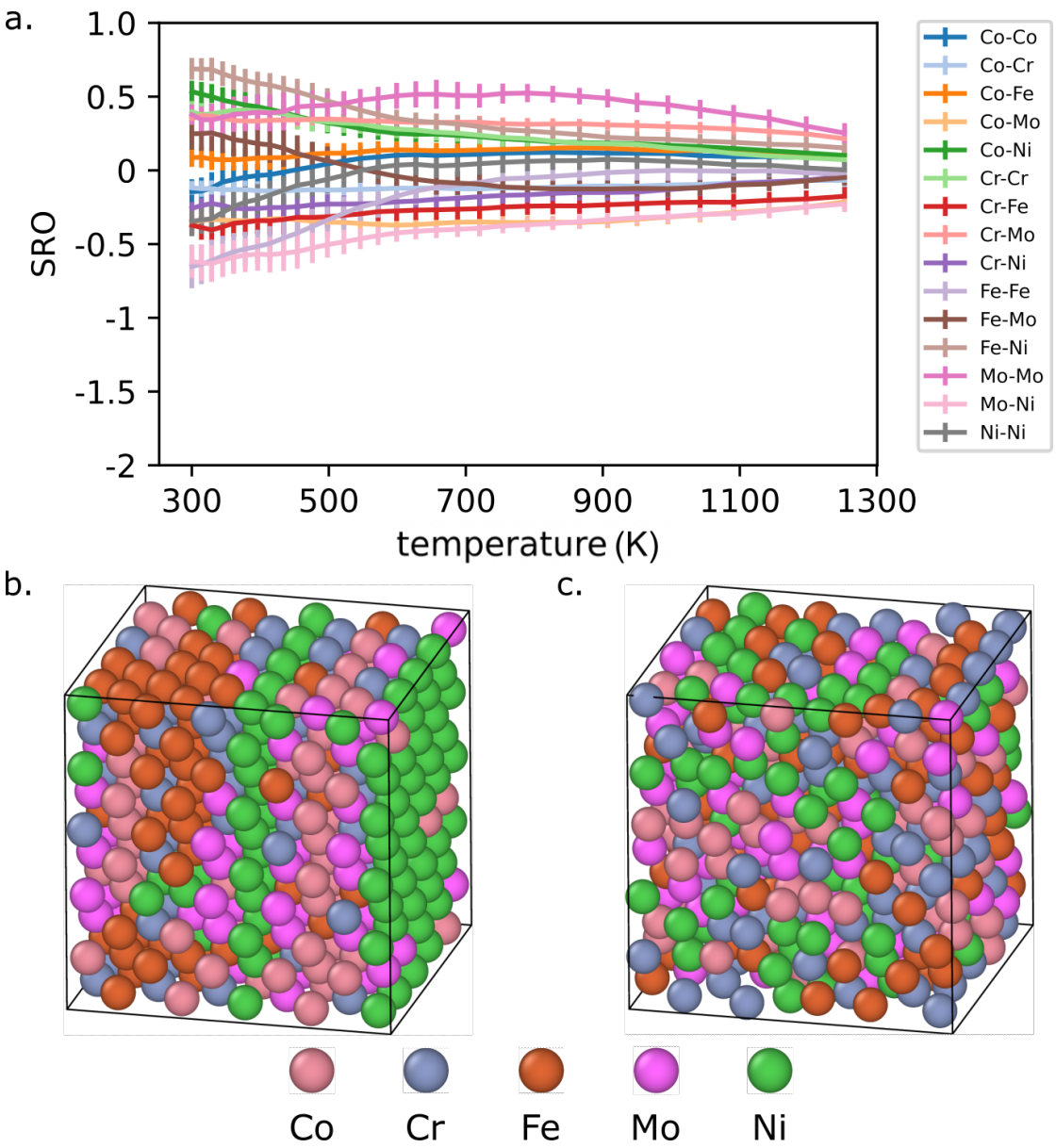


Figure 5.12: a. Cowley's short-range (SRO) parameters for the first shell in CoCrFeMoNi HEA, shown for the 10 replicas between 300 and 1253 K, averaged over the last 1000 steps and two independent runs. Good mixing of atomic species can be assumed due to the small values of SRO parameters. The y-axis is adjusted to the example shown in Fig. 5.13 to facilitate comparison. b,c. snapshot from MC/MD simulations at $T = 300\,\mathrm{K}$ and at $T = 1253\,\mathrm{K}$, respectively. In the 300 K snapshot, two planes of Ni can be seen.

phases with a complex structure, which, together with the low vacancy-mediated diffusivity[345], help explain the observed stability of HEAs that contain (Co,Fe,Cr,Ni). A tendency to develop short-range ordering is consistent with previous simulations in other classes of HEAs[346], and with observation of phase separation in equimolar CoCrFeMnNi in high-mobility environments such as grain boundaries[347] or under deformation[348].

While the leading effect in CoCrFeMnNi and CoCrFeMoNi is the appearance of partial ordering at low temperatures, in the case of IrPdPtRhRu we observe clear-cut phase separation betwen a (Pd,Pt) and a (Ru,Ir,Rh) phase, with Rh accumulating preferentially at the interface between the two phases (see Fig. 5.13b,c). The strong tendency to segregate is already evident in the high-temperature regime, where the system is visually well-mixed, but with large SRO parameters. This is in contrast to the experimental observation that this HEA forms a complex solid solution with random atom distribution[128, 342]. As shown in the Appendix of Ref. [307], the large enthalpic gain arising from demixing is not an artefact of HEA25-4-NN, and the ML error on the free-energy change upon ordering is of the order of 3 meV/atom. These observations suggest that kinetic trapping, or finite-size effects associated with the synthesis in the form of nanoparticles, might be key to stabilize a homogeneous phase.

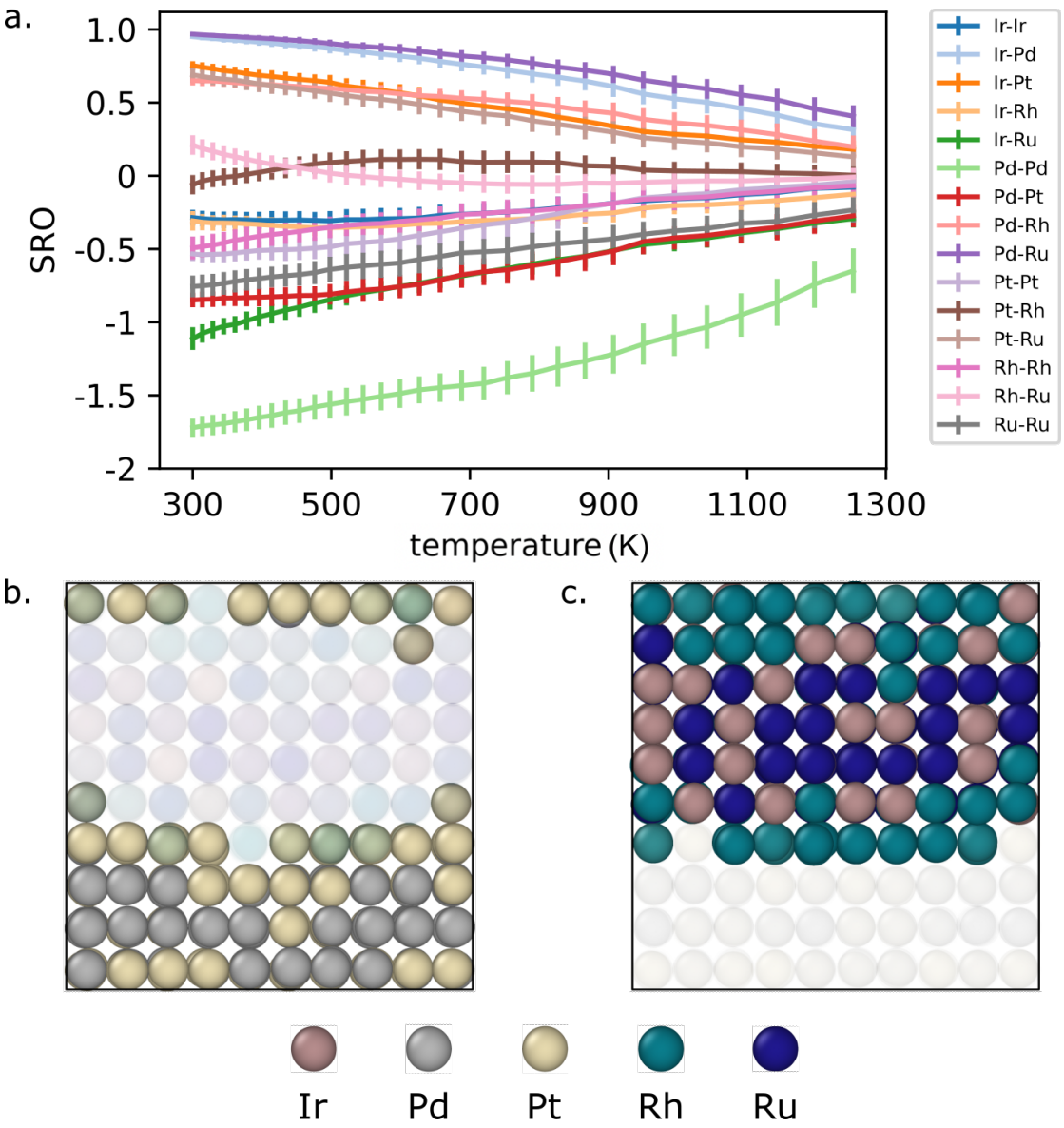


Figure 5.13: a. Cowley's short-range parameters for the first shell in IrPdPtRhRu HEA, shown for the 10 replicas between 500 and 933 K, averaged over the last 1000 frames and with an error estimation from independent repetition runs. The most pronounced local order can be seen for the Pd-Pd atom pair (light green line, mathematically smallest SRO). Demonstration of the phase segregation tendency by highlighting the b. PdPt and c. IrRhRu atoms in an MC/MD snapshot.

6 Conclusions

Grasping the implications of finite temperature effects is essential for a precise description of heat transport, mechanical response, microstructure, and phase formation, which are all important factors in guiding the design and synthesis of materials. Traditional methods of studying materials at finite temperatures can be prohibitively expensive for studying phenomena which require large time and length scales, such as interfacial properties. To overcome these constraints, this thesis has tackled two primary challenges: 1) building a workflow based on an ML model to compute finite temperature properties accurately and 2) presenting a model capable of generalizing and transferring the knowledge learned from different chemistries to model materials of real-life interest.

By addressing the problems of finite temperature description and chemical complexity separately, this work has made substantial progress in both areas. To tackle the first problem, we focused on a simple, single-element system – nickel, which is an important component in several industrial alloys. We began by constructing a high-dimensional neural network potential based on the approach proposed by Behler and Parinello. For this purpose, we generated a dataset using an exploitation/exploration strategy, encompassing bulk and liquid phases, as well as surfaces and defects. The developed MLIP achieved accuracy comparable to DFT for pure nickel across a wide range of properties.

Using a simple system as an example, we demonstrated a general workflow independent of the system's choice combining machine learning models with statistical mechanics techniques to account for quantum effects and electronic excitations. Additionally, we provided an example of how MLIP enables accurate sampling of free energy surfaces with enhanced sampling techniques to compute interfacial properties. Moreover, we calculated the stability of defects using MLIP for thermodynamic integration, significantly reducing the computational cost compared to ab initio calculations. This approach presented promising prospects for

estimating defect formation free energies in more complex defects and materials with a wider range of chemistry and crystallography.

Building on the achievements in addressing finite temperature descriptions, this thesis also tackles the challenge of the transferability of ML models across different chemistries by adopting the method of alchemical learning. This approach is based on the principle that chemical elements close to each other in the periodic table often exhibit similar behaviour, a fundamental concept in chemistry that plays a vital role in the design of new materials. To this end, an ML framework was developed that incorporated this principle through linear compression of chemical space. This framework was applied to regress the potential energy surface of a dataset containing 25 d-block elements. To ensure the necessary level of phase space sampling, a training set was generated using a protocol specifically designed for quasi-random sampling in the high-dimensional phase space.

The ML framework successfully trained a potential capable of describing bulk phases of arbitrary combinations of represented chemical elements with semi-quantitative accuracy. The intuitive, functional form of the contraction facilitated a critical analysis of the model's performance, revealing that only 3-4 dimensions were necessary for capturing the diverse behaviour of transition metals.

The optimized combination weights of the model unveil relationships between elements that correspond to their arrangement in the periodic table. This insight allows the prediction of properties for missing elements with only a moderate loss in accuracy. This capability demonstrates the potential of the ML framework in expanding our understanding of complex chemical systems.

Utilizing the potential, the thesis explores an ambitious experiment involving the equilibration of an equimolar mixture of all 25 elements. This process results in a disordered structure characterized by strong element segregation. The observed affinity between elements is consistent with known high-entropy alloys, providing a foundation for defining a data-driven version of the Hume-Rothery rules. These rules can be adapted to subsets of elements relevant to specific applications, further enhancing the utility of the developed framework.

The thesis investigates three specific compositions: the Cantor alloy, one Mn$\rightarrow$Mo substitution alloy, and one noble metal alloy. Each of these compositions exhibits unique phase behaviors, showcasing the versatility and applicability of the developed ML framework in studying diverse material systems.

As the field of materials science advances, there are several directions for future work that can build upon the achievements of this thesis. One such direction involves extending the dataset

to include a more diverse range of compounds and structures, such as molten and defective configurations. Additionally, expanding the dataset to encompass not only bulk phases but also surfaces would offer valuable insights into surface segregation and surface reactivity of multi-component materials, which are relevant for corrosion and catalytic applications.

Considering the generality of the presented model, another impactful extension would be to tackle the organic space of chemistry. By applying this approach to systems of biomolecular relevance, researchers could gain a deeper understanding of complex biological systems and their interactions with materials invaluable in drug discovery, biomaterials, and biotechnology.

Further improvements to the methods employed in this thesis are also possible. For instance, the compression scheme could be enhanced by optimizing the chemical embedding independently for different central species of the atom-centred representation. Exploring the transferability of the method to other datasets is another intriguing research direction. By assessing if the compressed matrix can be effectively applied to various datasets, with only the need to fit model weights, new opportunities for investigating a wide range of materials might be unveiled.

Application-focused research, such as studying 4 and 5-element high-entropy alloys, can also build upon the work presented in this thesis. The HEA25-4-NN model will provide valuable insights into the stability range of multi-principal-component alloys and guide synthetic efforts in developing novel materials with desirable properties.

To conclude, this thesis highlights the benefits of combining machine learning predictions with physics and chemistry-based methods, creating a versatile hybrid modelling approach. By using the strengths of both machine learning and sampling techniques, more accurate and efficient models can be developed to study the stability of multi-component alloys. This approach has the potential to lead the production of a new generation of materials with enhanced properties, finding applications that substantially impact industries such as aerospace, biomedical, energy, chemical processing and electronics.

Bibliography

[1] Sidney Yip, ed. *Handbook of Materials Modeling*. Dordrecht ; New York: Springer, 2005.

[2] Gerbrand Ceder, Y-M Chiang, DR Sadoway, MK Aydinol, Y-I Jang, and Biying Huang. "Identification of cathode materials for lithium batteries guided by first-principles calculations". In: *Nature* 392.6677 (1998), pp. 694–696.

[3] Flemming Besenbacher, Ib Chorkendorff, BS Clausen, Bjørk Hammer, AM Molenbroek, Jens Kehlet Nørskov, and Ivan Stensgaard. "Design of a surface alloy catalyst for steam reforming". In: *Science* 279.5358 (1998), pp. 1913–1915.

[4] Jeff Greeley, Thomas F Jaramillo, Jacob Bonde, IB Chorkendorff, and Jens K Nørskov. "Computational high-throughput screening of electrocatalytic materials for hydrogen evolution". In: *Nature materials* 5.11 (2006), pp. 909–913.

[5] Jun Yan, Prashun Gorai, Brenden Ortiz, Sam Miller, Scott A Barnett, Thomas Mason, Vladan Stevanović, and Eric S Toberer. "Material descriptors for predicting thermoelectric performance". In: *Energy & Environmental Science* 8.3 (2015), pp. 983–994.

[6] Kee-Joo Chang and Marvin L Cohen. "Structural and electronic properties of the high-pressure hexagonal phases of Si". In: *Physical Review B* 30.9 (1984), p. 5376.

[7] Artem R Oganov and Mario Valle. "How to quantify energy landscapes of solids". In: *The Journal of chemical physics* 130.10 (2009), p. 104504.

[8] Chris J Pickard and RJ Needs. "Ab initio random structure searching". In: *Journal of Physics: Condensed Matter* 23.5 (2011), p. 053201.

[9] K Kunc and Richard M Martin. "Ab initio force constants of GaAs: A new approach to calculation of phonons and dielectric properties". In: *Physical Review Letters* 48.6 (1982), p. 406.

[10] K Parlinski, ZQ Li, and Y Kawazoe. "First-principles determination of the soft mode in cubic ZrO 2". In: *Physical Review Letters* 78.21 (1997), p. 4063.

[11] PAT Olsson, AR Massih, Jakob Blomqvist, A-M Alvarez Holston, and Christina Bjerkén. "Ab initio thermodynamics of zirconium hydrides and deuterides". In: *Computational materials science* 86 (2014), pp. 211–222.

[12] Yi Zhang, Xuezhi Ke, Changfeng Chen, Jihui Yang, and PRC Kent. "Thermodynamic properties of PbTe, PbSe, and PbS: First-principles study". In: *Physical review B* 80.2 (2009), p. 024304.

[13] DV Minakov and PR Levashov. "Melting curves of metals with excited electrons in the quasiharmonic approximation". In: *Physical Review B* 92.22 (2015), p. 224102.

[14] Abdallah Khellaf, Alfred Seeger, and Roy M Emrick. "Quenching studies of lattice vacancies in high-purity aluminium". In: *Materials transactions* 43.2 (2002), pp. 186–198.

[15] Andrew Ian Duff, Theresa Davey, Dominique Korbmacher, Albert Glensk, Blazej Grabowski, Jörg Neugebauer, and Michael W Finnis. "Improved method of calculating ab initio high-temperature thermodynamic properties with application to ZrC". In: *Physical Review B* 91.21 (2015), p. 214311.

[16] Albert Glensk, Blazej Grabowski, Tilmann Hickel, and Jörg Neugebauer. "Understanding anharmonicity in fcc materials: from its origin to ab initio strategies beyond the quasiharmonic approximation". In: *Physical review letters* 114.19 (2015), p. 195901.

[17] Blazej Grabowski, Yuji Ikeda, Prashanth Srinivasan, Fritz Körmann, Christoph Freysoldt, Andrew Ian Duff, Alexander Shapeev, and Jörg Neugebauer. "Ab initio vibrational free energies including anharmonicity for multicomponent alloys". In: *npj Computational Materials* 5.1 (2019), pp. 1–6.

[18] Ying Zhou, Prashanth Srinivasan, Fritz Körmann, Blazej Grabowski, Roger Smith, Pooja Goddard, and Andrew Ian Duff. "Thermodynamics up to the melting point in a TaVCrW high entropy alloy: Systematic a b i n i t i o study aided by machine learning potentials". In: *Physical Review B* 105.21 (2022), p. 214302.

[19] B Grabowski, P Söderlind, T Hickel, and J Neugebauer. "Temperature-driven phase transitions from first principles including all relevant excitations: The fcc-to-bcc transition in Ca". In: *Physical Review B* 84.21 (2011), p. 214107.

[20] Jun Jiang, Weifu Sun, and Ning Luo. "Molecular dynamics study of microscopic deformation mechanism and tensile properties in AlxCoCrFeNi amorphous high-entropy alloys". In: *Materials Today Communications* 31 (2022), p. 103861.

[21] Kuan-Ting Chen, Ting-Ju Wei, Guo-Chi Li, Mei-Yi Chen, Yi-Shiang Chen, Shu-Wei Chang, Hung-Wei Yen, and Chuin-Shan Chen. "Mechanical properties and deformation mechanisms in CoCrFeMnNi high entropy alloys: A molecular dynamics study". In: *Materials Chemistry and Physics* 271 (2021), p. 124912.

[22] Hong Yang, YongJun Lü, Min Chen, and ZengYuan Guo. "A molecular dynamics study on melting point and specific heat of Ni3Al alloy". In: *Science in China Series G: Physics, Mechanics and Astronomy* 50.4 (2007), pp. 407–413.

[23] S Özdemir Kart, A Erbay, H Kılıç, T Cagin, and M Tomak. "Molecular dynamics study of Cu-Pd ordered alloys". In: *Journal of Achievements in Materials and Manufacturing Engineering* 31.1 (2008), pp. 41–46.

[24] J Davoodi and F Katouzi. "High pressure molecular dynamics simulation of Au-x% Ni alloys". In: *Journal of Applied Physics* 115.9 (2014), p. 094905.

[25] DY Sun, MI Mendelev, CA Becker, K Kudin, Tomorr Haxhimali, Mark Asta, JJ Hoyt, Alain Karma, and David J Srolovitz. "Crystal-melt interfacial free energies in hcp metals: A molecular dynamics study of Mg". In: *Physical Review B* 73.2 (2006), p. 024116.

[26] F Calvo. "Molecular dynamics determination of the surface tension of silver-gold liquid alloys and the Tolman length of nanoalloys". In: *The Journal of Chemical Physics* 136.15 (2012), p. 154701.

[27] Mark Tuckerman. *Statistical mechanics and molecular simulations*. 2008.

[28] B. Grabowski, L. Ismer, T. Hickel, and J. Neugebauer. "Ab initio up to the melting point: Anharmonicity and vacancies in aluminum". In: *Phys. Rev. B* 79.13 (Apr. 2009).

[29] Raynol Dsouza, Liam Huber, Blazej Grabowski, and Jörg Neugebauer. "Approximating the impact of nuclear quantum effects on thermodynamic properties of crystalline solids by temperature remapping". In: *Physical Review B* 105.18 (2022), p. 184111.

[30] Jin Yang, Jianping Long, and Lijun Yang. "First-principles investigations of the physical properties of lithium niobate and lithium tantalate". In: *Physica B: Condensed Matter* 425 (2013), pp. 12–16.

[31] M Parrinello and A Rahman. "Study of an F Center in Molten KCl". In: *J. Chem. Phys.* 80 (1984), p. 860.

[32] R P Feynman and A R Hibbs. *Quantum Mechanics and Path Integrals*. New York: McGraw-Hill, 1964.

[33] Mark Tuckerman. *Statistical Mechanics and Molecular Simulations*. Oxford University Press, 2008.

[34] Walter Kohn and Lu Jeu Sham. “Self-consistent equations including exchange and correlation effects”. In: *Physical review* 140.4A (1965), A1133.

[35] Lu J Sham and Michael Schlüter. “Density-functional theory of the energy gap”. In: *Physical review letters* 51.20 (1983), p. 1888.

[36] Murray S Daw and Michael I Baskes. “Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals”. In: *Physical Review B* 29.12 (1984), p. 6443.

[37] MW Finnis and JE Sinclair. “A simple empirical N-body potential for transition metals”. In: *Philosophical Magazine A* 50.1 (1984), pp. 45–55.

[38] MI Baskes. “Application of the embedded-atom method to covalent materials: a semiempirical potential for silicon”. In: *Physical review letters* 59.23 (1987), p. 2666.

[39] Michael I Baskes. “Modified embedded-atom potentials for cubic materials and impurities”. In: *Physical review B* 46.5 (1992), p. 2727.

[40] R Pasianot, Diana Farkas, and EJ Savino. “Empirical many-body interatomic potential for bcc transition metals”. In: *Physical Review B* 43.9 (1991), p. 6952.

[41] Frank H Stillinger and Thomas A Weber. “Computer simulation of local order in condensed phases of silicon”. In: *Physical review B* 31.8 (1985), p. 5262.

[42] J Tersoff. “New empirical model for the structural properties of silicon”. In: *Physical review letters* 56.6 (1986), p. 632.

[43] Wojciech J. Szlachta, Albert P. Bartók, and Gábor Csányi. “Accuracy and Transferability of Gaussian Approximation Potential Models for Tungsten”. In: *Phys. Rev. B* 90.10 (Sept. 2014), p. 104108. DOI: 10.1103/PhysRevB.90.104108.

[44] Daniel Marchand, Abhinav Jain, Albert Glensk, and W. A. Curtin. “Machine Learning for Metallurgy I. A Neural-Network Potential for Al-Cu”. In: *Phys. Rev. Materials* 4.10 (Oct. 2020), p. 103601. DOI: 10.1103/PhysRevMaterials.4.103601.

[45] Markus Stricker, Binglun Yin, Eleanor Mak, and W. A. Curtin. “Machine Learning for Metallurgy II. A Neural-Network Potential for Magnesium”. In: *Phys. Rev. Materials* 4.10 (Oct. 2020), p. 103602. DOI: 10.1103/PhysRevMaterials.4.103602.

[46] Daniele Dragoni, Thomas D. Daff, Gábor Csányi, and Nicola Marzari. “Achieving DFT Accuracy with a Machine-Learning Interatomic Potential: Thermomechanics and Defects in Bcc Ferromagnetic Iron”. In: *Phys. Rev. Materials* 2.1 (Jan. 2018), p. 013808. DOI: 10.1103/PhysRevMaterials.2.013808.

[47] Bingqing Cheng, Edgar A. Engel, Jörg Behler, Christoph Dellago, and Michele Ceriotti. "Ab Initio Thermodynamics of Liquid and Solid Water". In: *Proc. Natl. Acad. Sci. U. S. A.* 116.4 (Jan. 2019), pp. 1110–1115. DOI: 10.1073/pnas.1815117116.

[48] Ryosuke Jinnouchi, Jonathan Lahnsteiner, Ferenc Karsai, Georg Kresse, and Menno Bokdam. "Phase transitions of hybrid perovskites simulated by machine-learning force fields trained on the fly with Bayesian inference". In: *Physical review letters* 122.22 (2019), p. 225701.

[49] Chiheb Ben Mahmoud, Andrea Anelli, Gábor Csányi, and Michele Ceriotti. "Learning the electronic density of states in condensed matter". In: *Physical Review B* 102.23 (2020), p. 235130.

[50] O Isayev, D Fourches, EN Muratov, C Oses, K Rasch, A Tropsha, and S Curtarolo. "Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints". In: *Chemical Materials* 27 (2015), pp. 735–743.

[51] B Sanchez-Lengeling and A Aspuru-Guzik. "Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering". In: *Science* 361 (2018), pp. 360–365.

[52] J Wang, S Olsson, C Wehmeyer, A P'erez, NE Charron, G De Fabritiis, F No'e, and C Clementi. "Machine Learning of Coarse-Grained Molecular Dynamics Force Fields". In: *ACS Central Science* 5 (2019), pp. 755–767.

[53] J Behler and M Parrinello. "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces". In: *Phys. Rev. Lett.* 98 (2007), p. 146401.

[54] BJ Braams and JM Bowman. "Permutationally invariant potential energy surfaces in high dimensionality". In: *International Reviews in Physical Chemistry* 28 (2009), pp. 577–606.

[55] AP Bart'ok, MC Payne, R Kondor, and G Cs'anyi. "Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons". In: *Physical Review Letters* 104.13 (2010), p. 136403.

[56] GC Sosso, G Miceli, S Caravati, J Behler, and M Bernasconi. "Neural network interatomic potential for the phase change material GeTe". In: *Physical Review B* 85.17 (2012), p. 174103.

[57] J Behler. "Perspective: Machine learning potentials for atomistic simulations". In: *The Journal of Chemical Physics* 145.17 (2016), p. 170901.

[58] Volker L Deringer and Gábor Csányi. "Machine learning based interatomic potential for amorphous carbon". In: *Physical Review B* 95.9 (2017), p. 094203.

[59] Y Mishin. "Machine-learning interatomic potentials for materials science". In: *Acta Materialia* 214 (2021), p. 116980.

[60] Daniel Marchand, Abhinav Jain, Albert Glensk, and WA Curtin. "Machine learning for metallurgy I. A neural-network potential for Al-Cu". In: *Physical Review Materials* 4.10 (2020), p. 103601.

[61] Gabriele C Sosso, Volker L Deringer, Stephen R Elliott, and Gábor Csányi. "Understanding the thermal properties of amorphous solids using machine-learning-based interatomic potentials". In: *Molecular Simulation* 44.11 (2018), pp. 866–880.

[62] Felix C Mocanu, Konstantinos Konstantinou, Tae Hoon Lee, Noam Bernstein, Volker L Deringer, Gábor Csányi, and Stephen R Elliott. "Modeling the phase-change memory material, Ge2Sb2Te5, with a machine-learned interatomic potential". In: *The Journal of Physical Chemistry B* 122.38 (2018), pp. 8998–9006.

[63] Yunxing Zuo, Chi Chen, Xiangguo Li, Zhi Deng, Yiming Chen, Jörg Behler, Gábor Csányi, Alexander V Shapeev, Aidan P Thompson, Mitchell A Wood, et al. "Performance and cost assessment of machine learning interatomic potentials". In: *The Journal of Physical Chemistry A* 124.4 (2020), pp. 731–745.

[64] Giulio Imbalzano, Andrea Anelli, Daniele Giofré, Sinja Klees, Jörg Behler, and Michele Ceriotti. "Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials". In: *The Journal of Chemical Physics* 148.24 (June 2018), p. 241730. DOI: 10.1063/1.5024611.

[65] Michael W Mahoney and Petros Drineas. "CUR matrix decompositions for improved data analysis". In: *Proceedings of the National Academy of Sciences* 106.3 (2009), pp. 697–702.

[66] Justin S Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E Roitberg. "Less is more: Sampling chemical space with active learning". In: *The Journal of chemical physics* 148.24 (2018), p. 241733.

[67] Troy D Loeffler, Tarak K Patra, Henry Chan, Mathew Cherukara, and Subramanian KRS Sankaranarayanan. "Active learning the potential energy landscape for water clusters from sparse training data". In: *The Journal of Physical Chemistry C* 124.8 (2020), pp. 4907–4916.

[68] Max Hodapp and Alexander Shapeev. "Machine-learning potentials enable predictive and tractable high-throughput screening of random alloys". In: *Physical Review Materials* 5.11 (2021), p. 113802.

[69] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[70] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015.

[71] Jörg Behler. "Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations". In: *Physical Chemistry Chemical Physics* 13.40 (2011), p. 17930. DOI: 10.1039/c1cp21668f.

[72] Albert P. Bartók, Risi Kondor, and Gábor Csányi. "On representing chemical environments". In: *Physical Review B* 87.18 (May 28, 2013). DOI: 10.1103/PhysRevB.87.184115.

[73] W J Szlachta, A P Bartok, and G Csa nyi. "Accuracy and Transferability of Gaussian Approximation Potential Models for Tungsten". In: *Phys. Rev. B: Condens. Matter Mater. Phys.* 90 (2014), p. 104108.

[74] Daniele Dragoni, Thomas D Daff, Gábor Csányi, and Nicola Marzari. "Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron". In: *Physical Review Materials* 2.1 (2018), p. 013808.

[75] A Thompson, L Swiler, C Trott, S Foiles, and G Tucker. "Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials". In: *J. Comput. Phys.* 285 (2015), p. 316.

[76] C Chen, Z Deng, R Tran, H Tang, I-H Chu, and S P Ong. "Accurate Force Field for Molybdenum by Machine Learning Large Materials Data". In: *Phys. Rev. Mater.* 1 (2017), p. 043603.

[77] X-G Li, C Hu, C Chen, Z Deng, J Luo, and S P Ong. "Quantum-Accurate Spectral Neighbor Analysis Potential Models for Ni-Mo Binary Alloys and Fcc Metals". In: *Phys. Rev. B: Condens. Matter Mater. Phys.* 98 (2018), p. 094104.

[78] Z Deng, C Chen, X-G Li, and S P Ong. “An Electrostatic Spectral Neighbor Analysis Potential for Lithium Nitride”. In: *npj Comput. Mater.* 5 (2019), p. 75.

[79] A V Shapeev. “Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials”. In: *Multiscale Model. Simul.* 14 (2016), pp. 1153–1173.

[80] E V Podryabinkin and A V Shapeev. “Active Learning of Linearly Parametrized Interatomic Potentials”. In: *Comput. Mater. Sci.* 140 (2017), pp. 171–180.

[81] K Gubaev, E V Podryabinkin, G L W Hart, and A V Shapeev. “Accelerating high-throughput searches for new alloys with active learning of interatomic potentials”. In: *Comput. Mater. Sci.* 156 (2019), pp. 148–156.

[82] V Botu and R Ramprasad. “Adaptive Machine Learning Framework to Accelerate Ab Initio Molecular Dynamics”. In: *Int. J. Quantum Chem.* 115 (2015), pp. 1074–1083.

[83] M Rupp, A Tkatchenko, K-R Müller, and O A von Lilienfeld. “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning”. In: *Phys. Rev. Lett.* 108 (2012), p. 058301.

[84] L Zhang, J Han, H Wang, R Car, and W E. “Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics”. In: *Phys. Rev. Lett.* 120 (2018), p. 143001.

[85] T Bereau, R A DiStasio, A Tkatchenko, and O A von Lilienfeld. “Non-Covalent Interactions across Organic and Biological Subsets of Chemical Space: Physics-Based Potentials Parametrized from Machine Learning”. In: *J. Chem. Phys.* 148 (2018), p. 241706.

[86] J S Smith, O Isayev, and A E Roitberg. “ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost”. In: *Chem. Sci.* 8 (2017), pp. 3192–3203.

[87] K Yao, J E Herr, D W Toth, R Mckintyre, and J Parkhill. “The TensorMol-0.1 Model Chemistry: A Neural Network Augmented with Long-Range Physics”. In: *Chem. Sci.* 9 (2018), pp. 2261–2269.

[88] V Botu and R Ramprasad. “Learning Scheme to Predict Atomic Forces and Accelerate Materials Simulations”. In: *Phys. Rev. B: Condens. Matter Mater. Phys.* 92 (2015), p. 094306.

[89] V Botu, R Batra, J Chapman, and R Ramprasad. “Machine Learning Force Fields: Construction, Validation, and Outlook”. In: *J. Phys. Chem. C* 121 (2017), pp. 511–522.

[90] Z Li, J R Kermode, and A De Vita. “Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces”. In: *Phys. Rev. Lett.* 114 (2015), p. 096405.

[91] K T Schütt, P Kessel, M Gastegger, K A Nicoli, A Tkatchenko, and K-R Müller. "SchNetPack: A Deep Learning Toolbox for Atomistic Systems". In: *J. Chem. Theory Comput.* 15 (2019), pp. 448–455.

[92] F Brockherde, L Vogt, L Li, M E Tuckerman, K Burke, and K-R Müller. "Bypassing the Kohn-Sham Equations with Machine Learning". In: *Nat. Commun.* 8 (2017), p. 872.

[93] M Rupp. "Machine Learning for Quantum Mechanics in a Nutshell". In: *Int. J. Quantum Chem.* 115 (2015), pp. 1058–1073.

[94] Y Huang, J Kang, W A Goddard, and L-W Wang. "Density Functional Theory Based Neural Network Force Fields from Energy Decompositions". In: *Phys. Rev. B: Condens. Matter Mater. Phys.* 99 (2019), p. 064103.

[95] K Hansen, F Biegler, R Ramakrishnan, W Pronobis, O A von Lilienfeld, K-R Müller, and A Tkatchenko. "Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space". In: *J. Phys. Chem. Lett.* 6 (2015), pp. 2326–2331.

[96] Jörg Behler. "Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials". In: *The Journal of Chemical Physics* 134.7 (Feb. 2011), p. 074106. DOI: 10.1063/1.3553717.

[97] Michael J. Willatt, Félix Musil, and Michele Ceriotti. "Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements". In: *Physical Chemistry Chemical Physics* 20.47 (2018), pp. 29661–29668. DOI: 10.1039/c8cp05921g.

[98] Max Veit, David M. Wilkins, Yang Yang, Robert A. DiStasio, and Michele Ceriotti. "Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles". In: *The Journal of Chemical Physics* 153.2 (July 2020), p. 024113. DOI: 10.1063/5.0009106.

[99] Andrew E Sifain, Nicholas Lubbers, Benjamin T Nebgen, Justin S Smith, Andrey Y Lokhov, Olexandr Isayev, Adrian E Roitberg, Kipton Barros, and Sergei Tretiak. "Discovering a transferable charge assignment model using machine learning". In: *The journal of physical chemistry letters* 9.16 (2018), pp. 4495–4501.

[100] David M. Wilkins, Andrea Grisafi, Yang Yang, Ka Un Lao, Robert A. DiStasio, and Michele Ceriotti. "Accurate Molecular Polarizabilities with Coupled Cluster Theory and Machine Learning". In: *Proc. Natl. Acad. Sci. U. S. A.* 116.9 (Feb. 2019), pp. 3401–3406. DOI: 10.1073/pnas.1816132116.

[101] Chaoqiang Feng, Jin Xi, Yaolong Zhang, Bin Jiang, and Yong Zhou. "Accurate and Interpretable Dipole Interaction Model-Based Machine Learning for Molecular Polarizability". In: *Journal of Chemical Theory and Computation* (2023).

[102] Andrea Grisafi, Alberto Fabrizio, Benjamin Meyer, David M. Wilkins, Clemence Corminboeuf, and Michele Ceriotti. "Transferable Machine-Learning Model of the Electron Density". In: *ACS Cent. Sci.* 5.1 (Jan. 2019), pp. 57–64. DOI: 10.1021/acscentsci.8b00551.

[103] Jigyasa Nigam, Michael J Willatt, and Michele Ceriotti. "Equivariant representations for molecular Hamiltonians and N-center atomic-scale properties". In: *The Journal of Chemical Physics* 156.1 (2022), p. 014115.

[104] Edoardo Cignoni, Lorenzo Cupellini, and Benedetta Mennucci. "Machine Learning Exciton Hamiltonians in Light-Harvesting Complexes". In: *Journal of Chemical Theory and Computation* 19.3 (2023), pp. 965–977.

[105] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.1 (2016), pp. 1–9.

[106] Leopold Talirz, Snehal Kumbhar, Elsa Passaro, Aliaksandr V Yakutovich, Valeria Granata, Fernando Gargiulo, Marco Borelli, Martin Uhrin, Sebastiaan P Huber, Spyros Zoupanos, et al. "Materials Cloud, a platform for open computational science". In: *Scientific data* 7.1 (2020), pp. 1–12.

[107] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation". In: *APL materials* 1.1 (2013), p. 011002.

[108] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. "Open catalyst 2020 (OC20) dataset and community challenges". In: *ACS Catalysis* 11.10 (2021), pp. 6059–6072.

[109] Colin R Groom, Ian J Bruno, Matthew P Lightfoot, and Suzanna C Ward. "The Cambridge structural database". In: *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* 72.2 (2016), pp. 171–179.

[110] Lorenz C. Blum and Jean-Louis Reymond. "970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13". In: *J. Am. Chem. Soc.* 131.25 (July 2009), pp. 8732–8733. DOI: 10.1021/ja902302h.

[111] Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus Robert Müller, and O. Anatole Von Lilienfeld. "Machine Learning of Molecular Electronic Properties in Chemical Compound Space". In: *New J. Phys.* 15 (2013), p. 095003. DOI: 10.1088/1367-2630/15/9/095003.

[112] Noam Bernstein, Gábor Csányi, and Volker L Deringer. "De novo exploration and self-guided learning of potential-energy surfaces". In: *npj Computational Materials* 5.1 (2019), pp. 1–9.

[113] Chris J Pickard and RJ Needs. "High-pressure phases of silane". In: *Physical review letters* 97.4 (2006), p. 045504.

[114] Ivan S Novikov, Konstantin Gubaev, Evgeny V Podryabinkin, and Alexander V Shapeev. "The MLIP package: moment tensor potentials with MPI and active learning". In: *Machine Learning: Science and Technology* 2.2 (2020), p. 025002.

[115] Jonathan Vandermause, Steven B Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M Kolpak, and Boris Kozinsky. "On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events". In: *npj Computational Materials* 6.1 (2020), pp. 1–11.

[116] J.-W. Yeh, S.-K. Chen, S.-J. Lin, J.-Y. Gan, T.-S. Chin, T.-T. Shun, C.-H. Tsau, and S.-Y. Chang. "Nanostructured High-Entropy Alloys with Multiple Principal Elements: Novel Alloy Design Concepts and Outcomes". In: *Adv. Eng. Mater.* 6.5 (May 2004), pp. 299–303. DOI: 10.1002/adem.200300567.

[117] B. Cantor, I.T.H. Chang, P. Knight, and A.J.B. Vincent. "Microstructural Development in Equiatomic Multicomponent Alloys". In: *Materials Science and Engineering: A* 375–377 (July 2004), pp. 213–218. DOI: 10.1016/j.msea.2003.10.257.

[118] B. Cantor. "Multicomponent High-Entropy Cantor Alloys". In: *Progress in Materials Science* 120 (July 2021), p. 100754. DOI: 10.1016/j.pmatsci.2020.100754.

[119] E.P. George, W.A. Curtin, and C.C. Tasan. "High Entropy Alloys: A Focused Review of Mechanical Properties and Deformation Mechanisms". In: *Acta Materialia* 188 (Apr. 2020), pp. 435–474. DOI: 10.1016/j.actamat.2019.12.015.

[120] Weidong Li, Di Xie, Dongyue Li, Yong Zhang, Yanfei Gao, and Peter K. Liaw. "Mechanical Behavior of High-Entropy Alloys". In: *Progress in Materials Science* 118 (May 2021), p. 100777. DOI: 10.1016/j.pmatsci.2021.100777.

[121] Yifan Sun and Sheng Dai. "High-entropy materials for catalysis: A new frontier". In: *Science Advances* 7.20 (May 2021). DOI: 10.1126/sciadv.abg1600.

[122] Bing Wang, Yingfang Yao, Xiwen Yu, Cheng Wang, Congping Wu, and Zhigang Zou. "Understanding the enhanced catalytic activity of high entropy alloys: from theory to experiment". In: *Journal of Materials Chemistry A* 9.35 (2021), pp. 19410–19438. DOI: 10.1039/d1ta02718b.

[123] Wen-Yi Huo, Shi-Qi Wang, Wen-Han Zhu, Ze-Ling Zhang, Feng Fang, Zong-Han Xie, and Jian-Qing Jiang. "Recent progress on high-entropy materials for electrocatalytic water splitting applications". In: *Tungsten* 3.2 (May 2021), pp. 161–180. DOI: 10.1007/s42864-021-00084-8.

[124] Yiqiong Zhang, Dongdong Wang, and Shuangyin Wang. "High-Entropy Alloys for Electrocatalysis: Design, Characterization, and Applications". In: *Small* 18.7 (Nov. 2021), p. 2104339. DOI: 10.1002/smll.202104339.

[125] Xiaoran Huo, Huishu Yu, Bowei Xing, Xiaojiao Zuo, and Nannan Zhang. "Review of High Entropy Alloys Electrocatalysts for Hydrogen Evolution, Oxygen Evolution, and Oxygen Reduction Reaction". In: *The Chemical Record* (Sept. 2022). DOI: 10.1002/tcr.202200175.

[126] Guoliang Zhang, Kaisheng Ming, Jianli Kang, Qin Huang, Zhijia Zhang, Xuerong Zheng, and Xiaofang Bi. "High entropy alloy as a highly active and stable electrocatalyst for hydrogen evolution reaction". In: *Electrochimica Acta* 279 (July 2018), pp. 19–23. DOI: 10.1016/j.electacta.2018.05.035.

[127] Kang Huang, Bowei Zhang, Junsheng Wu, Tianyuan Zhang, Dongdong Peng, Xun Cao, Zhan Zhang, Zhong Li, and Yizhong Huang. "Exploring the impact of atomic lattice deformation on oxygen evolution reactions based on a sub-5 nm pure face-centred cubic high-entropy alloy electrocatalyst". In: *Journal of Materials Chemistry A* 8.24 (2020), pp. 11938–11947. DOI: 10.1039/d0ta02125c.

[128] Dongshuang Wu, Kohei Kusada, Tomokazu Yamamoto, Takaaki Toriyama, Syo Matsumura, Ibrahima Gueye, Okkyun Seo, Jaemyung Kim, Satoshi Hiroi, Osami Sakata, Shogo Kawaguchi, Yoshiki Kubota, and Hiroshi Kitagawa. "On the electronic structure and hydrogen evolution reaction activity of platinum group metal-based high-entropy-alloy nanoparticles". In: *Chemical Science* 11.47 (2020), pp. 12731–12736. DOI: 10.1039/d0sc02351e.

[129] Martin Bondesgaard, Nils Lau Nyborg Broge, Aref Mamakhel, Martin Bremholm, and Bo Brummerstedt Iversen. "General Solvothermal Synthesis Method for Complete Solubility Range Bimetallic and High-Entropy Alloy Nanocatalysts". In: *Advanced Functional Materials* 29.50 (Oct. 2019), p. 1905933. DOI: 10.1002/adfm.201905933.

[130] Matthew W. Glasscott, Andrew D. Pendergast, Sondrica Goines, Anthony R. Bishop, Andy T. Hoang, Christophe Renault, and Jeffrey E. Dick. "Electrosynthesis of high-entropy metallic glass nanoparticles for designer, multi-functional electrocatalysis". In: *Nature Communications* 10.1 (June 2019). DOI: 10.1038/s41467-019-10303-z.

[131] Zeyu Jin, Juan Lv, Henglei Jia, Weihong Liu, Huanglong Li, Zuhuang Chen, Xi Lin, Guoqiang Xie, Xingjun Liu, Shuhui Sun, and Hua-Jun Qiu. "Nanoporous Al-Ni-Co-Ir-Mo High-Entropy Alloy for Record-High Water Splitting Activity in Acidic Environments". In: *Small* 15.47 (Oct. 2019), p. 1904180. DOI: 10.1002/smll.201904180.

[132] Steven D. Lacey, Qi Dong, Zhennan Huang, Jingru Luo, Hua Xie, Zhiwei Lin, Dylan J. Kirsch, Vivek Vattipalli, Christopher Povinelli, Wei Fan, Reza Shahbazian-Yassar, Dunwei Wang, and Liangbing Hu. "Stable Multimetallic Nanoparticles for Oxygen Electrocatalysis". In: *Nano Letters* 19.8 (July 2019), pp. 5149–5158. DOI: 10.1021/acs.nanolett.9b01523.

[133] Miaomiao Liu, Zihao Zhang, Francis Okejiri, Shize Yang, Shenghu Zhou, and Sheng Dai. "Entropy-Maximized Synthesis of Multimetallic Nanoparticle Catalysts via a Ultrasonication-Assisted Wet Chemistry Method under Ambient Conditions". In: *Advanced Materials Interfaces* 6.7 (Feb. 2019), p. 1900015. DOI: 10.1002/admi.201900015.

[134] Hua-Jun Qiu, Gang Fang, Jiaojiao Gao, Yuren Wen, Juan Lv, Huanglong Li, Guoqiang Xie, Xingjun Liu, and Shuhui Sun. "Noble Metal-Free Nanoporous High-Entropy Alloys as Highly Efficient Electrocatalysts for Oxygen Evolution Reaction". In: *ACS Materials Letters* 1.5 (Oct. 2019), pp. 526–533. DOI: 10.1021/acsmaterialslett.9b00414.

[135] Hua-Jun Qiu, Gang Fang, Yuren Wen, Pan Liu, Guoqiang Xie, Xingjun Liu, and Shuhui Sun. "Nanoporous high-entropy alloys for highly stable and efficient catalysts". In: *Journal of Materials Chemistry A* 7.11 (2019), pp. 6499–6506. DOI: 10.1039/c9ta00505f.

[136] Shaojie Gao, Shaoyun Hao, Zhennan Huang, Yifei Yuan, Song Han, Lecheng Lei, Xingwang Zhang, Reza Shahbazian-Yassar, and Jun Lu. "Synthesis of high-entropy alloy nanoparticles on supports by the fast moving bed pyrolysis". In: *Nature Communications* 11.1 (Apr. 2020). DOI: 10.1038/s41467-020-15934-1.

[137] Xiaoting Chen, Conghui Si, Yulai Gao, Jan Frenzel, Junzhe Sun, Gunther Eggeler, and Zhonghua Zhang. "Multi-component nanoporous platinum–ruthenium–copper–osmium–iridium alloy with enhanced electrocatalytic activity towards methanol oxidation and oxygen reduction". In: *Journal of Power Sources* 273 (Jan. 2015), pp. 324–332. DOI: 10.1016/j.jpowsour.2014.09.076.

[138] Tobias Löffler, Hajo Meyer, Alan Savan, Patrick Wilde, Alba Garzón Manjón, Yen-Ting Chen, Edgar Ventosa, Christina Scheu, Alfred Ludwig, and Wolfgang Schuhmann. "Discovery of a Multinary Noble Metal-Free Oxygen Reduction Catalyst". In: *Advanced Energy Materials* 8.34 (Oct. 2018), p. 1802269. DOI: 10.1002/aenm.201802269.

[139] Shiyin Li, Xiaowei Tang, Henglei Jia, Huanglong Li, Guoqiang Xie, Xingjun Liu, Xi Lin, and Hua-Jun Qiu. "Nanoporous high-entropy alloys with low Pt loadings for high-performance electrochemical oxygen reduction". In: *Journal of Catalysis* 383 (Mar. 2020), pp. 164–171. DOI: 10.1016/j.jcat.2020.01.024.

[140] J. Barranco and A.R. Pierna. "On the enhancement of methanol and CO electro-oxidation by amorphous (NiNb)PtSnRu alloys versus bifunctional PtRu and PtSn alloys". In: *Journal of Non-Crystalline Solids* 354.47-51 (Dec. 2008), pp. 5153–5155. DOI: 10.1016/j.jnoncrysol.2008.04.053.

[141] Chih-Fang Tsai, Kung-Yu Yeh, Pu-Wei Wu, Yi-Fan Hsieh, and Pang Lin. "Effect of platinum present in multi-element nanoparticles on methanol oxidation". In: *Journal of Alloys and Compounds* 478.1-2 (June 2009), pp. 868–871. DOI: 10.1016/j.jallcom.2008.12.055.

[142] An-Liang Wang, Hao-Chuan Wan, Han Xu, Ye-Xiang Tong, and Gao-Ren Li. "Quinary PdNiCoCuFe Alloy Nanotube Arrays as Efficient Electrocatalysts for Methanol Oxidation". In: *Electrochimica Acta* 127 (May 2014), pp. 448–453. DOI: 10.1016/j.electacta.2014.02.076.

[143] Kirill V. Yusenko, Sephira Riva, Patricia A. Carvalho, Maria V. Yusenko, Serena Arnaboldi, Aleksandr S. Sukhikh, Michael Hanfland, and Sergey A. Gromilov. "First hexagonal close packed high-entropy alloy with outstanding stability under extreme conditions and electrocatalytic activity for methanol oxidation". In: *Scripta Materialia* 138 (Sept. 2017), pp. 22–27. DOI: 10.1016/j.scriptamat.2017.05.022.

[144] Jien-Wei Yeh. "Alloy Design Strategies and Future Trends in High-Entropy Alloys". In: *JOM* 65.12 (Oct. 2013), pp. 1759–1771. DOI: 10.1007/s11837-013-0761-6.

[145] E. J. Pickering and N. G. Jones. "High-entropy alloys: a critical assessment of their founding principles and future prospects". In: *International Materials Reviews* 61.3 (Apr. 2016), pp. 183–202. DOI: 10.1080/09506608.2016.1180020.

[146] Sandipan Sen, Xi Zhang, Lukasz Rogal, Gerhard Wilde, Blazej Grabowski, and Sergiy V Divinski. "'Anti-sluggish'Ti diffusion in HCP high-entropy alloys: Chemical complexity vs. lattice distortions". In: *Scripta Materialia* 224 (2023), p. 115117.

[147] Jiashi Miao, CE Slone, TM Smith, C Niu, Hongbin Bei, M Ghazisaeidi, GM Pharr, and Michael J Mills. "The evolution of the deformation substructure in a Ni-Co-Cr equiatomic solid solution alloy". In: *Acta Materialia* 132 (2017), pp. 35–48.

[148] Fuyang Tian, Lajos Károly Varga, Jiang Shen, and Levente Vitos. "Calculating elastic constants in high-entropy alloys using the coherent potential approximation: Current issues and errors". In: *Computational materials science* 111 (2016), pp. 350–358.

[149] IA Balyakin, AA Yuryev, BR Gelchinski, and AA Rempel. "Ab initio molecular dynamics and high-dimensional neural network potential study of VZrNbHfTa melt". In: *Journal of Physics: Condensed Matter* 32.21 (2020), p. 214006.

[150] Konstantin Gubaev, Yuji Ikeda, Ferenc Tasnádi, Jörg Neugebauer, Alexander V Shapeev, Blazej Grabowski, and Fritz Körmann. "Finite-temperature interplay of structural stability, chemical complexity, and elastic properties of bcc multicomponent alloys from ab initio trained machine-learning potentials". In: *Physical Review Materials* 5.7 (2021), p. 073801.

[151] Mehdi Jafary-Zadeh, Khoong Hong Khoo, Robert Laskowski, Paulo S Branicio, and Alexander V Shapeev. "Applying a machine learning interatomic potential to unravel the effects of local lattice distortion on the elastic properties of multi-principal element alloys". In: *Journal of Alloys and Compounds* 803 (2019), pp. 1054–1062.

[152] Jigyasa Nigam, Sergey Pozdnyakov, and Michele Ceriotti. "Recursive evaluation and iterative contraction of N-body equivariant features". In: *The Journal of Chemical Physics* 153.12 (Sept. 2020), p. 121101. DOI: 10.1063/5.0021116.

[153] James P Darby, Dávid P Kovács, Ilyes Batatia, Miguel A Caro, Gus LW Hart, Christoph Ortner, and Gábor Csányi. "Tensor-reduced atomic density representations". In: *arXiv preprint arXiv:2210.01705* (2022).

[154] M Attarian Shandiz and R Gauvin. "Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries". In: *Computational Materials Science* 117 (2016), pp. 270–278.

[155] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žıdek, Anna Potapenko, et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589.

[156] Zach Jensen, Edward Kim, Soonhyoung Kwon, Terry ZH Gani, Yuriy Román-Leshkov, Manuel Moliner, Avelino Corma, and Elsa Olivetti. "A machine learning approach to zeolite synthesis enabled by automatic literature data extraction". In: *ACS central science* 5.5 (2019), pp. 892–899.

[157] Edesio Alcobaca, Saulo Martiello Mastelini, Tiago Botari, Bruno Almeida Pimentel, Daniel Roberto Cassar, André Carlos Ponce de Leon Ferreira, Edgar Dutra Zanotto, et al. "Explainable machine learning algorithms for predicting glass transition temperatures". In: *Acta Materialia* 188 (2020), pp. 92–100.

[158] Prabudhya Roy Chowdhury, Colleen Reynolds, Adam Garrett, Tianli Feng, Shashishekar P Adiga, and Xiulin Ruan. "Machine learning maximized Anderson localization of phonons in aperiodic superlattices". In: *Nano Energy* 69 (2020), p. 104428.

[159] Yong-Jie Hu, Ge Zhao, Mingfei Zhang, Bin Bin, Tyler Del Rose, Qian Zhao, Qun Zu, Yang Chen, Xuekun Sun, Maarten de Jong, et al. "Predicting densities and elastic moduli of SiO2-based glasses by machine learning". In: *Npj Computational Materials* 6.1 (2020), p. 25.

[160] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. "A general-purpose machine learning framework for predicting properties of inorganic materials". In: *npj Computational Materials* 2.1 (2016), pp. 1–7.

[161] Volker L Deringer, Miguel A Caro, and Gábor Csányi. "Machine learning interatomic potentials as emerging tools for materials science". In: *Advanced Materials* 31.46 (2019), p. 1902765.

[162] Dylan M Anstine and Olexandr Isayev. "Machine Learning Interatomic Potentials and Long-Range Physics". In: *The Journal of Physical Chemistry A* (2023).

[163] George Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314.

[164] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5 (1989), pp. 359–366.

[165] Harris Drucker, Christopher J Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. "Support vector regression machines". In: *Advances in neural information processing systems* 9 (1996).

[166] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[167] Svante Wold, Kim Esbensen, and Paul Geladi. "Principal component analysis". In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.

[168] Michele Ceriotti, Gareth A Tribello, and Michele Parrinello. "Demonstrating the transferability and the descriptive power of sketch-map". In: *Journal of chemical theory and computation* 9.3 (2013), pp. 1521–1532.

[169] Michael J Willatt, Félix Musil, and Michele Ceriotti. "Atom-density representations for machine learning". In: *The Journal of chemical physics* 150.15 (2019), p. 154110.

[170] Felix Musil, Andrea Grisafi, Albert P Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. "Physics-inspired structural representations for molecules and materials". In: *Chemical Reviews* 121.16 (2021), pp. 9759–9815.

[171] Yoav Goldberg. "A primer on neural network models for natural language processing". In: *Journal of Artificial Intelligence Research* 57 (2016), pp. 345–420.

[172] Ray J Frank, Neil Davey, and Stephen P Hunt. "Time series prediction and neural networks". In: *Journal of intelligent and robotic systems* 31 (2001), pp. 91–103.

[173] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. "Speech recognition using deep neural networks: A systematic review". In: *IEEE access* 7 (2019), pp. 19143–19165.

[174] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[175] Michael W. Mahoney and Petros Drineas. "CUR Matrix Decompositions for Improved Data Analysis". In: *Proc. Natl. Acad. Sci. U. S. A.* 106.3 (Jan. 2009), pp. 697–702. DOI: 10.1073/pnas.0803205106.

[176] Giulio Imbalzano, Andrea Anelli, Daniele Giofré, Sinja Klees, Jörg Behler, and Michele Ceriotti. "Automatic Selection of Atomic Fingerprints and Reference Configurations for Machine-Learning Potentials". In: *J. Chem. Phys.* 148.24 (June 2018), p. 241730. DOI: 10.1063/1.5024611.

[177] Michael J. Willatt, Félix Musil, and Michele Ceriotti. "Atom-Density Representations for Machine Learning". In: *J. Chem. Phys.* 150.15 (Apr. 2019), p. 154110. DOI: 10.1063/1.5090481.

[178] Alexander V. Shapeev. "Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials". In: *Multiscale Model. Simul.* 14.3 (Jan. 2016), pp. 1153–1173. DOI: 10.1137/15M1054183.

[179] Ralf Drautz. "Atomic Cluster Expansion for Accurate and Transferable Interatomic Potentials". In: *Phys. Rev. B* 99.1 (Jan. 2019), p. 014104. DOI: 10.1103/PhysRevB.99.014104.

[180] Michael J. Willatt, Michele Ceriotti, and Stuart C. Althorpe. "Approximating Matsubara Dynamics Using the Planetary Model: Tests on Liquid Water and Ice". In: *J. Chem. Phys.* 148.10 (Mar. 2018), p. 102336. DOI: 10.1063/1.5004808.

[181] "International Year of the Periodic Table from a Physical Chemistry Perspective". In: *J. Phys. Chem. Lett.* 10.19 (Oct. 2019). Ed. by George C. Schatz, pp. 5956–5956. DOI: 10.1021/acs.jpclett.9b02147.

[182] Nongnuch Artrith, Alexander Urban, and Gerbrand Ceder. "Efficient and Accurate Machine-Learning Interpolation of Atomic Energies in Compositions with Many Species". In: *Phys. Rev. B* 96.1 (July 2017), p. 014112. DOI: 10.1103/PhysRevB.96.014112.

[183] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand. "wACSF—Weighted Atom-Centered Symmetry Functions as Descriptors in Machine Learning Potentials". In: *J. Chem. Phys.* 148.24 (June 2018), p. 241709. DOI: 10.1063/1.5019667.

[184] James P. Darby, James R. Kermode, and Gábor Csányi. "Compressing Local Atomic Neighbourhood Descriptors". In: *npj Comput Mater* 8.1 (Aug. 2022), p. 166. DOI: 10.1038/s41524-022-00847-y.

[185] Alexander Goscinski, Félix Musil, Sergey Pozdnyakov, Jigyasa Nigam, and Michele Ceriotti. "Optimal radial basis for density-based atomic representations". In: *The Journal of Chemical Physics* 155.10 (2021), p. 104106.

[186] Ivan S. Novikov and Alexander V. Shapeev. "Improving Accuracy of Interatomic Potentials: More Physics or More Data? A Case Study of Silica". In: *Materials Today Communications* 18 (Mar. 2019), pp. 74–80. DOI: 10.1016/j.mtcomm.2018.11.008.

[187] Jörg Behler. "Constructing high-dimensional neural network potentials: A tutorial review". In: *International Journal of Quantum Chemistry* 115.16 (Aug. 15, 2015), pp. 1032–1050. DOI: 10.1002/qua.24890.

[188] Stefano Curtarolo, Gus LW Hart, Marco Buongiorno Nardelli, Natalio Mingo, Stefano Sanvito, and Ohad Levy. "The high-throughput highway to computational materials design". In: *Nature materials* 12.3 (2013), pp. 191–201.

[189] Artem R. Oganov, Chris J. Pickard, Qiang Zhu, and Richard J. Needs. "Structure Prediction Drives Materials Discovery". In: *Nat Rev Mater* 4.5 (May 2019), pp. 331–348. DOI: 10.1038/s41578-019-0101-8.

[190] Anubhav Jain, Yongwoo Shin, and Kristin A Persson. "Computational predictions of energy materials using density functional theory". In: *Nature Reviews Materials* 1.1 (2016), pp. 1–13.

[191] Zhi Wei Seh, Jakob Kibsgaard, Colin F Dickens, IB Chorkendorff, Jens K Nørskov, and Thomas F Jaramillo. "Combining theory and experiment in electrocatalysis: Insights into materials design". In: *Science* 355.6321 (2017), eaad4998.

[192] William C Swope, Hans C Andersen, Peter H Berens, and Kent R Wilson. "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters". In: *The Journal of chemical physics* 76.1 (1982), pp. 637–649.

[193] Hans C Andersen. "Molecular dynamics simulations at constant pressure and/or temperature". In: *The Journal of chemical physics* 72.4 (1980), pp. 2384–2393.

[194] Shuichi Nosé. "A unified formulation of the constant temperature molecular dynamics methods". In: *The Journal of chemical physics* 81.1 (1984), pp. 511–519.

[195] Michele Ceriotti, Giovanni Bussi, and Michele Parrinello. "Langevin Equation with Colored Noise for Constant-Temperature Molecular Dynamics Simulations". In: *Phys. Rev. Lett.* 102.2 (Jan. 2009), p. 020601. DOI: 10.1103/PhysRevLett.102.020601.

[196] Giovanni Bussi, Tatyana Zykova-Timan, and Michele Parrinello. "Isothermal-Isobaric Molecular Dynamics Using Stochastic Velocity Rescaling". In: *J. Chem. Phys.* 130.7 (Feb. 2009), p. 074101. DOI: 10.1063/1.3073889.

[197] Wei Zhang, Chun Wu, and Yong Duan. "Convergence of replica exchange molecular dynamics". In: *The Journal of chemical physics* 123.15 (2005), p. 154105.

[198] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. "Equation of state calculations by fast computing machines". In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[199] Massimo Bocus, Ruben Goeminne, Aran Lamaire, Maarten Cools-Ceuppens, Toon Verstraelen, and Veronique Van Speybroeck. "Nuclear quantum effects on zeolite proton hopping kinetics explored with machine learning potentials and path integral molecular dynamics". In: *Nature Communications* 14.1 (2023), p. 1008.

[200] Dmitriy Borodin, Nils Hertl, G Barratt Park, Michael Schwarzer, Jan Fingerhut, Yingqi Wang, Junxiang Zuo, Florian Nitz, Georgios Skoulatakis, Alexander Kandratsenka, et al. "Quantum effects in thermal reaction rates at metal surfaces". In: *Science* 377.6604 (2022), pp. 394–398.

[201] David Chandler and Peter G Wolynes. "Exploiting the isomorphism between quantum theory and classical statistical mechanics of polyatomic fluids". In: *The Journal of Chemical Physics* 74.7 (1981), pp. 4078–4095.

[202] Fritz Körmann, Alexey Dick, Blazej Grabowski, B Hallstedt, T Hickel, and J Neugebauer. "Free energy of bcc iron: Integrated ab initio derivation of vibrational, electronic, and magnetic contributions". In: *Physical Review B* 78.3 (2008), p. 033102.

[203] Roman Martoňák, Davide Donadio, Artem R Oganov, and Michele Parrinello. "Crystal structure transformations in SiO2 from classical and ab initio metadynamics". In: *Nature materials* 5.8 (2006), pp. 623–626.

[204] Roman Martoňák, Alessandro Laio, Marco Bernasconi, Chiara Ceriani, Paolo Raiteri, Federico Zipoli, and Michele Parrinello. "Simulation of structural phase transitions by metadynamics". In: *Zeitschrift für Kristallographie-Crystalline Materials* 220.5-6 (2005), pp. 489–498.

[205] Bernd Ensing and Michael L Klein. "Perspective on the reactions between F–and CH3CH2F: The free energy landscape of the E2 and SN2 reaction channels". In: *Proceedings of the National Academy of Sciences* 102.19 (2005), pp. 6755–6759.

[206] Scott Habershon and David E Manolopoulos. "Thermodynamic Integration from Classical to Quantum Mechanics." In: *J. Chem. Phys.* 135.22 (Dec. 2011), p. 224111. DOI: 10.1063/1.3666011.

[207] Mariana Rossi, Piero Gasparotto, and Michele Ceriotti. "Anharmonic and Quantum Fluctuations in Molecular Crystals: A First-Principles Study of the Stability of Paracetamol". In: *Phys. Rev. Lett.* 117.11 (Sept. 2016), p. 115702. DOI: 10.1103/PhysRevLett.117.115702.

[208] Bingqing Cheng, Anthony T. Paxton, and Michele Ceriotti. "Hydrogen Diffusion and Trapping in α -Iron: The Role of Quantum and Anharmonic Fluctuations". In: *Phys. Rev. Lett.* 120.22 (May 2018), p. 225901. DOI: 10.1103/PhysRevLett.120.225901.

[209] Kyu-Kwang Han and Hyeon S. Son. "On the isothermal-isobaric ensemble partition function". In: *The Journal of Chemical Physics* 115.16 (Oct. 2001), pp. 7793–7794. DOI: 10.1063/1.1407295.

[210] Ulf R. Pedersen. "Direct calculation of the solid-liquid Gibbs free energy difference in a single equilibrium simulation". In: *The Journal of Chemical Physics* 139.10 (Sept. 14, 2013), p. 104102. DOI: 10.1063/1.4818747.

[211] G.M. M Torrie and J.P. P Valleau. "Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling". In: *J. Comput. Phys.* 23.2 (Feb. 1977), pp. 187–199. DOI: 10.1016/0021-9991(77)90121-8.

[212] A Laio and M Parrinello. "Escaping Free-Energy Minima". In: *Proc. Natl. Acad. Sci.* 99.20 (2002), pp. 12562–12566.

[213] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. "Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method". In: *Phys. Rev. Lett.* 100.2 (2008), p. 20603.

[214] Nataliya Lopanitsyna, Chiheb Ben Mahmoud, and Michele Ceriotti. "Finite-temperature materials modeling from the quantum nuclei to the hot electron regime". In: *Physical Review Materials* 5.4 (2021), p. 043802.

[215] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, et al. "QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials". In: *Journal of physics: Condensed matter* 21.39 (2009), p. 395502.

[216] John P Perdew, Kieron Burke, and Matthias Ernzerhof. "Generalized gradient approximation made simple". In: *Physical review letters* 77.18 (1996), p. 3865.

[217] Kari Laasonen, Roberto Car, Changyol Lee, and David Vanderbilt. "Implementation of ultrasoft pseudopotentials in ab initio molecular dynamics". In: *Physical Review B* 43.8 (1991), p. 6796.

[218] Gianluca Prandini, Antimo Marrazzo, Ivano E. Castelli, Nicolas Mounet, and Nicola Marzari. "Precision and Efficiency in Solid-State Pseudopotential Calculations". In: *npj Comput Mater* 4.1 (Dec. 2018), p. 72. DOI: 10.1038/s41524-018-0127-2.

[219] Hendrik J Monkhorst and James D Pack. "Special points for Brillouin-zone integrations". In: *Physical review B* 13.12 (1976), p. 5188.

[220] M. Methfessel and A. T. Paxton. "High-precision sampling for Brillouin-zone integration in metals". In: *Physical Review B* 40.6 (Aug. 15, 1989), pp. 3616–3621. DOI: 10.1103/PhysRevB.40.3616.

[221] Y. Wang, Z.-K. Liu, and L.-Q. Chen. "Thermodynamic properties of Al, Ni, NiAl, and Ni3Al from first-principles calculations". In: *Acta Materialia* 52.9 (May 2004), pp. 2665–2671. DOI: 10.1016/j.actamat.2004.02.014.

[222] Fritz Körmann, Pui-Wai Ma, Sergei L Dudarev, and Jörg Neugebauer. "Impact of magnetic fluctuations on lattice excitations in fcc nickel". In: *J. Phys.: Condens. Matter* 28.7 (Jan. 2016), p. 076002. DOI: 10.1088/0953-8984/28/7/076002.

[223] Yilun Gong, Blazej Grabowski, Albert Glensk, Fritz Körmann, Jörg Neugebauer, and Roger C. Reed. "Temperature dependence of the Gibbs energy of vacancy formation of fcc Ni". In: *Physical Review B* 97.21 (June 28, 2018), p. 214106. DOI: 10.1103/PhysRevB.97.214106.

[224] Ryo Kobayashi, Daniele Giofré, Till Junge, Michele Ceriotti, and William A. Curtin. "Neural Network Potential for Al-Mg-Si Alloys". In: *Phys. Rev. Mater.* 1.5 (2017), p. 053604. DOI: 10.1103/PhysRevMaterials.1.053604.

[225] Albert P Bartók, James Kermode, Noam Bernstein, and Gábor Csányi. "Machine learning a general-purpose interatomic potential for silicon". In: *Physical Review X* 8.4 (2018), p. 041048.

[226] G. P. Purja Pun and Y. Mishin. "Development of an interatomic potential for the Ni-Al system". In: *Philosophical Magazine* 89.34 (Dec. 1, 2009), pp. 3245–3267. DOI: 10.1080/14786430903258184.

[227] Yuji Sugita and Yuko Okamoto. "Replica-exchange molecular dynamics method for protein folding". In: *Chemical Physics Letters* 314.1-2 (Nov. 1999), pp. 141–151. DOI: 10.1016/s0009-2614(99)01123-9.

[228] Changjun Chen, Yi Xiao, and Yanzhao Huang. "Improving the replica-exchange molecular-dynamics method for efficient sampling in the temperature space". In: *Physical Review E* 91.5 (May 18, 2015). DOI: 10.1103/PhysRevE.91.052708.

[229] Venkat Kapil, Mariana Rossi, Ondrej Marsalek, Riccardo Petraglia, Yair Litman, Thomas Spura, Bingqing Cheng, Alice Cuzzocrea, Robert H. Meißner, David M. Wilkins, Benjamin A. Helfrecht, Przemysław Juda, Sébastien P. Bienvenue, Wei Fang, Jan Kessler, Igor Poltavsky, Steven Vandenbrande, Jelle Wieme, Clemence Corminboeuf, Thomas D. Kühne, David E. Manolopoulos, Thomas E. Markland, Jeremy O. Richardson, Alexandre Tkatchenko, Gareth A. Tribello, Veronique Van Speybroeck, and Michele Ceriotti. "I-PI 2.0: A Universal Force Engine for Advanced Molecular Simulations". In: *Comput. Phys. Commun.* 236 (Mar. 2019), pp. 214–223. DOI: 10.1016/j.cpc.2018.09.020.

[230] Nataliya Lopanitsyna, Chiheb Ben Mahmoud, and Michele Ceriotti. *Dataset: Finite-Temperature Materials Modeling from the Quantum Nuclei to the Hot Electrons Regime.* https://archive.materialscloud.org/record/2020.64. 2021. DOI: 10.24435/materialscloud:vk-qd..

[231] Jörg Behler and Michele Parrinello. "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces". In: *Phys. Rev. Lett.* 98.14 (Apr. 2007), p. 146401. DOI: 10.1103/PhysRevLett.98.146401.

[232] Nongnuch Artrith and Jörg Behler. "High-Dimensional Neural Network Potentials for Metal Surfaces: A Prototype Study for Copper". In: *Phys. Rev. B* 85.4 (Jan. 2012), p. 045439. DOI: 10.1103/PhysRevB.85.045439.

[233] Nongnuch Artrith, Tobias Morawietz, and Jörg Behler. "High-Dimensional Neural-Network Potentials for Multicomponent Systems: Applications to Zinc Oxide". In: *Phys. Rev. B* 83.15 (Apr. 2011), p. 153101. DOI: 10.1103/PhysRevB.83.153101.

[234] Jörg Behler. "Neural Network Potential-Energy Surfaces in Chemistry: A Tool for Large-Scale Simulations." In: *Phys. Chem. Chem. Phys. PCCP* 13.40 (Oct. 2011), pp. 17930–55. DOI: 10.1039/c1cp21668f.

[235] Andreas Singraber, Tobias Morawietz, Jörg Behler, and Christoph Dellago. "Parallel multi-stream training of high-dimensional neural network potentials". In: *Journal of chemical theory and computation* (2019).

[236] Andreas Singraber. *N2P2*. DOI: 0.5281/zenodo.1344447.

[237] Felix Brockherde, Leslie Vogt, Li Li, Mark E. Tuckerman, Kieron Burke, and Klaus Robert Müller. "Bypassing the Kohn-Sham Equations with Machine Learning". In: *Nat. Commun.* 8.1 (Dec. 2017), p. 872. DOI: 10.1038/s41467-017-00839-3.

[238] Alberto Fabrizio, Andrea Grisafi, Benjamin Meyer, Michele Ceriotti, and Clemence Corminboeuf. "Electron Density Learning of Non-Covalent Systems". In: *Chem. Sci.* 10 (2019), p. 9424. DOI: 10.1039/C9SC02696G.

[239] Chiheb Ben Mahmoud, Andrea Anelli, Gábor Csányi, and Michele Ceriotti. "Learning the electronic density of states in condensed matter". In: *arxiv:2006.11803* (2020).

[240] Félix Musil, Max Veit, Alexander Goscinski, Guillaume Fraux, Michael J Willatt, Markus Stricker, Till Junge, and Michele Ceriotti. "Efficient implementation of atom-density representations". In: *The Journal of Chemical Physics* 154.11 (2021), p. 114109.

[241] Félix Musil, Michael J. Willatt, Mikhail A. Langovoy, and Michele Ceriotti. "Fast and Accurate Uncertainty Estimation in Chemical Machine Learning". In: *J. Chem. Theory Comput.* 15.2 (Feb. 2019), pp. 906–915. DOI: 10.1021/acs.jctc.8b00959.

[242] Chiheb Ben Mahmoud, Federico Grasselli, and Michele Ceriotti. "Predicting hot-electron free energies from ground-state data". In: *Physical Review B* 106.12 (2022), p. L121116.

[243] Benedict Leimkuhler and Charles Matthews. "Robust and Efficient Configurational Molecular Sampling via Langevin Dynamics". In: *J. Chem. Phys.* 138.17 (2013). DOI: 10.1063/1.4802990.

[244] G Bussi, D Donadio, and M Parrinello. "Canonical Sampling through Velocity Rescaling". In: *J. Chem. Phys.* 126.1 (2007), p. 14101.

[245] Michele Ceriotti, Giovanni Bussi, and Michele Parrinello. "Colored-Noise Thermostats à La Carte". In: *J. Chem. Theory Comput.* 6.4 (Apr. 2010), pp. 1170–1180. DOI: 10.1021/ct900563s.

[246] Andreas Singraber, Tobias Morawietz, Jörg Behler, and Christoph Dellago. "Parallel Multistream Training of High-Dimensional Neural Network Potentials". In: *J. Chem. Theory Comput.* 15.5 (May 2019), pp. 3075–3092. DOI: 10.1021/acs.jctc.8b01092.

[247] Steve Plimpton. "Fast Parallel Algorithms for Short-Range Molecular Dynamics". In: *J. Comput. Phys.* 117.1 (Mar. 1995), pp. 1–19. DOI: 10.1006/jcph.1995.1039.

[248] Michele Ceriotti, Joshua More, and David E. Manolopoulos. "I-PI: A Python Interface for Ab Initio Path Integral Molecular Dynamics Simulations". In: *Comput. Phys. Commun.* 185.3 (Nov. 2014), pp. 1019–1026. DOI: 10.1016/j.cpc.2013.10.027.

[249] Tsuneyasu Okabe, Masaaki Kawata, Yuko Okamoto, and Masuhiro Mikami. "Replica-exchange Monte Carlo method for the isobaric–isothermal ensemble". In: *Chemical Physics Letters* 335.5-6 (Mar. 2001), pp. 435–439. DOI: 10.1016/s0009-2614(01)00055-0.

[250] Riccardo Petraglia, Adrien Nicolaï, Matthew D. Wodrich, Michele Ceriotti, and Clemence Corminboeuf. "Beyond static structures: Putting forth REMD as a tool to solve problems in computational organic chemistry". In: *J. Comput. Chem.* 37.1 (July 2015), pp. 83–92. DOI: 10.1002/jcc.24025.

[251] LJ Swartzendruber, VP Itkin, and CB Alcock. "The Fe-Ni (iron-nickel) system". In: *Journal of phase equilibria* 12.3 (1991), pp. 288–312.

[252] Francis Birch. "Finite elastic strain of cubic crystals". In: *Physical review* 71.11 (1947), p. 809.

[253] X Zhang, PR Stoddart, JD Comins, and AG Every. "High-temperature elastic properties of a nickel-based superalloy studied by surface Brillouin scattering". In: *Journal of Physics: Condensed Matter* 13.10 (2001), p. 2281.

[254] Y. Wang, Z. -K. Liu, and L. -Q. Chen. "Thermodynamic properties of Al, Ni, NiAl, and Ni3Al from first-principles calculations". In: *Acta Materialia* 52.9 (May 17, 2004), pp. 2665–2671. DOI: 10.1016/j.actamat.2004.02.014.

[255] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. "The atomic simulation environment—a Python library for working with atoms". In: *Journal of Physics: Condensed Matter* 29.27 (2017), p. 273002.

[256] Yi Wang, JJ Wang, WY Wang, ZG Mei, SL Shang, LQ Chen, and ZK Liu. "A mixed-space approach to first-principles calculations of phonon frequencies for polar materials". In: *Journal of Physics: Condensed Matter* 22.20 (2010), p. 202201.

[257] Dario Alfè. "PHON: A program to calculate phonons using the small displacement method". In: *Computer Physics Communications* 180.12 (Dec. 2009), pp. 2622–2633. DOI: 10.1016/j.cpc.2009.03.010.

[258] Sergei Yur'evich Glazkov. "Formation of point-defects and thermo-physical properties of nickel a high-temperatures". In: *Teplofizika vysokikh temperatur* 25.1 (1987), pp. 59–64.

[259] Charles George Broyden. "The convergence of a class of double-rank minimization algorithms 1. general considerations". In: *IMA Journal of Applied Mathematics* 6.1 (1970), pp. 76–90.

[260] Roger Fletcher. "A new approach to variable metric algorithms". In: *The computer journal* 13.3 (1970), pp. 317–322.

[261] Donald Goldfarb. "A family of variable-metric methods derived by variational means". In: *Mathematics of computation* 24.109 (1970), pp. 23–26.

[262] David F Shanno. "Conditioning of quasi-Newton methods for function minimization". In: *Mathematics of computation* 24.111 (1970), pp. 647–656.

[263] Binglun Yin, Zhaoxuan Wu, and W.A. Curtin. "Comprehensive first-principles study of stable stacking faults in hcp metals". In: *Acta Materialia* 123 (Jan. 2017), pp. 223–234. DOI: 10.1016/j.actamat.2016.10.042.

[264] Lawrence Eugene Murr. "Interfacial phenomena in metals and alloys". In: (1975).

[265] Ryo Kobayashi, Daniele Giofré, Till Junge, Michele Ceriotti, and William A. Curtin. "Neural network potential for Al-Mg-Si alloys". In: *Physical Review Materials* 1.5 (Oct. 30, 2017). DOI: 10.1103/PhysRevMaterials.1.053604.

[266] M Yousuf, P C Sahu, H K Jajoo, S Rajagopalan, and K Govinda Rajan. "Effect of magnetic transition on the lattice expansion of nickel". In: *J. Phys. F: Met. Phys.* 16.3 (Mar. 1986), pp. 373–380. DOI: 10.1088/0305-4608/16/3/015.

[267] J. Bandyopadhyay and K.P. Gupta. "Low temperature lattice parameter of nickel and some nickel-cobalt alloys and Grüneisen parameter of nickel". In: *Cryogenics* 17.6 (June 1977), pp. 345–347. DOI: 10.1016/0011-2275(77)90130-8.

[268] Masao Shimizu. "Forced Magnetostriction, Magnetic Contributions to Bulk Modulus and Thermal Expansion and Pressure Dependence of Curie Temperature in Iron, Cobalt and Nickel". In: *J. Phys. Soc. Jpn.* 44.3 (Mar. 1978), pp. 792–800. DOI: 10.1143/jpsj.44.792.

[269] Venkat Kapil, Jörg Behler, and Michele Ceriotti. "High Order Path Integrals Made Easy". In: *J. Chem. Phys.* 145.23 (Dec. 2016), p. 234103. DOI: 10.1063/1.4971438.

[270] M Suzuki. "Hybrid Exponential Product Formulas for Unbounded Operators with Possible Applications to Monte Carlo Simulations". In: *Phys. Lett. A* 201.5-6 (1995), pp. 425–428.

[271] Siu A. Chin. “Symplectic Integrators from Composite Operator Factorizations”. In: *Phys. Lett. A* 226.6 (1997), pp. 344–348.

[272] Fritz Körmann, Alexey Dick, Tilmann Hickel, and Jörg Neugebauer. “Role of spin quantization in determining the thermodynamic properties of magnetic transition metals”. In: *Physical Review B* 83.16 (2011), p. 165114.

[273] Takeshi M Yamamoto. “Path-Integral Virial Estimator Based on the Scaling of Fluctuation Coordinates: Application to Quantum Clusters with Fourth-Order Propagators”. In: *J. Chem. Phys.* 123.10 (Sept. 2005), p. 104101. DOI: 10.1063/1.2013257.

[274] Arnaud Metsue, Abdelali Oudriss, Jamaa Bouhattate, and Xavier Feaugas. “Contribution of the entropy on the thermodynamic equilibrium of vacancies in nickel”. In: *The Journal of Chemical Physics* 140.10 (Mar. 2014), p. 104705. DOI: 10.1063/1.4867543.

[275] Wolfram Wycisk and Monika Feller-Kniepmeier. “Quenching experiments in high purity Ni”. In: *Journal of Nuclear Materials* 69.1-2 (1978), pp. 616–619.

[276] Heinz-Peter Scholz. *Messungen der absoluten leerstellenkonzentration in nickel und geordneten intermetallischen nickel-legierungen mit einem differentialdilatometer*. Cuvillier, 2001.

[277] G Michot and B Deviot. “Influence de l’oxydation sur la trempe des lacunes dans le nickel—durcissement du a la trempe”. In: *Revue de Physique Appliquée* 12.12 (1977), pp. 1815–1817.

[278] Luca M. Ghiringhelli, Jan Vybiral, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler. “Big Data of Materials Science: Critical Role of the Descriptor”. In: *Phys. Rev. Lett.* 114.10 (Mar. 2015), p. 105503. DOI: 10.1103/PhysRevLett.114.105503.

[279] Bingqing Cheng and Michele Ceriotti. “Computing the Absolute Gibbs Free Energy in Atomistic Simulations: Applications to Defects in Solids”. In: *Phys. Rev. B* 97.5 (Feb. 2018), p. 054102. DOI: 10.1103/PhysRevB.97.054102.

[280] Andrew Ian Duff, Theresa Davey, Dominique Korbmacher, Albert Glensk, Blazej Grabowski, Jörg Neugebauer, and Michael W. Finnis. “Improved method of calculating ab initio high-temperature thermodynamic properties with application to ZrC”. In: *Phys. Rev. B* 91.21 (June 2015).

[281] MW Johnson, NH March, B McCoy, SK Mitra, DI Page, and RC Perrin. “Structure and effective pair interaction in liquid nickel”. In: *Philosophical Magazine* 33.1 (1976), pp. 203–206.

[282] A Meyer, S Stüber, D Holland-Moritz, O Heinen, and T Unruh. “Determination of self-diffusion coefficients by quasielastic neutron scattering measurements of levitated Ni droplets”. In: *Physical Review B* 77.9 (2008), p. 092201.

[283] S Mavila Chathoth, A Meyer, MM Koza, and F Juranyi. "Atomic diffusion in liquid Ni, NiP, PdNiP, and PdNiCuP alloys". In: *Applied physics letters* 85.21 (2004), pp. 4881–4883.

[284] Martin Walbrühl, Andreas Blomqvist, and Pavel A Korzhavyi. "Atomic diffusion in liquid nickel: First-principles modeling". In: *The Journal of chemical physics* 148.24 (2018), p. 244503.

[285] T Iida and RIL Guthrie. *The Physical Properties of Liquid Metals.[Sl]: Oxford University Press*. 1988.

[286] Burkhard Dünweg and Kurt Kremer. "Molecular dynamics simulation of a polymer chain in solution". In: *The Journal of Chemical Physics* 99.9 (Nov. 1993), pp. 6983–6997. DOI: 10.1063/1.465445.

[287] In-Chul Yeh and Gerhard Hummer. "System-Size Dependence of Diffusion Coefficients and Viscosities from Molecular Dynamics Simulations with Periodic Boundary Conditions". In: *The Journal of Physical Chemistry B* 108.40 (Oct. 2004), pp. 15873–15879. DOI: 10.1021/jp0477147.

[288] Jürgen Brillo and I Egry. "Surface tension of nickel, copper, iron and their binary alloys". In: *Journal of materials science* 40.9-10 (2005), pp. 2213–2216.

[289] J.P.R.B. Walton, D.J. Tildesley, J.S. Rowlinson, and J.R. Henderson. "The pressure tensor at the planar surface of a liquid". In: *Molecular Physics* 48.6 (Apr. 1983), pp. 1357–1368. DOI: 10.1080/00268978300100971.

[290] Y. Cai, H. A. Wu, and S. N. Luo. "Cavitation in a metallic liquid: Homogeneous nucleation and growth of nanovoids". In: *The Journal of Chemical Physics* 140.21 (June 2014), p. 214317. DOI: 10.1063/1.4880960.

[291] MW Chase Jr. "NIST-JANAF thermochemical tables fourth edition". In: *J. Phys. Chem. Ref. Data, Monograph* 9 (1998).

[292] Stefano Angioletti-Uberti, Michele Ceriotti, Peter D. Lee, and Mike W. Finnis. "Solid-Liquid Interface Free Energy through Metadynamics Simulations". In: *Phys. Rev. B - Condens. Matter Mater. Phys.* 81.12 (Mar. 2010), p. 125416. DOI: 10.1103/PhysRevB.81.125416.

[293] PD Desai. "Thermodynamic properties of nickel". In: *International journal of thermophysics* 8.6 (1987), pp. 763–780.

[294] T Saito and Y Sakuma. "Thermodynamic properties of nickel". In: *Sci. Rep. Res. Inst.,Ser.A* 22.57 (1970).

[295] Li-Fang Zhu, Fritz Körmann, Andrei V. Ruban, Jörg Neugebauer, and Blazej Grabowski. "Performance of the Standard Exchange-Correlation Functionals in Predicting Melting Properties Fully from First Principles: Application to Al and Magnetic Ni". In: *Phys. Rev. B* 101.14 (Apr. 2020), p. 144108. DOI: 10.1103/PhysRevB.101.144108.

[296] Monica Pozzo and Dario Alfè. "Melting Curve of Face-Centered-Cubic Nickel from First-Principles Calculations". In: *Phys. Rev. B* 88.2 (July 2013), p. 024111. DOI: 10.1103/PhysRevB.88.024111.

[297] J. J. Hoyt, Mark Asta, and Alain Karma. "Method for Computing the Anisotropy of the Solid-Liquid Interfacial Free Energy". In: *Physical Review Letters* 86.24 (June 11, 2001), pp. 5530–5533. DOI: 10.1103/PhysRevLett.86.5530.

[298] Jeremy Q Broughton and George H Gilmer. "Molecular dynamics investigation of the crystal–fluid interface. VI. Excess surface free energies of crystal–liquid systems". In: *The Journal of chemical physics* 84.10 (1986), pp. 5759–5768.

[299] Xian-Ming Bai and Mo Li. "Calculation of solid-liquid interfacial free energy: A classical nucleation theory based approach". In: *The Journal of chemical physics* 124.12 (2006), p. 124707.

[300] Bingqing Cheng, Gareth A. Tribello, and Michele Ceriotti. "Solid-Liquid Interfacial Free Energy out of Equilibrium". In: *Phys. Rev. B* 92.18 (Nov. 2015), p. 180102. DOI: 10.1103/PhysRevB.92.180102.

[301] Bingqing Cheng and Michele Ceriotti. "Communication: Computing the Tolman Length for Solid-Liquid Interfaces". In: *J. Chem. Phys.* 148.23 (June 2018), p. 231102. DOI: 10.1063/1.5038396.

[302] Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provasi, Paolo Raiteri, Davide Donadio, Fabrizio Marinelli, Fabio Pietrucci, Ricardo A Broglia, and Michele Parrinello. "PLUMED: A Portable Plugin for Free-Energy Calculations with Molecular Dynamics". In: *Comput. Phys. Commun.* 180.10 (Oct. 2009), pp. 1961–1972. DOI: 10.1016/j.cpc.2009.05.011.

[303] Massimiliano Bonomi, Giovanni Bussi, Carlo Camilloni, Gareth Tribello, Pavel Bonas, Alessandro Barducci, Mattia Bernetti, Peter G Bolhuis, Sandro Bottaro, Davide Branduardi, et al. "Promoting transparency and reproducibility in enhanced molecular simulations". In: *Nature methods* 16.8 (2019), pp. 670–673.

[304] J. Hoyt, Mark Asta, and Alain Karma. "Method for Computing the Anisotropy of the Solid-Liquid Interfacial Free Energy". In: *Phys. Rev. Lett.* 86.24 (June 2001), pp. 5530–5533. DOI: 10.1103/PhysRevLett.86.5530.

[305] R. E. Rozas and J. Horbach. “Capillary wave analysis of rough solid-liquid interfaces in nickel”. In: *EPL(Europhysics Letters)* 93.2 (Jan. 2011), p. 26006. DOI: 10.1209/0295-5075/93/26006.

[306] JJ Hoyt, Alain Karma, MA Asta, and DY Sun. “From atoms to dendrites”. In: *JOM* 56.4 (2004), pp. 49–54.

[307] Nataliya Lopanitsyna, Guillaume Fraux, Maximilian A Springer, Sandip De, and Michele Ceriotti. “Modeling high-entropy transition metal alloys with alchemical compression”. In: *Physical Review Materials* 7.4 (2023), p. 045802.

[308] G Kresse and J Furthmüller. “Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set”. In: *Phys. Rev. B* 54.16 (1996), pp. 11169–11186.

[309] Gábor I. Csonka, John P. Perdew, Adrienn Ruzsinszky, Pier H. T. Philipsen, Sébastien Lebègue, Joachim Paier, Oleg A. Vydrov, and János G. Ángyán. “Assessing the performance of recent density functionals for bulk solids”. In: *Physical Review B* 79.15 (Apr. 2009). DOI: 10.1103/physrevb.79.155107.

[310] Georg Kresse and Daniel Joubert. “From ultrasoft pseudopotentials to the projector augmented-wave method”. In: *Physical review b* 59.3 (1999), p. 1758.

[311] H J Monkhorst and J D Pack. “Special Points for Brillouin-Zone Integrations”. In: *Phys. Rev. B* 13.12 (1976), pp. 5188–5192. DOI: 10.1103/PhysRevB.13.5188.

[312] Ivan Novikov, Blazej Grabowski, Fritz Körmann, and Alexander Shapeev. “Magnetic Moment Tensor Potentials for Collinear Spin-Polarized Materials Reproduce Different Magnetic States of Bcc Fe”. In: *npj Comput Mater* 8.1 (Dec. 2022), p. 13. DOI: 10.1038/s41524-022-00696-9.

[313] C. S. Wang, R. E. Prange, and V. Korenman. “Magnetism in Iron and Nickel”. In: *Phys. Rev. B* 25.9 (May 1982), pp. 5766–5777. DOI: 10.1103/PhysRevB.25.5766.

[314] J Kubler, K -H Hock, J Sticht, and A R Williams. “Density Functional Theory of Non-Collinear Magnetism”. In: *J. Phys. F: Met. Phys.* 18.3 (Mar. 1988), pp. 469–483. DOI: 10.1088/0305-4608/18/3/018.

[315] Heather J. Kulik, Matteo Cococcioni, Damian A. Scherlis, and Nicola Marzari. “Density Functional Theory in Transition-Metal Chemistry: A Self-Consistent Hubbard U Approach”. In: *Phys. Rev. Lett.* 97.10 (Sept. 2006), p. 103001. DOI: 10.1103/PhysRevLett.97.103001.

[316] Zhiming Li, Alfred Ludwig, Alan Savan, Hauke Springer, and Dierk Raabe. “Combinatorial metallurgical synthesis and processing of high-entropy alloys”. In: *Journal of Materials Research* 33.19 (2018), pp. 3156–3169.

[317] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, 1994.

[318] Dong C. Liu and Jorge Nocedal. "On the Limited Memory BFGS Method for Large Scale Optimization". In: *Mathematical Programming* 45.1-3 (Aug. 1989), pp. 503–528. DOI: 10.1007/BF01589116.

[319] Hans C. Andersen. "Molecular Dynamics Simulations at Constant Pressure and/or Temperature". In: *J. Chem. Phys.* 72.4 (1980), pp. 2384–2393. DOI: 10.1063/1.439486.

[320] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. "Equation of State Calculations by Fast Computing Machines". In: *J. Chem. Phys.* 21.6 (1953), pp. 1087–1092.

[321] David J Earl and Michael W Deem. "Parallel Tempering: Theory, Applications, and New Perspectives". In: *Phys. Chem. Chem. Phys.* 7.23 (Dec. 2005), p. 3910. DOI: 10.1039/b509983h.

[322] Giulio Imbalzano and Michele Ceriotti. "Modeling the Ga/As Binary System across Temperatures and Compositions from First Principles". In: *Phys. Rev. Materials* 5.6 (June 2021), p. 063804. DOI: 10.1103/PhysRevMaterials.5.063804.

[323] Riccardo Petraglia, Adrien Nicolaï, Matthew D. Wodrich, Michele Ceriotti, and Clemence Corminboeuf. "Beyond Static Structures: Putting Forth REMD as a Tool to Solve Problems in Computational Organic Chemistry". In: *J. Comput. Chem.* 37.1 (Jan. 2016), pp. 83–92. DOI: 10.1002/jcc.24025.

[324] Felix A. Faber, Alexander Lindmaa, O. Anatole Von Lilienfeld, and Rickard Armiento. "Machine Learning Energies of 2 Million Elpasolite (ABC2D6) Crystals". In: *Phys. Rev. Lett.* 117.13 (Sept. 2016), p. 135502. DOI: 10.1103/PhysRevLett.117.135502.

[325] Volker L. Deringer and Gábor Csányi. "Machine Learning Based Interatomic Potential for Amorphous Carbon". In: *Phys. Rev. B* 95.9 (Mar. 2017), p. 094203. DOI: 10.1103/PhysRevB.95.094203.

[326] Volker L. Deringer, Albert P. Bartók, Noam Bernstein, David M. Wilkins, Michele Ceriotti, and Gábor Csányi. "Gaussian Process Regression for Materials and Molecules". In: *Chem. Rev.* 121.16 (Aug. 2021), pp. 10073–10141. DOI: 10.1021/acs.chemrev.1c00022.

[327] Glenn T Seaborg. *Plutonium: The Ornery Element*. 1964.

[328] Daniel Sheppard, Graeme Henkelman, and O. Anatole lilienfeldvon Lilienfeld. "Alchemical Derivatives of Reaction Energetics". In: *The Journal of Chemical Physics* 133.8 (Aug. 2010), p. 084104. DOI: 10.1063/1.3474502.

[329] Chi Chen and Shyue Ping Ong. *A Universal Graph Deep Learning Interatomic Potential for the Periodic Table*. 2022. DOI: 10.48550/ARXIV.2202.02450.

[330] 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". In: *APL Mater.* 1.1 (July 2013), p. 011002. DOI: 10.1063/1.4812323.

[331] J. M. Cowley. "An Approximate Theory of Order in Alloys". In: *Physical Review* 77.5 (Mar. 1950), pp. 669–675. DOI: 10.1103/physrev.77.669.

[332] Uichiro Mizutani. "Hume-Rothery rules for structurally complex alloy phases". In: *MRS Bulletin* 37.2 (2012), pp. 169–169. DOI: 10.1557/mrs.2012.45.

[333] Friedrich Waag, Yao Li, Anna Rosa Ziefuß, Erwan Bertin, Marius Kamp, Viola Duppel, Galina Marzun, Lorenz Kienle, Stephan Barcikowski, and Bilal Gökce. "Kinetically-controlled laser-synthesis of colloidal high-entropy alloy nanoparticles". In: *RSC Advances* 9.32 (2019), pp. 18547–18558. DOI: 10.1039/c9ra03254a.

[334] Hailong Peng, Yangcenzi Xie, Zicheng Xie, Yunfeng Wu, Wenkun Zhu, Shuquan Liang, and Liangbing Wang. "Large-scale and facile synthesis of a porous high-entropy alloy CrMnFeCoNi as an efficient catalyst". In: *Journal of Materials Chemistry A* 8.35 (2020), pp. 18318–18326. DOI: 10.1039/d0ta04940a.

[335] Ren He, Linlin Yang, Yu Zhang, Xiang Wang, Seungho Lee, Ting Zhang, Lingxiao Li, Zhifu Liang, Jingwei Chen, Junshan Li, Ahmad Ostovari Moghaddam, Jordi Llorca, Maria Ibanez, Jordi Arbiol, Ying Xu, and andreu cabot. "A Crmnfeconi High Entropy Alloy Boosting Oxygen Evolution/Reduction Reactions and Zinc-Air Battery Performance". In: *SSRN Electronic Journal* (2022). DOI: 10.2139/ssrn.4289857.

[336] J. Tang, J.L. Xu, Z.G. Ye, X.B. Li, and J.M. Luo. "Microwave sintered porous CoCrFeNiMo high entropy alloy as an efficient electrocatalyst for alkaline oxygen evolution reaction". In: *Journal of Materials Science and Technology* 79 (July 2021), pp. 171–177. DOI: 10.1016/j.jmst.2020.10.079.

[337] Simon Schumacher, Sabrina Baha, Alan Savan, Corina Andronescu, and Alfred Ludwig. "High-throughput discovery of hydrogen evolution electrocatalysts in the complex solid solution system Co–Cr–Fe–Mo–Ni". In: *Journal of Materials Chemistry A* 10.18 (2022), pp. 9981–9987. DOI: 10.1039/d2ta01652d.

[338] Gang Cui, Bin Han, Ying Yang, Yong Wang, and Hu Chunyang. "Microstructure and tribological property of CoCrFeMoNi High entropy alloy treated by ion sulfurization". In: *Journal of Materials Research and Technology* 9.2 (Mar. 2020), pp. 2598–2609. DOI: 10.1016/j.jmrt.2019.12.090.

[339] Zhu Wang, Jie Jin, Guo-Hui Zhang, Xue-Hua Fan, and Lei Zhang. "Effect of temperature on the passive film structure and corrosion performance of CoCrFeMoNi high-entropy alloy". In: *Corrosion Science* 208 (Nov. 2022), p. 110661. DOI: 10.1016/j.corsci.2022.110661.

[340] Thomas A.A. Batchelor, Jack K. Pedersen, Simon H. Winther, Ivano E. Castelli, Karsten W. Jacobsen, and Jan Rossmeisl. "High-Entropy Alloys as a Discovery Platform for Electrocatalysis". In: *Joule* 3.3 (Mar. 2019), pp. 834–845. DOI: 10.1016/j.joule.2018.12.015.

[341] Nils L. N. Broge, Martin Bondesgaard, Frederik Søndergaard-Pedersen, Martin Roelsgaard, and Bo Brummerstedt Iversen. "Autocatalytic Formation of High-Entropy Alloy Nanoparticles". In: *Angewandte Chemie* 132.49 (Sept. 2020), pp. 22104–22108. DOI: 10.1002/ange.202009002.

[342] Jack K. Pedersen, Christian M. Clausen, Olga A. Krysiak, Bin Xiao, Thomas A. A. Batchelor, Tobias Löffler, Vladislav A. Mints, Lars Banko, Matthias Arenz, Alan Savan, Wolfgang Schuhmann, Alfred Ludwig, and Jan Rossmeisl. "Bayesian Optimization of High-Entropy Alloy Compositions for Electrocatalytic Oxygen Reduction∗∗". In: *Angewandte Chemie* 133.45 (Oct. 2021), pp. 24346–24354. DOI: 10.1002/ange.202108116.

[343] Gilhwan Lee, Ngoc-Anh Nguyen, Van-Toan Nguyen, Liudmila L. Larina, Enkhjin Chuluunbat, Eunhee Park, Jeongseon Kim, Ho-Suk Choi, and Michael Keidar. "High entropy alloy electrocatalyst synthesized using plasma ionic liquid reduction". In: *Journal of Solid State Chemistry* 314 (Oct. 2022), p. 123388. DOI: 10.1016/j.jssc.2022.123388.

[344] Jakub Šebesta, Karel Carva, and Dominik Legut. "Role of Magnetism in the Stability of the High-Entropy Alloy CoCrFeMnNi and Its Derivatives". In: *Phys. Rev. Materials* 3.12 (Dec. 2019), p. 124410. DOI: 10.1103/PhysRevMaterials.3.124410.

[345] K.-Y. Tsai, M.-H. Tsai, and J.-W. Yeh. "Sluggish Diffusion in Co–Cr–Fe–Mn–Ni High-Entropy Alloys". In: *Acta Materialia* 61.13 (Aug. 2013), pp. 4887–4897. DOI: 10.1016/j.actamat.2013.04.058.

[346] Shuai Chen, Zachary H. Aitken, Subrahmanyam Pattamatta, Zhaoxuan Wu, Zhi Gen Yu, David J. Srolovitz, Peter K. Liaw, and Yong-Wei Zhang. "Simultaneously Enhancing the Ultimate Strength and Ductility of High-Entropy Alloys via Short-Range Ordering". In: *Nat Commun* 12.1 (Aug. 2021), p. 4953. DOI: 10.1038/s41467-021-25264-5.

[347] Marcel Glienke, Mayur Vaidya, K. Gururaj, Lydia Daum, Bengü Tas, Lukasz Rogal, K.G. Pradeep, Sergiy V. Divinski, and Gerhard Wilde. "Grain Boundary Diffusion in CoCrFeMnNi High Entropy Alloy: Kinetic Hints towards a Phase Decomposition". In: *Acta Materialia* 195 (Aug. 2020), pp. 304–316. DOI: 10.1016/j.actamat.2020.05.009.

[348] Kap Ho Lee, Soon-Ku Hong, and Sun Ig Hong. “Precipitation and Decomposition in CoCrFeMnNi High Entropy Alloy at Intermediate Temperatures under Creep Conditions”. In: *Materialia* 8 (Dec. 2019), p. 100445. DOI: 10.1016/j.mtla.2019.100445.

Nataliya LOPANITSYNA

Avenue de Cour 15, 1007 Lausanne, Switzerland
nataliya.lopanitsyna@epfl.ch
+41 78 7400817
linkedin.com/in/nataliya-lopanitsyna/

Computational materials scientist, fusing machine-learning and quantum chemistry to model metal alloys. Experience developing machine learning potentials (MLP) and integrating them with statistical sampling methods to predict electronic properties. Optimised representation to enhance MLP performance for multi-component systems crucial for identifying novel catalyst materials in close collaboration with the chemical industry.

Professional experience

Doctoral Assistant **Lausanne, Switzerland**
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE *2018-2023*

- co-developed a pytorch package to compress the computational cost due of atom density representation
- published a machine learning model to improve accuracy of metal alloy properties relevant for additive manufacturing at elevated temperatures
- simulated and analysed multicomponent metal alloy systems using various statistical mechanics methods
- implemented optimization algorithms in a large open source project

All my results were presented at conferences and published in peer reviewed journals

Research Intern **Lausanne, Switzerland**
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE *2017-2018*

- worked on building a neural network potential for Ni

Research Fellow **Moscow, Russia**
JOINT INSTITUTE OF HIGH TEMPERATURES OF THE RUSSIAN ACADEMY OF SCIENCE *2014-2017*

- fitted a spline-based model of an interatomic potential for Au-Si system
- built analytical descriptors to analyse crystalline structure and track concentration profiles of alloy compounds under cooling process
- introduced an empirical formula to estimate tensile strength of metal alloys

Education

PhD, Materials Science **Lausanne, Switzerland**
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE *2018-2023*
Supervisor: Prof Michele Ceriotti

M.S., Applied Mathematics and Physics (with honours) **Moscow, Russia**
MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY *2015-2017*
"Atomistic simulation of metal disruption and nanoparticle formation under the action of subpicosecond laser pulses
Supervisor: Dr Alexey Kuksin

B.S., Applied Mathematics and Physics **Moscow, Russia**
MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY *2011-2015*
"Atomistic simulation and theoretical description of nucleation in liquid metals under tension"
Supervisor: Dr Alexey Kuksin

Skills & Tools

OS environments Linux/Unix (CentOS, Ubuntu, Fedora) , MacOS, Windows
Computational chemistry VASP, Quantum Espresso, Gaussian, DFTB+, xTB, LAMMPS, Plumed, I-Pi, ASE, RDKit, Materials Project, OpenBabel
Programming Python(NumPy, SciPy, Matplotlib, Seaborn, Pandas, Tensorflow, PyTorch), Wolfram Mathematica, C/C++, Bash, LaTeX, git, Slurm
Visualisation & Presentation Office Suite, Google Suite, Blender/, Adobe, Dash, Adobe CC (Photoshop, Illustrator), Tableau, Blender, Inkscape, VMD, Ovito

Publications

Modeling high-entropy transition-metal alloys with alchemical compression
***Lopanitsyna N.**, Fraux G., Springer, M. A., De, S. and Ceriotti M.*
Phys. Rev. Materials, V.7, №4, 2023
Finite-temperature materials modeling from the quantum nuclei to the hot electron regime
***Lopanitsyna N.**, Ben Mahmoud C., and Ceriotti M.*
Phys. Rev. Materials, V.5, №4, 2021
Atomistic simulation of Si-Au melt crystallization with novel interatomic potential
*Starikov V., **Lopanitsyna N.**, Smirnova D., Makarov S.*
Computational Material Science, V.142, 2018
Efficient Second-Harmonic Generation in Nanocrystalline Silicon Nanoparticle
*Makarov S., Petrov M., Zywietz U., Milichko V., Zuev D., **Lopanitsyna N.**, Kuksin A., Mukhin I., Zograf G., Ubyivovk E., Smirnova D., Starikov* S., Chichkov B., Kivshar Y.
Nano Letters, V.17, №5, 2017

Fellowships, Awards & Grants

Exploring the Space of High Entropy Alloys, from DFT to Machine Learning (*100k node hours*)
HPC grant at CSCS Swiss National Supercomputing centre *2021-2022*
Grant for a peer-mentoring group (5000 CHF) *Fix the Leaky Pipeline, Switzerland* *2020*
Best pitch talk *EDMX Research days EPFL, Lausanne, Switzerland* *2019*
INSPIRE Potentials Marvel Master Fellowship (12'000 CHF) *NCCR Marvel, Switzerland* *2017*
Outstanding report award *International Conference for Young Scientists "Lomonosov"* *2017*
Increased State Academic Scholarship *MIPT for research achievements* *2017 & 2016*
Scholarship *from JIHT RAS Found for Supporting Young Scientists* *2016 & 2015*
Award for the best oral report *XIV Russian Conference «Structure and Properties of Metals»* *2015*
Best Student of the year *one-year scholarship for studying in Moscow, Yekaterinburg, Russia* *2011*

Additional Experience

Teaching assistant **Lausanne, Switzerland**

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE *2018-2021*

- assisted in preparation and implementation of the courses “Statistical Mechanics”,“Surface Science”, “Advanced Metallurgy”

Organiser of an international workshop **Moscow, Russia**

SKOLTECH *2019*

- made a website and tentative program of the workshop
- contacted all the parties and established communication between them
- managed bureaucratic details of the workshop

Languages

English(C2) Italian(B2) French(B1) German(A1) Russian(native)