EPFL

Thèse n° 8457

Transferable machine-learning models of complex materials: the case of GaAs

Présentée le 19 novembre 2021

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

Giulio IMBALZANO

Acceptée sur proposition du jury

Dr A. Hessler-Wyser, présidente du jury
Prof. M. Ceriotti, directeur de thèse
Prof. F. Montalenti, rapporteur
Prof. A. Bartok-Partay, rapporteur
Prof. W. Curtin, rapporteur

École polytechnique fédérale de Lausanne

2021

Acknowledgements

As the PhD comes to an end, there are a number of people that I want to thank for allowing me to be here, and helping me become the person that I am today.
First of all, I would like to thank my supervisor, Michele Ceriotti, for providing me with the opportunity to embark on this journey and helping along the way. Thank you for our meetings and discussions, as I always left knowing more than before, and thank you for showing how to carry out high-level scientific research. I would also like to thank the Ermenegildo Zegna Founder's Foundation for sharing my objectives and providing the opportunity to begin this PhD at EPFL.
Then, I would like to thank the people in the laboratory of COSMO, with whom I shared most of these years. Thanks to Piero and Edoardo for welcoming me in the group. I hope I have managed to welcome in the same way the new recruits, such as Natasha and Jigyasa, who have a bright future ahead. I have been happy to begin my PhD at a similar time as Andrea, Felix, and Andrea, as we could share the joys and sorrows of our work. Sometimes all you need is someone lending a sympathetic ear while you complain about the world being unfair, only to find out that a mistyped variable was the source of your anguish. Similarly, I would like to thank Venkat, who has also been a great technical help and has taught me a lot of what I know today about (PI)MD. A big thank you to Kevin and Federico Gi., for the many discussions and coffee breaks, and to Federico Gr. for your infinite patience and detail-oriented approach. It has been a great experience to work with you! Last but not least, thanks to Anne, for your friendliness, while making sure that the bureaucracy would not overwhelm us. Beyond the group, I would also like to thank Mahdi, Anna, and Yongbin, for the work we did together and the things I learnt from you.
Thanks to Nicolò, Riccardo, Sergio, and Stefano, who have been close, despite the physical distance. I am happy that in our ever-changing world we are still the same as we were in high-school.
Thanks to my family, who has been a constant source of strength and support in these years.
Finally, thanks to my love, Elisabetta. With you at my side I feel like I can face anything. I have thoroughly enjoyed remote working from the same "office", as we could more easily push each other through even the hardest days. And as the hard days passed, we then shared also the good ones, our happiness and achievements. Without you, I would be half of the person that I am today. Thank you, my dear.

G. I.

Abstract

Molecular simulations allow to investigate the behaviour of materials at the atomistic level, shedding light on phenomena that cannot be directly observed in experiments. Accurate results can be obtained with *ab initio* methods, while simulations of large-scale systems are usually possible only with coarse approximations of the molecular interactions. Machine learning interatomic potentials (MLIP) combine the strengths of the two methods in a framework that allows iterative refinement, opening the doors to the investigation of complex systems.
Currently, the training of a MLIP is still human-centered. The success or failure is often dictated by the complexity of the system and by the experience of the user with the software. Therefore, in this thesis, we want to provide some methods that would make the training and validation of the potentials easier and more general, even for complex, heterogeneous systems.
We begin by comparing the learning ability of three widely adopted frameworks that have been developed by the community, which make use of different representations, as well as different algorithms, proving that a well-constructed set of input features allow to learn at a similar accuracy datasets of water dimers and trimers.
Then, we compare heuristic methods based on the intrinsic correlations of the dataset to automatically identify the "best" inputs out of a larger set of candidates, which results in an accurate description of the system at a low computational cost. This allows to simplify the construction of potentials that use symmetry functions as inputs, as well as reduce the computational cost of gaussian approximated potentials based on the smooth overlap of atomic positions.
Finally, we introduce and implement a method to cheaply compute the uncertainty of the thermodynamic properties obtained through MD simulations with MLIPs. This method can be used either to assess the confidence of a given result obtained with a MLIP –necessary when we make quantitative predictions of properties– or to safely explore the phase space of interest, with the aid of a fall-back potential that takes over when the MLIP cannot be trusted.
We showcase these methods with a real example, in which we train a potential for the complex Ga_xAs_{1-x} system. The MLIP that we have developed is able to accurately predict the behavior across the whole phase diagram, spanning liquid and solid, metallic and semiconducting phases. In this endeavour we investigate a variety of methods to obtain a comprehensive dataset of structures that are fed into the MLIP.
To demonstrate the transferability of the potential, we compute multiple properties, some of which (e.g. the liquid surface tension) are well beyond the limits of *ab initio* methods.

We compare these results to our reference calculations and to experiments, finding a good agreement, within the limits of the selected level of theory (density functional theory at the generalized gradient approximation level).
Finally, we use our Ga_xAs_{1-x} MLIP to investigate the behaviour of liquid gallium in contact with the polar [111] surface of solid GaAs. Recent experimental findings assign an important role to the pre-ordering of the liquid at the interface during the growth of GaAs nanowires, pointing to the polarity as one of the main drivers for the correct growth. Our simulations allow to investigate this pre-ordering with increased detail, supporting and complementing the experimental observations. Furthermore, we explore the free energy of As atoms in the liquid Ga, which allows to understand the behaviour of As atoms during the growth to help identifying the ideal growth conditions.

Key words: Machine learning potentials, CUR selection, Uncertainty estimation, DFT, gallium arsenide, GaAs Nanowires

Sommario

Le simulazioni molecolari ci permettono di studiare le proprietà dei materiali a partire dal comportamento dei singoli atomi, osservando fenomeni che non sono accessibili agli esperimenti. I metodi *ab initio* ci permettono di ottenere predizioni accurate, ma possiamo studiare sistemi di grandi dimensioni solo con l'uso di formulazioni approssimate delle interazioni molecolari. I potenziali machine learning (PML) uniscono i punti di forza dei due metodi in un framework che può essere iterativamente perfezionato, aprendo le porte all'indagine di sistemi complessi.
Finora, l'addestramento dei PML è molto legato al fattore umano. Il successo o il fallimento è spesso dettato alla complessità del sistema e all'esperienza dell'utente con il software usato. Pertanto, in questa tesi vogliamo proporre alcuni metodi che renderebbero l'addestramento e la validazione dei potenziali più facile e più generale, anche per sistemi complessi ed eterogenei.
Iniziamo la trattazione confrontando la capacità di apprendimento di tre framework comunemente usati dalla comunità, che fanno uso di rappresentazioni e algoritmi ML diversi, dimostrando che un set ben costruito di input permette di ottenere risultati simili nell'apprendimento di un dataset di dimeri e trimeri d'acqua.
In secondo luogo, confrontiamo algoritmi euristici che sfruttano la correlazione dei dati appartenenti a un dataset per selezionare automaticamente i migliori input partendo da un set di candidati, ottenendo così una descrizione accurata a un costo computazionale ridotto. Questo permette di semplificare la costruzione dei potenziali basati sulle *symmetry functions* e ridurre il costo computazionale dei potenziali basati sui *smooth overlap of atomic positions*.
Infine, introduciamo un metodo per ottenere con pochi calcoli aggiuntivi l'incertezza delle proprietà termodinamiche ottenute attraverso simulazioni di dinamica molecolare con PML. Questo metodo può essere usato sia per stimare l'incertezza di un valore ottenuto con un PML, che è necessario quando facciamo previsioni quantitative delle proprietà, sia per esplorare in sicurezza lo spazio delle fasi di interesse, con l'aiuto di un potenziale di ripiego che subentra quando il PML non è ritenuto affidabile.
Nella seconda parte della tesi, dimostriamo la qualità di questi metodi con un caso reale, in cui addestriamo un potenziale per il sistema binario Ga_xAs_{1-x} . Il PML che abbiamo addestrato è in grado di predire accuratamente l'intero diagramma binario di fase, che comprende fasi liquide e solide, metalliche e semiconduttrici. In questo lavoro confrontiamo anche diversi metodi che possono essere usati per generare un set completo di strutture usate per addestrare il PML.

Per dimostrare la trasferibilità del potenziale, calcoliamo varie proprietà, alcune delle quali (ad esempio la tensione superficiale del Ga e GaAs liquido) sono ben oltre le possibilità dei metodi ab initio. Confrontiamo questi risultati con i calcoli *ab initio* di riferimento e con gli esperimenti disponibili, con risultati soddisfacenti, pur limitati dal livello di teoria utilizzato (teoria funzionale di densità con pseudopotenziali GGA).
In ultimo, usiamo il nostro PML per studiare il comportamento del gallio liquido in contatto con la superficie polare [111] del GaAs solido. Risultati sperimentali recenti ritengono che il pre-ordine del liquido all'interfaccia abbia un ruolo importante durante l'accrescimento dei nanofili di GaAs, indicando la polarità come uno dei driver principali per la crescita corretta. Le nostre simulazioni permettono di indagare questo pre-ordine con maggiore dettaglio, supportando e completando le osservazioni sperimentali. Inoltre, esploriamo l'energia libera degli atomi di As nel Ga liquido, che permette di capire il comportamento degli atomi di As durante la crescita per aiutare a identificare le condizioni di crescita ideali.

Parole chiave: potenziali machine learning, selezione CUR, stima dell'incertezza, DFT, arseniuro di gallio, nanofili di GaAs

Contents

List of Figures

List of Tables

1 Introduction

1.1 Molecular simulations

Molecular simulations have become an invaluable tool to predict the behaviour of materials, complementing experiments in our effort to understand and tailor their properties. On one side, we can use simulations to sample spatial and temporal resolutions that are not available to experiments, on the other side we can use our computational predictions to screen the pool of optimal candidates for a certain task before synthesizing and testing them.

Focusing on the first task, molecular dynamics (MD) is one of the most widespread method used for sampling thermodynamic observables in condensed matter and molecules. In this method we explore the phase space by moving particles according to Newton's equations of motion, after determining the mutual interactions. The accuracy of the predictions depends both on the physical approximation used to compute the interatomic forces and on the dimensions of the simulations. In general, the duration and physical size of the simulations determine both the type of problems that we can study and the accuracy by which we can determine their average properties.

Therefore, since the dawn of the field, there have been massive advancements to allow more accurate and faster sampling. These advancements have mostly happened in two directions: the increase in computational power (and the subsequent development of software able to exploit it) and the refinement of the algorithms used to run simulations[1]. In addition to them, the last decade has seen the disruptive rise of machine learning, which promises to boost the accuracy of the simulations[2], reduce their cost[3], and aid with the extraction and analysis of the vast amount of data produced[4].

Thanks to these advancements, we are now able to accurately model large-scale systems, allowing us to peek into the atomistic origin of complex phenomena.

1.2 Machine learning the potential energy surface

In the last decades, machine learning (ML) has heavily influenced all the fields of science where we produce or harvest large amount of data. In the case of materials science, the early applications tried to leverage experimental data to train predictive models of quantitative properties, based on other, simpler, measurements. A review on these early efforts is, for example, Ref. 5.

On the modelling side, the early adopters focused on learning the high-dimensional landscape of interactions, i.e. the potential energy surface (PES) of atoms and molecules[6–8]. Normally, there is a trade-off between the accuracy of the method used to compute the forces in a MD simulation and its computational cost. Large-scale simulations are possible only using empirical interatomic potentials (i.e. parametrizations of the PES based on a handful of parameters), which are computationally cheap but accurate only in limited parts of the phase space. At the other end of the spectrum, we can obtain accurate results using a transferable framework by solving one of the many approximations to the electronic time-independent Schrödinger equation. These models, also called *ab initio* methods, can usually be used only for systems up to a few hundred atoms. Machine learning allows to strike a balance between the two, training on few accurately computed structures and predicting over new configurations at a fraction of the original cost.

Comparing the early works to today, there have been massive advancements in the theory and practice of training machine learning interatomic potentials (MLIP). However, despite the experience that we have accumulated, the training of each potential is often a story on its own. From the choice of the inputs to the construction of the dataset, the human side still plays a major role in the development and refinement of a MLIP.

In general, we need three main ingredients to train a potential:

- a representation of the atomistic structure that provides a machine-efficient description of the system;
- an algorithm, borrowed from the field of supervised machine learning methods, that finds the optimal solution to minimize the loss function of the predictions vs the known energies;
- a dataset of structure-energy pairs that covers the phase space of interest.

Although the general framework is clear, there is still a lot of development ongoing. One of the objectives of current research is to automatize the construction of the potentials, making it as effortless and efficient as possible. This would reduce the amount of man-time needed for the task, which could be better spent on running and analyzing simulations. Some works in this directions include a framework that allows to run *ab initio* MD simulations while training

on-the-fly (i.e. while the simulations is running) a MLIP to reduce the number of calls to the DFT code[9, 10], or a data-driven construction of the training dataset[11–13].

Regarding the inputs, there have been major developments over the last decade, which have also led to a unified theory of the atom-centered representations (see Ref. 14 for a comprehensive review). Today, we know that a successful description must satisfy some constraints, such as being additive (the structure is identified as a sum of local fingerprints) and consistent with the symmetries of the property that we are learning[15–17]. However, there are still some challenges, depending on the representation and framework used to train a potential.

The symmetry functions (SF)[18], one of the earliest successful descriptors, are heavily tied by construction to the system that is investigated. Their functional form makes them sensitive to atoms at specific distances, which means that we need to use completely different sets of SFs to describe a molecular system such as water[19] or crystalline bulk sodium[20]. The choice of symmetry functions for a given system is usually done through a mix of chemical intuition and human experimentation to find an optimal[18, 19, 21], which can be very time consuming. At the other end of the spectrum, the smooth overlap of atomic positions (SOAP)[16], as well as other systematically built descriptors, are not inherently tied to the system, although they still possess a number of hyperparameters that must be tuned. On top of this, their memory footprint and computational cost can grow quickly to impractical number of features when accurately describing systems with multiple species.

In chapter II in this thesis, we first compare these two prototypical descriptor by assessing their ability to learn a given dataset, then compare some feature selection methods, outlining their usefulness for both type of descriptors.

The choice of the ML algorithm is often tied to the descriptor, as some software packages provide the two together, such as symmetry functions and neural networks in RuNNer[22] and ANI[23], or SOAP and Gaussian process regression in QUIP[24]. However, nothing prohibits us from using some specific input together with any supervised-learning method available from a given ML library. As we often borrow our tools from other fields of ML, we will not cover this part in detail in this thesis.

The construction of the dataset used to train the MLIP is another situation where the human choice can make a difference. In general, we distinguish potentials that are trained to sample only small regions of the phase space, such as studies on the stability of alloys around specific stoichiometries and conditions[25, 26], and general-purpose MLIP, that try to include multiple phases of pure elements[27–29], binary compounds[30], or nanoclusters[31]. Whatever the case, we aim to produce an *ad hoc* dataset that contains as much information as possible about the region of interest. Particularly in the case of general purpose potentials, we need to span a very broad phase space, and then compute at *ab initio* level these points. Therefore, there has been a lot of interest to find optimal strategies for the choice and generation of configurations that are used in the training set[11, 12].

Finally, the results obtained with the aid of MLIPs need verification. As we move towards the applications of these potentials to make quantitative predictions of properties, we need methods that allow for a rapid evaluation of the uncertainty of simulations based on MLIPs. Although we often publish numbers (e.g. the prediction error over the training and test set) that try to define the fidelity of the potential with respect to the underlying level of theory, we then use our potentials to run simulations that are not accessible to *ab initio* methods. Since the algorithms that we use ultimately rely on the interpolation of the large number of data provided, we need to understand when the predictions go beyond the sampled phase space, falling into the extrapolation regime where predictions cannot be trusted.

We investigate these issues in chapter III, where we present a method to quantify the uncertainty of thermodynamic quantities obtained through MD simulations with MLIPs. We also show how this can effectively be used to generate new training points over large parts of the phase space, even with inaccurate MLIPs, without generating unphysical configurations.

All of the work done in the recent years has allowed to transform a purely human-driven effort to a more standardized one. Figure 1.1 shows the workflow associated with the construction of a modern MLIP, that allows online control of the accuracy and the possibility to use a fall-back potential (that is used whenever the MLIP is found to be too uncertain) to safely explore larger regions of the phase space. In the end, the results are collected and analyzed to compute a thermodynamic property, together with the uncertainty derived from the use of a MLIP.

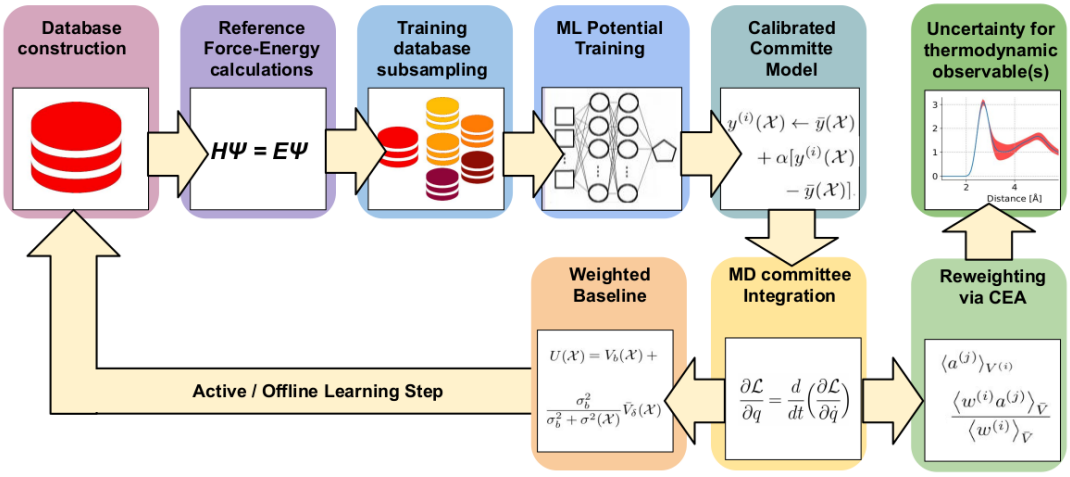

Figure 1.1 – Graphical summary describing the essential steps to train and validate a generic MLIP. We consider here a committee of potentials, which are re-scaled to produce energies and forces in agreement with a Gaussian distribution. This allows for the quantification of the uncertainty for both the energies and the thermodynamic observables computed during the MD simulation

In chapter IV we apply all these methods that we have described this far to fit a potential for

the Ga_xAs_{1-x} system. The automatic selection of the inputs allows to quickly find the optimal description for the very heterogeneous phase space, which encompasses solid and liquid, metallic and semiconducting phases. The uncertainty estimation is useful to measure the confidence of the potential, as well as find new candidate structures for the training set.

It should be mentioned here that there are a number of other applications that we have not discussed. Using the same representation that we introduced for learning the PES, we can also explore the data produced, identifying recurring patterns, both spatial and temporal[4, 32, 33]. These tasks fall within the large family of unsupervised machine learning methods and are not going to be treated in the context of this thesis. Some examples in this category include the automatic detection of ion diffusion in solid-state ionic conductors[34] or the identification of candidate structures for the synthesis of materials[35]. Further recent advancements include the possibility of generating completely new configurations with a set of target properties by working in the feature space and then finding the corresponding structure in the real space[36].

1.3 Bridging the gap between theory and experiments

MLIPs allow us to address some long-standing limitations of *ab initio* MD. Thanks to the low computational cost and its linear scaling with respect to the number of atoms, we can run simulations that are usually out of reach to *ab initio* accuracy. Some examples include:

- Thermodynamic properties computed at finite temperature (although in some cases this can be done with *ab initio* methods[37, 38]);
- The quantum behaviour of the nuclei[39];
- Systems whose length or time-scale are too large.

Thanks to these improvements, we can compute more accurately many properties, that can be compared to the ones measured in the experiments. This allows our models to complement the experimental results, explaining the origin of the observed behaviour, or to guide the future searches by providing accurate predictions.

Regarding the first point, it is straightforward to compute at *ab initio* level the formation energy for a set of (geometrically optimized) candidate phases of a molecular crystal to find the most stable configuration. However, at non-zero temperatures, the dynamics of the structure comes into play, which might stabilize a different phase[40]. A complete picture of the finite temperature effects can be obtained, for example, by running long MD simulations of all the competing systems to compare their free energies. However, this can be achieved only with empirical potentials due to the large computational cost, without the guarantee that the potential is accurate over the whole phase space. Another possibility is to approximate the physical behaviour of the system, limiting the accessible degrees of freedom, thus reducing

the number of calculations needed. For example, as a first approximation, we can study only the harmonic motion of the bonds around the stable configuration, which allows to compute the approximate free energy of the system with only a handful of configurations, even at *ab initio* level. However, these approximations can still fail for complex structures[41], even when we try to add anharmonic contributions to the motion. MLIPs allow us to overcome the limits of both methods and run the full MD trajectories, obtaining accurate free energies that can directly be compared to experiments[42].

The same approach can be used for the case of vibrational frequencies. Instead of computing the vibrational spectrum on a handful of configurations, we can now run accurate MD trajectories to obtain the average spectrum at the given temperature, to compare to the experimental results. This has been done for both infrared[43] and Raman[44, 45] spectroscopy, as well as nuclear magnetic resonance shifts[46].

Another limit that we can overcome with the aid of MLIPs is the classical treatment of the atoms, irrespective of their mass and the temperature of the simulations. By using Newton's equations of motion, we implicitly assume that the atoms are classical particles. However, this approximation does not hold for light nuclei, such as hydrogen, or for properties that are computed at low temperature, where the quantum behaviour of the nuclei becomes relevant, such as the heat capacity.

There are several methods that allow to properly account for these nuclear quantum effects (NQE) by mapping the quantum problem onto a system of multiple classical dynamics trajectories, although this ultimately introduces additional degrees of freedom and therefore additional calculations. Here we consider path integral molecular dynamics (PIMD), where atoms are replaced by chains of P beads connected by springs. The computational effort to run a simulation within the PIMD formalism is therefore P times the cost of the original simulation, making it impossible to reach with *ab initio* methods. In the past, this limited our studies of NQEs to few systems, mostly water, for which accurate empirical potentials have been fitted[47]. Today, we can run accurate PIMD simulations of any system after fitting an appropriate MLIP, thus quantifying the contribution of the NQEs to a given property, reducing the distance between our predictions and the experimental values (see, for example, the case of methane[48], aspirin[49], and fullerene[50]).

In chapter IV we show how MLIPs allow to run finite temperature simulations with the inclusion of NQEs, achieving accurate results in good agreement with the experiments. Furthermore, we run the same simulations for the available empirical potentials, demonstrating their lack of transferability when moving away from the conditions for which they have been fitted.

Finally, experiments and simulations are converging towards the study of the same systems, both in size and timescales, thanks to the major advancements in both fields. For example, recent works have shown qualitative and quantitative agreement in the measurement of the scattering angle of hydrogen atoms on graphene nanosheets[51], where the models helped to elucidate the different contributions to the final result. MLIPs have also uncovered the origin of

the structural and electronic transitions in disordered silicon, running massive 100000-atoms big simulations for 200 ps, complementing the experimental findings that are not able to observe these local arrangements in such detail[52]. Other examples where the use of MLIPs has allowed to study systems that were not accessible before include the study of complex catalytical interfaces[53, 54] and accurate studies of the radiation damage in bulk silicon[55].

In chapter V we provide an example of a system of experimental interest that we can now investigate with our simulations. Using the potential introduced in chapter IV, we simulate the complex Ga_{liq}-$GaAs_{sol}$ interface and compare it to the experimental observations. Our results agree with the experimental scanning tunneling electron microscopy images, allowing an in-depth analysis of the effect of the pre-ordering of the liquid at the interface. We follow up with further calculations of the free energy of the As atoms in the liquid, which can provide some direction for the future experimental endeavours.

2 Representations for Machine Learning[1]

2.1 Introduction

To identify a structure, we generally define it through the list of Cartesian coordinates $\{r_i\}$ of its constituent atoms. However, this is not effective when it comes to interatomic potentials. For example, it is more straightforward to define the forces acting on a pair of atoms using the relative distance between the two particles rather than using their absolute positions, as the relative distance incorporates the notion that total energy will not change if the whole system is translated. Although these representations can be used as an input for both classic (with this term we refer to potentials built and optimized for a set of given parameters) and machine learning potentials, a different approach has become preponderant in the last decade for MLIPs.

For the MLIPs we favour a representation of the system built on "local environments", i.e. the description of the surroundings of an atom through a many-body mathematical function defined within a cut-off. This allows us, for instance, to describe the energy as a sum of local contributions. Similarly to the case of classical potentials, we must ensure that the representation that we use is invariant with respect to trivial symmetries, such as translations and rotations of the system, and permutation of the atoms. Much work has been done in the recent years to obtain a more complete understanding of these representations, with a focus on their relationship to each other[58, 59], their completeness[60, 61], their sensitivity to small variations in the structures[62], and their computational efficiency[63].

In this chapter we provide a general introduction to these atom-centered representations, first in general and then focusing on some specific descriptors. Then, we compare their ability to learn a dataset of water dimers and trimers used for fitting the many-body water potential

[1] Sections 2.4 and 2.5.4 in this chapter are adapted from Ref. 56. The author of the thesis has contributed in this work by producing and analysing the figures, and by comparing the effect of selecting training points on the learning, presented in Sec. 2.5.4 The fitting and testing of the various frameworks has been done by the other authors. Sections 2.3 and 2.5 have been adapted from Ref. 57. The author has contributed in the paper by writing parts of the code used, testing the method on all the systems described, and writing the paper.

MB-pol[64]. Finally, we study how we can effectively reduce the number of features used in the learning task without compromising the quality of the regression.

2.2 Atom centered representations

In this section we provide only a quick and (hopefully) intuitive description of the theory of atom centered representations. More comprehensive discussions on the topic can be found in specialized articles and reviews[14, 58, 59].

The starting idea is to represent the set of atom positions $\{r_i\}$ with localized functions g placed on top of each atom. Usually, either Dirac-δ functions or Gaussians are used to represent the position of the atoms. This allows us to express the system as a density field

$$\langle \mathbf{x}|A;\rho\rangle \equiv \sum_{i\in A}\langle \mathbf{x}|\mathbf{r}_i;g\rangle, \tag{2.1}$$

where $\langle \mathbf{x}|\mathbf{r}_i;g\rangle \equiv g(\mathbf{x}-\mathbf{r}_i)$. Here we use a notation that mimics the Dirac bra-ket formalism, where the bra indicate the entity that is being represented (i.e. the positions $\mathbf{x}$ of the atoms) and the ket the representation target (the structure A) and the nature of the representation (the density field ρ and the function g). Sometimes, in the ket we omit the explicit reference to a target structure when we discuss the general theoretical development.

The first step is to distinguish the different atomic species of the system, which is done by decorating the positions with a function a, which can also be used to describe other properties of the atoms, such as polarization. Then, we make the representation translationally invariant by centering on an atom i and describing only the surrounding environment with the density field introduced above. We limit our description only to entities within a given cut-off radius with the aid of a cut-off function $f_{\text{cut}}(r_{ij})$, where r_{ij} is the distance between the central atom i and other atoms j. The density field around i can be written now as,

$$\langle a\mathbf{x}|A;\rho_i\rangle = \sum_{j\in A_i}\delta_{aa_j}\langle \mathbf{x}|\mathbf{r}_{ij};g\rangle f_{\text{cut}}(r_{ij}). \tag{2.2}$$

The rotational invariance is achieved in two steps. First we rewrite the field representation to include $(\nu+1)$ body order correlations, then we perform Haar integration over the rotation group and over inversion. This leaves us with

$$\langle a_1\mathbf{x}_1;\ldots a_\nu\mathbf{x}_\nu|\overline{\rho_i^{\otimes\nu}}\rangle = \sum_{k=0,1}\int_{SO^3} d\hat{R}\langle a_1\mathbf{x}_1|\hat{R}\hat{i}^k|\rho_i\rangle\ldots\langle a_\nu\mathbf{x}_\nu|\hat{R}\hat{i}^k|\rho_i\rangle, \tag{2.3}$$

where $\overline{\rho_i^{\otimes\nu}}$ is a tensor product of ν atom-centered fields averaged over all possible improper rotations, $\hat{i}$ is the inversion operator and $\hat{R}$ the rotation operator.

Most of the known representations used today for machine learning purposes can be recovered by choosing an appropriate basis and by expanding up to a certain body order $\nu+1$. In the

next section we show how this can be done for the case of the SOAP representation.

2.2.1 Smooth Overlap of Atomic Positions

The SOAP representation introduced by Bartók *et al.*[16] can be obtained from Eq. 2.3 by expanding the translationally invariant ket of Eq. 2.2 into a basis of orthonormal radial basis functions $R_n(r) \equiv \langle x|n\rangle$ and spherical harmonics $Y_l^m(\hat{\mathbf{x}}) \equiv \langle \mathbf{x}|lm\rangle$,

$$\langle anlm|\rho_i\rangle = \sum_{j\in A_i} \delta_{aa_j} \int d\mathbf{x} \langle nl|x\rangle \langle lm|\hat{\mathbf{x}}\rangle \langle \mathbf{x}-\mathbf{r}_{ji}|g\rangle \tag{2.4}$$

Then, by fixing the body order expansion to $\nu = 2$, we obtain the power spectrum as in Ref. [16],

$$\langle a_1 n_1; a_2 n_2; l|\overline{\rho_i^{\otimes\nu}}\rangle = \frac{1}{\sqrt{2l+1}} \sum_m (-1)^m \langle a_1 n_1 lm|\rho_i\rangle \langle a_2 n_2 l(-m)|\rho_i\rangle. \tag{2.5}$$

By increasing the number of radial basis functions and spherical harmonics, we can converge the discrete representation of the system to its limit. If the expansion contains $n_{\max}$ radial functions, and maximum angular momentum channel $l_{\max}$, the power spectrum contains $n_{\max}^2 l_{\max}$ elements. In the case of a system with multiple species, this comes at a considerable computational cost, since tens of thousands of power spectrum elements have to be computed and processed.

2.2.2 Behler-Parrinello Symmetry Functions

Following the same notation, the atom-centered symmetry functions (SF) can be obtained from Eq. 2.3 by projecting the SO^3 invariant ket onto a suitable test function G, either for the 2-body SFs or the 3-body ones[58]. The difference between this representation and the "historical" Behler-Parrinello SFs lies in the fact that the original ones had not been built to converge to the limit of Eq. 2.3, and only a limited number of carefully selected SFs are used every time.

From here on, we describe briefly the nature of the SFs as they were originally intended, leaving a more thorough treatment of the topic to the many reviews available[18, 65–67]. We limit our scope to the two families of Behler-Parrinello SFs which we use in this work.

The first functional form, called G_2 following the convention used in previous works [18, 67, 68], provides information about pair correlations between a central atom i and its neighbours,

$$G_2^i = \sum_j e^{-\eta(r_{ij}-r_s)^2} \cdot f_c(r_{ij}), \tag{2.6}$$

where the parameters η and r_s control the width and the position of the Gaussian with respect

to the central atom and $f_c(r_{ij})$ is a cutoff function that ensures that the symmetry function smoothly decreases to 0 in value and slope at a fixed cutoff r_c. The sum is over all neighboring atoms being closer than r_c. The second type of symmetry functions, called G_3, provides information about angular correlations, and has the form

$$G_3^i = 2^{1-\zeta} \sum_j \sum_{k \neq j} (1 + \lambda \cdot \cos\theta_{ijk})^\zeta \cdot e^{-\eta(r_{ij}^2 + r_{ik}^2 + r_{jk}^2)} \cdot f_c(r_{ij}) f_c(r_{ik}) f_c(r_{jk}), \tag{2.7}$$

where ζ, η, and λ are the three parameters that determine the shape of this type of symmetry function, and θ_{ijk} is the angle among the triplets of atoms considered. The indices j and k run over all the atoms in the neighbourhood of the tagged atom i. The cutoff function that we have used has the form

$$f_c(r_{ij}) = \begin{cases} \tanh^3\left[1 - \frac{r_{ij}}{r_c}\right] & \text{for} \quad r_{ij} \leq r_c \\ 0.0 & \text{for} \quad r_{ij} > r_c \end{cases}. \tag{2.8}$$

Since in section 2.3 we discuss a method to sparsify a large set of these SF fingerprints, a first preparatory step involves the determination of a thorough yet manageable pool of candidate SFs. The generation is done spanning over all of the meaningful sets of parameters, using simple heuristic rules to represent most of the possible correlations within the cutoff distance. We generate two separate sets of radial symmetry functions, G_2. The first group is centered on the reference atom (i.e. $r_s = 0$) and the width varies as

$$\eta_m = \left(\frac{n^{m/n}}{r_c}\right)^2, \tag{2.9}$$

where n is the number of intervals in which we have chosen to divide the space and $m = \{0, 1, ..., n\}$. The second group is centered along the path between the central atom and its neighbours, at increasing distances following

$$r_{s,m} = \frac{r_c}{n^{m/n}}, \tag{2.10}$$

while the Gaussian widths are chosen as

$$\eta_{s,m} = \frac{1}{(r_{s,n-m} - r_{s,n-m-1})^2} \tag{2.11}$$

in order to have narrow Gaussians close to the central atom and wider ones as the distances increases. This effectively creates a finer grid closer to the central atom, where small variations in the position have a larger effect on the potential (see Fig. 2.1).

The G_3 symmetry functions were generated with a similar procedure, choosing values for η according to Eq. 2.9, setting λ to both values $\{-1, 1\}$ that were originally proposed and choosing a few values of ζ on a logarithmic scale. For instance, in the examples below we use $\{1, 4, 16\}$.

By increasing the cutoff radius and the number N of symmetry functions that are generated,

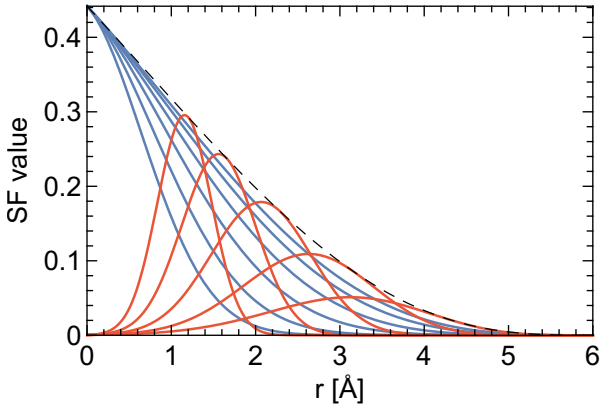

Figure 2.1 – Examples of radial symmetry functions generated using $N = 5$ and $r_c = 6$ Å. The blue curves are the symmetry functions centered in the origin ($r_s = 0$) and η varying as in Eq. 2.9, while the red ones have their center shifted using r_s as described in Eq. 2.10 and η is described by Eq. 2.11. The black dashed curve is the cutoff function for $r_c = 6$ Å.

one can make the description of the environment more and more complete. This comes however at the expense of greater computational costs, since a large number of SFs would then have to be generated at each potential evaluation. Less obviously, using too many, strongly correlated symmetry functions could lead to overfitting and difficulties in the regression process.

2.2.3 Permutationally invariant polynomials

Similar to the SFs, but more systematic in their construction are the permutationally invariant polynomials (PIPs) that have been introduced soon after the SFs[69]. A thorough description of the 2-body and 3-body expansion of Eq. 2.3 is obtained by enumerating all the bonds and angles within a cutoff. These bonds and angles are then used as variables for linear or exponential functions that are used to fit the PES. This approach has been shown to work in the context of small molecules[70–72], but its inherent scaling have made it too expensive for large systems[73].

Thanks to its accurate results it has still been used in the past to fit the short range interactions in the MB-pol water potential[64]. In Sec. 2.4 we use it as a basis to compare the SOAP and the SFs that have already introduced.

It should be noted that the issue of the scaling has been mitigated recently, showing promising results in the fitting of bulk materials[73].

2.2.4 Other representations

Although we are not discussing in this thesis all the other representations that have been proposed in the last decade, it is important to mention that many groups have undertaken a huge effort in the advancement of the field.

From the general formulation of Sec. 2.2 we can recover other representations, which come from using different basis sets (e.g. the atomic cluster expansion[59] and the moment tensor potentials[74]), expanding to different body-orders (the bispectrum[16]), or by using a different symmetrization procedure (e.g. the DeepMD inputs[75] and the FHCL features[76]).

In general, they have all been developed to retain the largest amount of information of the system, to learn a given property (the energy, in the context of MLIPs) with the lowest number of training points. Another, equally important, target is the computational efficiency, since running MD for large systems require as many calculations as there are atoms in the system.

2.3 Feature selection

Despite being based on very different premises, most representations discussed in Sec. 2.2 can lead to an arbitrarily high-dimensional feature space. This can be a bottleneck in the MD simulations, where we strive for efficient calculations of forces and energies, or for large databases with structures containing several chemical species. A possible solution to lessen the computational cost and memory footprint is to reduce the number of features used for learning, discarding those that provide a low informational content or are redundant.

For example, the "systematic" generation of symmetry functions that we have proposed in the second part of Sec. 2.2.2 yields a certain number of redundant functions, that probe very similar regions of the space. A simple selection method could be, in a more or less automatic fashion, the empirical evaluation of the accuracy of a ML model based on various subsets of SFs. Alternatively, genetic algorithms have been recently proposed as a method to generate an optimal selection [77], similar to what had been done in the past to select an optimal set of reference structures [78].

Here we focus on unsupervised approaches that rely only on knowledge of the geometries of the reference structures, without using information on energy and forces, nor on the performance of the ML model that results from a given choice of input features.

The first approach we discuss is based on a relatively simple idea: given a set of M structures $\{A\}$ that are representative of the system of interest, and a large number N of fingerprints $\{\Phi_j\}$, one can build the $M \times N$ matrix $\mathbf{X}$ such that $X_{ij} = \Phi_j(A_i)$ (where A_i are the environments of structure A). The most effective fingerprints can then be chosen by using standard linear algebra techniques to approximate $\mathbf{X}$. Unless otherwise specified, we consider the local environments, rather than the entire structure, as the core of our discussion. The elements of $\mathbf{X}$ refer to the fingerprints defining these environments A_i, which we consider, in order to

simplify the notation, without explicit reference to the structure they are part of.

Therefore, we aim to find the optimal $M \times N'$ feature matrix $\mathbf{X}'$, where $N' \ll N$, that still provides a satisfactory representation of the space while reducing the computational load of the ML scheme. This is essentially a dimensionality reduction problem, that can be interpreted in terms of the construction of a low-rank approximation $\tilde{\mathbf{X}}$ of the feature matrix. Most of the dimensionality techniques available for this task, such as singular value decomposition (SVD), generate new features that are a linear combination of the initial set and cannot be used for our current purpose, as they would still require the evaluation of all the N features and, only as a second step, project them onto a lower-dimensional space. We have therefore considered methods that strive to obtain a low-rank approximation of the feature matrix or its associated covariance using only rows and columns of $\mathbf{X}$. We discuss in particular two approaches, namely CUR decomposition and farthest point sampling (FPS).

2.3.1 CUR Decomposition

CUR decomposition [79] is a feature selection method that has been developed to deal with data where the information provided by the singular vectors cannot be properly interpreted, such as gene expression data. In analogy with the low-rank approximation obtained with a singular value decomposition, one writes

$$\mathbf{X} \approx \tilde{\mathbf{X}} = \mathbf{C}\,\mathbf{U}\,\mathbf{R} \tag{2.12}$$

where $\mathbf{C}$ and $\mathbf{R}$ are actual rows and columns of the original matrix. The objective is still to find the best low-rank approximation to $\mathbf{X}$, but in this case only actual elements of the matrix are used, which implies that $\tilde{\mathbf{X}}$ can be obtained without having to compute all N fingerprints.

We discuss in particular the procedure for selecting a reduced number of columns (i.e. fingerprints), but the method can also be used to reduce the number of rows (i.e. reference structures) [80]. An intuitive representation of the method is provided in Fig. 2.2. Each column c of the initial feature matrix is given an "importance score" calculated as

$$\pi_c = \sum_{j=1}^{k} (v_c^{(j)})^2, \tag{2.13}$$

where $v_c^{(j)}$ is the c-th coordinate of the j-th right singular vector, and k is the number of features that have yet to be selected and runs from N' to 1. We also observe that a very effective selection can be obtained by using a fixed number of singular vectors $k = 1$ at each iteration in the procedure (CUR($k = 1$)). Not only this makes the method numerically more stable and significantly faster, but it makes the selection independent on the target number of symmetry functions, so that one can effectively perform a single selection with a large N', obtaining a list of SF that is sorted from the most important to the least important. The importance score can also be weighted by a factor if one wants to prioritize the selection of a

certain type of features, e.g. if the cost of evaluating different fingerprints varies greatly, and one would rather take several "cheap" fingerprints than a single "expensive" one.

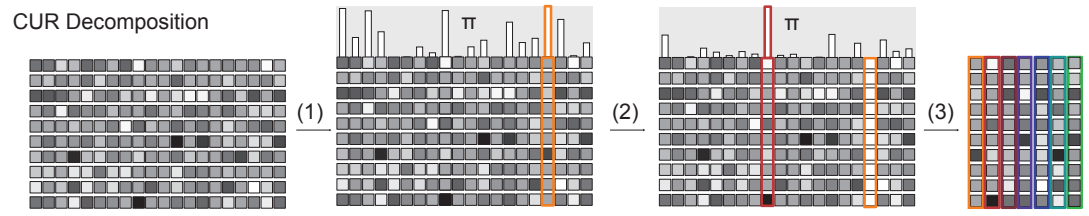

Figure 2.2 – Schematic representation of the deterministic CUR selection used to select fingerprints. (1)Compute the importance score π for each column; select the feature that maximises π.(2)Orthogonalise the matrix with respect to selected feature; recompute π; select the feature which maximises π;(3)Repeat step (2) until the target number of features is obtained. Figure reproduced with permission from Ref. [81].

Most CUR schemes employ a probabilistic criterion for feature selection, to guarantee e.g. that if several nearly-identical features are present, any of them will have approximately the same probability of being selected. To obtain a deterministic selection, we pick at each step the column with the highest score, and avoid selecting multiple nearly-identical features with an orthogonalization procedure. After having selected the l-th column with the highest importance score, every remaining column in $\mathbf{X}$ is orthogonalized relative to it

$$X_j \leftarrow X_j - X_l \, (X_l \cdot X_j) / \, |X_l|^2 \, . \tag{2.14}$$

The SVD is then re-computed based on the orthogonalized matrix, and the column weights are re-evaluated. The procedure is iterated until all N' features have been chosen to build the $\mathbf{C}$ matrix, that corresponds to the reduced feature matrix $\mathbf{X}'$. Since in this application we are only interested in reducing the number of fingerprints, $\mathbf{R} = \mathbf{X}$, and we can compute $\mathbf{U} = \mathbf{C}^{+}\mathbf{X}\mathbf{X}^{+}$, where $\mathbf{A}^{+}$ indicates the pseudoinverse. One can then compute the accuracy of the approximation as

$$\epsilon = \|\mathbf{X} - \mathbf{C}\mathbf{U}\mathbf{R}\|_F \, / \, \|\mathbf{X}\|_F \tag{2.15}$$

The total number of features to be selected, can either be fixed a priori, or increased until ϵ becomes smaller than a prescribed threshold.

2.3.2 Farthest Point Sampling

Alternatively, one can select the features using a farthest-point sampling (FPS) approach[82]. This is analogous to the strategy with which one can select uniformly-spaced reference points (see e.g. [83]), but here we apply it to the *columns* of $\mathbf{X}$, so as to select fingerprints that are as diverse as possible for the data set being investigated. In a FPS scheme, successive points are chosen so as to maximize the Euclidean distance between them. After arbitrarily selecting the

first fingerprint, each subsequent one is chosen as

$$k = \text{argmax}(\min_j |X_k - X_j|), \tag{2.16}$$

where j refers to all of the features that have already been selected. The procedure is repeated until all N' features have been chosen.

2.3.3 Global Fingerprints and Training Set Selection

As mentioned before, one could use the CUR or FPS methods to sparsify the training set, that is to reduce the number of reference structures rather than the number of fingerprints. This can be useful to reduce the cost of evaluating a ridge regression model, or to minimize the number of property evaluations that need to be performed in order to train the model [80, 84, 85]. In order to do so, it is useful to construct a set of fingerprints associated with the whole structure, rather than with individual atomic environments. A straightforward definition of a "global" fingerprint associated with a structure A, $\bar{\Phi}(A)$ is the average of all the local fingerprints for the environments that compose the structure A, i.e.

$$\bar{\Phi}_j(A) = \sum_{A_i \in A} \Phi_j(A_i) / N_{\text{at}}(A). \tag{2.17}$$

In the case of Behler-Parrinello symmetry functions, that are defined separately for each chemical species, we consider that the global fingerprint is composed by concatenating sections corresponding to each element. In other terms, one can see this as a sparse representation for a larger fingerprint vector that is padded with zeros in all sections but the relevant one, even though in a practical NN implementation one only computes symmetry functions associated with the identity of the central atom. The fingerprint vector for the entire structure can then be built according to (2.17), summing these zero-padded vectors over all atoms in the structure.

In general, this global representation of the structure can also be used to fit quantities that are related to the state of the whole structure, such as thermodynamic quantities, in the same way that local representations are used to fit extensive quantities. However, in this thesis we will only use global descriptors to compare different structures, to identify configurations that are representative of specific regions of the phase space, discarding the redundant ones.

2.4 Comparing representations on a real dataset

Having set up the theoretical background that we need, we can move to the more practical concerns related to the representations.

It is natural, when considering the plethora of different descriptors and relative ML algorithms that are available, to question what is the "best" one for regression. It has been shown that the quality of the representation has a great impact on the regression[86]. However, it is hard to

draw an accurate comparison among the many representations available today, as the quality of the fit can depend on many factors, such as the dataset in use, the implementation, and even the experience of the user.

Here we present a comparison between different frameworks that are available to fit a MLIP. It should be clear that it is not a direct comparison of the different representations, but rather an analysis of the performances of the frameworks (i.e. the combination of representations with a certain ML algorithm, usually packed in a single software) on a specific dataset. The three frameworks that we compare are SOAP with Gaussian approximation potentials (GAP)[24], SFs with neural networks (NN)[22], and the PIPs, which are used as inputs for a set of linear and exponential functions[64]. For the sake of brevity, we omit the details on the hyperparameters used for each representation, which can be found in the original paper and S.I.[56].

2.4.1 The dataset

The comparison is done on a "real-world" dataset of water dimers and trimers, which aims to map all the relevant short range interactions among water molecules to fit the MB-pol water potential[64].

The base idea of the MB-pol potential is to express the energy of N interacting water molecules as a sum of body-order expansions, where the n-th order is defined iteratively as the energy of a cluster of n molecules from which we subtract every lower body-order, as

$$V^{\mathrm{nB}}(1,..,n) = E_n(1,..,n) - \sum_{i=1}^{N} V^{\mathrm{1B}}(i) - \sum_{i<j}^{N} V^{\mathrm{2B}}(i,j) - \cdots - \sum_{i<j<\cdots<n-1}^{N} V^{(\mathrm{n\text{-}1})\mathrm{B}}(i,j,...,(n-1)) \quad (2.18)$$

where $V^{\mathrm{1B}}(i) = E(i) - E_{eq}(i)$ corresponds to the 1B (one-body) energy required to deform an individual water molecule.

In this learning exercise, we use the datasets of water dimers and trimers that are used to define V^{2B} and V^{3B} respectively. However, we do not learn the full V^{2B} and V^{3B} terms, but only the short-range correlations $V^{\mathrm{2B}}_{\mathrm{short}}$ and V^{3B}_{short}, since the ML frameworks and representations that we use are defined only up to a given cut-off. The short range energies are defined by subtracting the long-range interactions that can be explicitly accounted using the classical expressions for electrostatics, induction and dispersion, i.e.

$$V^{\mathrm{2B}}_{\mathrm{short}}(i,j) = V^{\mathrm{2B}}(i,j) - V^{\mathrm{2B}}_{\mathrm{TTM,elec}}(i,j) - V^{\mathrm{2B}}_{\mathrm{TTM,ind}}(i,j) - V^{\mathrm{2B}}_{\mathrm{disp}}(i,j) \quad (2.19)$$

and

$$V^{\mathrm{3B}}_{\mathrm{short}}(i,j,k) = V^{\mathrm{3B}}(i,j,k) - V^{\mathrm{3B}}_{\mathrm{TTM,ind}}(i,j,k) \quad (2.20)$$

where TTM refers to a modified Thole-type scheme originally used in the TTM4-F model of water[87].

2.4.2 Results

In Table 2.1 we report the root mean squared errors (RMSEs) obtained with PIPs, NNs, and GAPs for the 2B and 3B datasets. For the 2B term, all three methods achieve similar accuracy:

	2B			3B		
	training	validation	test	training	validation	test
PIP	0.0349	0.0449	0.0494	0.0262	0.0463	0.0465
NN	0.0493	0.0784	0.0792	0.0318	0.0658	0.0634
GAP	0.0176	0.0441	0.0539	0.0052	0.0514	0.0517

Table 2.1 – RMSE (in kcal/mol) per isomer on the provided training, validation, and test sets in the PIP, NN, GAP short range interaction two-body (2B) and three-body (3B) energy fitting.

the error on the training set is less than 0.050 kcal/mol per dimer while the errors on the validation and test sets are less than 0.080 kcal/mol per dimer. These errors demonstrate a high level of accuracy since the average value of the target energies in the dataset is 3 kcal/mol. Among the three, the 2B PIP model appears to perform better on the validation and test sets and suffers less from overfitting. The difference in RMSEs for the training set and the test set are below 0.02 kcal/mol with PIP, but around 0.03 kcal/mol with NN and 0.04 kcal/mol with GAP. The GAP model gets a slightly lower error for the training set, but overfitting prevents to achieve a similar accuracy for the test set.

In order to investigate in more detail the performance of the different regression schemes for predicting the 2B and 3B energies over the MB-pol dimer and trimer data sets, we use a dimensionality reduction scheme to obtain a 2D representation of the structure of the training set. We follow a procedure similar to that used in Ref. 88 to map a database of oligopeptide conformers. We assess the similarity between reference conformations of dimers or trimers with a metric based on SOAP descriptors [84]. We obtain a 2D map that best preserves the similarity between 1000 reference configurations selected by farthest point sampling [82] using the sketch-map algorithm [83, 89]. All other configurations (training and testing) are then assigned 2D coordinates (x_i, y_i) by projecting them on the same reference sketch-map. We then compute the histogram of configurations $h(x,y)$, the averages of the properties of the different configurations, and of the test RMSE for the various methods, conditional on the position on the 2D map, e.g.

$$\begin{aligned} h(x,y) &= \langle \delta(x-x_i)\delta(y-y_i) \rangle \\ V^{\mathrm{2B}}_{\mathrm{short}}(x,y) &= \frac{\langle V^{\mathrm{2B}}_{\mathrm{short}}(i)\delta(x-x_i)\delta(y-y_i)\rangle}{h(x,y)}. \end{aligned} \tag{2.21}$$

Figure 2.3 demonstrates the application of this analysis to the dimer dataset. One of the sketch-map coordinates correlates primarily with O-O distance, while different relative orientations and internal monomer deformations are mixed in the other direction. Conformational space

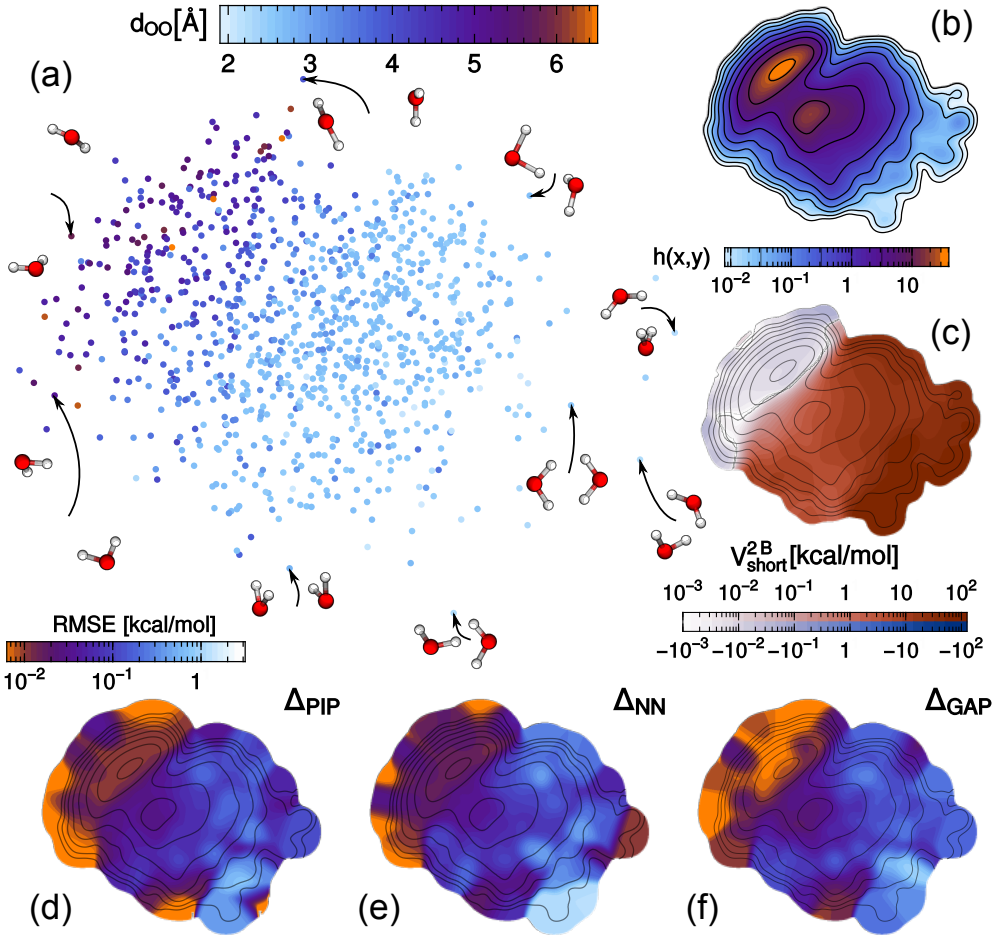

Figure 2.3 – (a) Sketch-map representation for the training data set for dimer configurations. Points are colored according to O-O distance, and a few reference configurations are also shown. (b) Histogram of the training point positions on the sketch-map. The training set density is also reported on other plots as a reference for comparison. (c) Conditional average of the 2B energies for different parts of the training set. (d-f) Conditional average RMSE for the PIP, NN, GAP fits of the 2B energy in different parts of the test set.

is very non-uniformly sampled (Fig. 2.3b), with a large number of configurations at large O-O distance – which correspond to V^{2B}_{short} of less than 0.01 kcal/mol – and at intermediate distances, with sparser sampling in the high-energy, repulsive region (Fig. 2.3c). It is interesting to see that the three regression schemes we consider exhibit very similar performance in the various regions, with tiny errors < 0.01 kcal/mol for far-away molecules, and much larger errors, as large as 1 kcal/mol, for configurations in the repulsive region. These large errors are not only due to the high energy scale of V^{2B}_{short} in this region: the largest errors appear in the portion of the map which is characterized by both large V^{2B}_{short} and low density of sample points.

Figure 2.4 shows a similar analysis for the case of the trimer data and V^{3B}_{short}. 3B energies

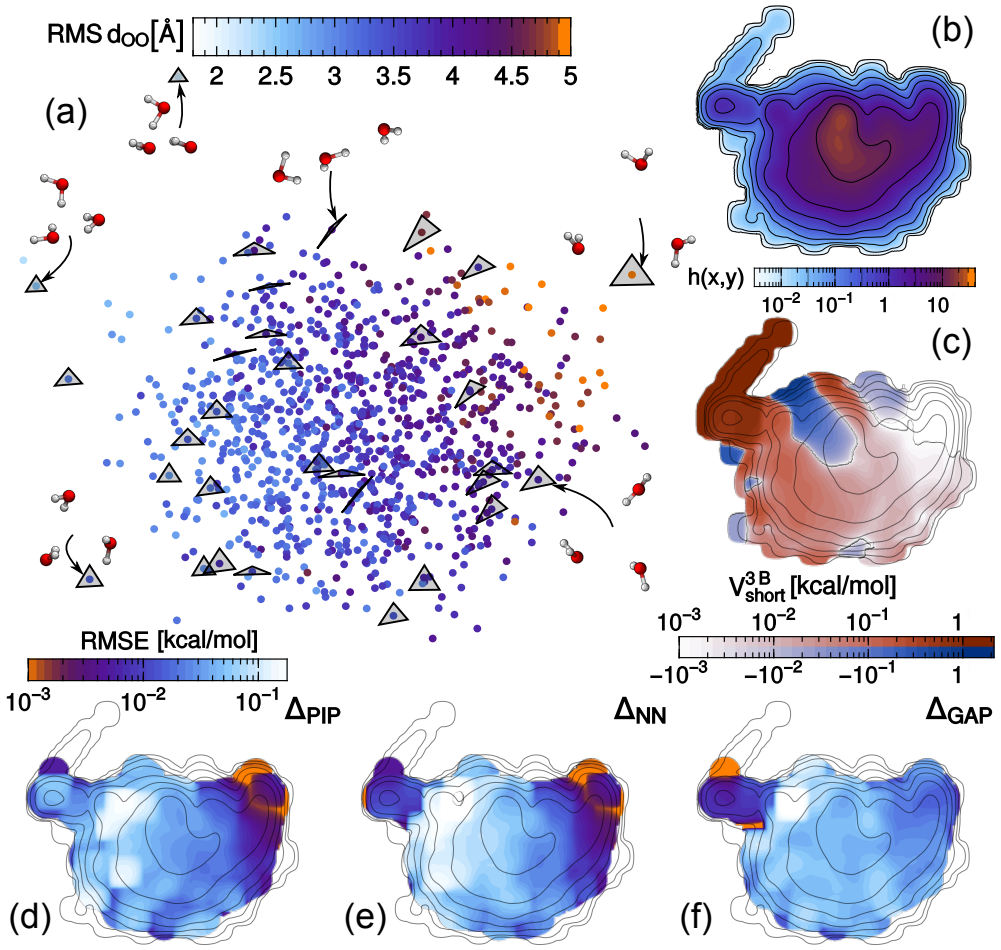

Figure 2.4 – (a) Sketch-map representation for the training data set for trimer configurations. Points are colored according to the root mean square of the three O-O distances; trimer geometries are also represented as triangles, together with a few structures for which a snapshot is shown. (b) Histogram of the training point positions on the sketch-map. The training set density is also reported on other plots as a reference for comparison. (c) Conditional average of the 3B energies for different parts of the training set. (d-f) Conditional average RMSE for the PIP, NN, GAP fits of the 3B energy in different parts of the test set.

span a smaller range than the 2B component, that includes most of the core repulsion. The higher dimensionality of the problem, however, makes this a harder regression problem, as is apparent from the irregular correlations between energy and position on the map, that reveals an alternation of regions of positive and negative contributions.

As a result, the absolute RMSE accuracy of the regression models is comparable to that for the 2B terms, with PIP and GAP yielding comparable accuracy (RMSE ≈ 0.05 kcal/mol), followed closely by NN (RMSE ≈ 0.06 kcal/mol). As in the case of 2B energy contributions, an analysis of the error distribution shows that improving the sampling density and uniformity for the training set is likely to be the most effective strategy to further improve the model. Errors are

concentrated at the periphery of the data set. The good performance of the GAP model can be traced to the fact that it provides a very good description of the short RMS d_{OO} region, even if only a few reference structures are available, even though it performs less well than PIP or NN for configurations that involve far away molecules.

Overall, we can see that these three representations are more or less equivalent, achieving similar results for both datasets. Therefore, one can argue that the choice should be dictated by other metrics, such as the computational cost, or the complexity of the code used for training.

It should be noted that more recently, other works have tried to systematically investigate these frameworks and the underlying representations. A recent work has compared the GAP and the NN with other frameworks, investigating not only the quality of the fit, but also the computational cost associated[90]. Other works have investigated more directly the representations and their ability to retain information and capture small changes across different structures[62, 91].

2.5 Applications of feature selection

A second practical concern that we discuss is the possibility to reduce the computational cost associated with a certain representation by reducing the number of fingerprints that are evaluated and kept in memory. As we mentioned in Sec. 2.3, the feature vectors that we generate can become impractical for complex systems. We show here that we can retain the accuracy of the full representation while reducing the features, even down to 3% of the initial number.

The first two examples that we discuss refer to the case of SFs and NNs, the third one is done on the SOAP representation, while the last one shows the advantage of systematically selecting the structures used for training.

2.5.1 A Potential for Liquid Water

As a first example, we consider the case of liquid water. For this system we can compare our approaches to the SFs of a previously published NN potential that has been built out of carefully-chosen fingerprint functions, selected based on a combination of physical intuition and trial-and-error. This potential, that has been trained on a DFT reference data set [19], and that has been applied to study a variety of properties of liquid water, provides a remarkably concise description of water environments, consisting of only 32 G_2 and 25 G_3 functions. Using the same or similar symmetry functions, also alternative parameterizations have provided excellent results for water [92, 93], electrolytes [94] and even solid-liquid interfaces [95, 96].

In order to identify automatically suitable sets of fingerprints for water, we start by taking the same data set that was used in Ref. [92], and selected by FPS a set of 1000 structures that we use for symmetry function selection and training. We generate an initial pool of 768 SF combining

three sets of G_2 functions obtained following the protocol discussed in Section 2.2.2, with $N = 8$ and cutoffs $r_c = 4, 8, 12$ bohr, and two sets of G_3 SF generated with $N = 8$ – one with $r_c = 4$ bohr and $\zeta = 1, 2, 4, 8, 16$ and one $r_c = 8$ bohr and $\zeta = 1, 2, 4$. Final results are not sensitive to these choices, that we only made to have intermediate files of manageable size. We removed duplicate SFs and those with a length scale smaller than 0.75 Å. We weighted the importance scores (2.13) by a factor proportional to $\rho_A \rho_B r_c^3$ for G_2 functions between atoms A and B and $\rho_A \rho_B \rho_C r_c^6$ for G_3 functions between atoms A, B and C, to reflect the cost associated with evaluating them. We note that these importance scores do not enter the functional form of the SFs finally used in the fit.

$N'_{\mathrm{O}}, N'_{\mathrm{H}}$	$\epsilon_{\mathrm{O}}, \epsilon_{\mathrm{H}}$ $\times 10^{-4}$	RMSE(E) [meV/at.]	RMSE(f) [eV/Å]	Runtime [s/step]
CUR selection				
16,16	51,63	1.55	0.147	0.35
32,32	2.5,6.2	1.18	0.126	0.43
64,64	0.1,0.3	0.99	0.114	0.52
$\mathrm{CUR}_{k=1}$ selection				
16,16	51,63	1.49	0.145	0.35
32,32	2.6,7.6	1.23	0.123	0.42
64,64	0.1,0.3	1.02	0.113	0.52
FPS selection				
16,16	56,132	3.89	0.251	0.34
32,32	7.1,12	1.62	0.150	0.40
64,64	0.3,0.9	1.19	0.128	0.51
Default SF set				
36,36		1.62	0.238	0.85
SFs of Ref. 19				
30,27	-	0.98	0.115	0.69

Table 2.2 – The table reports, for different numbers of SF selected from a pool of 768 candidates using different strategies, the error in the approximation of the feature matrix, and the RMSE for energies and forces from a test set, averaged over four NNs trained starting from different random weights. The spread between results of the 4 independent training runs for each choice of SF is of the order of 2-4%. Results from the SF used in Ref. 19 and of a "default set" are also shown for comparison.

For assessing the performance of the optimized SFs, we selected SF sets containing $N' = 16$, 32, and 64 symmetry functions for each element using a CUR and FPS procedure. Additionally, for comparison we include a "default" symmetry function set in our benchmark, which is used for first preliminary potentials. For a binary system like water this default set contains 6 G_2 functions for each element pair with parameters η chosen such that the turning points of the terms in the summation in Eq. 2.6 are equidistantly arranged between the minimum interatomic distance and the function with maximum spatial extension ($\eta = 0$). r_s is set to zero and $r_c = 12$ bohr. For the angular functions G_3 (Eq. 2.7) we use for each possible

element combination the parameter sets $\zeta = 1, 2, 4, 16$ along with $\lambda = \pm 1$, $\eta = 0$ and $r_c = 12$ bohr. Therefore, the default set contains each 36 SFs for the oxygen and hydrogen atoms. Finally, also the SFs of Ref. [19] have been tested with our reference data set.

For each set of symmetry functions we train 4 NN potentials based on atomic NNs with two hidden layers and 20 neurons per hidden layer using the RuNNer code [22], with random initial weights and a 3:1 random split of train:test points using the same 1000 FPS subset. Table 2.2 reports the average test error for energy and forces obtained using the CUR and FPS SFs sets as well as of the "default" set and the SF set of Ref. [19]. The table also shows the CUR approximation errors for O and H fingerprints for each number of symmetry functions, and the execution time per MD step for a simulation with 216 water molecules ran using the LAMMPS RuNNer plugin [19, 97] on a single Intel Xeon 2.60GHz core.

All the different strategies to automatically select fingerprints show that it is possible to progressively improve the test set accuracy by making the selection more inclusive. CUR gives by far the best performance, both in terms of error in approximating $\mathbf{X}$ and in terms of the energy and force test RMSE, followed by FPS. All the automatic selection protocols perform better than the "default" SF set, dramatically so in the case of CUR.

However, manual optimization of symmetry functions, taking into account the physical parameters of the system, and the actual accuracy of the training, seems to provide an advantage. The selection from Ref. [19] achieves with only 57 SF the same accuracy as a CUR selection of 128. The automatic selection, however, requires a lower computational effort, since the estimated cost of evaluating a SF is taken into account when generating the selection. It would be possible to further improve the performance of the automatic selection by considering also the correlations between the SF values and the target property, such as energies or forces – so as to select the descriptors that are not only structurally uncorrelated, but also strongly coupled to the stability of the system.

2.5.2 A Potential for Aluminum Alloys

Water is a two-component system, but its molecular nature means that the number of possible environments is affected less dramatically by the number of species. A NN potential for Al-Si-Mg alloys has been recently demonstrated [26], that instead deals with a ternary system, where all of the interactions among the different species and defects must be accounted for to obtain accurate predictions across the full range of relevant compositions. The presence of multiple interactions at different length scales makes the manual selection of SF a particularly cumbersome task. In the previous work [26], the problem was circumvented by restricting the SF pool to the 2-body G_2 components, making it possible to obtain a systematic - if not optimal - selection. The automatic selection procedure makes it much easier to automatically determine an efficient feature set that includes both G_2 and G_3 SFs, which makes it possible to take into account the angular dependence of the atomic interactions explicitly.

The reference data set we use as a starting point is composed of the 10551 structures used by Kobayashi et al. [26], supplemented by 609 structures of β''-phase precipitates and interfaces that have been generated in a previous DFT study of the alloy [98]. Given that many of the resulting 11160 structures are taken from short MD runs and are highly correlated, we selected 2000 structures with FPS, that have been used both for the selection of SF and the training/testing procedure. This sparser selection leads to a larger absolute magnitude of the fit error, but does not affect the quality of the fit, while making the optimization procedure faster and more stable.

The initial generation of SF is done similarly to the case of water. Six sets of G_2 SF have been generated using $N = 4, 12$ and $r_c = 8, 16, 20$ bohr, and two sets of G_3 SF have been generated using $N = 8$ - one with $r_c = 8$ bohr and $\zeta = 1, 2, 4, 8, 16$ and the other with $r_c = 12$ bohr and $\zeta = 1, 2, 4$. Duplicate SF have been eliminated, together with those that had a width smaller than 1.06 Å for the radial ones and smaller than 1.32 Å for the angular ones. The same weighting described for water has been used here when selecting the SF. The details of the fingerprints can be found in the S.I.of the original paper[57], and the performance of the resulting NN potentials can be seen in Table 2.3. The test set RMSE decreases systematically as the number of selected SF increases, up to 64 SFs per species. We also compare the results with those obtained with the SF selection from Ref. 26; we verify that the accuracy of the re-trained NN for the properties we test here is comparable of better than that of the original potential. To ensure a fair comparison we re-optimize and test the potential using the RuNNer [22] software and the same FPS selection we discuss above. Already at $N' = 96$ (32 SFs per species) the automatic selection that includes 3-body SFs leads to a better test set error than the systematic selection of 120 G_2 SFs.

$N'_{\mathrm{Al}}, N'_{\mathrm{Mg}}, N'_{\mathrm{Si}}$	$\epsilon_{\mathrm{Al}}, \epsilon_{\mathrm{Mg}}, \epsilon_{\mathrm{Si}}$ $\times 10^4$	RMSE(E) [meV/at.]	RMSE(f) [eV/Å]
CUR selection			
16,16,16	79,99,101	16.22	0.084
32,32,32	7.9,14,10	4.08	0.052
64,64,64	0.9,1.3,0.8	2.47	0.022
SFs of Ref. 26			
40,40,40	-	9.2	0.069

Table 2.3 – The table reports, for different numbers of SF, the error in the approximation of the feature matrix, and the RMSE for energies and forces from a test set. Results from the SF used in Ref. [26] are also shown for comparison.

While the test set RMSE is a good measure of the quality of a potential, it is important to also verify the stability of the NN when computing a property for which configurations had not been explicitly included in the training set. As an example of the behaviour of the different potentials, that is very relevant for the potential application of this NN in the description of the early stages of precipitation in Al-6xxx alloys [98], we compute the configuration energy along the minimum energy pathways for the vacancy-assisted migration of Al, Si, Mg atoms in

a matrix of 256 Al atoms. Atomic configurations along the pathway between the minimum energy states are obtained by linear interpolation, and by local optimization using the nudged elastic band (NEB) method [99] with the climbing image algorithm [100] as implemented in QUANTUM ESPRESSO [101]. The details of the DFT calculations are the same as described in Refs. [26, 98]. 7 images have been used for Mg and 13 have been used for Al and Si, and lead to relaxed vacancy migration barriers that are consistent with previous DFT calculations [102]. Keeping the configurations fixed, we compute the energy along the migration barrier for both the linear transition path between the initial and final configurations and the corresponding relaxed positions.

As shown in Figure 2.5, there is a considerable improvement in the quality of the fit when going from 16 to 32 SFs per species, whereas the improvement is less dramatic when using a larger number of SFs, and actually in the case of the vacancy-assisted diffusion of Si the 64-SF NN performs worse than the 32-SF NN. This observation underscores the fact that refining the SF selection does systematically improve the accuracy in the interpolative regime, as probed by cross-validation, but not necessarily to a systematic improvement in the extrapolative regime. For all of the vacancy-assisted diffusion processes we consider, however, NN potentials reproduce the correct qualitative behaviour. Excluding the case with 16 SF per element, which is clearly insufficient for this system, the error in the relaxed barrier is below 0.1 eV, which is comparable to the typical DFT error. Automatic SF selections that include 3-body terms perform better than the G_2-only choice of Ref. 26, that nevertheless predicts diffusion barriers with a remarkably small error.

2.5.3 Learning Molecular Energies

To provide a very different example of the application of dimensionality reduction strategies to sparsify the feature matrix, we turn to the case of SOAP fingerprints, and to the GPR of atomization energies for a molecular data set composed of 7211 small organic molecules, containing up to 7 heavy atoms (N, C, O, Cl, S) [103]. As we discussed in Section 2.2, the SOAP framework provides a very systematic method to describe a chemical environment, but can easily lead to thousands of descriptors. In this case, which involves 6 chemical species, and for which we used an environment cutoff of $r_c = 3.0$ Å, $n_{\max} = 9$ and $l_{\max} = 9$, one has to deal with a total of $N = 14852$ rotationally-invariant fingerprints. This huge number of features is in stark contrast with the handful of symmetry functions that are used in the NN scheme to generate accurate interatomic potentials. It is reasonable to speculate that a small fraction of the initial features could also provide a satisfactory description of the chemical environments, and therefore an accurate prediction of properties. To test this idea, we apply the same framework we discussed for the BP symmetry functions to the power spectrum $\mathbf{p}_{\alpha\beta}(A)$.

There is however an important difference compared to the previous case. In the case of the NN, the feature vectors are subject to a linear transformation before being fed to the first

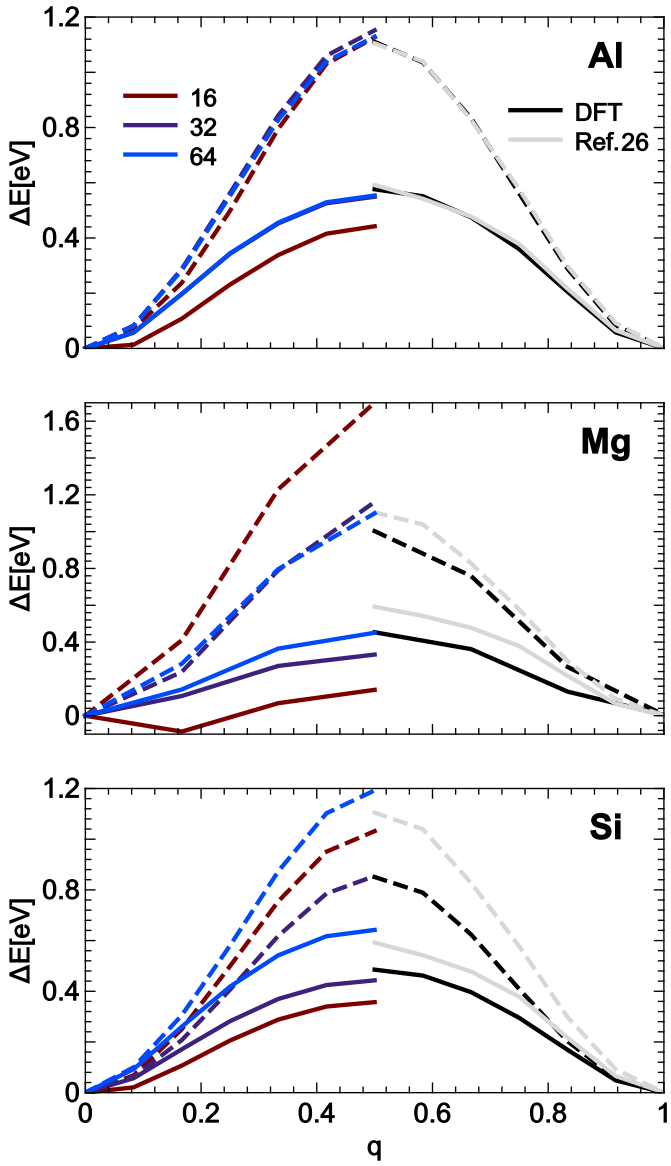

Figure 2.5 – The energy barrier for the vacancy-assisted migration of Al, Mg, and Si using an increasing number of symmetry functions are presented on the left, compared to DFT and the choice of SF from Ref. 26, presented on the right. Dashed lines correspond to the unrelaxed configurations, solid lines to the minimum energy pathway. The energies are shown as a difference from the minimum energy structure.

layer of non-linear activation functions. SOAP fingerprints are typically used to compute a kernel for Gaussian process regression, that in its simplest form corresponds to the scalar product between features, without an optimization step to determine the most effective linear combination of the inputs. For this reason, in order to reduce the size of the input vectors without compromising the regression accuracy, it is necessary to introduce an additional ingredient. The original kernel is calculated as $\mathbf{K} = \mathbf{X}\mathbf{X}^T$, whereas now we intend to compute it using the approximate form of $\mathbf{X}$, i.e. we intend to find $\tilde{\mathbf{K}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$, where $\tilde{\mathbf{X}}$ is shown in eq 2.12.

As explained in section 2.2.1, given that we only aim to reduce the number of features, $\mathbf{UR} = \mathbf{C}^+\mathbf{X}$. The approximate kernel can then be written as

$$\tilde{\mathbf{K}} = \mathbf{C}\mathbf{C}^+\mathbf{X}\mathbf{X}^T(\mathbf{C}^+)^T\mathbf{C}^T. \tag{2.22}$$

Computing the approximate kernel also involves the $N' \times N'$ matrix $\mathbf{W} = \mathbf{C}^+\mathbf{X}\mathbf{X}^T(\mathbf{C}^+)^T$. Since this matrix is symmetric and positive-definite, it can be decomposed as $\mathbf{W} = \mathbf{A}\mathbf{A}^T$. Finally, we see that the kernel can be written in terms of scalar products of the reduced-dimensionality features, provided we define $\mathbf{X}' = \mathbf{C}\mathbf{A}$, since

$$\tilde{\mathbf{K}} = (\mathbf{C}\mathbf{A})(\mathbf{C}\mathbf{A})^T \quad . \tag{2.23}$$

Therefore, after using the previously described schemes to select features from $\mathbf{X}$, we also have to compute the $\mathbf{A}$ matrix in order to scale adequately the selected features. It should be noted that the matrix $\mathbf{A}$, although computed only once during the fingerprint selection stage, must be stored and applied to the selected components of the power spectrum when performing training or predictions.

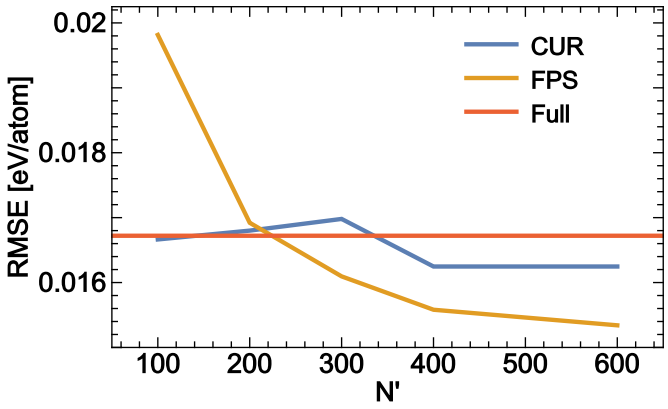

Figure 2.6 – The RMSE of the GPR for 1442 randomly chosen structures in the test set, with a varying number of elements of the power spectrum, chosen for both CUR and FPS, compared to the result of the GPR with the full power spectrum. The training set is composed by 500 FPS structures.

Let us now turn to discuss the performance of different feature selection strategies for the prediction of the atomization energies on the QM7b data set. All the results we present are

tested using the same set, composed of 1442 randomly selected structures (which correspond to roughly 20% of the full QM7b data). From the overall training set, containing 5769 structures, we select 500 structures with a FPS strategy that we use to construct the initial feature matrix. We then apply both the CUR and the FPS methods to perform feature selection, and use the reduced dimensionality set of descriptors to train a GPR model on the same 500 FPS structures. Figure 2.6 shows the RMSE in the prediction of the atomization energies of the test-set structures. It is remarkable to see that using only 100 CUR-selected elements of the power spectrum it is possible to match the prediction accuracy obtained with the original kernel based on more than 14,000 features. Interestingly, increasing the number of features to 400 leads to *lower* test error, suggesting that for this small training set the use of a smaller set of fingerprints helps to combat overfitting. FPS selection also performs remarkably well, and at $N' = 400$ it yields a test RMSE which is 5% lower than the baseline SOAP result.

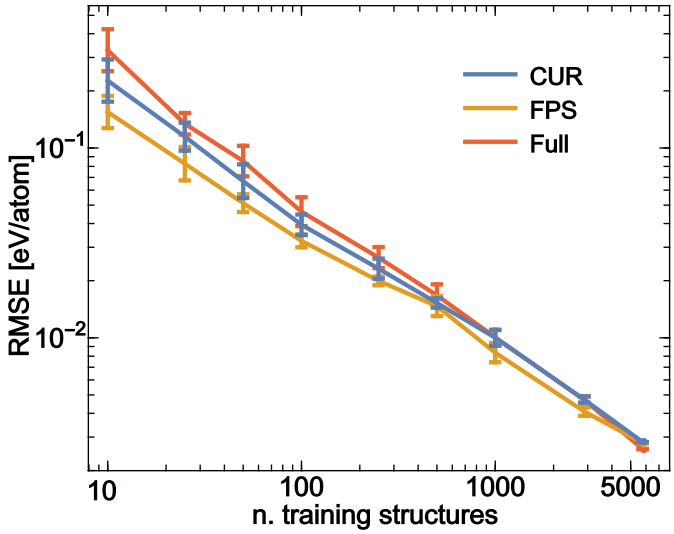

Figure 2.7 – Learning curves for the QM7b atomization energies [103], when using the full SOAP power spectrum, 400 features selected with FPS, and 400 selected with CUR. The results shown for each training set size are the average and standard deviation from 10 different models trained on random selections extracted from the overall training set

The question is of course whether this reduced-dimensionality description is sufficient to further improve the prediction accuracy, when more structures are used for training. As seen in the learning curves in Fig. 2.7, using 400 features is enough to obtain errors that are comparable to the reference value, or even lower. It is only when considering the full 5769 structures in the training set that the baseline kernel reaches a marginally better accuracy than the reduced-dimensionality model, that discards as much as 97% of the elements of the SOAP power spectrum.

2.5.4 Selecting configurations for the training set

Finally, we explore the advantages of using the feature selection methods to select training points, as explained in Sec.2.3.3. We go back to the example of the water dimers, where we saw that the non-uniform sampling of the dimer space configuration led to higher errors in the less explored regions.

Figure 2.8 compares the test RMSE obtained by NN fits constructed on subsets of the overall training set. We notice that the error can be reduced by up to a factor of five by choosing the subset with a FPS strategy, rather than at random. This observation is consistent with recent observations made using SOAP-GAP in a variety of other systems [85, 104].

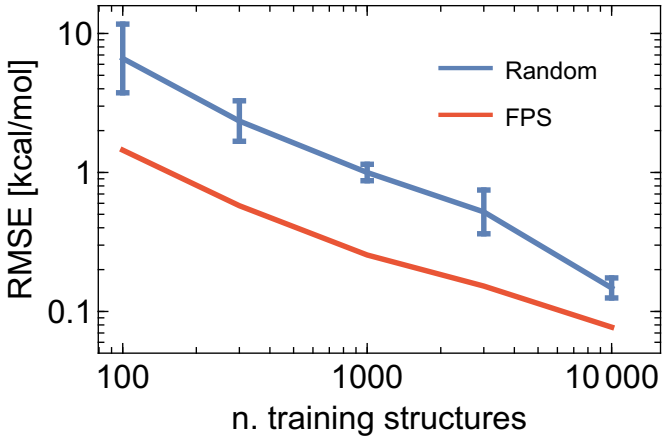

Figure 2.8 – TEST RMSE as a function of the size of the training set for the 2B energy contribution, using a NN for the regression. Training configurations are selected at random (5 independent selections, average and standard deviation shown) or by farthest point sampling.

Therefore, selecting training configurations from a larger database of potential candidates using FPS is a viable strategy to reduce the number of high-end calculations that have to be performed when building the dataset. We will see how this can effectively be applied on a real MLIP in Sec.4.2.2.

3 Uncertainty estimation for molecular dynamics[1]

3.1 Introduction

As the use of ML models to compute atomic-scale properties becomes more common, we naturally end up questioning of how much we can trust the predictions of a purely inductive, data-driven approach when using it on systems that are not part of the training set. The regression techniques that underlie ML models are inherently interpolative, and their ability to make predictions on new systems relies on the possibility of decomposing the target property into a sum of atom-centred contributions. Thus, a ML prediction is only reliable if all the local environments that appear in the system of interest are properly represented in the training set.

In the recent years, many methodological frameworks have been proposed that yield a measure of the uncertainty in the prediction of a machine learning model.[106] Within Bayesian schemes, such as Gaussian process regression, the uncertainty quantification is naturally encoded in the regression algorithm – although computing the error is substantially more demanding than evaluating the prediction.[107] Sub-sampling approaches constitute an alternative. The uncertainty is estimated on the basis of the spread of the predictions of an ensemble (committee) of independently trained ML models, which yields a qualitative information of the reliability of the ML predictions, which can later be used for online or offline addition of new training points [10, 68, 108–115]. On the other hand, a quantitative measure of the uncertainty can be obtained by appropriate rescaling of the committee results[116], which can be readily propagated to estimate the error in properties that are obtained indirectly from the ML predictions such as vibrational spectra [45].

In this chapter we consider how to best exploit the availability of machine-learning models that include an error estimation in the context of molecular dynamics simulations. First, we construct a *weighted baseline* ML scheme, in which the uncertainty is used to ensure that whenever the simulation enters an extrapolative regime, the potential falls back to a reliable

[1] The majority of the chapter has been extracted from Ref. 105. The author has contributed in the theoretical development, the writing, and the validation of the work, by applying the method to the majority of the examples provided, with the exception of the Phe-Gly-Phe tripeptide and methanesulphonic simulations.

(if not very accurate) baseline. Second, we use errors computed for individual configurations to estimate the ML uncertainty associated with static *thermodynamic averages* from MD trajectories computed using a single potential.

3.2 Theory

We consider a machine-learning model that can predict, for a structure A, the value of a property $y(A)$ as well as its uncertainty $\sigma^2(A)$. We focus our derivations on committee models, that are easy to implement and allow for straightforward error propagation. However, most of the results we derive can be applied to any scheme that provide a differentiable uncertainty estimate for each property prediction.

3.2.1 Committee model and single-point uncertainty estimation

Here we summarize the uncertainty estimation, while the full discussion can be found in the original paper[116]. In a nutshell, the full training set of N input-observation pairs $(A, y_{\mathrm{ref}}(A))$ is sub-sampled (without replacement) into M training subsets of size $N_s < N$. M models are then trained independently on this ensemble of resampled data sets, inducing a fully non-parametric estimate of the distribution $P(y|A)$ of the prediction y, given an input A. The moments of such distribution can be readily computed, so that, for instance, the first (mean value) and second (variance) moments are

$$\bar{y}(A) = \frac{1}{M}\sum_{i=1}^{M} y^{(i)}(A) \tag{3.1}$$

$$\sigma^2(A) = \frac{1}{M-1}\sum_{i=1}^{M} \left| y^{(i)}(A) - \bar{y}(A) \right|^2 . \tag{3.2}$$

Here, $y^{(i)}(A)$ is the prediction of the $i-$th model, while the mean value $\bar{y}(A)$ will be dubbed in the following as the *committee* prediction. The advantage of this machinery is that the ensemble $\{y^{(i)}(A)\}_{i=1,\dots,M}$ of model predictions provides an immediate estimate of the single-point uncertainty $\sigma^2(A)$, since it fully characterises the error statistics.

The reduced size N_s of the set of input-observation pairs on which the sub-sampled models are trained implies that the conditional probability distribution $P(y_{\mathrm{ref}}(A)|A)$ may deviate from the ideal Gaussian behaviour. We assume that such deviation only affects the width of the distribution, which may be too broad or (usually) too narrow, an effect that can also be seen as a consequence of the fact that training points cannot be considered to be independent identically distributed samples. We incorporate this deviation through a linear re-scaling factor α of the width σ of the distribution. We further assume that α is independent of A, and that any two true values $y_{\mathrm{ref}}(A)$ and $y_{\mathrm{ref}}(A')$ are uncorrelated if $A \neq A'$, so that the predictive

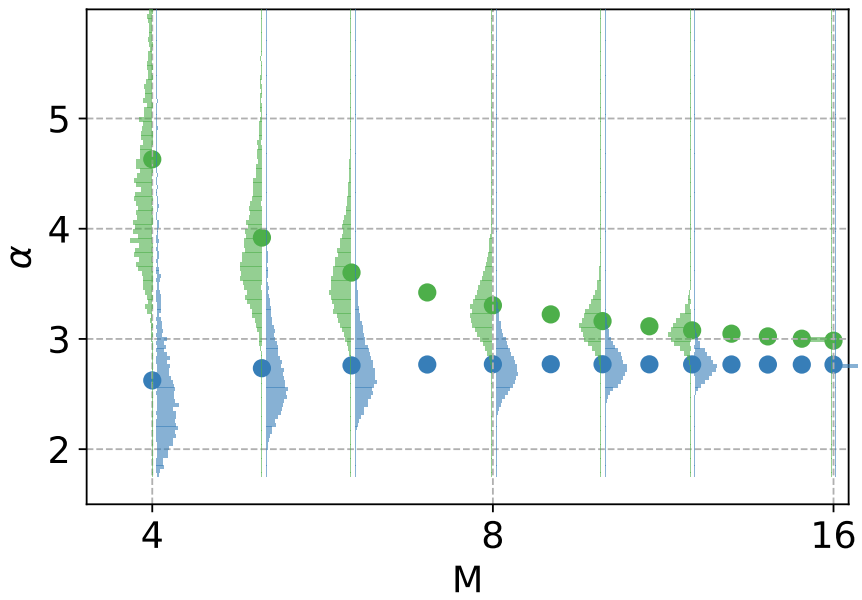

Figure 3.1 – Violin plot of biased (green) and unbiased (blue) estimators for the correction factor α, as a function of the number of models in the committee, M.

distribution has the following form:

$$P(\mathbf{y}_{\text{ref}}|\{A\},\alpha) = \prod_A \frac{1}{\sqrt{2\pi\alpha^2\sigma^2(A)}} \exp\left[-\frac{\left|y_{\text{ref}}(A)-\bar{y}(A)\right|^2}{2\alpha^2\sigma^2(A)}\right] \tag{3.3}$$

The parameter α is then fixed by maximizing the log-likelihood of this distribution,

$$LL(\alpha) = \frac{1}{N_{\text{val}}} \sum_{A\in\text{val}} \log P(y_{\text{ref}}(A)|A,\alpha) \tag{3.4}$$

over a set of N_{val} validation configurations, giving the optimal

$$\alpha^2 \equiv \frac{1}{N_{\text{val}}} \sum_{A\in\text{val}} \frac{\left|y_{\text{ref}}(A)-\bar{y}(A)\right|^2}{\sigma^2(A)}. \tag{3.5}$$

In practice, the explicit construction of a validation set can be avoided by means of a scheme where the validation points still belong to the training set, yet they are absent from a given number of sub-sampled models, as discussed in depth in Ref. 116.

Unfortunately, Eq. (3.5) is a biased estimator when the number of committee members M is small, as it can be seen in Fig. 3.1, where the biased estimator is shown in green. In the paper from which this chapter is adapted[105], we discuss the issue in more detail, and show that

the bias can be corrected by computing

$$\alpha^2 \equiv -\frac{1}{M} + \frac{M-3}{M-1}\frac{1}{N_{\text{val}}}\sum_{A\in\text{val}}\frac{\left|y_{\text{ref}}(A)-\bar{y}(A)\right|^2}{\sigma^2(A)} \tag{3.6}$$

which leads to an unbiased estimator (blue) shown in Fig. 3.1. As it is clear from the $M-3$ term at numerator, this method can be applied only when we use at least 4 potentials for our committee.

The determination of the optimal α also allows us to properly re-scale the predictions of the models to be consistent with Eqs. (3.1) and (3.2) and the optimized distribution:

$$y^{(i)}(A) \leftarrow \bar{y}(A) + \alpha[y^{(i)}(A) - \bar{y}(A)]. \tag{3.7}$$

The committee prediction $\bar{y}$ is invariant under rescaling, and the spread of the predictions is adjusted according to $\sigma \leftarrow \alpha\sigma$. The rescaled predictions can be used to compute arbitrarily-complicated non-linear functions of y, and the mean and spread of the transformed predictions are indicative of the distribution of the target quantities. In what follows, we always assume that the committee predictions have been subject to this calibration procedure.

3.2.2 Using errors for robust sampling and active learning

Let us consider the following *baselined model*

$$V^{(i)}(A) = V_b(A) + V_\delta^{(i)}(A) \tag{3.8}$$

where the training of the $i-$th model potential $V_\delta^{(i)}$ is on the (set of) differences between a target, say DFT-accurate, potential $\{V_{\text{ref}}(A)\}$ and a baseline potential $\{V_b(A)\}$. Splitting a potential in a cheap-to-compute but inaccurate, and an accurate-but-expensive parts has been part of the molecular dynamics toolkit for a long time [92, 117, 118], and has proven very effective in the context of machine-learning models [85, 119]. Let us define the full committee potential

$$\bar{V}(A) = V_b(A) + \bar{V}_\delta(A), \tag{3.9}$$

the committee average of the correction potentials

$$\bar{V}_\delta(A) = \frac{1}{M}\sum_{i=1}^{M} V_\delta^{(i)}(A), \tag{3.10}$$

and its uncertainty

$$\sigma^2(A) = \frac{1}{M-1}\sum_{i=1}^{M}\left|V_\delta^{(i)} - \bar{V}_\delta(A)\right|^2, \tag{3.11}$$

as in Eqs. (3.1) and (3.2). This uncertainty estimate, as well as any other similarly accurate and differentiable measure of the error, can be used as an indication of the reliability of the ML predictions, and incorporated in an active-learning framework [114, 120–122]: during a molecular dynamics simulation, whenever the trajectory enters a region in which the model exhibits an extrapolative behaviour, the uncertainty σ increases, and one can gather new configurations for an improved model [113]. Unfortunately, trajectories entering an extrapolative region often become unstable very quickly, leading to sampling of unphysical configurations or the complete failure of the simulation. Crucially, when using a baseline potential, one can stabilize the simulation by dynamically switching to using only V_b. This automatic fall-back mechanism can be realized by performing MD using the weighted-baseline potential

$$U(A) = \left[\frac{1}{\sigma_b^2} + \frac{1}{\sigma^2(A)}\right]^{-1} \left[\frac{1}{\sigma_b^2} V_b(A) + \frac{1}{\sigma^2(A)} \bar{V}(A)\right] = V_b(A) + \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2(A)} \bar{V}_\delta(A), \tag{3.12}$$

where the baseline uncertainty σ_b is estimated as the variance of the difference between baseline and reference

$$\sigma_b^2 \equiv \frac{1}{N-1}\left[\sum_A |V_b(A) - V_{\mathrm{ref}}(A)|^2 \; - \frac{1}{N}\left(\sum_A V_b(A) - V_{\mathrm{ref}}(A)\right)^2\right], \tag{3.13}$$

the sum running on the full training set, and $V_{\mathrm{ref}}(A)$ being the target energy for configuration A. This definition explicitly takes into account the fact that the baseline and reference often differ by a huge constant. Eq. (3.12) corresponds to the weighted sum of the baseline potential $V_b(A)$ and the full committee potential $\bar{V}(A)$, consistent with a minimization of the combined error. The forces (and higher derivatives) can be defined straightforwardly, paying attention to the A-dependence of $\sigma^2(A)$ when the derivatives of $U(A)$ are taken. Note also that in many cases – including Behler-Parrinello neural networks [15] and SOAP-GAP models [24] – the ML energy is computed as a sum of atom-centred contributions

$$\bar{V}_\delta(A) = \sum_{k \in A} \bar{V}_\delta(A_k), \tag{3.14}$$

where A_k indicates the environment centred on the k-th atom in structure A. Thus, it is possible to compute uncertainty estimates at the level of individual atomic contributions, and evaluate Eq. (3.12) as

$$U(A) = V_b(A) + \sum_{k \in A} \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2(A_k)} \bar{V}_\delta(A_k). \tag{3.15}$$

This expression can be used even if the baseline does not entail a natural atom-centred decomposition, although in such a case one needs to re-define σ_b so that it corresponds to the estimated error *per atom*. This can be beneficial when the error is not spread equally across the system, e.g. when an unexpected chemical reaction occurs in an otherwise homogeneous system.

By monitoring the weight of the ML correction one can determine whether the simulation remains largely in the low-uncertainty region, or whether it enters the extrapolative regime too frequently, requiring further training. Finally, it is worth mentioning that a similar strategy could be used to combine multiple ML potentials with different levels of accuracy, for instance one based on short-range/two-body interactions, that is more resilient but inaccurate, and one based on a long-range and high-body-order parameterization, which is likely to be more accurate, but requires large amounts of data for training, and is therefore more likely to enter high-uncertainty regions.

3.2.3 On-the-fly uncertainty of thermodynamic averages

The machinery discussed so far paves the way for reliable estimates of the uncertainty of single-point calculations, i.e. of the value an observable quantity assumes when evaluated at a specific point in phase-space. It also allows computing the uncertainty of predictions averaged over several samples, assuming that the only source of error is that associated with the ML model of the target property [123]. However, the uncertainty in predictions also propagates to thermodynamic averages of target properties. Estimating how such uncertainty propagates is particularly straightforward in the case of a committee-based estimate. Computing the mean of an observable a over a trajectory sampling e.g. the mean potential $\bar{V}$ from a committee of M potential models (PMs) $V^{(i)}$ yields

$$\bar{\bar{a}} \equiv \langle \bar{a} \rangle_{\bar{V}} = \frac{1}{M'} \sum_{j=1}^{M'} \langle a^{(j)} \rangle_{\bar{V}}, \tag{3.16}$$

where $a^{(j)}$ indicates the member of a committee of M' observable models (OMs), and $\langle a \rangle_V$ the mean of an observable over the ensemble defined by the potential V.

When computing thermodynamic averages, one should therefore also include the uncertainty in the ensemble of configurations. A naïve (but very time-consuming) way to estimate the full uncertainty relies on running M simulations, each driven by the (re-scaled) force field of a specific PM, and computing the averages $\langle a^{(j)} \rangle_{V^{(i)}}$ of the target observable $a^{(j)}$ for each OM, and finally the average

$$\tilde{a} \equiv \frac{1}{MM'} \sum_{i=1}^{M} \sum_{j=1}^{M'} \langle a^{(j)} \rangle_{V^{(i)}} \tag{3.17}$$

and variance over both OMs and PMs. While trivially parallelizable, this strategy is inconvenient, as it prevents exploiting the considerable computational savings that can be achieved by computing multiple committee members over the same atomic configuration.

The need for different trajectories can be avoided by employing an on-the-fly re-weighting

strategy [124]. For a canonical distribution at temperature $T = 1/(\beta k_B)$,

$$\langle a^{(j)} \rangle_{V^{(i)}} \equiv \frac{1}{Z^{(i)}} \int a^{(j)}(\mathbf{q}) e^{-\beta V^{(i)}(\mathbf{q})} d\mathbf{q}, \tag{3.18}$$

where $\mathbf{q} = (\mathbf{q}_1, \ldots, \mathbf{q}_{N_p})$ is the set of positions of the N_p particles,

$$Z^{(i)} \equiv \int e^{-\beta V^{(i)}(\mathbf{q})} d\mathbf{q} \tag{3.19}$$

is the configurational partition function and $V^{(i)}(\mathbf{q})$ is the potential energy of the $i-$th model. By introducing the *weights*

$$w^{(i)}(\mathbf{q}) \equiv e^{-\beta[V^{(i)}(\mathbf{q}) - \bar{V}(\mathbf{q})]}, \tag{3.20}$$

where $\bar{V}$ is the mean committee potential energy, we find

$$\langle a^{(j)} \rangle_{V^{(i)}} = \frac{\int w^{(i)}(\mathbf{q}) a^{(j)}(\mathbf{q}) e^{-\beta \bar{V}(\mathbf{q})} d\mathbf{q}}{\int w^{(i)}(\mathbf{q}) e^{-\beta \bar{V}(\mathbf{q})} d\mathbf{q}} \tag{3.21}$$

or, in shorthand notation,

$$\langle a^{(j)} \rangle_{V^{(i)}} = \frac{\left\langle w^{(i)} a^{(j)} \right\rangle_{\bar{V}}}{\left\langle w^{(i)} \right\rangle_{\bar{V}}}. \tag{3.22}$$

This means that, under the ergodic hypothesis, the re-weighting technique allows us to run *a single trajectory* driven by the force field of the committee, and yet to obtain estimates for the averages as computed via the different models. Thus, it is possible to compute the full uncertainty, including both the error on the OMs and the PMs, by using the reweighting formula to evaluate

$$\tilde{\sigma}^2 \equiv \frac{1}{MM' - 1} \sum_{i=1}^{M} \sum_{j=1}^{M'} \left| \langle a^{(j)} \rangle_{V^{(i)}} - \tilde{a} \right|^2 \tag{3.23}$$

This reweighing approach has further important implications to molecular dynamics simulations: for instance, in on-the-fly learning it is customary to correct (re-train) the ML force-field from time to time along a molecular dynamics simulation so to include new configurations in the training set:[120, 125] an operation which can introduce systematic errors on the estimation of canonical averages, due to the different potential-energy fields along the trajectory. By simply storing the model-dependent potential energies along the simulation alongside the corresponding configurations, one can at any time compute a set of weights based on the most recent value of the potential, to obtain averages that use the entire trajectory and yet are consistent with the most accurate model available.

Equation (3.22) is in principle exact. However, from a computational standpoint, the efficiency in sampling the probability measure of the $i-$th model through reweighing is in general lower

than what it would be by direct sampling as in Eq. (3.18), with an error growing exponentially with the variance of $h^{(i)} \equiv -\ln w^{(i)} = \beta(V^{(i)} - \bar{V})$, that inevitably increases with system size. Given that we are only interested in computing an estimate of the uncertainty, we can use an approximate (but statistically more stable) expression introduced in Ref. 126, based on a cumulant expansion. Assuming that $a^{(j)}$ and $h^{(i)}$ are correlated Gaussian variates (all with respect to the committee phase-space probability measure), we have

$$\langle a^{(j)} \rangle_{V^{(i)}} \approx \langle a^{(j)} \rangle_{\bar{V}} - \beta[\langle a^{(j)}(V^{(i)} - \bar{V}) \rangle_{\bar{V}} - \langle a^{(j)} \rangle_{\bar{V}} \langle V^{(i)} - \bar{V} \rangle_{\bar{V}}]. \tag{3.24}$$

In order to compare the different definitions given so far for a physical example, we consider a simple thermodynamic average, i.e. the radial pair correlation function $g(r)$ between H atoms in water. We refer to Sec. 3.3 for the specific details of the simulation. The top panel in Fig. 3.2 displays $\bar{\bar{g}}(r)$ determined, as in Eq. (3.16), by averaging over a significant number of atomic configurations sampled from a trajectory driven by a committee of $M = 4$ models (neural network potentials, NNPs). The middle panel displays the differences $\Delta g^{(i)}(r) = g^{(i)}(r) - \bar{\bar{g}}(r)$, with $g^{(i)}(r)$ obtained after sampling structures from separate trajectories driven by each NNP model. In the bottom panel, we focus on one of the models, and we compare the deviation of the pair distribution function, with respect to $\bar{\bar{g}}(r)$, computed according to: an independent trajectory driven by NNP 3 (orange, same as in the central panel); the direct re-weighting of the sampling from the trajectory driven by the committee as in Eq. (3.22) (purple); and within the cumulant expansion approximation (CEA), Eq. (3.24) (dark green). The match between the three curves shows that the re-weighting procedure, both in its exact form and using the CEA, is capable of reproducing the result obtained from an independent trajectory generated by a specific NNP without the need of explicitly running it.

For this example, which entails a relatively small simulation cell and low discrepancy between the committee average and the individual NNPs, there is no substantial difference between the exact and CEA reweighing. We recommend using the CEA over the direct estimator, not only because of its improved stability and statistical efficiency, but also because the linearized form emphasizes the different sources of error associated with the single-trajectory average (3.16), and has several desirable formal implications. First, using the CEA the mean over the trajectories is consistent with the average computed over the trajectory driven by $\bar{V}$ – whereas in general Eq. (3.17) would yield a different value from (3.16):

$$\tilde{a} \approx \bar{\bar{a}} + \frac{\beta}{M} \sum_i [\langle \bar{a}(V^{(i)} - \bar{V}) \rangle_{\bar{V}} - \langle \bar{a} \rangle_{\bar{V}} \langle V^{(i)} - \bar{V} \rangle_{\bar{V}}] = \bar{\bar{a}}. \tag{3.25}$$

Second, one sees that

$$\tilde{\sigma}^2 \approx \frac{M(M'-1)}{MM'-1} \sigma_a^2 + \frac{M'(M-1)}{MM'-1} \sigma_{aV}^2 \underset{M,M' \to \infty}{=} \sigma_a^2 + \sigma_{aV}^2 \tag{3.26}$$

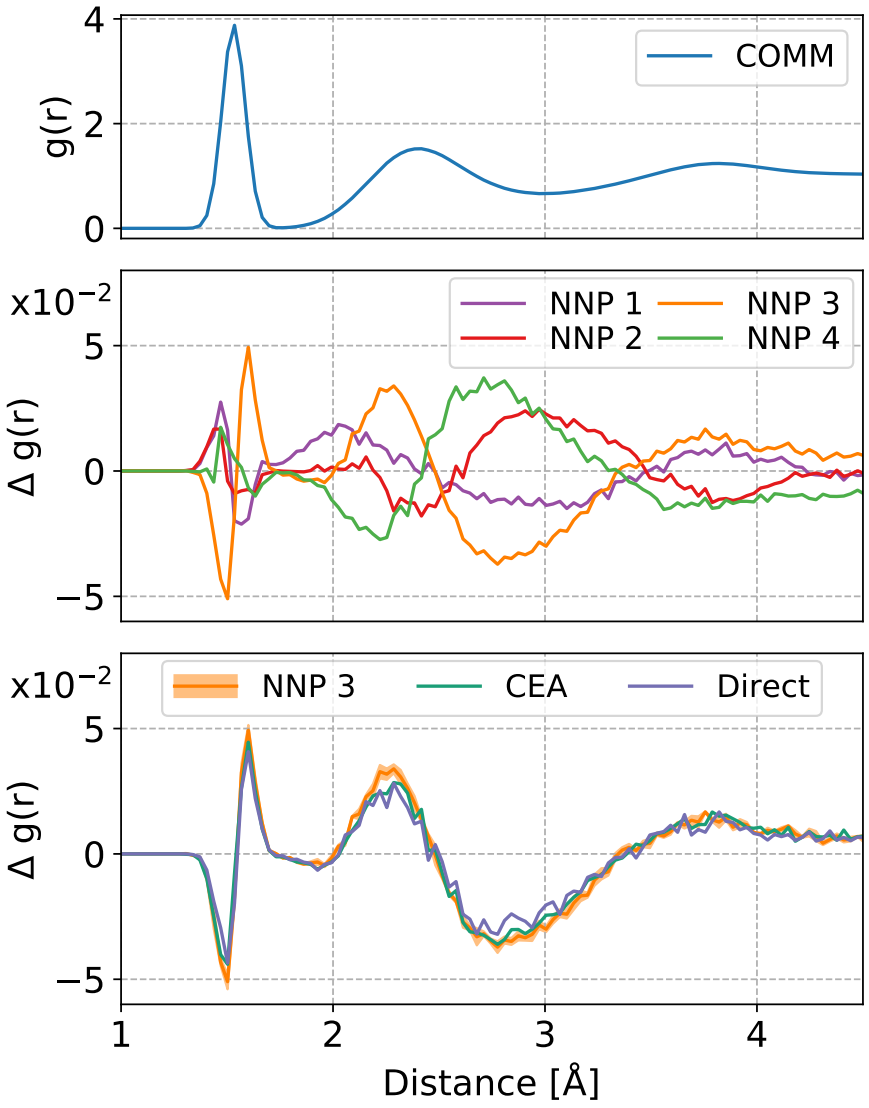

Figure 3.2 – Hydrogen-hydrogen radial pair correlation function in water. (Top) pair distribution function computed for a simulation driven by the committee average; (middle) deviations, from the plot in the top panel, of the pair distribution functions extracted from $M = 4$ independent trajectories (one for each NNP, displayed in different colours); (bottom) comparison between the result from an independent trajectory driven by NNP 3 (orange), and the pair correlation obtained from the committee-driven trajectory by direct re-weighting, Eq. (3.22) and the cumulant expansion approximation (CEA), Eq. (3.24).

where

$$\sigma_a^2 \equiv \frac{1}{M'-1}\sum_{j=1}^{M'}\left|\langle a^{(j)}\rangle_{\bar{V}} - \bar{\bar{a}}\right|^2 \tag{3.27}$$

indicates the uncertainty arising from the OMs, and

$$\begin{aligned}\sigma_{aV}^2 &\equiv \frac{1}{M'}\sum_{j=1}^{M'}\sigma_{aV}^{2\ (j)},\\ \sigma_{aV}^{2\ (j)} &\equiv \frac{1}{M-1}\sum_{i=1}^{M}\left|\langle a^{(j)}\rangle_{V^{(i)}} - \frac{1}{M}\sum_{i=1}^{M}\langle a^{(j)}\rangle_{V^{(i)}}\right|^2\\ &\approx \frac{\beta^2}{M-1}\sum_{i=1}^{M}\left|\langle a^{(j)}(V^{(i)}-\bar{V})\rangle_{\bar{V}} - \langle a^{(j)}\rangle_{\bar{V}}\langle V^{(i)}-\bar{V}\rangle_{\bar{V}}\right|^2\end{aligned} \tag{3.28}$$

indicates the uncertainty that arises due to the sampling of the different PMs. In the general case of an uncertainty estimation that is *not* based on a committee model, where only the "best values", $\bar{a}(\mathbf{q})$ and $\bar{V}(\mathbf{q})$, and their uncertainties, $\sigma_{\bar{a}}(\mathbf{q})$ and $\sigma_{\bar{V}}(\mathbf{q})$, are available, the reweighting technique so far described becomes inapplicable. The error-propagation formula for the uncertainty $\tilde{\sigma}^2$ on the canonical average $\langle\bar{a}\rangle_{\bar{V}}$ cannot be straightforwardly implemented either, since it requires the off-diagonal elements of the covariance matrix, and not only $\sigma_{\bar{a}}^2(\mathbf{q})$ and $\sigma_{\bar{V}}^2(\mathbf{q})$. Nonetheless even in this case, a simple upper bound for $\tilde{\sigma}^2$ can be obtained:

$$\tilde{\sigma} \leq \langle\sigma_{\bar{a}}\rangle + \beta\left\langle\left|\langle\bar{a}\rangle - \bar{a}\right|\sigma_{\bar{V}}\right\rangle, \tag{3.29}$$

which corresponds, at least in spirit, to the results we obtain for the committee model, Eqs. (3.26), (3.27), and (3.28), and its implementation shows no hurdles. The theoretical details that lead to this formula are provided in the appendix of the original paper[105].

3.3 Applications

Fig. 1.1 summarizes how the weighted baseline scheme, and the on-the-fly estimation of errors for statistical averages, can be integrated with a calibrated committee model, in the context of a molecular dynamics simulation. After the construction of a suitable database on which reference values (say of energies and forces) are computed, the database is randomly sub-sampled into M smaller training sets on which a committee of M ML models are trained. Depending on the specific physical system/quantity analyzed we adopt two alternative but equally correct approaches to construct a validation set, in order to calibrate the uncertainty of the committee and estimate the re-scaling factor α. The first strategy consists in extracting N_{val} decorrelated configurations from short committee MD trajectories, calculating forces and energies with the reference method, and employing these as the validation set. In the second strategy, instead, the ensemble of N_{val} validation structures was gathered by selecting, in the original training database, those structures that do not appear in at least n of the

training subset. Following the α calibration step, MD simulation are driven by the committee model. The weighted-baseline numerical integration of the equations of motion is based on Eq. (3.12), which reduces to a non-baselined model by setting $V_b = 0$. During the MD simulation driven by the committee model, all the (re-scaled) model-dependent quantities of interest are stored for a significant set of (uncorrelated) configurations, eventually leading to re-weighting and, therefore, to uncertainty estimation of the chosen thermodynamic averages. Any configuration encountered along the trajectory that is associated with an error higher than a set threshold can be used to improve the reference database, in an offline (or online) active learning scheme.

In the next subsections we describe how we applied this routine to weighted baseline integration (Sec. 3.3.1), as well as to compute thermodynamic average and the related ML uncertainty for different observables in different physico-chemical environments (Secs. 3.3.2, 3.3.3, 3.3.4). All the simulations are run with the molecular dynamics engine i-PI[127] interfaced with the massively parallel molecular dynamics code LAMMPS[97] with the n2p2 plugin [128] to evaluate the neural network potentials.

3.3.1 Weighted baseline integration

We begin by performing and analyzing a 120 ps temperature replica-exchange molecular dynamics (REMD) [129] simulation of the Phe-Gly-Phe tripeptide, using the weighted baseline method. The i-PI energy and force engine [127] is used to simulate 12 Langevin-thermostatted replicas with temperatures between 300 K and 2440 K using a time-step of 0.5 fs. Baseline density-functional-based tight binding energies and forces are evaluated using the DFTB+ [130] package and the DFTB3/3OB [131, 132] parametrisation with a D3BJ [133] dispersion correction (3OB+D3BJ). An ensemble of $M = 4$ Behler-Parrinello artificial neural networks (NN) [15] is then used to promote this baseline to a first-principles density-functional-theory (DFT) level of theory. The DFT calculations are performed using the GAMESS-US [134, 135] code and the PBE density functional [136] with a dDsC dispersion correction [137–139] and the def2-TZVP basis set [140]. The NNs are trained to reproduce the differences between the DFTB+ baseline and the target DFT energies and forces. The NNs differ only in the initialisation of the NN weights and the internal cross-validation splits of the reference data into 90% training and 10% test data. The reference data underlying the NNs is constructed by farthest-point sampling configurations from 1.5 ns long REMD simulations of 26 aminoacids, each composed of 16 Langevin-thermostatted replicas with logarithmically-spaced temperatures between 300 K and 1000 K. The resultant set of configurations is enriched with 3,380 geometry-optimised dimers from the BioFragment Database[141]. Note that the aminoacids are simulated at less than half the maximum temperature, at which the tripeptide is simulated. The uncertainties associated with the ensemble predictions are estimated using the scheme of Ref. 116, using a scaling correction of $\alpha = 1.0$, computed on the tripeptide validation data. The uncertainty of the ML model is used, together with a baseline uncertainty of DFTB $\sigma_b = 7 \times 10^{-3}$ meV/atom, estimated according to Eq. (3.13), to build a weighted baseline model

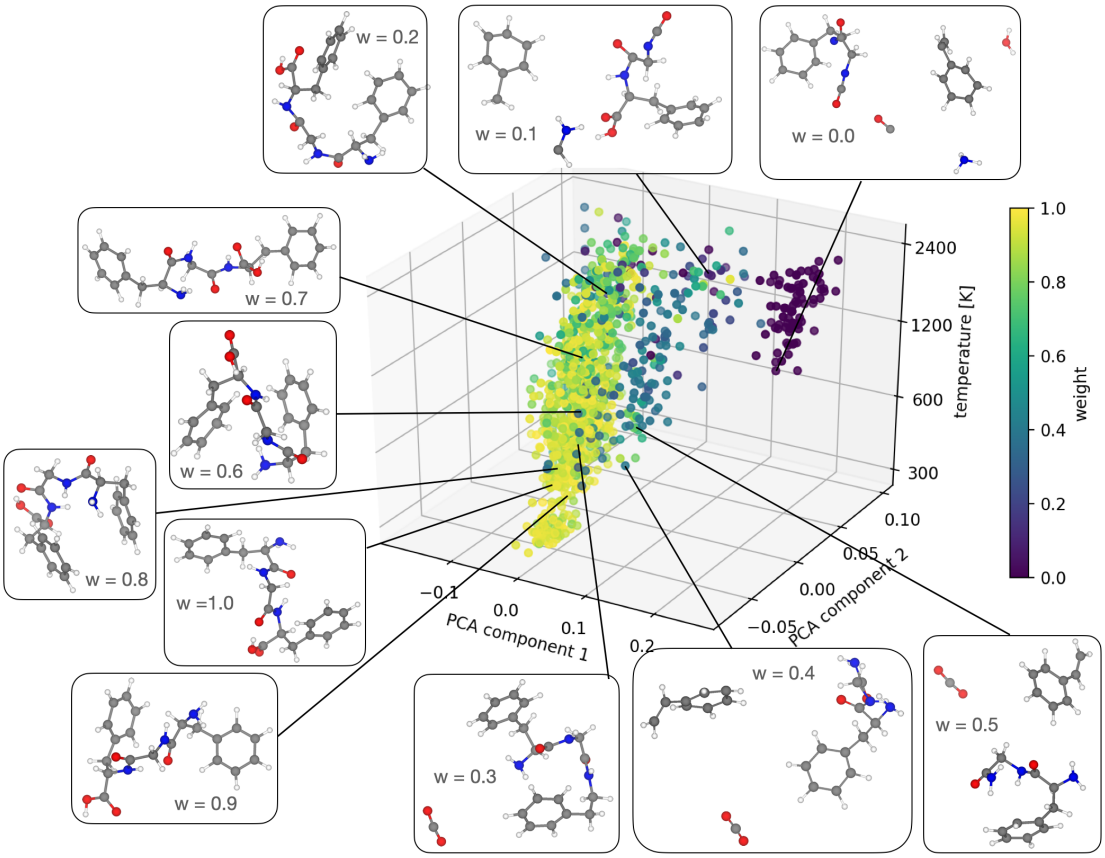

Figure 3.3 – A visualization of the results of the replica-exchange MD simulation of the Phe-Gly-Phe tripeptide, using a weighted-baseline scheme. Central scatter-plot: a set of 2,000 atomic configurations collected from all replicas is classified according to the first two principal components of their SOAP features (x and y axes), and the replica temperature (z axis, in logarithmic scale). The SOAP representation employs a cut-off radius of 4 Å, a basis of $n = 6$ radial and $l = 4$ angular functions, and a Gaussian width of 0.3 Å. Each point corresponds to one configuration, colour-coded according to the weight of the ML correction to the baseline potential, see Eq. (3.12). Examples of typical configurations that are representative of the different temperatures and ML correction weights are displayed in the panels surrounding the scatter plot.

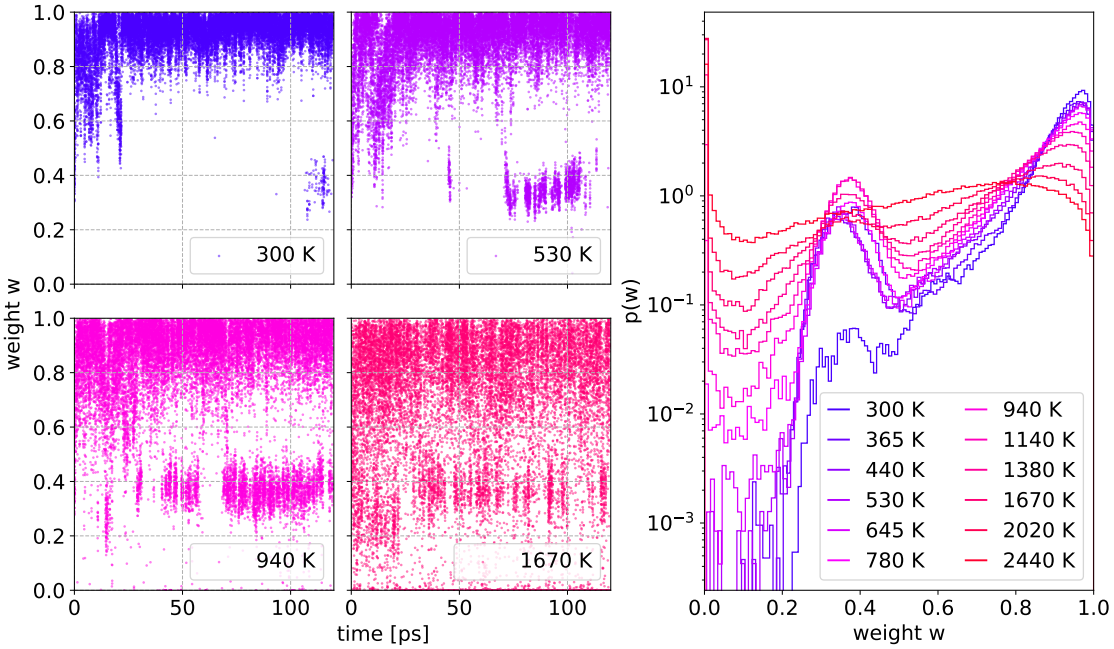

Figure 3.4 – Weights for the ML correction in the weighted-baseline scheme for the Phe-Gly-Phe tripeptide discussed in the text. In the left panels the weights w are displayed at different temperatures for a segment of the REMD trajectory. The rightmost panel shows the log-histogram of the occurrences of the weights at different temperatures.

following Eq. (3.12).

The results of the REMD simulation of the Phe-Gly-Phe tripeptide are portrayed in Fig. 3.3. The central scatter-plot shows 2,000 atomic configurations, drawn at constant stride from all REMD target ensemble temperatures. The configurations are classified according to the first two principal components (x and y axes), obtained from a principal component analysis (PCA) of their SOAP features, and temperature (z axis). Each configuration A is coloured according to the weight $w(A) = \sigma_b^2/[\sigma_b^2 + \sigma^2(A)]$ of the ML correction applied to the baseline potential during the simulation (see Eq. (3.12)). Examples of configurations with very low ($0 \leq w \leq 0.2$), modest ($0.3 \leq w \leq 0.5$), and large weights ($0.6 \leq w \leq 1$) are grouped at the top, bottom and left of the scatter plot, respectively. The figure shows that at low temperature the simulation samples exclusively different conformations of the polypeptide chain, that are well-represented in the training set and that are therefore associated with low ML uncertainty and high values of $w(A)$. At temperatures above $\approx$ 500K, the polypeptide starts decomposing, releasing first CO_2 and, at temperatures above $\approx$ 1000K, NH_3, H_2O, as well as larger fragments. None of these highly energetic reactions are represented in the training set, which is reflected in the sharp decrease of the weight. Upon entering the extrapolative regime, the NN correction to the baseline, $\bar{V}_\delta$, is suppressed by the vanishing weight w, thereby ensuring numerical stability of the simulation subject to the baseline potential.

A quantitative analysis of weight distributions is shown in Fig. 3.4. Higher temperatures are

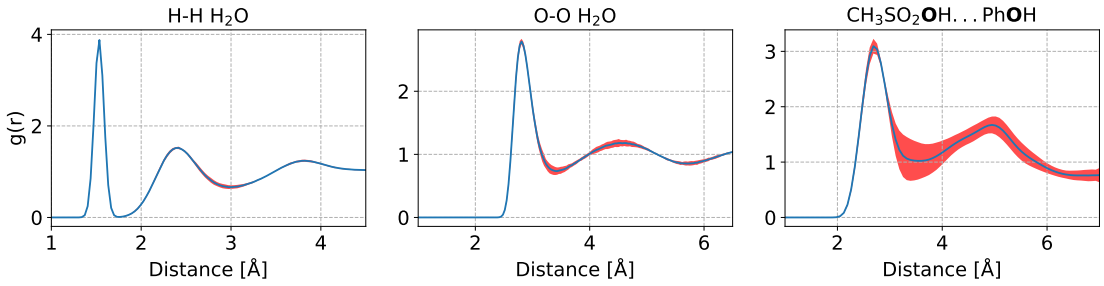

Figure 3.5 – Pair correlation function in water (left, middle panels) and phenol-solvated methanesulphonic acid (right panel). The committee value (blue solid line) and its uncertainty (shaded red area) as estimated from Eq. (3.28) are displayed.

displayed in warmer colours. The left panels show the weights w along the REMD trajectory. These values are collected in the rightmost histogram which displays, in semi-log scale, the distribution $p(w)$ of weights at different temperatures. We see that at intermediate T, an "island" at $w \approx 0.4$ – or a peak in $p(w \approx 0.4)$ – emerges, which corresponds to the tripeptide dissociation and the release of a CO_2 molecule. At even larger temperatures the probability $p(w = 0)$ grows, while the peak at $w \approx 0.4$ is levelled out by the increase in the number of low-w snapshots and the $p(1)$ decreases by more than an order of magnitude due to the persistence of the extrapolative regime at $T \gg 1000$ K. This simulation provides a compelling example of how a weighted baseline scheme allows exploring all parts of configuration space without incurring in unphysical behaviour and instability due to extrapolations of the NNs – which typically occur within the first 100 ps of a similar REMD simulation using a non-weighted baseline correction. Quite obviously, the configurations collected in the extrapolative regime do not reach the level of accuracy of the high-end electronic structure method, but only that afforded by the baseline potential. Nonetheless, simulations based on this scheme can be used whenever extrapolation occurs only over brief stretches of the trajectory, or when (as it is often the case) one is only interested in the low-temperature portion of a REMD simulation, with the high temperature replicas used only to accelerate sampling. Furthermore, one can store configurations characterised by a large $\sigma(A)$ in order to add them to the training database, which simplifies greatly the implementation of online and offline active learning schemes.

3.3.2 Pair distribution function

The radial distribution function represents a simple and insightful structural observable to test the method developed in Sec. 3.2 to estimate the uncertainty on thermodynamic averages. Computationally, $g(r)$ is usually determined *i)* by sampling a significant number of atomic configurations from a thermodynamic ensemble; *ii)* by computing the minimum image separations $\mathbf{r}_i - \mathbf{r}_j$ of all the atomic pairs, for each sampled configuration, and *iii)* by sorting these separations into an histogram h whose bins extend in the interval $[r, r + \delta r]$. When the reweighting procedure is considered point *i)* is performed by running a MD trajectory driven by the committee model alone, and the model-dependent phase-space sampling is accounted

by the weights, Eq. (3.20). Notice that the calculation of the radial distribution function $g^{(i)}(r)$ of the i-th member of the ML committee depends on i through the weights alone, i.e. through the calibrated potential energy estimate for each member.

Water

A committee of $M = 4$ NNP models was trained via the n2p2 code [142] over a dataset of 1593 64-molecule bulk liquid water structure whose total energy and the full set of interatomic-force components were computed at the revPBE0-D3 level with CP2K [143]. The atomic environments are described within a cutoff radius of 12.0 a.u. using the symmetry function sets for H atoms (27 functions) and O atoms (30 functions), as selected in Ref. 19. The hydrogen and oxygen atomic NNs consist of two hidden layers with 20 nodes each. We refer to Ref. 144 for further details on the training set.

We run an NVT MD trajectory, driven by the committee, at $T = 300$ K for 2 ns on a system of 64 water molecules inside an equilibrated cubic box of side 23.86 Å. We obtain an unbiased estimate for the correction factor $\alpha = 2.1$, using Eq. 3.6. Note that without applying the correction for the estimator bias, would lead to substantial over-estimation of the correction factor, in this case $\alpha = 3.75$. Figure 3.5 displays the hydrogen-hydrogen (left) and oxygen-oxygen (middle) pair distribution function $g(r)$. The ML uncertainty, computed as in Eq. (3.28), is shown as a shaded area. The error on position and height of the first peak is minuscule, while slighlty larger uncertainty is predicted on the longer-range features for the O–O correlations. This analysis demonstrates, with a simple post-processing of a single trajectory, that the accuracy of the NNP is sufficient to describe quantitatively the $g(r)$ – a useful verification of the reliability of the model.

Methanesulphonic acid in phenol

As a second example, we consider the solvation of methanesulfonic acid (CH_3SO_2OH) in phenol (C_6H_6O), a system that was studied in Ref. 113 because of its relevance to the synthesis of commodity chemicals such as hydroquinone and catechol, in which methanesulfonic acid acts as a catalyst for the reaction between H_2O_2 and phenol. We use an ensemble of $M = 5$ neural network (NN) machine learning potentials to simulate one acid molecule dissolved in 20 phenol molecules at $T = 363$ K. The technical details and the resulting potentials are identical to those presented in Ref. 113, that are available from Ref. 145. Note that in the original publication the calibration factor was estimated to be $\alpha = 5.8$. Using the unbiased estimator introduced here, Eq. (3.6), yields a corrected value of $\alpha = 4.1$.

An understanding of the solvation of CH_3SO_2OH by phenol is a necessary preliminary step towards rationalizing the regio-selectivity of this acid in the catalytic hydroxylation of phenol to form catechol or hydroquinone. Methanesulfonic acid acts both as a hydrogen bond acceptor through its sulfonil oxygen atoms, and as a donor through the methanesulfonic hydroxyl

group. The strength and population of hydrogen bonds can be inferred by a quantitative analysis of the pair correlation function $g(r)$ between the protonated O in the hydroxyl group of methanesulfonic acid ($CH_3SO_2\mathbf{O}H$) and the O atom in phenol ($C_6H_5\mathbf{O}H$). We compute the pair correlation function from 16 independent MD simulation runs for a total of about 1.6 ns. A thorough discussion of the MD integration set up and the related technical details can be found in Ref. 113.

The uncertainty in the $g(r)$ obtained by a CEA reweighing of the committee members, as in Eq. (3.28), is considerably larger than what observed for the case of water (right panel of Fig. 3.5), together with its uncertainty calculated as in Eq. (3.28) (shaded area). This can be ascribed in part to the slightly larger test error computed for the ML potential (which is unsurprising given the considerably more complex composition), but also in part to poorer statistics due to the presence of just a single acid molecule in the simulation cell. The statistical uncertainty on the committee $g(r)$ obtained via a block analysis is indeed comparable to the one due estimated by the committee reweighing. Similar to what we observe for the O-O $g(r)$ in water, the uncertainty is not constant, but is largest at the minimum between the first and second coordination shell. The fact that the first coordination shell is affected by a small error is reassuring, suggesting that the geometry and population of hydrogen-bonded configuration is predicted reliably. Overall, this example demonstrates how the estimates we introduce for the effects of the ML error on sampling make it possible to assess the reliability of structural observables, particularly in difficult cases in which the model exhibits a substantial error, and so it is important to determine precisely whether such error does or does not (as in this case) affect the qualitative interpretation of simulations results.

3.3.3 Free energy landscapes

Combining ML potentials and enhanced sampling techniques makes it possible to explore computationally free-energy landscapes that involve activated events, such as chemical reactions and phase transitions. In this Section, we show how on-the-fly reweighing can straightforwardly applied to the calculation of free-energy differences and enhanced sampling simulations.

Melting point of water

We begin by demonstrating the calculation of the free energy difference between hexagonal ice and liquid water, $\Delta\mu = \mu^{Ih} - \mu^{L}$, at 8 different temperatures, using the interface pinning (IP) technique.[146] The basic idea of IP involves performing a biased simulation in which the system is forced to retain a solid-liquid interface whose position fluctuates around an average value. This is practically achieved by including an additional pinning potential

$$W(A) = \frac{\kappa}{2}[Q(A) - a]^2 \tag{3.30}$$

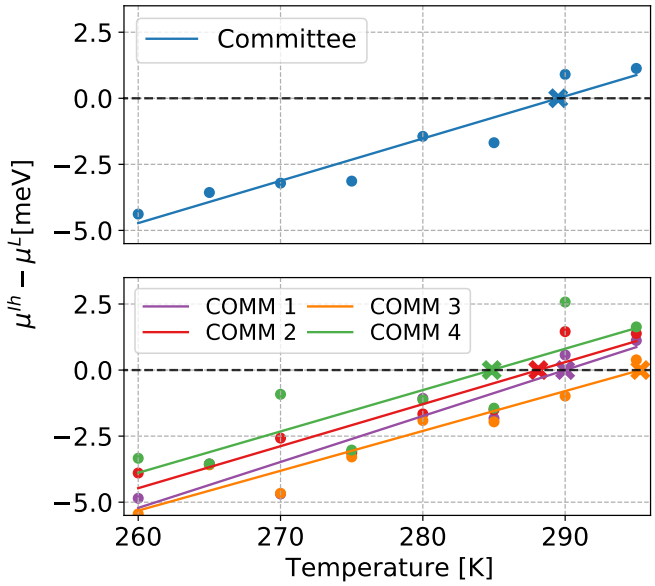

Figure 3.6 – Chemical potential difference between hexagonal ice and liquid water as a function of temperature. Upper panel: the fit obtained for the trajectory driven by the committee mean. Lower panel: individual fits for each committee model.

where $Q(A)$ is an order parameter which identifies the phase of the system (local Q6, defined as in Ref.[147, 148]), κ is a spring constant dictating the amplitude of interface fluctuations, and a is the reference value for the collective variable (usually taken as the value of Q at which half the system is in the solid phase). The chemical potential difference at the simulation temperature T can then be estimated by

$$\Delta\mu(T) = -\kappa(\langle Q\rangle' - a) \tag{3.31}$$

where $\langle\cdot\rangle'$ indicates $NP_z\kappa T$-ensemble averages with the additional term W defined in Eq. (3.30). In the present work, simulations are driven by the same committee of $M = 4$ NNP models discussed in Sec. 3.3.2, using PLUMED[149] to constrain the order parameter to the target value $a = 165$. A total of 336 water molecules are simulated in a supercell with an elongated side to allow probing the coexistence of the two phases, separated by the planar interface; in particular we employed an orthorhombic supercell of size $15.93 \times 13.79 \times 52.47$ Å^3.

We compute the value of $\Delta\mu(T)$ at different temperatures, and perform a linear fit from which we determine the melting temperature T_m as the intercept with the abscissa, $\Delta\mu(T_\mathrm{m}) = 0$. We also obtain the entropy of melting per molecule, $\Delta s_\mathrm{m} = \left.\frac{\partial\Delta\mu}{\partial T}\right|_{T_\mathrm{m}}$, as the slope of the fit, and the latent heat of melting per molecule $\Delta h_\mathrm{m} = T_\mathrm{m}\Delta s_\mathrm{m}$. As shown in the top panel of Fig. 3.6, even though the individual points are somewhat scattered due to statistical errors, it is possible to determine a clear linear trend resulting in $T_\mathrm{m} = 290$ K, $\Delta s_\mathrm{m} = 0.16$ meV/K/molecule, and $\Delta h_\mathrm{m} = 46$ meV/molecule. It should be noted that these values deviate from those that have been computed with a similarly trained potential[144] due to the presence of substantial finite-

size effects in the present simulations, which are only meant to demonstrate the application of this uncertainty quantification approach, and not to provide size and sampling-converged values of the averages.

In order to estimate the uncertainty due to the MLPs, we combine Eq. (3.31) with the CEA, to compute the model-dependent chemical potential differences $\Delta\mu^{(i)}$ using

$$\langle Q\rangle'_{V^{(i)}} = \langle Q\rangle'_{\bar{V}} - \beta[\langle Q(V^{(i)}-\bar{V})\rangle'_{\bar{V}} - \langle Q\rangle'_{\bar{V}}\langle V^{(i)}-\bar{V}\rangle'_{\bar{V}}]. \tag{3.32}$$

In line with the uncertainty propagation framework developed in Sec. 3.2, we compute four different fits, one for each model, and from them four different melting temperatures $T_{\mathrm{m}}^{(i)}$, indicated by the coloured crosses in the lower panel of Fig. 3.6. By taking the average and standard deviation of the model-dependent $T_{\mathrm{m}}^{(i)}$, as well as the associated $\Delta s_{\mathrm{m}}^{(i)}$ and $\Delta h_{\mathrm{m}}^{(i)}$, we can determine the mean values and the ML uncertainty intervals, namely $\overline{T_{\mathrm{m}}} = 290 \pm 5$ K, $\overline{\Delta s_{\mathrm{m}}} = 0.16 \pm 0.01$ meV/K/molecule, and $\overline{\Delta h_{\mathrm{m}}} = 46 \pm 3$ meV/molecule. In view of the linear nature of the CEA, the values of the molar entropy and latent heat of melting computed from the mean of the committee estimates match exactly those computed directly from the committee estimates. In principle, the two estimates $\overline{T_{\mathrm{m}}}$ and T_{m} differ, even if in this case they are equal within the confidence interval. Whenever a non-linear procedure is involved in the calculation of the property of interest, results may change based on the way the committee estimates are combined. Comparing different approaches is then a useful check to assess the robustness of the error estimation.

Deprotonation of methanesulfonic acid

We use the committee model discussed in Sec. 3.3.2 and the same metadynamics protocol described in Ref. 113 to compute the free energy profile for the deprotonation of CH_3SO_2OH in phenol, a key quantity to rationalize the activity of methanesulfonic acid in catalyzing the hydroxylation of phenol. We define the free energy as a function of the coordination, s^{O}, of the oxygen atoms in the acid with respect to the hydrogen atoms in the system. The free energy at s^{O} is by definition kT times the negative of the logarithm of the population fraction $p(\mathrm{s}^{\mathrm{O}})$ of the configurations with a given s^{O}.

To obtain an unbiased estimate of $p(\mathrm{s}^{\mathrm{O}})$ from a trajectory with a time-dependent bias $\tilde{v}(t)$, we weight the configurations by $u(A(t)) = e^{\beta(\tilde{v}(t)-c(t))}$, where the time-dependent offset $c(t)$ is computed using the Iterative Trajectory Reweighting (ITRE) algorithm.[150] The population fraction for the committee, $\bar{p}(\mathrm{s}^{\mathrm{O}})$, is computed as the ITRE-reweighted normalized histogram of the occurrences of configurations A with a given $\mathrm{s}^{\mathrm{O}}(A)$:

$$\bar{p}(s) = \langle \delta(\mathrm{s}^{\mathrm{O}}(A)-s)\,u(A)\rangle_{\bar{V}} \tag{3.33}$$

where the average is over the metadynamics trajectory, and $\delta(\mathrm{s}^{\mathrm{O}}(A)-s)$ selects structures with a prescribed value of the coordination number. In turn, the model-dependent population can

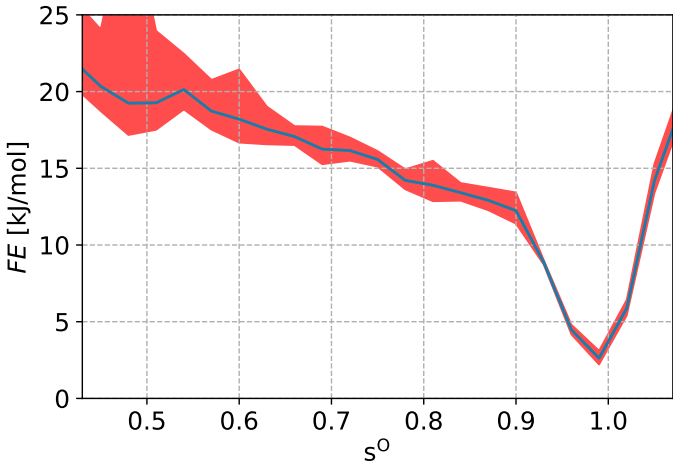

Figure 3.7 – Projection of the free energy along the proton transfer reaction s^O for a system of one methanesulfonic acid molecule dissolved in 20 phenol molecules. $s^O \approx 1$ indicates the neutral state, while $s^O < 1$ a deprotonated state of the acid. The shaded area represents the ML uncertainty obtained from Eq. (3.26).

be readily obtained, through the CEA, as

$$\begin{aligned} p^{(i)}(s) &= \bar{p}(s) - \Delta p^{(i)}(s) \\ \Delta p^{(i)}(s) &= \beta \langle\, \delta(s^O(A) - s)\; u(A)\; (V^{(i)}(A) - \bar{V}(A)) \,\rangle_{\bar{V}} \\ &\quad - \beta \langle\, \delta(s^O(A) - s)\; u(A) \,\rangle_{\bar{V}} \langle V^{(i)}(A) - \bar{V}(A) \,\rangle_{\bar{V}} \end{aligned} \tag{3.34}$$

Finally, the uncertainty in the population, Δp, is obtained as the standard deviation of $\Delta p^{(i)}$ over the M models, as in Eq. (3.28). The symmetric uncertainty on the population results in a confidence range on the free energy which is *asymmetric* about $-kT\log(\bar{p})$, spanning values from $-kT\log(\bar{p}+\Delta p)$ to $-kT\log(\bar{p}-\Delta p)$.

As shown in Fig. 3.7, the uncertainty between the models is very small around the minimum corresponding to the neutral state of the acid, but grows substantially in the deprotonated state – which is consistent with the qualitative observation made in Ref. 113 of the increase in the uncertainty on the NNP predictions for dissociated configurations, that are less represented in the training set. Interpreting the configurations with $s^O \approx 0.5$ as the deprotonated state, the free energy cost for the dissociation of CH_3SO_2OH in phenol can be estimated to be 20^{+5}_{-2} kJ/mol. Even though in this specific instance other errors, e.g. those due to finite-size effects and reference energetics, are likely to be comparable with that obtained from the spread of the committee members, the substantial uncertainty computed by on-the-flight reweighting underscores the importance of error estimation when using machine learning models.

3.3.4 Finite-temperature density of states

As a last example, that we use to highlight the interplay between sampling and model uncertainties, we consider the finite-temperature density of states (DOS) of gallium in its metallic liquid phase. The sampling of configurations is performed through MD simulations driven by a committee of $M = 4$ NNPs, based on the potential introduced in Ref. 151, that is available from Ref. 152. We consider a system of 384 Ga atoms in the NVT ensemble, sampled at a temperature $T = 1800$ K using a combination of a generalized Langevin[153] and stochastic velocity rescaling thermostats,[154] as implemented in i-PI. We employ a timestep of 4 fs to integrate the equations of motion for a total of 400 ps. The DOS model is based on the framework developed in Ref. 123, which we briefly summarise. For a given configuration A, the DOS is defined as

$$\mathrm{DOS}(E,A) = \frac{2}{N_b N_{\mathbf{k}}} \sum_n \sum_{\mathbf{k}} \delta(E - E_n(\mathbf{k},A)), \tag{3.35}$$

where $E_n(\mathbf{k})$ is the energy for the (doubly-degenerate) n-th band and wavevector $\mathbf{k}$. The DOS is normalized to the number of electronic states, $N_b N_{\mathbf{k}}$, where N_b and $N_{\mathbf{k}}$ are the number of bands and $\mathbf{k}$-points considered, respectively. We adopt a ML approach based on a local-environments decomposition to predict $\mathrm{DOS}(E,A)$, and train a committee of observable models (OM, see Sec. 3.2), in order to estimate a ML uncertainty. The predicted DOS of a given structure A, and the j-th model reads

$$\mathrm{DOS}^{(j)}(E,A) = \sum_{k\in A} \mathrm{LDOS}^{(j)}(E,A_k), \quad j = 1,\dots,M'. \tag{3.36}$$

The training set for each OM is represented by 150 random structures extracted from a total of 274 Ga training configurations, including mostly liquid structures at various temperatures and pressures and a few solid ones. For this training set, we compute reference DFT calculations for the $\mathrm{DOS}_{\mathrm{ref}}(E,A)$ as the convolution of the Kohn-Sham eigenvalues $E_{\mathrm{ref},n}(\mathbf{k})$ with a Gaussian smearing of width 0.5 eV.[123] The reference DFT calculations are performed at the level of the PBE functional [136] via the QUANTUM ESPRESSO code,[101, 155] with a Monkhorst-Pack k-point grid that ensures a density of at least 6.5 $\mathbf{k}$-points Å. In order to compare DOS belonging to the different structures of the training set, we align the DOS at the Fermi level. The latter, $E_F(A,T)$, is defined as the solution of the charge-neutrality constraint $N_e = \sum_E f(E,E_F,T)\mathrm{DOS}(E,A)$, where $f(E,E_F,T)$ is the Fermi-Dirac distribution and $N_e = 2$ due to spin degeneracy. The featurization is done using a SOAP kernel with $n = 12$, $l = 9$, $g_s = 0.5$, $r_c = 6$Å, $c = 1$, $m = 5$, $r_0 = 6.0$ (the parameters follow the notation in Ref. 123). Given the small train set size, and that commitee predictions for sparse kernel models add negligible overhead on top of a single prediction, we use a large committee with $M' = 64$ members. According to Eqs. (3.26), (3.27), and (3.28), the total ML uncertainty σ on $\langle \mathrm{DOS}(E)\rangle_T$ derives from both the uncertainty on individual DOS predictions, σ_a and the uncertainty on the phase space sampling associated with the committee of MLPs driving the dynamics, σ_{aV}.

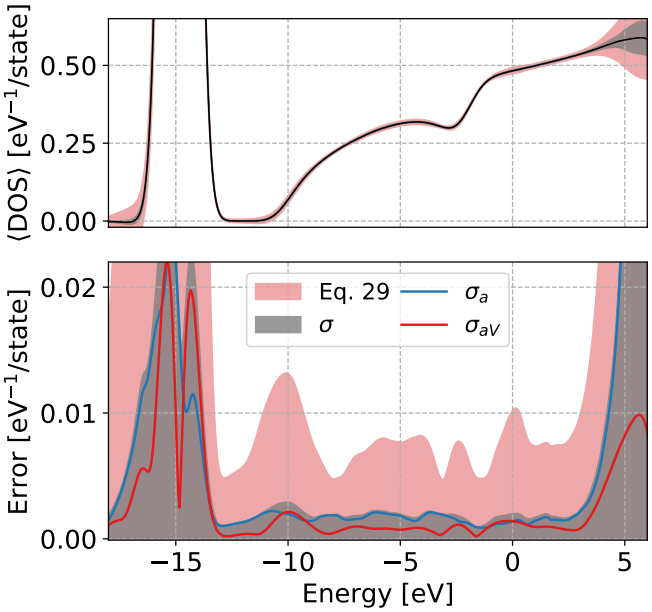

Figure 3.8 – Machine-learned average density of states, $\langle \mathrm{DOS}(E) \rangle_T$, computed for a simulation of liquid gallium at T = 1800 K. The zero is set at the Fermi energy, to which the single-configuration DOS entering the average were align. The average $\langle \mathrm{DOS}(E) \rangle_T$ (solid line) is reported together with its statistical uncertainty (shaded gray area).The red shaded area represents the upper bound of the uncertainty, computed as in Eq. (3.29)

The results of these calculations are displayed in Fig. 3.8: in the upper panel the average $\langle \mathrm{DOS}(E) \rangle_T$ is reported together with its total ML uncertainty, σ, as computed by in Eq. (3.26). In the lower panel we show the individual contributions of the uncertainty on the property, σ_a, and that associated with sampling, σ_{aV}, to the total σ, together with the upper bound estimate of the uncertainty. The absolute error on the DOS is small, and is dominated by σ_a. The contribution σ_{aV} associated with sampling is sizeable, and in some energy range it dominates the uncertainty. The coupling between the potential energy and the observable property cannot be neglected. Notice that the upper bound given by Eq. 3.29 (shaded red area) largely overestimates the uncertainty based on the committee model. The full characterization of the error statistics that is enabled by a committee model provides a substantial improvement of the quality of the error bound, in comparison to an uncertainty estimation that is limited to the standard deviation of individual predictions.

4 Fitting a potential for the GaAs phase diagram[1]

4.1 Introduction

Till now, we have discussed how we can streamline the generation of a MLIP through an automatic selection of training points and features, and how to quantify with few additional calculations the uncertainty due to the use of a MLIP to compute thermodynamical properties. From here on, we will apply these methods, together with other state-of-the-art techniques, to train a new MLIP for the Ga_xAs_{1-x} system.

The choice of the system is dictated on one side by the scientific interest, since both Ga and GaAs are systems of technological relevance[157–161] and their atomistic structure plays a role in their unique properties[162–165], and on the other side by the complexity of the system, as it consists of semiconducting phases mixed with metallic ones. Moreover, there are a number of empirical potentials that have been widely used in the past that we can use for comparison, demonstrating their limited transferability. Overall, we believe this to be an excellent testing ground to understand the limits in the training of a potential.

We begin this chapter by describing the technical side of the training of the potential, i.e. the atom-centered representation used and the ML model. Then, we detail the generation of the database, trying to highlight the necessary steps in the selection of training points to thoroughly cover the binary phase diagram. We continue by predicting a number of properties with the trained potential, starting from static lattice properties, to solid and liquid ones. For the solid phase, we present the results for the heat capacity and thermal expansion from cryogenic temperatures up to melting. For the liquid, we show the density, diffusion, and radial pair distribution functions of Ga, As, and GaAs. Most of these quantities are also computed for the two most widely used empirical potentials for GaAs[166, 167]. At last, we conclude with the binary phase diagram predicted by our potential.

[1] The majority of this chapter is extracted from Ref. 156. The author of the thesis has done all of the work discussed in the chapter.

4.2 Methods

4.2.1 Architecture of the potential

Details of the MLP

For our potential, we choose to follow the work done by Behler and Parrinello[15], since it has already been thoroughly tested on a number of different materials and it has been shown to perform well on systems similar to ours[19, 30, 168]. We use the implementation by Singraber and Dellago[142]. In general, it should be noted that transforming this potential into a different one would be as easy as refitting the training set that we have generated using a different package.

As we discussed in Sec. 2.2.2 (and as is discussed elsewhere[18, 91]), the SFs tend to be very correlated to each other. Therefore, we use the CUR selection[79] introduced in Sec. 2.3.1 to obtain an optimal set of uncorrelated SFs. We begin by generating a set of 604 viable SFs with the method detailed in Sec 2.2.2 and use the CUR selection to identify 128 SFs (64 for each species) that optimally describe the system. We repeat the selection after the addition of every set of new training structures, to ensure that we are able to capture all the novel relevant correlations. However, we also observe that late additions to the training set have little effect on the choice of the SFs, indicating the robustness of our method.

The regression scheme that we use in this work is a feed-forward neural network with 2 layers and 24 nodes per layer, for a total of 4370 parameters, 2185 for each species, that must be optimized. The optimization procedure is carried out minimizing the errors between the predicted energies and forces with respect to the known DFT values, using a parallel Kalman filter implementation[142].

Uncertainty estimation

As we discussed in Ch. 3, the estimation of the uncertainty deriving from the use of a ML model is necessary to be able to trust the predictions. Therefore, all the calculations in this chapter are performed using an ensemble of $M = 4$ potentials independently trained on different (but overlapping) subsets of the same training set and starting from different initial weights. The average of the forces and energies is used to drive the dynamics and provide numerical estimates of the confidence intervals for some properties. The set of structures used to estimate α (Eq. 3.6) is removed from the full dataset before starting the training procedure. More details about the dataset generation are presented in section 4.2.2. We demonstrate the use of the uncertainty estimation for thermodynamical averages for the pair distribution functions.

On top of the uncertainty deriving from the use of a ML model, we also take into account the statistical error due to the finite time of the MD simulations, computed using the block

averaging method.

4.2.2 Database generation and details

To generate a potential able to cover the full binary phase diagram, it is necessary to add training structures of all the various phases of GaAs, Ga, As and their relative interfaces. We use concepts that have already appeared in the literature to create a database that spans all of the phase space of interest. The structures that are contained in the database are obtained using three related but different approaches. We start with a potential limited to a small part of the phase space, we extend it to reproduce static properties of all of the phases of interest and we finally ensure its stability by using an active learning-like procedure on more challenging simulations.

A potential for the interface

The initial set of reference structures is generated following an iterative procedure, aiming to reduce the number of DFT calculations needed, while covering a part of the phase space that is relevant to the calculations that we want to be able to perform.

We begin by training a MLIP on short *ab initio* MD trajectories run on the geometry shown in fig. 4.1 at various temperatures and with unconverged DFT parameters to speed up the calculations (i.e. the k-point grid is limited to the Γ point only, as opposed to fully converged calculations where 3x3x1 grids are used). The structure is chosen to model the interface between solid GaAs (both zinc blende and wurtzite, along the [111] direction) and liquid gallium and include both the A and B surfaces. This was done to train a potential able to run the simulations related to the interface[151], which are described in detail in Ch. 5. Later, the potential was expanded to include other important regions of the phase space.

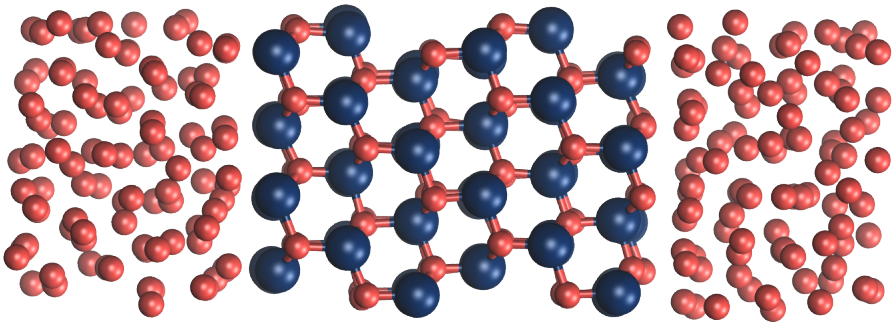

Figure 4.1 – A snapshot from a simulation, depicting the geometry used for all of the zincblende simulations that we run initially. For wurtzite simulations we exchange the solid part with an equal number of gallium and arsenic atoms, changing the unit cell to wurtzite.

The potential obtained at this point was used to run longer simulations of the same and other similar configurations with the aid of advanced integration schemes, allowing to explore a

larger part of the relevant phase space at a fraction of the original cost. We decompose the forces into a short-ranged and fast-varying part and a long-ranged, slow-varying correction, using the multiple time stepping integration scheme[117] as implemented in i-PI[127]. Since the MLIP is able to describe only interaction up to a predefined cut-off, we use it to compute the short-range interactions. Then, the remaining long-ranged correction is computed using the difference between DFT and MLIP. This method has the additional advantage of stabilizing the trajectories obtained with the MLIP at this stage, which have the tendency to become unreliable after a few hundred ps.

Simulations using the multiple time stepping scheme are run iteratively on the NVT ensemble for about 50 ps each for both ZB and WZ at 400 K, 600 K, 750 K, and 1300 K. Temperatures are controlled with a combination of a generalized Langevin[153] and stochastic velocity rescaling[154] thermostats. The inner timestep is set to 1 fs and the outer timestep to 20 fs, effectively allowing us to run *ab initio* quality calculations while reducing the original cost by a factor 20. Every new DFT calculation is used both as a testing point for the current iteration of the MLIP and later as a training point for the next iteration.

Continuous refinement of the potential allows to obtain a stable MLIP able to reproduce very accurately the DFT results during a MD simulation, as shown in fig 4.2. Although we have not explicitly computed it, we assume that the residual long range interactions, arising for example from the two polar interfaces, can be neglected as we consistently use a similarly sized supercell with the same number of layers in the solid phase of WZ and ZB GaAs. In general, we have observed that a variation of the number of layers contributed only with a constant to the total energy and has little effect on the local forces of the structure.

At this point, we choose a reduced set of structures, i.e. only those that contribute with new information using the FPS method detailed in Sec. 2.3.3, to recompute with converged DFT parameters.

This first iteration allow us to produce an initial set of 800 structures, most of which represent the interface between liquid Ga and solid GaAs, with the addition of some structures of bulk solid GaAs and bulk liquid Ga.

Complementing the potential

We extend this potential by explicitly including structures needed to compute known static properties, such as lattice constants, elastic constants, surface decohesion energies, surface reconstructions, point defects and selected plane defects for Ga, As, and GaAs.

For this purpose, we generate the structures needed to compute these properties, either as single-point calculations (e.g. lattice constants) or by relaxing the structure (e.g. defects and surfaces), thus obtaining a sequence of correlated structures. From the relaxations we choose to keep for training only a few out of all of the generated structures, making sure to include the initial, the final and some intermediate steps whose energies are found to be significantly

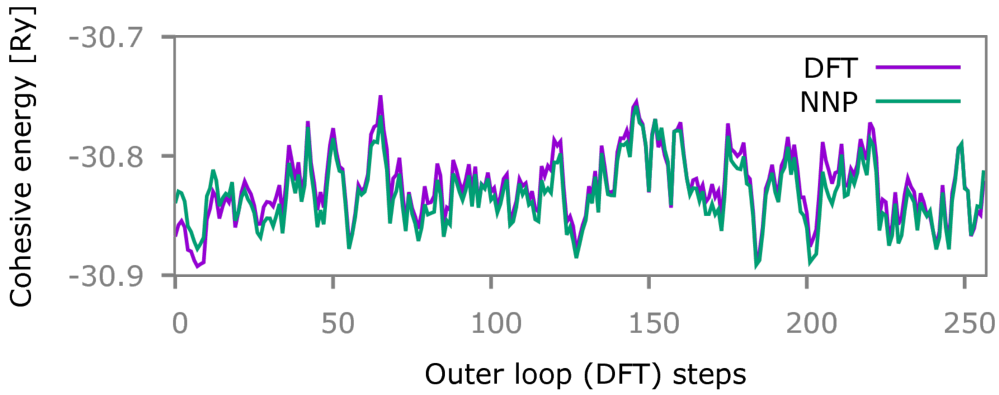

Figure 4.2 – The energy predicted by the NNP vs the energy computed with DFT for a multiple time stepping MD run. This comparison let us see the quality of the NNP predictions for structures generated during a simulation, thus outside of the training/validation set. The similarity between NNP and DFT also hints at the fact that the structures generated during the simulations are physically reasonable.

different compared to the initial and final configurations.

It should be noted that adding the discarded configurations to the training set would have an impact only on the training time, but not on the computational cost of the MLP in production. However, we prefer to keep a smaller and more efficient training set in order to reduce the future cost of recomputing the structures at a different level of theory.

This additional set of 557 structures yields a potential that is able to correctly reproduce these static properties across all these phases, but does not guarantee stability at high temperature or at intermediate stoichiometries.

Iterating over uncertain configurations

To complete the potential and ensure that it is reliable for all of the properties that we want to model, we use an offline active learning strategy, introducing in the dataset some of the structures generated throughout the validation process. Whenever we observe an uncertainty in the committee higher than a threshold (arbitrarily chosen to be 5 times the RMSE) during a simulation, we gather the structures that are poorly predicted and select a small and representative set of configurations for retraining. The structures are chosen either through FPS, or by iteratively adding those with the highest uncertainty, stopping when the predictions become accurate.

We observe that with this procedure we add many structures of liquid Ga_xAs_{1-x} with $0.05 < x < 0.45$ and $0.55 < x < 0.95$, which were initially found to be poorly predicted. This is an obvious consequence of the previous training procedures, where stoichiometries of $x = 0, 0.5, 1$ were favoured, leaving the other regions of the phase space poorly sampled. Similarly, we add

a number of structures of liquid Ga at high pressure, a region that we had not initially included in the training but that it is of great technological relevance.

Details of the DFT calculations

All the DFT calculations are run using QUANTUM ESPRESSO[101]. In order to ensure an absolute convergence of the calculation to below 1 meV/atom we use an energy cut-off of 50 Ry and a density of 6.5 k-points Å. The GGA approximation with PBE exhange-correlation function[136] is used, together with ultrasoft pseudopotentials[169] from the SSSP accuracy library (version 0.7)[170].

In order to minimize the errors arising from minute differences in the k-point grid, we use, as consistently as possible, similarly sized supercells, with average dimensions of 12x14x40 Å. The elongated shape and large size are chosen in order to accommodate two different bulk systems and their interface in a single supercell (e.g. the interface between liquid Ga_xAs_{1-x} and solid GaAs from the original dataset). This also helped to ensure that the cell was large enough to avoid interactions among periodic replicas for defect calculations.

4.2.3 Molecular dynamics

To test the potential beyond the properties that can be probed with single-point calculations, we run MD simulations for the system in its solid and liquid form, together with various interfaces. Since our investigation includes the evaluation of these properties at very low temperature, it is necessary to explicitly include the effects of the quantum motion of the nuclei to recover the correct properties.

Path integral molecular dynamics (PIMD) is a formalism needed to include nuclear quantum effects (NQE) into the simulation, which relies on the isomorphism between a quantum nucleus and a chain of P beads connected by springs, where P must be increased to ensure convergence to the quantum Boltzmann distribution. More details on the theory of PIMD can be found elsewhere[39, 171], whereas from our perspective it is important to mention that simulating a system of P beads has the same computational cost of running P parallel classical simulations of the same system.

All the MD and PIMD simulations are run using i-PI[127] to propagate the dynamics and LAMMPS[97] with the n2p2 plugin[128] to compute energies and forces at every step. Boxes of about 300 atoms are used in most cases for determining the properties, unless specified. The temperature is constrained using a combination of a generalized Langevin[153] and stochastic velocity rescaling thermostats[154], whereas the pressures, when needed, are constrained using an isotropic Bussi-Zykova-Parrinello barostat[172] as implemented in i-PI. A timestep of 4 fs is used to integrate the equations of motion.

4.3 Validating the potential

The final database generated as explained in Sec. 4.2.2 is composed of 1921 structures, out of which 100 are excluded from the training procedure and are used both as a final test set and to compute the α parameter from eq. 3.6. Each potential is trained on the remaining 1821 structure, 20% of which, randomly chosen for each potential, are used for internal validation.

Figure 4.3 illustrates the similarity among the structures that are present in the database. The colours represent the origin of the structures, following the methods detailed in section 4.2.2. The layout of the points is obtained with a KPCovR projection[173] and reflects the composition and stability of different configurations. It can be noticed that the initial configurations are limited to a small region at very precise stoichiometries and low relative energy, while the iterative sampling allows to fill the gaps between the regions and to incorporate defective, high energy structures.

Figure 4.4 shows the parity plots for energies (top) and forces (bottom). In these plots, we refer to "training set" to indicate the full 1821 structures that are used for training, even if not all of them appear in every potential, and the "test set" is the set of 100 structures initially removed from the database. The RMSE for the committee computed on the test set is found to be 2.4 meV/atom for the energies and 109 meV/Å for the forces. We correct for the intrinsic correlation in the dataset with a factor $\alpha = 2.2$. These values show a very accurate fitting, particularly when one considers the very diverse set of structures used in the training.

While these values provide a sense of the typical error for this potential, they do not necessarily reflect the ultimate accuracy when computing specific, physically and technologically relevant observables. To provide a compelling demonstration of the versatility and limits of the potential, we compute a selection of properties and compare them to DFT calculations or experimental values. The static lattice properties that we compute are closely related to structures that are part of the training set, and so they do not fully report on the transferability of the model but rather on the quality of the fit. Results on finite-temperature properties, discussed in section 4.4, provide a complementary perspective on the behaviour of the ML potential and the underlying DFT reference.

We also provide the results obtained for the same properties with two of the most successful empirical potentials that have been published in the past for GaAs, and are fitted to experimental data. The first is the so called ANNK potential, from the initials of the authors, which has the form of a modified Tersoff potential[166] and has seen wide use for the study of the effect of radiation on crystalline GaAs. The second is the bond-order potential (BOP) presented by Murdick *et al.* in 2006[167] to study the molecular beam epitaxy growth of GaAs MOSFETs. Single point calculations for the equation of state, plane decohesion and defect energies are run with the aid of the Atomic Simulation Environment (ASE) package[174].

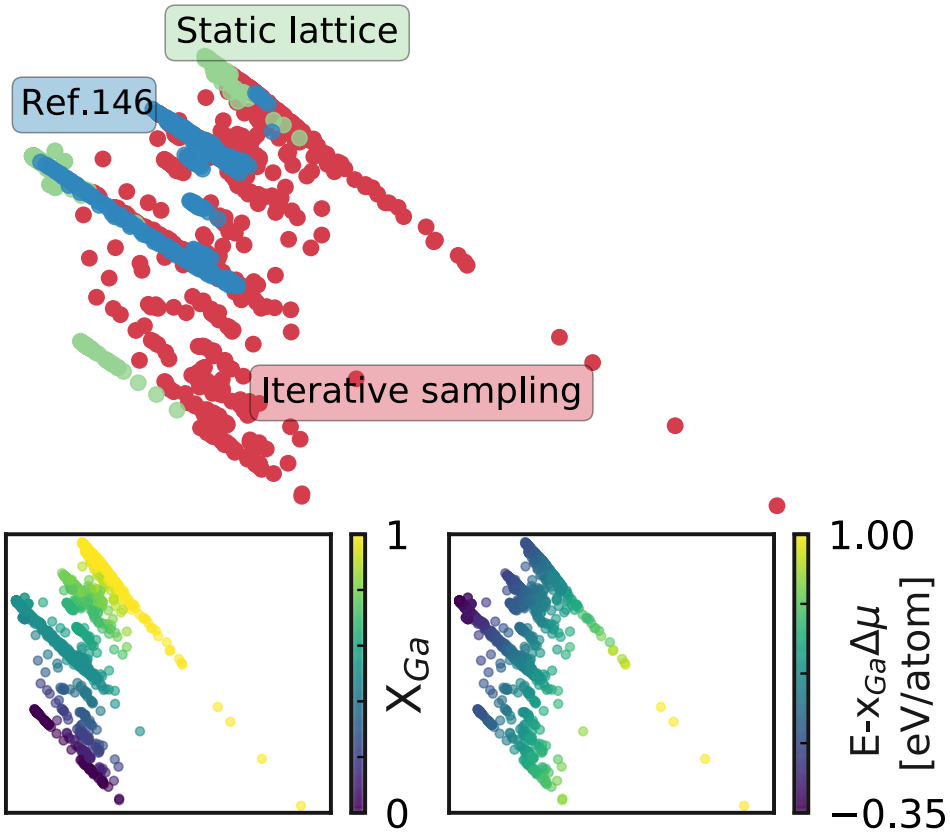

Figure 4.3 – KPCoVR map[173] of the configurations used to fit the final NNP. The map uses an equal mix of PCA and linear regression of the energy relative to the trivial combination $xE_{Ga} + (1-x)E_{As}$ ($\alpha = 0.5$, following the convention of Ref. 173) to illustrate the similarity among the structures. Different colours highlight the origin of the data, as presented in section 4.2.2. The subplots present the same map, coloured according to the stoichiometry (left) and the hull distance, the same quantity used for the map (right).

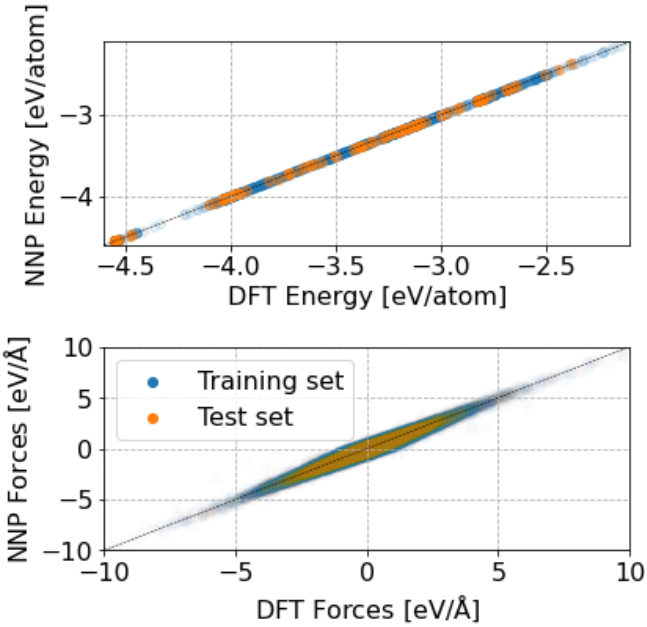

Figure 4.4 – Parity plot comparing the energies (top) and forces (bottom) predicted with the NNP against the reference values from DFT. The test set is an independent set of 100 structures which are excluded from the training procedure. The dashed line is added as a guide for the perfect match between prediction and reference.

4.3.1 Structural and mechanical properties

As a sanity check for our MLP, we compute the equation of state and the elastic constants for some stable phases of Ga, As, and GaAs. As a starting point, we use primitive cells obtained from the Materials Project[175], and optimize them separately for each potential, to provide a self-consistent reference.

The results are shown in table 4.1, together with the available experimental values, which the empirical potentials are fitted against. The same set of calculations is repeated for each potential, and the results of the ANNK and BOP potentials are in agreement with those presented in their respective original papers with the sole exception of the bulk modulus of the ANNK potential for α-Ga and Ga-II, which we find to be more than twice as large as the original value reported, a discrepancy whose origin we could not determine. As expected, our potential is in excellent agreement with the DFT data, while the ANNK and BOP potentials show good agreement with the experimental values of GaAs but are less accurate for single-species Ga and As phases, despite the fact that they were included in the fitting.

4.3.2 Defects

Structure and stability of defects are very important quantities for III/V semiconductors, because of the impact have on electronic properties and device performance. In this section

Property	DFT	NNP	ANNK BOP	Murdick BOP	Exp
			GaAs - ZB		
V_0 [Å^3]	23.82	23.76 ± 0.07	22.56	22.70	22.58
E_0 [eV/atom]	-4.04	-4.04 ± 0.00	-3.35	-3.37	-3.35
B [GPa]	58.82	59.75 ± 0.16	71.3	73.0	74.8
C_{11} [GPa]	98.72	96.22 ± 1.25	124.96	118.89	118.1
C_{12} [GPa]	41.23	46.44 ± 1.22	49.47	54.63	53.2
C_{44} [GPa]	50.92	44.09 ± 0.38	39.27	47.94	59.2
			GaAs - WZ		
V_0 [Å^3]	23.81	23.79 ± 0.06	22.56	22.70	
E_0 [eV/atom]	-4.03	-4.03 ± 0.00	-3.35	-3.37	
B [GPa]	58.73	59.01 ± 0.73	71.25	73.00	
			Ga - α		
V_0 [Å^3]	20.38	20.37 ± 0.02	19.27	20.88	19.58
E_0 [eV/atom]	-2.83	-2.83 ± 0.00	-2.83	-2.57	-2.810
B [GPa]	46.91	47.52 ± 1.80	90.75	49.1	61.3
			Ga - II		
V_0 [Å^3]	19.02	19.00 ± 0.07	16.53	16.71	
E_0 [eV/atom]	-2.81	-2.81 ± 0.00	-2.86	-2.60	
B [GPa]	47.57	48.25 ± 1.90	350.04	98.94	
			As		
V_0 [Å^3]	22.42	22.42 ± 0.03	19.25	19.86	21.51
E_0 [eV/atom]	-4.55	-4.55 ± 0.00	-2.91	-2.94	-2.9
B [GPa]	67.85	68.03 ± 0.37	69.03	103.20	55.6

Table 4.1 – Comparison of the structural properties between DFT, NNP, ANNK[166] and BOP[167]. It should be noted that the ANNK and BOP potentials are fitted to reproduce experimental data, while our potential is fitted on the DFT predictions.

we demonstrate the accuracy of MLP prediction for point and planar defects for the stable phases of As, Ga, and GaAs, while also showing the results obtained with the ANNK and BOP potentials. It should be noted that various works in the literature report that the most stable configuration of some of the defects that we present is charged[176–179]. However, we study them in their neutral state, because both the MLP and the empirical potentials do not have any information about the overall charge of the system, but rely only on the nuclear coordinates for their prediction. While we could, in principle, train the potential with charged defects instead of the neutral ones, this would be inconsistent with the rest of the bulk structures, that are neutral. This also limits the types of defects that we can study (e.g. surface reconstructions, that often involve macroscopic charge transfer).

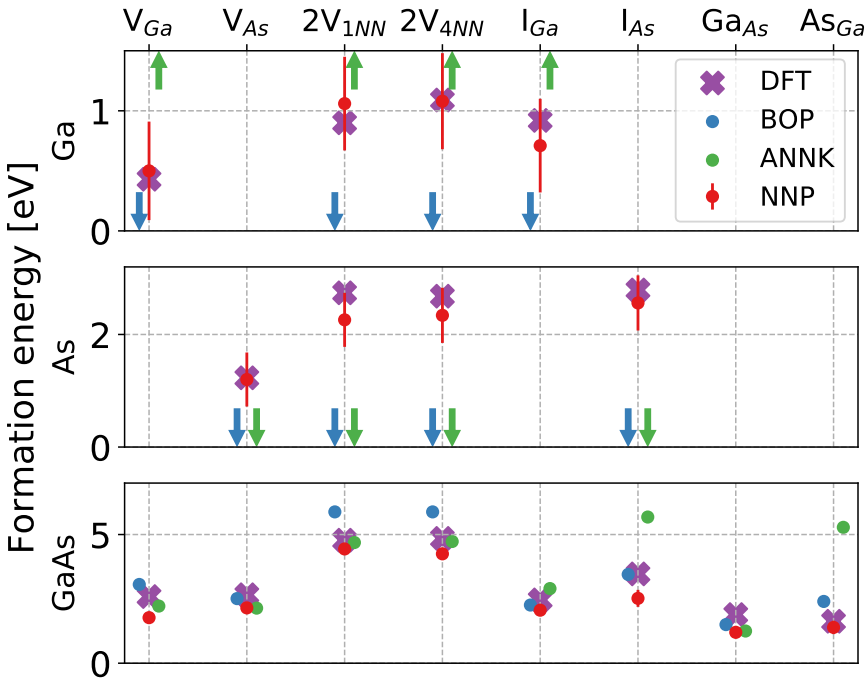

Figure 4.5 – Formation energy of selected defects in bulk Ga, As, GaAs. In the legend, V indicates a vacancy, 2V a divacancy, I an interstitial, and Ga_{As} is an antisite, where a Ga atom substitutes an As one (and viceversa for As_{Ga}). When multiple defects of the same kind were available, only the lowest-energy one is presented. The arrows indicate predictions that are far outside the range of reasonable values for the given defects. Numerical values are reported in the S.I.of the original paper[156]

Point defects

We compute the formation energies of vacancies, di-vacancies and interstitial atoms for Ga, As and GaAs. For the latter, we also include antisite configurations (substitution of an atom with the other chemical species, e.g. Ga instead of As). For each potential we generate the defective supercell at the corresponding equilibrium density, followed by relaxation of the

internal coordinates using the BFGS optimization algorithm. Therefore, when comparing the various "relaxed" configurations, we are effectively observing different minimum energy configurations, each obtained with the corresponding potential.

Since we could not find reference values for the geometry of interstitial atoms in crystalline Ga and As, we generate several possible configurations, and report here only the one that yields the lowest energy of formation with DFT, although we include in the training set all of those that have been created. Similarly, we compute all interstitial configurations that have been reported in GaAs[178, 180], but only discuss here what we find to be the most stable structure.

Results for all defects are summarized in Fig.4.5. The ANNK and BOP potentials both fail to produce meaningful results for defects in As and Ga, yielding extremely high, or negative formation energies – demonstrating the unphysical results that can be produced by an empirical forcefield outside of the range of configurations it is fitted for. On the other hand, the predictions for GaAs are closer to the DFT values. Our MLP can predict with a low error all the formation energies, although it tends to underestimates some particular defects. We also observe that, occasionally, the MLP geometry optimization converges to a structure having a small but significant distortion relative to the DFT geometry, which is associated with a further decrease in energy. When using the DFT-minimized structures for the comparison, the NNP is able to produce results closer to the DFT references, as shown in the S.I.of the original paper[156]. Given however that the overall error in terms of energy per atom is much smaller than the overall RMSE of the potential, we found that even adding more reference configurations could not improve the accuracy of the MLP, which underscores the need of including more specific training targets if one wants to achieve the ultimate accuracy in properties that depend on energy differences.

Surface energies and reconstructions

Figure 4.6 reports the rigid decohesion energies for all the stable surfaces of Ga, As, and GaAs, that are relevant to modelling fracture, and the surface-related phenomena that are relevant to modelling the synthesis of III/V nanostructures. Each supercell is computed at the equilibrium density of the corresponding potential. Even though in this case surface energies have reasonable values for all potentials, only the NNP reproduces quantitatively the DFT reference, and avoids an unphysical, near-discontinuous behaviour of the decohesion curve.

However, cleaved surfaces are usually not the most stable structure. The surfaces of semiconductors often undergo complex reconstructions, i.e. the atoms on the surface rearrange themselves and/or bind to one or more adatoms in the presence of a Ga or As atmosphere[181, 182]. Just as for silicon [183–185], surface reconstructions in GaAs have been subject of intense experimental and theoretical investigation, and many structures have been proposed and found for each of the high-symmetry orientations, i.e. [100], [110], and polar [111]. [181, 182, 186–188]

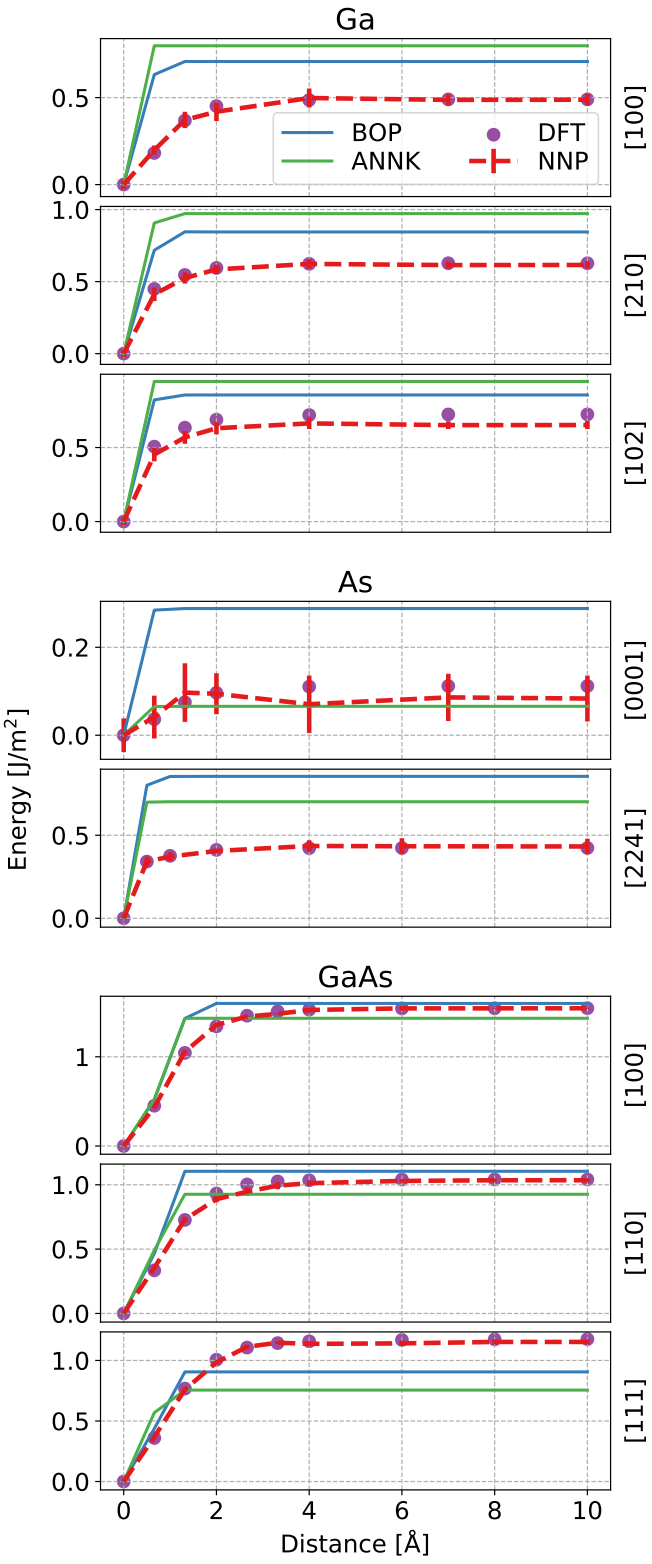

Figure 4.6 – Decohesion energy curves for the main high-symmetry surfaces in Ga, As, and GaAs. Results with BOP and ANNK potentials, our MLP and the DFT reference values are reported. Errorbars from the NNP reflect the distribution of estimates from the calibrated committee model.

When computing the surface free energy for the reconstructions, we have to account for the variation in stoichiometry of the configuration. We also assume that the surface is allowed to exchange atoms with a reservoir with a given chemical potential. The equilibrium free energy is obtained as

$$\gamma_{\text{surf}} A = E_{\text{surf}} - \sum_i \mu_i N_i \tag{4.1}$$

where N_i is the number of atoms of the species i in the system, and μ_i its chemical potential in the reservoir. The upper limit of the chemical potentials for each species is that of the respective condensed phase, as $\mu_i < \mu_{i(\text{bulk})}$. Since we know that in thermodynamic equilibrium the sum of chemical potentials of As and Ga must be equal to the bulk energy per GaAs pair

$$\mu_{\text{Ga}} + \mu_{\text{As}} = \mu_{\text{GaAs}} = \mu_{\text{Ga(bulk)}} + \mu_{\text{As(bulk)}} - \Delta H_f \tag{4.2}$$

where ΔH_f is the formation energy of GaAs from bulk Ga and As. Then, we can rewrite our range of chemical potentials in terms of variation of the chemical potential for a single species, which we choose to be As following the other references167182

$$-\Delta H_f < \mu_{\text{As}} - \mu_{\text{As(bulk)}} < 0. \tag{4.3}$$

Finally, we compute the surface free energy as

$$\gamma_{\text{surf}} A = E_{\text{surf}} - \mu_{\text{GaAs}} N_{\text{Ga}} - \mu_{\text{As}} (N_{\text{As}} - N_{\text{Ga}}) \tag{4.4}$$

In the case of a cleaved surface we have $(N_{\text{As}} - N_{\text{Ga}}) = 0$, thus leaving with a quantity that is independent of the chemical potential.

Most of these reconstructions involve charge redistribution between the surface and the bulk. In a typical slab supercell geometry, this requires introducing additional atoms to artificial balance the total charge (e.g. saturating dangling bonds with H atoms), and/or performing DFT simulations for charged systems. This poses a challenge for interatomic potentials, such as empirical forcefields and MLPs, whose parametrization relies only on the nuclear coordinates, and do not allow varying the overall charge. While it would be possible to compute MLP results for the [100] and polar [111] surfaces, and compare them with neutral-slab DFT simulations, the results would not be physically significant. As such, we compute and present results for the reconstruction of the [110] surface, the only one which is neutral in every case. As shown in Fig. 4.7, the MLP reproduces accurately the DFT results; the BOP also predicts qualitatively the correct ordering of surface reconstructions, while the ANNK potential incurs a large error in predicting the stability of the As-terminated reconstruction, and therefore incorrectly predicts the cleaved surface to be the most stable across all values of μ_{As} we consider.

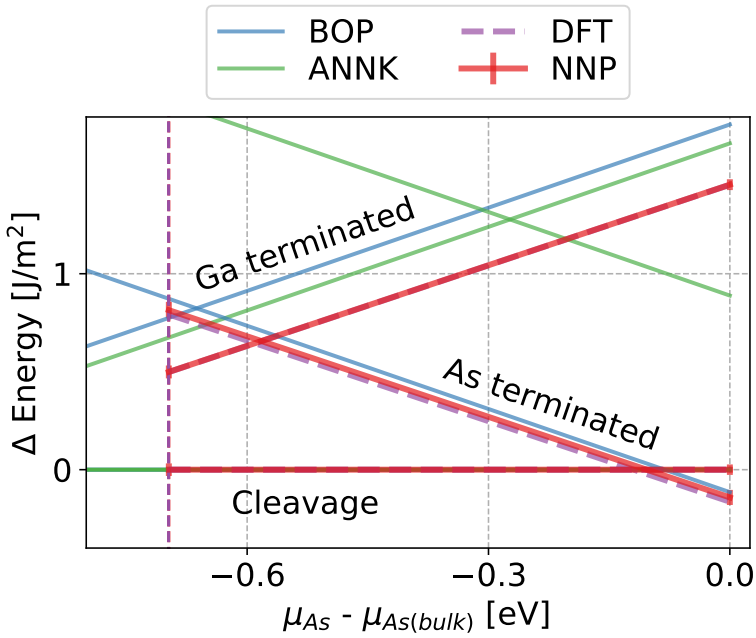

Figure 4.7 – Difference between the surface energy of the various GaAs [110] reconstructions and the cleaved surface plotted against the chemical potential of As. As indicated in Eq. (4.3), the physical values of μ_{As} vary within a range that depends on the potential, from zero down to $-\Delta H_f$, which is 0.7eV for the NNP and DFT, and approximately 0.9 for the two empirical potentials. Both the NNP and the BOP potentials recognize the correct stable structures in the observed range, while the ANNK potential finds the cleaved surface as the most stable across all values of the chemical potential.

Generalized stacking fault energy

Surface energies play an important role in the brittle fracture behaviour of a material. The generalized stacking fault (GSF) surface, instead, describes the energy cost associated with the sliding of two atomic planes, which is connected to plastic deformation, and the formation and dynamics of dislocations. We consider the [111] GSF surface gliding in the $\langle 11\bar{2} \rangle$ direction. We use a 24 atoms surface and the tilted supercell approach[189] to estimate the GSF energy profile (Figure 4.8). All the curves exhibit a similar overall glide barrier, but only the MLP reproduces qualitatively and quantitatively the DFT results. The ANNK potential predicts a flat-top, non-smooth GSF profile, while the BOP predicts an incorrect asymmetry of the path.

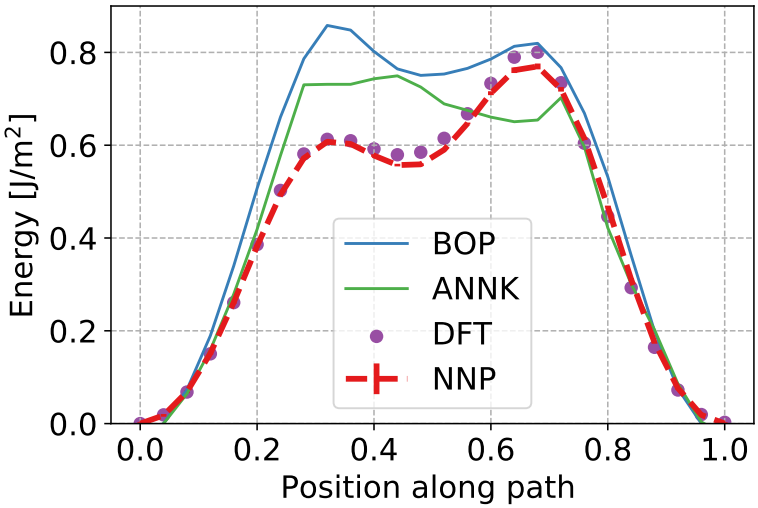

Figure 4.8 – Generalized stacking fault energy profile for the [111] surface gliding along the $\langle 11\bar{2} \rangle$ direction. The BOP, ANNK and MLP are compared to DFT reference calculations.

4.4 Finite-temperature properties

Having demonstrated the accuracy of the MLP for quantities that can be computed from static lattice calculations, and for which a direct comparison with DFT reference values is simple, we now move to consider finite-temperature properties, that require large simulation boxes and long sampling time, and that would be prohibitively demanding when performed with *ab initio* molecular dynamics. We investigate a broad temperature range, from 20 to 1600K, that covers both a cryogenic regime, which is well below the Debye temperature and requires a quantum mechanical treatment of the ionic degrees of freedom, up to the melting point of the highest-T_m phase, i.e. GaAs. Even though some of the quantities we compute can be obtained with approximate, perturbative methods at smaller computational cost, we report the fully anharmonic estimate using MD and path integral MD simulations, which are made feasible by the use of a MLP. Even though our results reflect accurately the thermodynamics of the MLP, which in light of the validation in Section 4.3 is likely to reproduce the DFT predictions,

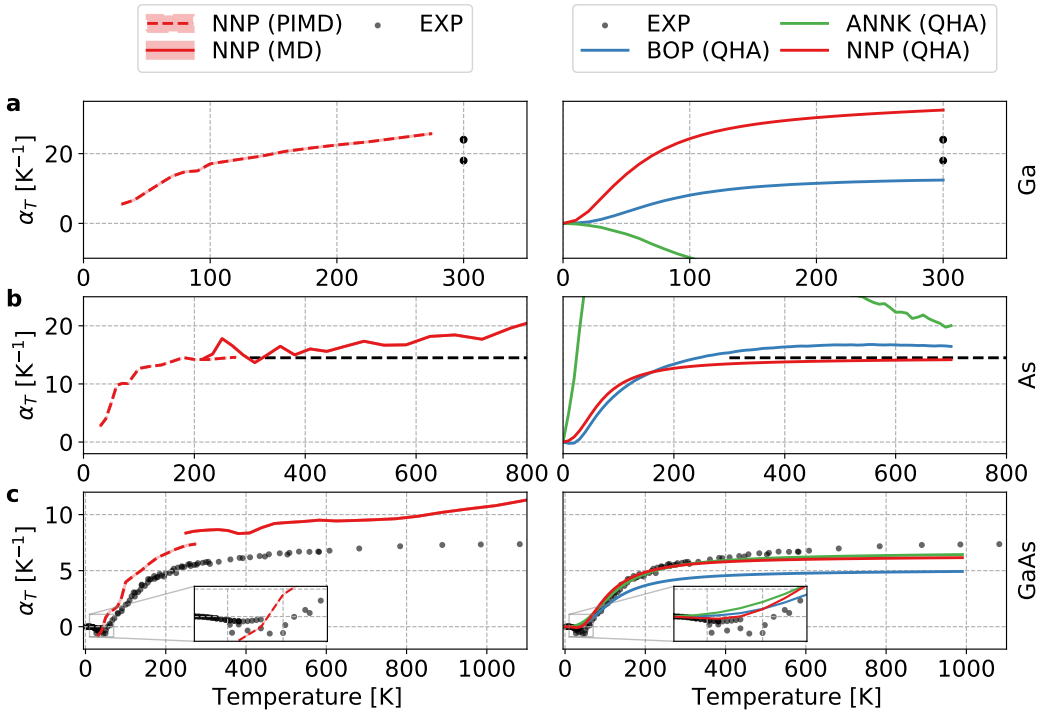

Figure 4.9 – The isotropic thermal expansion coefficient computed for Ga (a), As (b), and GaAs (c). On the left side we provide the results obtained with PIMD (up to 300 K) and MD (from 200 K onward) for the NNP, while on the right side we compare the three potentials at the QHA level. The inset in the (c) panel provides a clear view of the behaviour at low temperature of the three potentials. In this regime, our NNP is the only potential able to recover the negative thermal expansion coefficient. The experimental value of Ga is presented as the range between the maximum observed value and the minimum[190], while for As it is provided as an average over a large range of temperatures As[191]. Various sources are used for the experimental thermal expansion of GaAs[192–197]

we expect significant deviations from the experimental values, due to the shortcomings of the reference electronic structure methods. Still, the combination of a MLP and accurate finite-temperature sampling makes it possible to improve substantially the accuracy relative to existing empirical force fields. It should be noted that in the following sections the uncertainties presented for the properties arise from the finite time of the simulations and are computed by block averaging the simulations to account for the time correlation of the data.

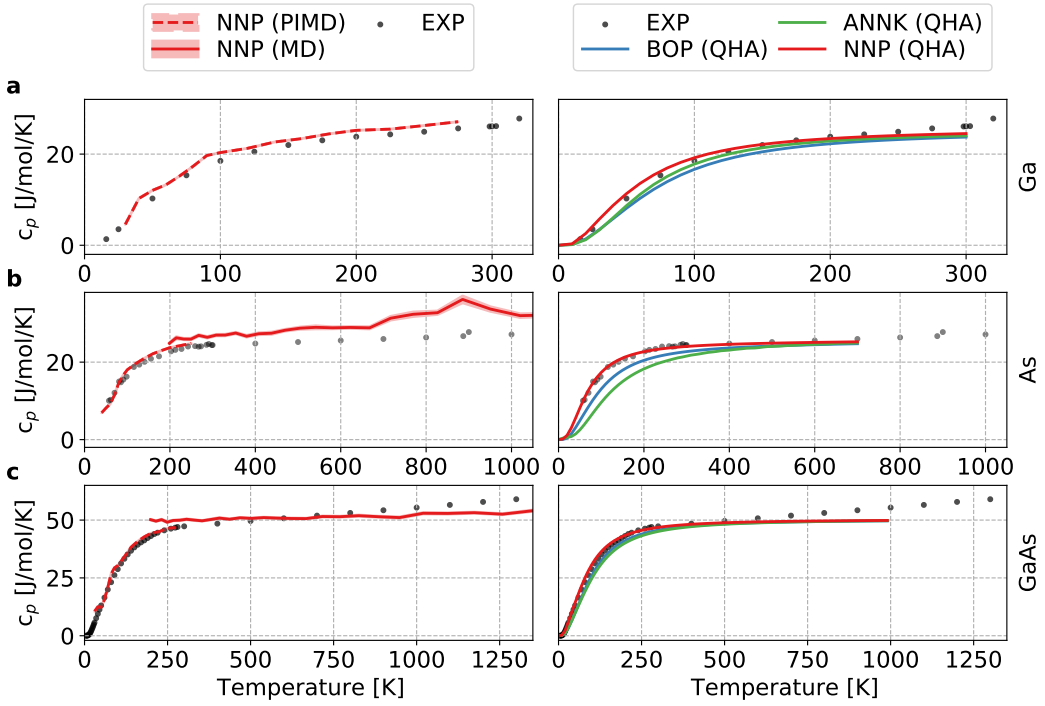

Figure 4.10 – The constant pressure heat capacity coefficient computed for Ga (a), As (b) and GaAs (c) with PIMD (up to 300 K) and MD (from 150 K onward) for the NNP and with the QHA for the BOP and ANNK potentials. MD simulations can only predict the classical value of the heat capacity, whereas with the explicit inclusion of NQEs we can recover the quantum behaviour. Experimental values are reported for Ga[198], As[199, 200], and GaAs[201–204]

4.4.1 Solid properties

We present the results of simulations of the solid phases for temperatures from 20 K up to the melting point for Ga, As, and GaAs, computed over a fine grid of temperature values. Based on this set of simulations, we compute and discuss bulk thermophysical properties such as heat capacity and thermal expansion for every phase which is stable at room temperature conditions.

The isotropic thermal expansion is computed by comparing the equilibrium volumes between simulations ran at subsequent temperatures (using PIMD and MD simulations separately), while we compute the heat capacity using the variation of the enthalpy with respect to the

temperature. The same quantities are also computed with the quasi-harmonic approximation (QHA) as implemented in Phonopy[205]. In the following figures, we will be presenting on the left side the results obtained with MD for our NNP committee, while on the right side the comparison with the empirical potentials at the QHA level.

The isotropic thermal expansion coefficients vs temperatures are presented for Ga, As, and GaAs in fig. 4.9 for all three potentials. The ANNK potential shows an unusual profile of the thermal expansion of bulk As and bulk Ga (figs 4.9a and 4.9b), while the BOP is able to follow more closely the experimental values. For the case of bulk As, the MD simulations run with the two potentials show that the ANNK potential is unstable when running beyond 800 K, while the BOP is unstable at 1400 K and never undergoes a spontaneous solid-liquid transition. Experimentally, a single result is found for the isotropic thermal coefficient that can be compared to our analysis, and is given as an average value for temperatures between 300 K and the melting point. Finally, the GaAs results are shown in fig. 4.9c. GaAs in its zincblende form has a negative thermal expansion coefficient at low temperature[206], which is predicted by our potential both in the MD simulations and the QHA, but not by the other models, neither for QHA nor for PIMD (which have been computed, but not presented to improve the clarity of the figure). At higher temperatures, our potential seems to be slightly overestimating the expansion of the solid in the MD simulations. The QHA results follow rather closely those obtained with MD simulations at lower temperatures, while slightly deviating at higher ones, when anharmonic contributions become relevant.

The results concerning the heat capacity converge, as expected, to the corresponding classical value. However, the BOP and the ANNK potentials deviate from the experimental values at low temperatures, particularly for Ga and As (Fig. 4.10a and Fig. 4.10b respectively). It should be noted that at higher temperatures we would observe a deviation for Ga of the MD results against the experimental ones, due to the electronic contribution to the heat capacity. However, since Ga melts at 303 K, this effect is not yet relevant, although for other systems this can actually be computed using an integrated ML model such as the one of Lopanitsyna *et al.*[207]. Moreover, as expected, classical MD is not able to reproduce the quantum behaviour of the heat capacity, that can be recovered only by using PIMD simulations, as it can be seen clearly in the calculations run with the NNP for all three phases.

4.4.2 Liquid properties

We turn now our analysis to properties related to the liquid part of the phase diagram, which are investigated using MD simulations of large supercells for long trajectories. In this section we present the results for the density of Ga, the radial pair distribution functions of liquid Ga, As, and GaAs, diffusion coefficients and viscosities of the liquid phases of Ga and GaAs. We also compare the values predicted by our potential with the ones that are reported experimentally, where available.

The density of liquid Ga is presented in fig 4.11, where it is clear that our potential qualita-

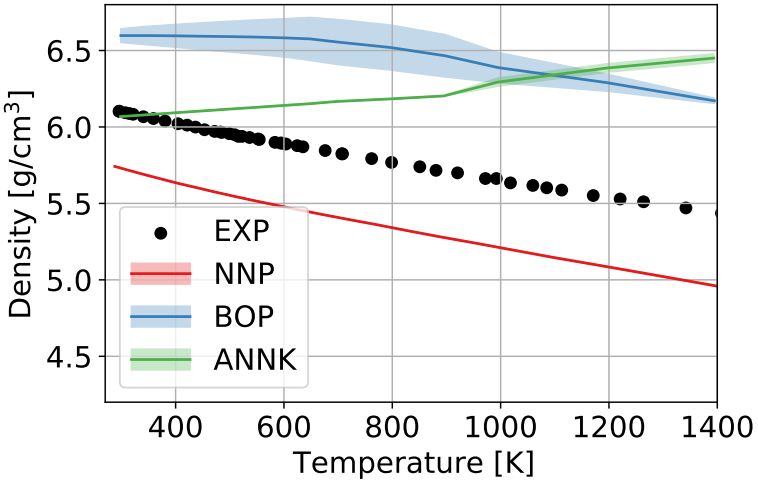

Figure 4.11 – The density of liquid gallium as predicted by the NNP compared to the experimental values. The values from the simulations with the BOP and ANNK potentials are added, but the density refers to the solid phase till 800 K, where a discontinuity is observed for both potentials.

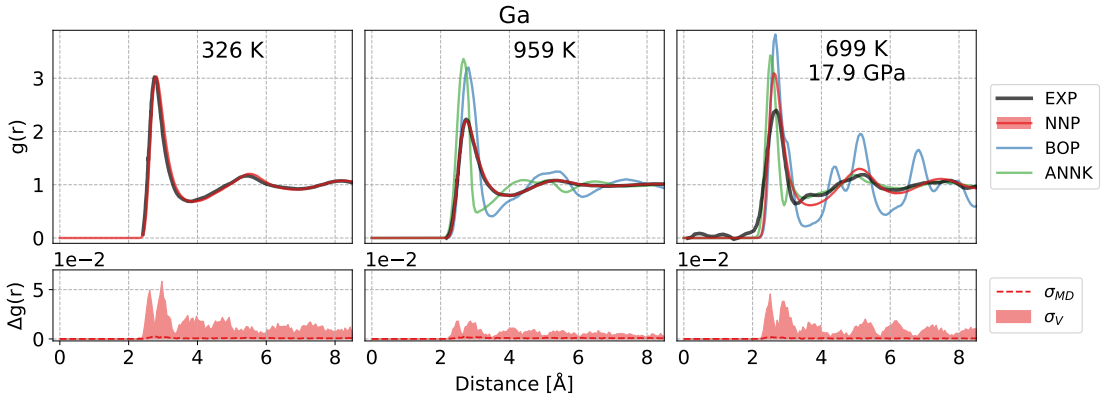

Figure 4.12 – The radial pair distribution function computed for the various potentials and compared to the experimental values at ambient pressure[208] and at high pressure[209]. The g(r) of the empirical potentials are not reported in the first panel because the structures remain solid at the reported temperature. The bottom panels present the uncertainty arising from the use of a ML model compared to the statistical uncertainty due to the finite time of the simulation.

tively reproduces the experimental values, although underestimating it by about 8%. This underestimation may be in part due to the lack of dispersion interactions, that have been found to play an important role in materials composed by row IV elements and above[210]. Investigating the role of dispersion by re-training the NNP against vdW-corrected functionals may be an interesting future line of research. Both the BOP and ANNK do not follow even qualitatively the experimental density. Both the empirical potentials are actually solid in the region T < 800 K and become liquid only afterwards. A discontinuity in the first derivative of the density can be observed around that temperature for both potentials. As predicted by the thermal expansion calculations of solid Ga (fig. 4.9a), the ANNK potential actually shows a compression of the box as the temperature increases.

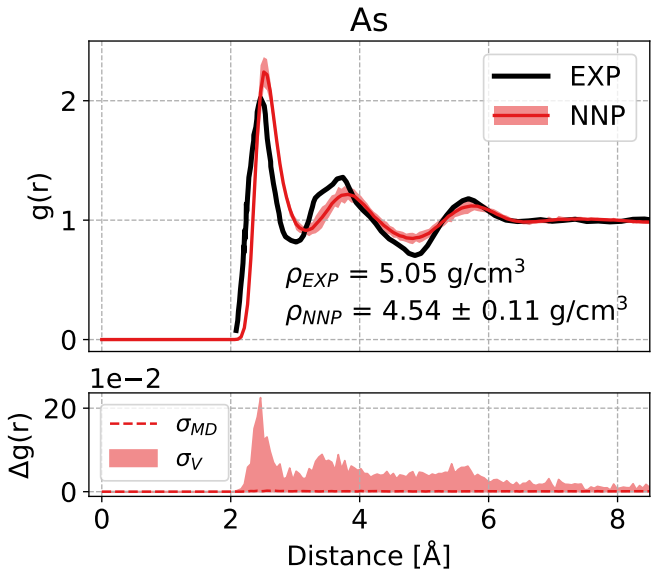

Figure 4.13 – The radial pair distribution function computed for the NNP at 1098 K and 4.8 MPa, compared to the corresponding experimental values available[211]. The bottom panel presents the uncertainty arising from the use of a ML model compared to the statistical uncertainty due to the finite time of the simulation.

Then, we report the radial pair distribution function, which we will refer to as g(r). We run simulations of liquid Ga at three different conditions in fig. 4.12, liquid As in fig. 4.13, and liquid GaAs in fig. 4.14 for a comprehensive view of the potential. For As and GaAs, we also provide the equilibrium density at the given temperature. We do not provide the results for the BOP and ANNK potentials in most cases because they are not liquid in the range of temperatures that we consider for the MD simulations (e.g. the BOP and ANNK melting points of GaAs are reported to be around 1950 K). In these figures we also provide a comparison between the thermodynamic uncertainty obtained by reweighting the trajectories for each potential in the committee (here called σ_V following the same notation as Ref. 105) and the statistical uncertainty due to the finite time of the simulations (indicated as σ_{MD}).

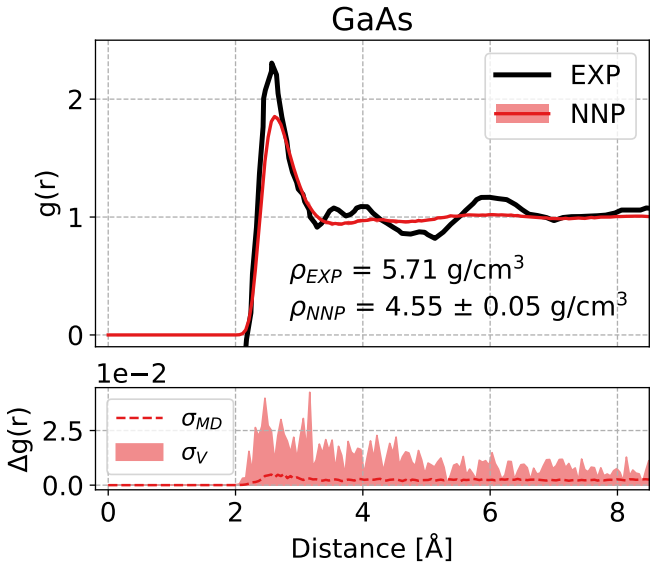

Figure 4.14 – The radial pair distribution function computed for the NNP at 1550 K and compared to the experimental values available[212]. The empirical potentials are omitted, since they are solid at this temperature. The bottom panel presents the uncertainty arising from the use of a ML model compared to the statistical uncertainty due to the finite time of the simulation.

In the case of liquid Ga, our potential is able to reproduce with striking accuracy the g(r) both at low and high temperature, similarly to the results of other *ab initio* studies run with GGA[209, 213] or LDA potentials[214]. At higher pressure, we obtain a good agreement with the experimental data, very similar to that of other studies with GGA potentials[209]. Both empirical potentials fail to provide a meaningful description of the liquid environment at 959 K, while the ANNK potential has a reasonable, but too ordered, g(r) at high pressure.

Arsenic does not undergo melting at atmospheric pressure, becoming directly gaseous at 887 K. Therefore, in fig. 4.13 we run simulations at T = 1098 K and p = 4.8 MPa, where it is liquid, to compare to the g(r) obtained experimentally in the same conditions[211]. Our prediction is less accurate compared to the Ga one, but we are still able to recover the position of the peaks in the liquid, although the shoulder in the second peak seems to be entirely missing. We are also slightly underestimating the density of the liquid.

The results obtained for liquid GaAs at 1550 K are presented in fig. 4.14, where we observe a reasonable agreement with the experimental data[212], although it is not entirely clear whether the splitting of the peaks in the experiments is a physical feature, possibly due to the undercooling of the sample, or due to the noise. Other *ab initio* simulations in literature also do not show the same splitting of the second peak[215–217].

The excessive smoothing of the $g(r)$ of both As and GaAs, and the underestimation of the density, are probably a reflection of the limitations of the ab initio reference rather than of the

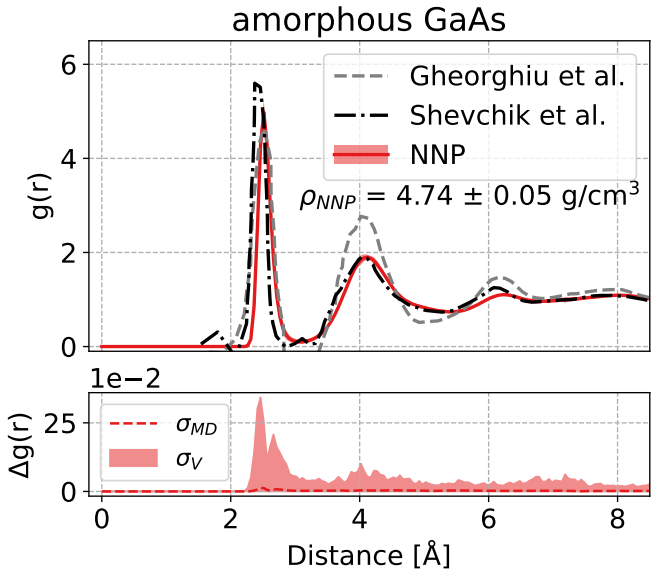

Figure 4.15 – The radial pair distribution function of amorphous GaAs computed for the NNP at 300 K after a slow quenching starting at 1800 K, compared to the experimental values of Gheorghiu *et al.*[218] and Shevchik *et al.*[219]. The bottom panel presents the uncertainty arising from the use of a ML model compared to the statistical uncertainty due to the finite time of the simulation.

NNP, as evidenced by the small estimated σ_V. As in the case of Ga, incorporating dispersion interactions might be a possible strategy to improve the accuracy of DFT energetics.

Finally, we provide predictions of the g(r) for amorphous GaAs, which is not included in the training set. We prepare the cell by first running 5 trajectories with different initial structures made of 1000 atoms where we quench the liquid from 1800 K to 300 K over 1 ns. Then, we compute the g(r) on 1 ns-long simulations of the final structure, at 300 K. The results presented in fig. 4.15 refer to the average g(r) of the 5 different simulations, compared to the experimental results of Gheorghiu *et al.*[218] and Shevchik *et al.*[219]. Overall, we find a good agreement with the experimental values, with very similar positions of the peaks. We also observe that the uncertainty over the energies is, on average, only twice as large as the same uncertainty computed for liquid GaAs, which translates in an uncertainty in the g(r) that is larger, but still negligible. In fact, the uncertainty of the g(r) of amorphous GaAs is comparable with the one we obtais for liquid As, which is explicitly included in the training set. Overall, we believe that the potential is able to produce reasonable results for the amorphous system, despite the lack of dedicated structures in the training set. For a study dedicated to the amorphous phases, however, we would recommend to extend the training set incorporating explicitly amorphous structures.

The third quantity that we compute is the surface tension, that we obtain by running 1 ns long simulations of the interface between bulk liquid and vacuum in a large orthorhombic

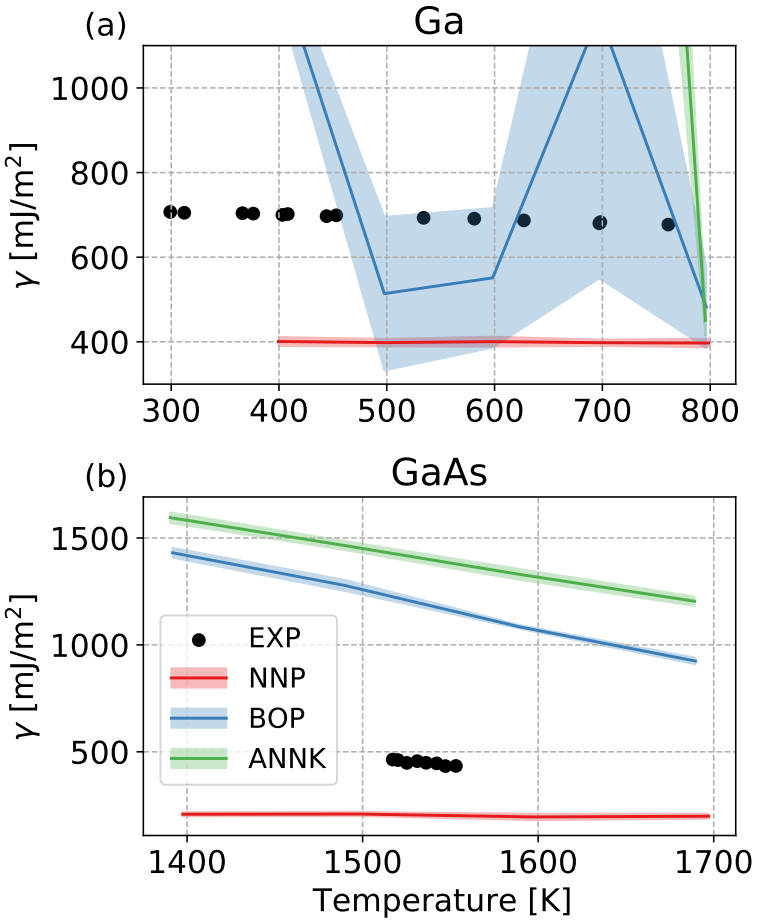

Figure 4.16 – The surface tension of Ga in panel (a) and GaAs in panel (b) for increasing temperature, compared to the experimental values[220, 221]. As previously mentioned, the empirical potentials tend to overestimate the melting point, so the results with these potentials refer to the solid phase

supercell with 1568 atoms for Ga and 1728 atoms for GaAs, with approximate dimensions of 32x32x100 Å at varying temperatures. To estimate the surface tension we use its relation to the diagonal elements of the stress tensor for the described box, as in eq. 4.5, where the 1/2 factor at the beginning accounts for the presence of two interfaces between liquid and vapour.

$$\gamma = \frac{1}{2} L_z [P_{zz} - \frac{1}{2}(P_{xx} + P_{yy})] \tag{4.5}$$

Our NNP seem underestimates the surface tension for both Ga and GaAs, as seen in fig. 4.16. To investigate the discrepancy, we check additional structures related to these trajectories and find errors of 1 to 2 meV/atom between our NNP and the DFT results, well below the overall RMSE of the potential, suggesting that the discrepancy might be due to the reference calculations and not to the accuracy of the fit.

The last properties that we present here are the diffusion coefficients and the viscosities for the liquid phases of Ga and GaAs. To obtain these, we run several simulations with cubic boxes with a side of 30 Å, relaxed at the equilibrium density. At each temperature we run 20 simulations starting from different initial configurations (at equilibrium density) for 200 ps each in the NVT ensemble using a weak SVR thermostat[154]. We compute the mean square displacement as an average over the 20 trajectories and obtain the diffusion coefficient for the finite-size system (which we indicate with the PBC subscript) with adequate statistics.

$$D_{\mathrm{PBC}} = \lim_{t\to\infty} \frac{1}{6t} \langle \sum_{j=1}^{N} (r_{j,i}(t) - r_{j,i}(0))^2 \rangle \tag{4.6}$$

Since the diffusion coefficient is known to be heavily affected by the size of the box[222], we determine the diffusion coefficient of the infinite bulk system by adding the correction factor computed by Yeh and Hummer[223], as

$$D_\infty = D_{\mathrm{PBC}} + \frac{\xi k_B T}{6\pi\eta L} \tag{4.7}$$

where ξ is a dimensionless constant equal to 2.837297 for cubic simulations boxes, η is the viscosity and L is the side of the box. The viscosity, which is independent from the box size[223–225], is obtained from the autocorrelation function of the off-diagonal elements of the stress tensor computed in the same simulation of the diffusion, as

$$\eta = \frac{V}{k_B T} \int_0^\infty \langle P_{\alpha\beta}(t) \cdot P_{\alpha\beta}(0) \rangle \mathrm{d}t. \tag{4.8}$$

An alternative method to compute the diffusion coefficient for the infinite bulk system is to compute the coefficient for supercells of increasing size, then extrapolate the value for an infinite supercell[225]. Therefore, we run additional calculations for smaller cells, to compare the values obtained with the two methods and found them to be in good agreement, as seen in Fig. 4.19.

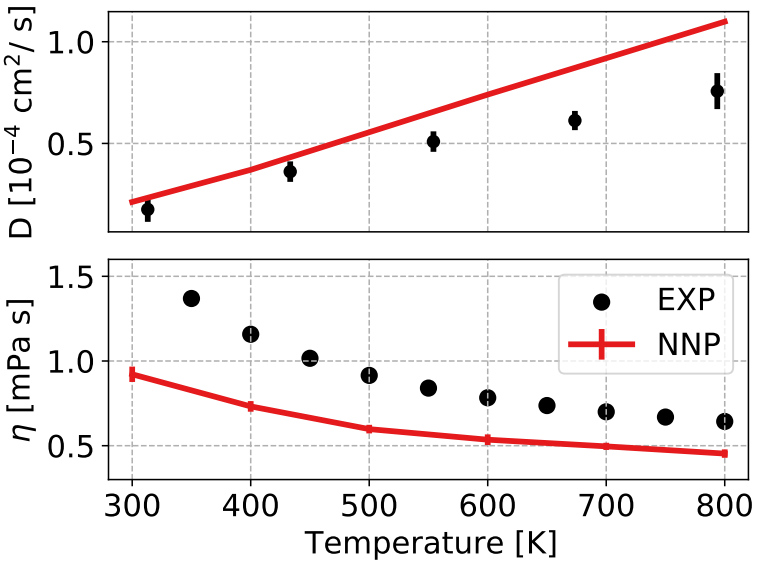

Figure 4.17 – The diffusion (top) and the viscosity (bottom) computed for liquid gallium at increasing temperatures. 20 simulations are run for each point and the spread in the predicted values is reported with the errorbars. Although the viscosity and the diffusion are related, we use experimental values reported from separate sources for the viscosity[226] and the diffusion[227]

While we are consistently underestimating the viscosity (and conversely overestimating the diffusion coefficient), we are able to recover the qualitative behaviour at lower temperature for gallium, as seen in fig. 4.17. The underestimation of the viscosity at a given temperature is to be expected given the lower value of the melting point, that we discuss next. A large underestimation of the viscosity is also observed in the case of GaAs (fig. 4.18), which also has a much lower melting point (1200 K against 1550 K observed experimentally). However, the fact that even at the lowest temperature we do not observe the sharp increase in viscosity that is observed in experiments when approaching the melting point suggests that our NNP should be used with care when investigating dynamical properties for molten GaAs.

4.4.3 Binary phase diagram

In the introduction we mentioned our aim to produce an accurate and transferable potential. Until now we have computed a number of properties with the purpose of showing the accuracy of this potential, albeit limited by the underlying DFT reference. Here we want to provide a compelling proof of the transferability of the potential, which is of utter importance when studying technologically relevant phenomena in varying conditions of temperatures, pressure, and stoichiometries.

Providing a full description of the phase diagram is a definitive test of reliability of the potential, since not only we are performing simulations at different stoichiometries, but every simulation

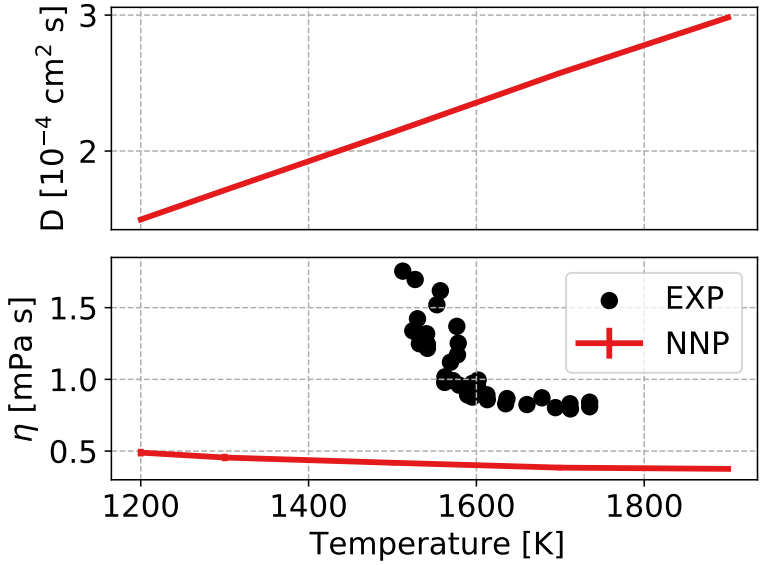

Figure 4.18 – The diffusion (top) and the viscosity (bottom) computed for liquid gallium arsenide at increasing temperatures. 20 simulations are run for each point and the spread in the predicted values is reported with the errorbars. The simulations are compared to the reported experimental values for the viscosity[228], whereas no direct measurement of the diffusion is found in literature.

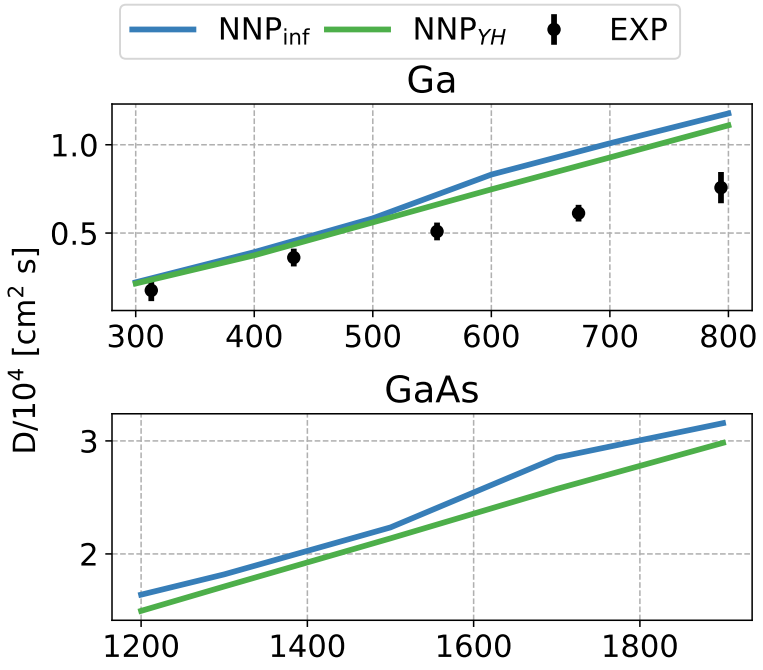

Figure 4.19 – A comparison between the diffusion computed with Eq. 4.7 and the one obtained by extrapolation to an infinitely sized cell[225]. The results are in good agreement between the two formulations, with the latter providing slightly higher values, mostly at the higher temperatures.

that we run contains both solid and liquid bulk, together with their interface.

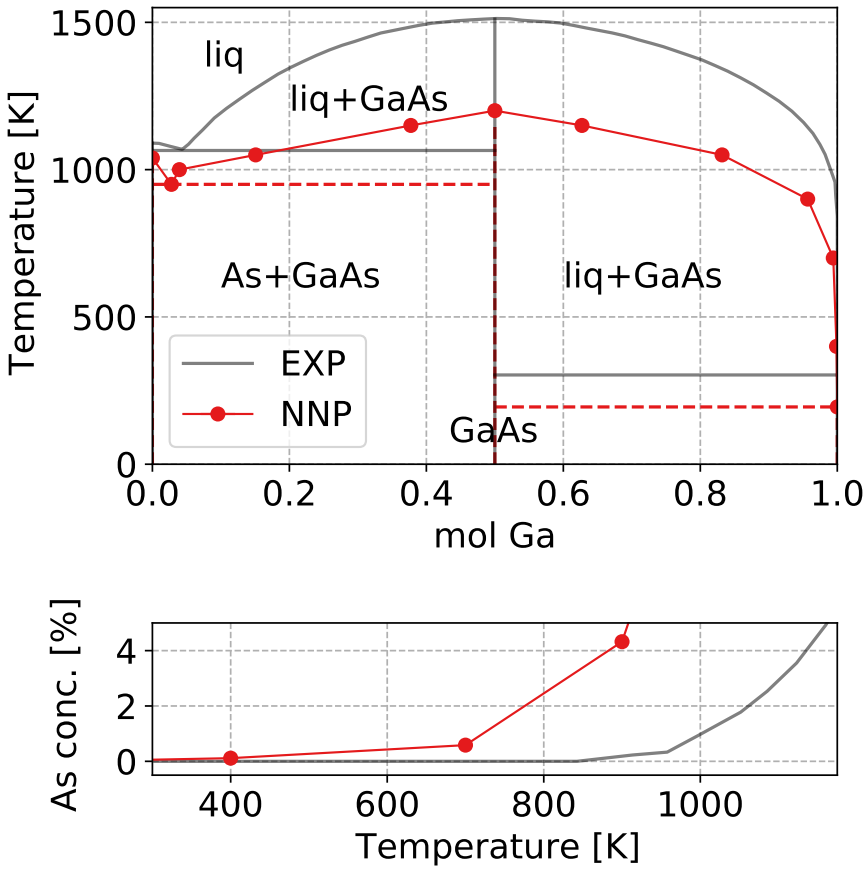

Figure 4.20 – Top: the binary phase diagram for GaAs as predicted from our NNP, compared to the experimental one. We used interface pinning simulations to find the melting point for the pure Ga, As, and GaAs cases, while the other points are measured using mixed Monte Carlo - MD simulations at different stoichiometries and temperatures. Bottom: saturation concentration of As in liquid Ga as a function of temperature in the region of low temperatures.

In figure 4.20 we see our predicted phase diagram, compared to the experimental one. At first glance we observe a good agreement in the shape of the two curves, with a low solubility of As predicted at low temperatures in the high-Ga region (highlighted in the bottom figure), followed by an almost flat central part. We also observe an eutectic point at T = 950 K and $x = 0.03$, not far from the experimental value of T = 1083 K and $x = 0.05$. The melting points are predicted to be 1039 K, 1200 K, and 195 K for As, GaAs, and Ga respectively, which are in relatively good agreement with the experimental values of 1090 K, 1511 K, and 303 K. It is important to stress that the discrepancy is probably due, in large part, to the underlying electronic-structure reference. In fig 4.21 we show the use of the thermodynamic uncertainty quantification scheme from Ref. 105 to determine the error due to the fit of the NNP. We find 1039 ± 51 K, $1200 \pm 5K$, $195 \pm 24K$ for As, GaAs and Ga: except for the case of As, the error

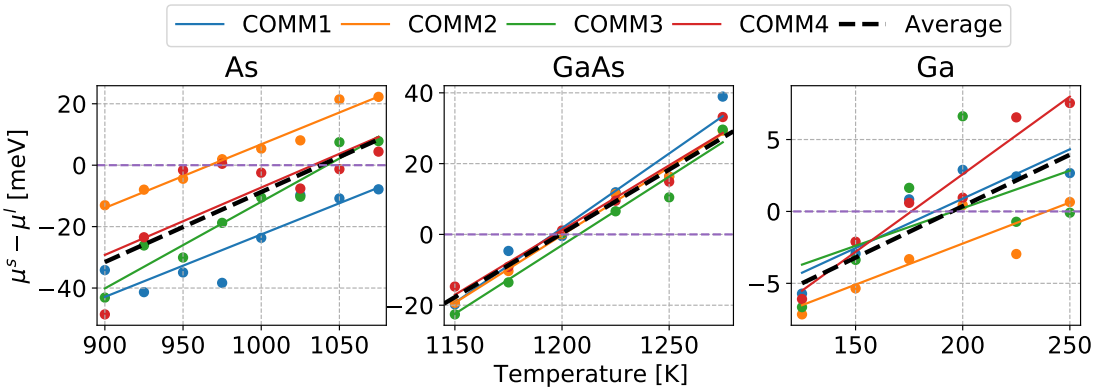

Figure 4.21 – Determination of the error in the melting point of As, GaAs, Ga associated with the NNP fit, using the uncertainty estimation scheme in Ref. [105]. The chemical potentials at each temperature are computed separately for each member of the NNP committee, using a reweighting scheme that makes it possible to obtain the four values by reweighting the trajectory driven by the committee average. This approach makes it possible to estimate the melting point of each potential, and determine the uncertainty in melting point by the spread in the four predictions.

associated with the machine-learning approximation is a small fraction of the discrepancy with experiments.

The points shown in figure 4.20 are computed with two different methods. The melting point of pure Ga, As, and GaAs is obtained with the interface pinning method, as described by Pedersen *et al.*[146]. The remaining points in the liquidus are obtained by running large supercells at various temperatures and stoichiometries, measuring the concentration of the two species in the liquid at equilibrium. In order to speed up the equilibration of the concentrations we add a Monte Carlo step on top of the MD calculation.

For the interface pinning simulations we first determine an optimal collective variable that can distinguish solid and liquid phases, and then run multiple simulations at regular temperature intervals for a large supercell in the Np$_z$T ensemble. To run these trajectories, we use the open source PLUMED library[149, 229] to add the bias potential, in addition to i-PI and LAMMPS. After obtaining the mean value of the collective variable at each temperature, we determine the melting point by fitting the chemical potentials to a line. The temperature at which we find a chemical potential of $\mu = 0$ is the melting point of the system[230]. To compute these trajectories we use the locally averaged Steinhardt parameters introduced by Lechner and Dellago[147, 148] q4 (for As and GaAs) and q6 (for Ga) as collective variables for the system.

The mixed Monte Carlo - MD simulations are run within i-PI. At every MD step, we attempt to swap 50 (on average) random Ga-As pairs in the system. The particle exchange is then accepted or rejected using a Metropolis criterion. The supercells used in this case are composed by 50% solid GaAs and 50% liquid Ga_xAs_{1-x} . The stoichiometry of the liquid is determined such that the total stoichiometry of the system varies between $0.25 < x < 0.75$. The simulations are

divided in a first NpT part, for 10 ps, to find the equilibrium density for the solid, and a second Np$_z$T part, run for 200 ps. In this second trajectory, we allow the system to equilibrate for the first 100 ps, and then measure the average concentration of As and Ga in the liquid for the remaining 100 ps.

This method works without the need to introduce an external potential to pin the interface because we are considering a binary mixture[231]. For an unary system, the chemical potential between the solid and liquid at the melting point is 0, thus the need to introduce the bias potential to avoid a random walk of the interface, which could result in complete freezing or melting. For the mixture, however, the curvature of the free energy at the interface depends on the composition of the two coexisting phases as

$$\left(\frac{\delta^2 G}{\delta f^2}\right)_{p,T,x} = \frac{(x_s - x_l)^3 \mu_l''(x_l)\mu_s''(x_s)}{(x_s - x)\mu_s''(x_s) + (x - x_l)\mu_l''(x_l)}, \tag{4.9}$$

where f represents the fraction of solid phase in the system, x_s, x_l and x are the compositions of the solid, the liquid, and the overall system, and $\mu_{l,s}$ is the chemical potential of the liquid and the solid, respectively. Thus, in any case in which solid and liquid have different equilibrium composition, there is a positive curvature that acts as a restoring force against fluctuations of the dividing surface, acting effectively as a pinning potential that keeps the solid fraction fluctuate around the value consistent with the lever rule. Measuring the mean composition of the two phases in equilibrium makes it possible to determine the position of the solidus and the liquidus. The derivation is provided in Ref. 231.

4.4.4 Beyond potentials

It is worth mentioning that the same transferability that is achieved for the potential also applies to other properties, such as those afforded by next-generation integrated ML models that also target predictions of the electronic structure of materials. As a proof of principle, we build a model of the DOS using the same protocol discussed in Sec. 3.3.4, using the Kohn-Sham eigenvalues from the same structures included in the training set for the potential, an additive decomposition of the DOS and a prediction of local contributions in terms of a multivariate Gaussian process regression and a description of atomic environments based on SOAP features[16], computed using the implementation in librascal[232]. As shown in Fig. 4.22, this DOS model gives accurate predictions of the single-particle energy states across the entirety of the phase diagram. Even though the limitations of DFT-PBE (which is known to underestimate the band gap in GaAs) make this preliminary model of limited utility, future work may build on our results to incorporate electronic-structure information at a higher level of theory, providing a full description of the stability and properties of the Ga$_x$As$_{1-x}$ system.

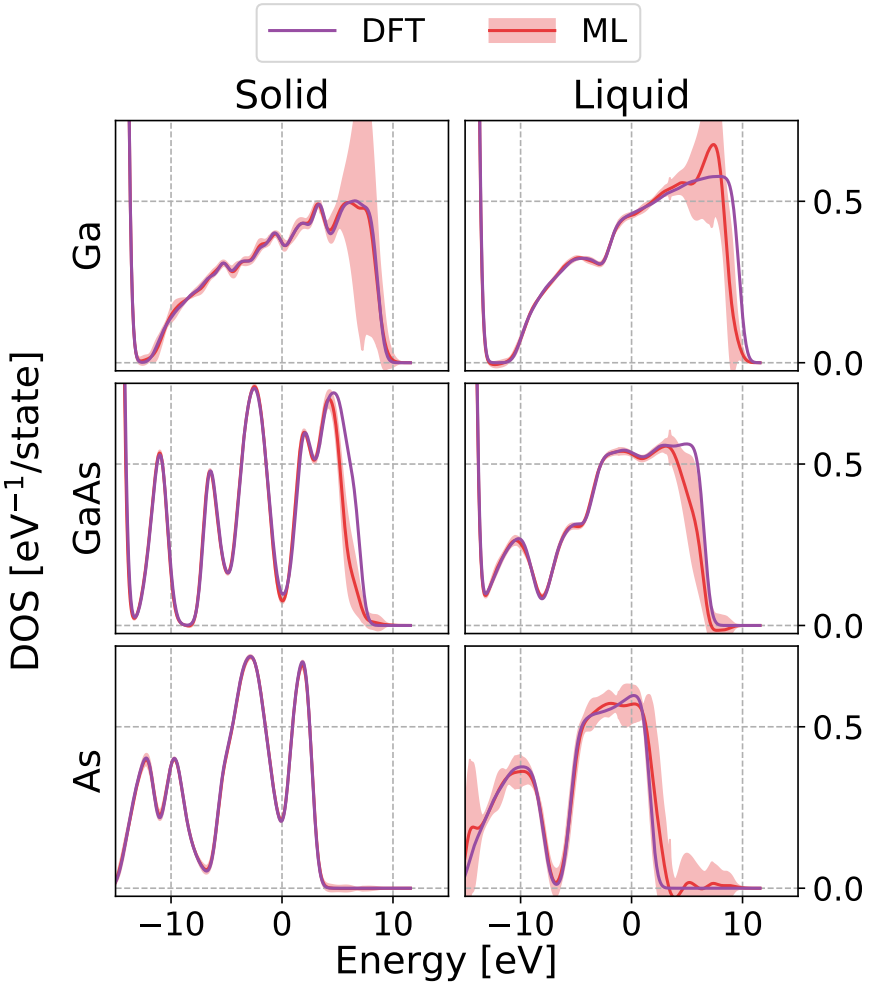

Figure 4.22 – The electronic density of states of liquid and solid Ga, As, and GaAs as predicted by a committee of 16 built following the same parameter choice as Sec. 3.3.4. All the curves are centered with respect to the Fermi energy, which represents the Energy=0 level.

5 Simulating the liquid-solid interface in GaAs nanowires[1]

5.1 Introduction

The use of MLIPs in our field is not restricted to theoretical methods or fitting of arbitrary databases, but has already shown that it can actually be used for simulations of systems of scientific and technological relevance. Some examples include potentials for silicon[29] and iron[27], that have been used to run simulations of tens and hundreds of thousands of atoms to understand the electronic transitions of disordered Si[52] and the mobility of screw dislocations in bcc Fe[233]. Other potentials have been fitted to reproduce and predict the experimental results of hydrogen (and deuterium) absorption on graphene[51] and to investigate the supercooled state of GeTe and $Ge_2Sb_2Te_5$ compounds[234, 235].

In this chapter we will make use of the potential that we have introduced in the last chapter to study a system of experimental interest, i.e. the liquid-solid interface of vapor-liquid-solid (VLS) grown nanowires (NWs)[236] (a schematic drawing of the system is presented in Fig. 5.1A). GaAs NWs are building blocks of future nanodevices, such as photovoltaic panels with high yield[237] and quantum dots[238, 239]. However, the conditions of their growth are not completely understood and it is common to observe the formation planar defects, such as alternating layers of different crystal structures during the growth[240, 241]. Therefore, being able to model this interface and the phenomena relevant to the growth, such as the contact angle[242, 243], would allow us to understand the best conditions to grow defect-free nanowires with the desired opto-electronic properties.

[1] The chapter is partially adapted from Ref. 151 and the second part is novel research not yet published. The author of the thesis has run and analyzed all of the simulations appearing in this chapter, with the help of two master thesis students, Sébastien Bienvenue for the first part and Tushar Thakur for the second part. The author of the thesis has not contributed on the experimental work, which has been entirely performed by Mahdi Zamani and others in the laboratories of profs. Fontcuberta i Morral and Cécile Hébert. The experimental set-ups used and a part of the quantitative analyses of the experimental findings is not shown here, and can be found in the original paper[151] and the relative S.I..

5.2 3D Ordering at the liquid-solid Polar Interface of Nanowires

We begin our investigation by focusing on the the bulk interface between liquid Ga and solid GaAs, in both ZB and WZ form. Recent advances in the experimental techniques allow to obtain clear images of the liquid-solid interface of VLS-grown nanowires, which we use as a guide for our simulations. Since the experimental images for this system are taken at ambient temperature, we expect no As atoms in the liquid due to the very low solubility of As in liquid Ga (as seen in the bottom panel of Fig. 4.20). Therefore, we also run our simulations in the case where no As atom can be found in the liquid Ga. In the next section, we will explore the results arising from the addition of an As atom in the liquid.

5.2.1 Experimental signatures of ordering at the liquid-solid interface

Usually, the liquid-solid boundary during the growth of GaAs nanowires through the VLS method is considered as a clear-cut, binary interface, with no regard for its structure. Modelling has mostly reasoned in terms of macroscopic parameters, including contact angle, the surface energies at the solid-vapor and liquid-solid interfaces and the chemical potentials[237, 244], while the atomistic nature of the participating parts has rarely been considered. However, it has already been observed in other systems that a solid surface can induce a local order to an adjacent liquid phase[245–247].

Scanning transmission electron microscopy images

In Fig. 5.1B and C, we show aberration-corrected high angular annular dark-field (HAADF) scanning transmission electron microscopy (STEM) micrographs of the interface of two NWs along the $[1\bar{1}0]$ zone-axis of zinc-blende (ZB) and $[11\bar{2}0]$ zone-axis of wurtzite (WZ). We mark the As and Ga atoms in blue and red, respectively.

The A-polar NW exhibits a pure ZB structure, whereas the B-polar NW exhibits a mixed phase structure, finishing with WZ. The interpretation of the chemical species is possible thanks to the dependence of the intensity to the atomic numbers. As expected, the solid shows strong peaks corresponding to the atomic columns, whereas, further from the interface, the liquid shows a uniform intensity because of its disordered nature. However, close to the interface, the liquid shows intensity fluctuations corresponding to an ordering. For the A-polar interface, this is limited to one additional layer, while several layers are observed for the B-polar interface. This is confirmed in the integrated HAADF intensity profiles shown on the right of panels 5.1B and 5.1C and we attribute it to a longer-range ordering on top of the B-polar interface. The first layer seems to be more clearly structured, with further layers becoming gradually more amorphous.

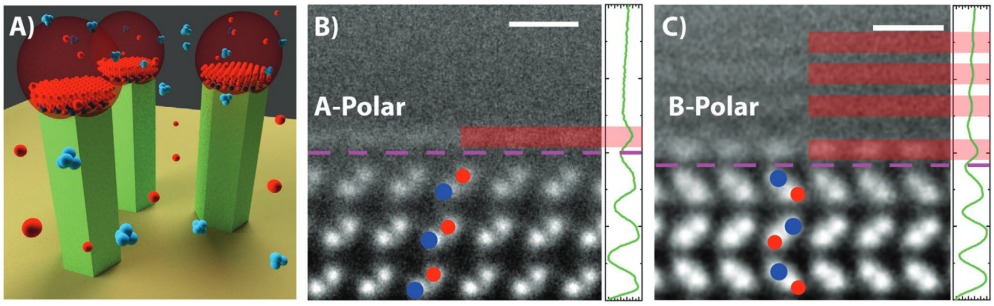

Figure 5.1 – A) Graphic representation of Ga-assisted growth of GaAs NWs: As_4 arrives on the Ga droplet and is dissolved. The Ga atoms at the interface with the GaAs are ordered following the underlying crystalline structure B-C) HAADF-STEM micrographs of the liquid-solid interface of the A and B-polar NWs observed along the $[1\bar{1}0]$ ZB and $[11\bar{2}0]$ WZ zone axes, respectively. The stacking for the A-polar GaAs is ABCABC, corresponding to the zinc-blende crystal structure, whereas the stacking in the tip of the B-polar NW is ABAB , which is indicative of the wurtzite crystal structure. The right-side of panels B and C show intensity profiles from the STEM images integrated along the direction parallel to the NW surfaces. The pink dashed line indicates the position of the interface, while the red rectangles highlight the position of the apparent layering. The white scale bar on the top right of panels B) and C) is 0.5 nm.

Electron energy loss spectroscopy analysis of the interface

We further analyze the liquid-solid interface using electron energy loss spectroscopy (EELS) hyperspectral mapping to probe the bulk plasmon response around the interface. Being related to the valence/free electron density and characteristics of a material, this technique can discriminate between different chemical/structural phases.

Low-loss EELS spectrum images of the interface were recorded using an atomically sized probe. As an example, Fig. 5.2A shows a HAADF-STEM image acquired simultaneously with the EELS signal from a spectrum image of an A-polar GaAs NW in contact with a Ga droplet. Fig. 5.2B depicts the corresponding spectral evolution integrated across a region of the interface, represented by the blue area shown in Fig. 5.2A.

The bulk plasmon excitation shifts smoothly from a broad peak in the GaAs to a sharper peak in the liquid Ga, over a spatial distance of ~10 nm. Reference spectra extracted away from the interface at positions deep in the Ga and GaAs phases (marked P1 and P3 on Fig. 5.2A), and a spectrum extracted from the interface (at position P2 on Fig. 5.2A) are presented in Fig.5.2 in panels C and D respectively. The spectrum at P2 might be assumed to be a linear combination of the contributions of the liquid and solid phases. However, when we perform an independent component analysis (ICA) of the system, a non-negligible residual is found. If a third ICA component is added, which is interpreted as the “ordered liquid” (OL) contribution, we are able to reconstruct the entire signal.

From this analysis, one can conclude that the ordered liquid has a distinct electronic nature, which can intuitively be correlated with its semi-structured nature. Considering this and

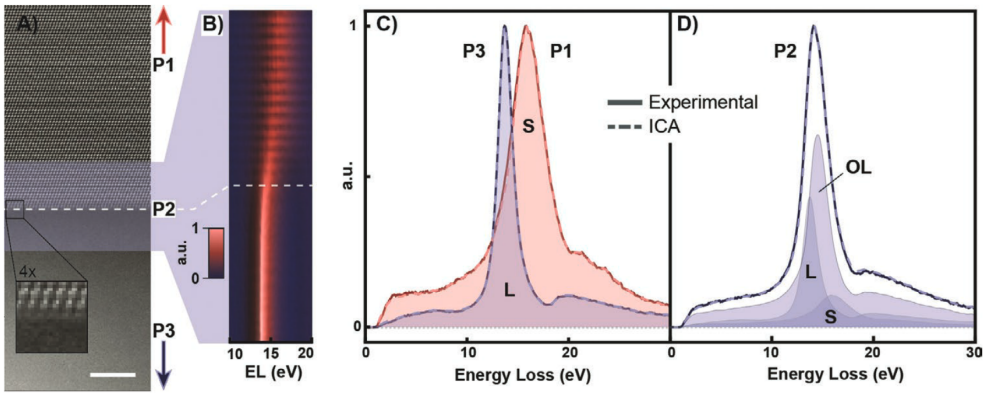

Figure 5.2 – EELS analysis of the liquid–solid interface on an A-polar NW. A) Atomic-resolution HAADF image acquired simultaneously with the EELS map. The interface is indicated by the dashed line and the scale bar is 5 nm. The inset shows a magnified view of the interface, where the ordered liquid region is visible. B) Variation in the plasmon peak position close to the interface. C,D) Comparison of EELS measurements (solid lines) with ICA model (dashed curves), and ICA decomposition corresponding to ordered liquid (OL), liquid phase (L), and solid (S) phase. Spectra are extracted from the ZB GaAs (P1), interface (P2), and liquid Ga (P3) regions of the map indicated in panel (A). The color plot and all curves show normalized spectra averaged on horizontal lines of the map.

given the fact that the ordered liquid cannot exist as a bulk phase outside of the interface, we propose that it can be considered as interfacial complexion[248–250].

5.2.2 Atomistic simulations of the liquid-solid interface

Details of the MD simulations

To investigate the ordering at the interface at the two (111) surfaces, we run MD simulations with the MLIP introduced in Chapter 4. The simulation box we use is the one shown in Fig. 4.1. We use an orthorhombic supercell composed of a central solid GaAs section (144 atoms, corresponding to 6 layers of 24 atoms), in contact with liquid Ga (192 atoms) on both of its surfaces, totaling 336 atoms.

The initial lattice parameter for the solid part is set to the one obtained from DFT calculations, whereas the initial density of liquid Ga is set to that obtained with independent simulations in a smaller (96 atoms) box at the objective temperature. All the simulations are run using i-PI[127] in combination with LAMMPS[97], and n2p2[128] to evaluate the NNP. First, the system is equilibrated in the $N\sigma T$ ensemble, allowing the cell degrees of freedom to change independently. After equilibration, production simulations have been run in the NVT ensemble, using the average lattice parameters, at the respective temperatures.

The temperatures are controlled using a combination of a generalized Langevin[153] and

stochastic velocity rescaling[154] thermostats. Pressures, where applicable, are constrained using an anisotropic barostat[251]. Simulations are run with a timestep of 4 fs (at 300 K) and 2 fs (at 900 K), for a total of 10 ns.

Analysis and comparison of the atomic density along the growth direction

We first compare in Fig. 5.3 the density of Ga along the z-axis between different simulations at 300 K, for both polarities and both crystal structure. We observe that the Ga density at the interface is similar between the two crystal structures, whereas it is clearly affected by the polarity. At the A polar surface we observe only a single major peak, followed by a smaller one and a liquid bulk. At the B polar surface, there are multiple peaks, deriving from a more ordered interface.

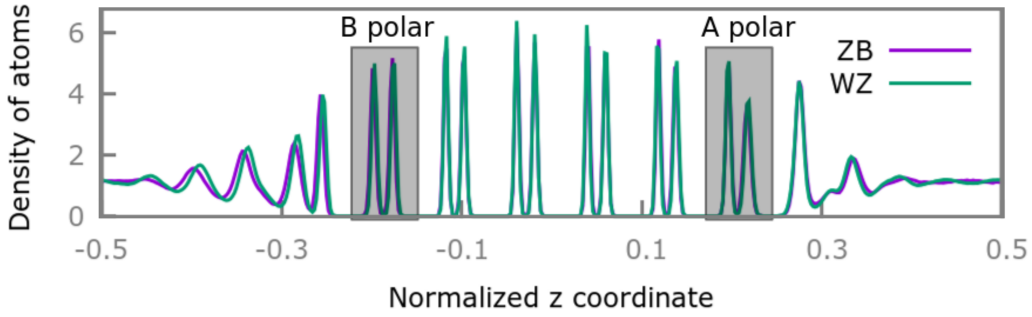

Figure 5.3 – The density of atoms along the normalized z-axis for both ZB and WZ at 300 K. The two curves seem to behave rather similarly, while the difference between the A and B surface is much more evident.

Then, we compare the projected linear density obtained across the simulation cell with intensity line profiles derived from experimental images for ZB-A and WZ-B surfaces respectively in panels A and B in Fig. 5.4. The experimental curves are obtained by projecting the intensity profiles from Fig. 5.1B-C along the axis normal to the surface. In the solid, we obtain regularly ordered peaks at the positions of the dumbbells. The liquid also exhibits some peaks in the density profile, characteristic of atomic-level ordering. The range of the ordering is different for the two polarities. As noted earlier, in STEM images, while the B-polar order is observed for four layers, it does not extend beyond the first layer for the A-polar case. Similar to the experimental observations, the simulations show that the ordering gradually diminishes when the distance from the crystalline phase is increased, which is in contrast to the conclusions of Ref. 252 for InP in contact with liquid InAu, where the first three layers had identical distances.

It should be noted that an exact correlation of the projected linear density obtained from the simulations with the HAADF-STEM integrated intensities is not expected: although the latter scales approximately with the former, a full quantum mechanical image simulation using the atomistic model derived from MD would be necessary for a quantitative comparison. Of

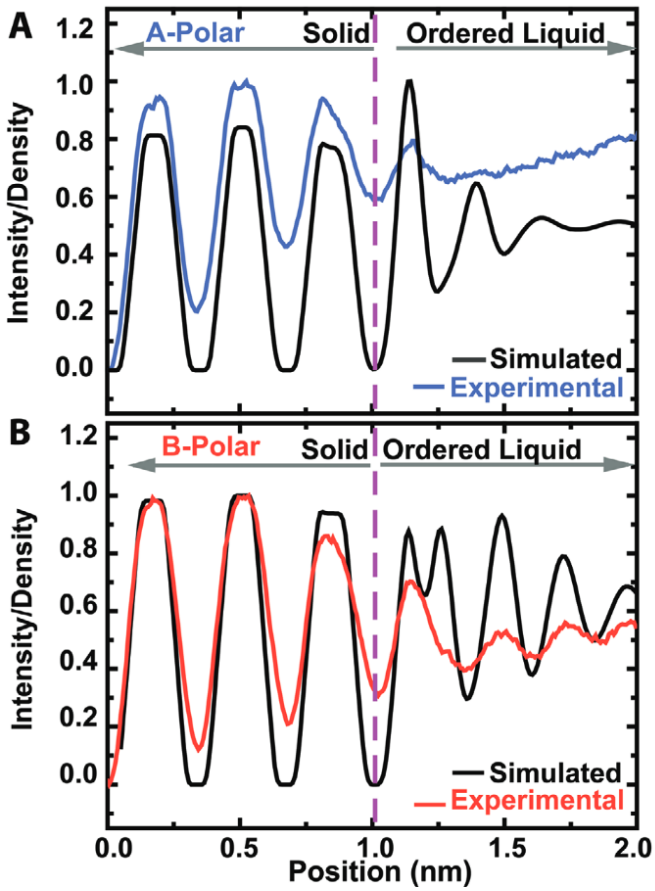

Figure 5.4 – Comparison of intensity/mean atom density profiles from experimental STEM and MD simulations. The ZB A-polar and WZ B-polar interfaces are demonstrate in panels A and B, respectively. The dashed pink line indicates the interface position. Note that the simulated Ga density appears different from the one in Fig. 5.3 because a different binning has been used. Here, we used a larger binning to observe the density, which results in a single peak for the GaAs dumbbell, to obtain results comparable with the experimental images.

	1st layer	2nd layer	3rd layer	4th layer
HAADF image	0.263	0.578	0.802	1.014
Simulations	0.264	0.576	0.799	1.013

Table 5.1 – The distances of ordered liquid layers from the last crystalline layer (nm) for a B polar interface from experimental observations and MD simulations

importance, however, is that there is a quantitative match between the MD predictions of spacing between ordered layers with the experimental observations. This is demonstrated in Tab. 5.1, in which the experimental observations and MD predictions of spacing between ordered layers are compared for a B-polar interface. We therefore conclude that MD based on the NNP correctly predicts the nature of the liquid ordering.

Three-dimensional visualisation of the Ga density

The simulations shown previously were performed at 300 K, corresponding to the temperature used to acquire the STEM images. To make the link with the growth process, we perform the same simulations at the growth temperature (900 K). Fig. 5.5 depicts the ordering of the interface at 900 K for both polarities where the spatial variation in the Ga atomic density is indicated by red isosurfaces. Fig. 5.5A depicts a 3D view of the ensemble for a B-polar WZ solid in contact with liquid Ga. Fig. 5.5B provides top and side views for the A- and B-polar solids. In both cases, we find ordering within the plane, however with significant differences in the arrangement of Ga atoms as a function of the polarity.

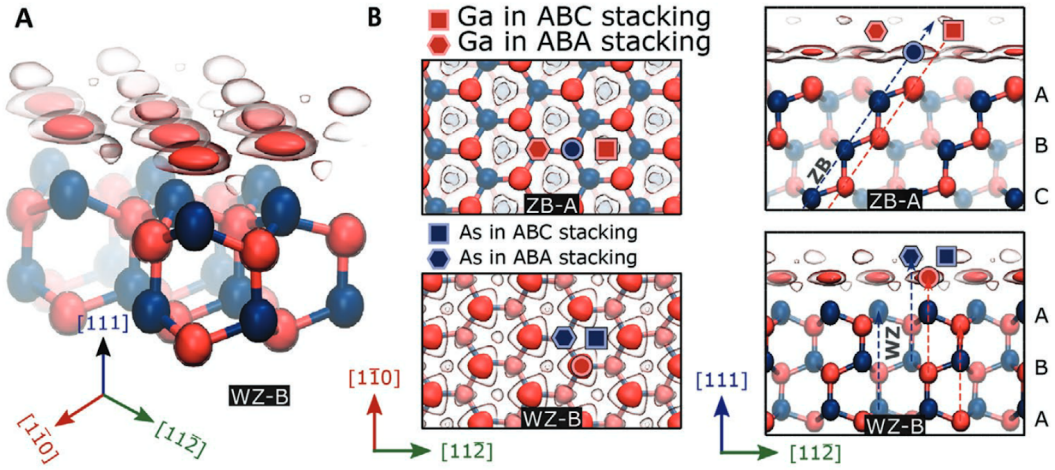

Figure 5.5 – Simulations of the Ga(l)-GaAs(s) interface at 900K. Isocontours of the Ga atom density, showing ordering at the liquid-Ga/GaAs interface – corresponding to density ρ = 0.11Bohr^{-3} (opaque) and ρ = 0.06Bohr^{-3} (translucent). As and Ga atoms are drawn in blue and red, respectively. Objects further from the viewer are represented with less contrast as a depth cue. (A) 3D view of WZ GaAs terminated with B-polarity; (B) views along the [111] (left) and [1$\bar{1}$0] ZB/ [11$\bar{2}$0] WZ (right) directions; in the case of A-polar ZB and B-polar WZ (top and bottom respectively). Similar isocontours presented for simulations at 300K in S.I.of the original paper. We indicate the positioning of the As and Ga atoms that are required to create a new Ga-As bilayer, taking the isocontours and the underlying structure as canvas.

On the B-polar surface, Ga is adsorbed right on top of the terminal As. This is consistent with the large electronegativity difference between As and Ga that is likely to induce strong

electrostatic interactions. In-plane ordering is also present at the A-polar surface. Here, instead, Ga atoms in the liquid pair with the terminal Ga atoms in the solid, consistent with the dimerization tendency of Ga. As expected, at the higher temperature, the range of the order has decreased relative to observations at 300 K, with simulations indicating the presence of at most one ordered layer on top of the solid. Most of the qualitative features of the liquid ordering, however, are preserved between the two temperatures, suggesting that experiments performed at 300 K can provide insights into the relevant mechanisms for growth at higher temperatures.

We move now to the microscopic picture of GaAs growth. Any new layer of GaAs forms new Ga–As pairs. Ga and As atoms are shown in red and blue in Fig. 5.5, while the positions leading to ZB and WZ configurations are indicated by squares and hexagons, respectively. The inclined dashed lines provide a guide to define the atomic positions that lead to the relevant crystalline structures in Fig. 5.5B,C. In the A-polar case, As must displace and occupy the position of ordered Ga in the liquid, with which it may form a dumbbell. The Ga atoms have two choices, as indicated in Fig. 5.5 by the square and hexagon. Depending on the position selected, ZB (ABCABC stacking) or a twin (ABA stacking) is formed. These two positions are not equivalent in terms of their first and second nearest neighbors configuration. In the ABC stacking, Ga is found at the middle of a projected hexagon, while in the ABAB it is at a vertex. To form a new bilayer, As atoms should first displace the ordered Ga layer on top of the NW, and this should increase the formation barrier. This is consistent with the difficulty in synthesizing A-polar GaAs NWs[253, 254], suggesting that the slower process may help the growing layer achieve the more thermodynamically stable ABC stacking.

During the formation of a new bilayer for the B-polar surface, the Ga atoms adsorbed on top of the As-terminated surface can stay at their position. The incoming As atoms occupy empty positions in the second row. The fact that the growth can proceed without displacing Ga atoms is consistent with the observation of a more facile growth for this polarity. The higher growth rate can also partly explain the higher propensity to introduce stacking defects and polytypism in B-polar GaAs NWs[253]. These results could explain why in polar semiconductors, growth in a certain polarity is preferred and how polarity determines the tendency for polytypism.

5.3 As free energy in the liquid Ga

Until now, we have discussed the case where pure liquid Ga is in contact with solid GaAs, as it allowed a direct comparison with the experimental images taken at ambient temperature. However, a complete understanding of the mechanisms underlying the growth on either surface requires to introduce As atoms in the liquid, to observe their behaviour during the growth. This is possible only in the simulations, as we have complete control over the choice of the parameters and the conditions and we can pinpoint and analyze only the As atoms in the liquid.

Since there are multiple possible competing effects, we begin with the analysis of the free

energy profile of a single As atom in the liquid Ga, along the z-axis and on the xy plane on the two polar surfaces.

5.3.1 Free energy profile along the z-axis

Enhanced sampling simulations of the system

To study the As free energy, we introduce a single As atom in the liquid Ga, in a supercell identical to the one used in Fig. 4.1. As we intend to study the free energy of As in the liquid Ga in the low solubility limit, we prefer to add a single atom instead of the As_4 unit used in the experimental set-up. As-As interactions in liquid Ga are investigated in the second part of this chapter. We study the free energy at three different temperatures, i.e. 350, 600, and 800 K, running 4 separate simulation for every temperature. For every cell we substitute a random Ga atom with an As one and let the simulation run for 4 ns.

As MD simulations yield meaningful results only when run long enough to get sufficient sampling of all available microstates, they cannot be used in systems where the metastable states are separated by large free energy barriers. Since we expect to observe a large free energy minimum at each interface, that would hinder the diffusion of the atom, we force the exploration of the phase space with the aid of a bias potential that penalizes the microstates that were already explored, a technique known as "metadynamics"[255].

More in detail, here we use the well-tempered version of metadynamics[256], where the bias potential introduced in the system decreases over simulation time to allow to converge the free energy, as implemented in PLUMED[149]. Since the trajectory obtained in this manner is perturbed by the bias potential, we recover the unbiased trajectory computing the time-dependent contribution of the external potential through an iterative procedure, as it has been recently proposed and implemented in ITRE[150].

Results at varying temperatures

The free energy of the system reconstructed by the statistical reweighting of the biased probability distribution and averaged over the four replicas at 350 K, 600 K and 800 K is shown in Fig. 5.6. Based on the left-side minimum, the As atom in the bulk liquid shows a preferential adsorption on the A-polar surface at all temperatures. The B-polar surface minimum is slightly further from the interface due to the formation of a layer of Ga atoms on the As terminating surface, onto which the extra As atom can then be adsorbed. Additional minima at both surfaces can be observed, of a far lower magnitude than the main one at the interface. The overall trend of the free energy is the same across the temperatures albeit with a flatter profile at higher temperatures which is to be expected due to higher thermal energy of the atoms.

Through this set of simulations, we tried to observe relative differences in the adsorption of As on the two surfaces. On the A surface we observe a small barrier near the interface at low

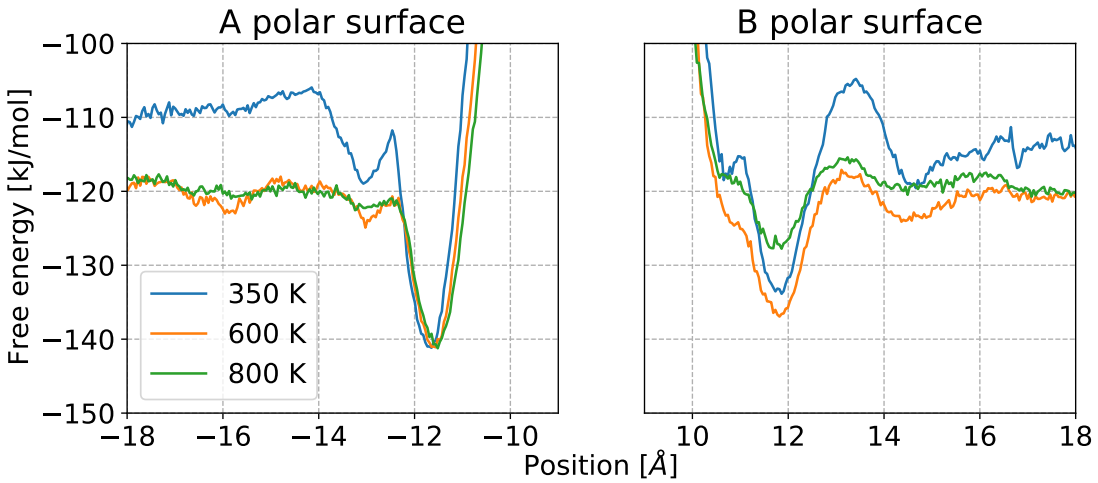

Figure 5.6 – Free energy profiles with respect to the distance of As atom from the bulk solid at 350 K, 600 K and 800 K, that show a slight preferential adsorption of As on A-polar surface, while B-polar surface shows a comparatively structured profile. The overall trend is the same across temperatures with a flatter profile at higher temperatures.

temperature, which disappears completely as we raise the temperature to 800 K. This suggests that the As atom moves virtually freely in the liquid Ga and can easily get to the surface. On the B surface, on the other hand, there is always a small barrier between the bulk liquid Ga and the interface. Moreover, the minimum at the interface becomes shallower as the temperature increases. Although this is only a preliminary study, it could be related to the tendency of As to form a new layer on the A surface at a greater speed compared to the B surface[243].

5.3.2 As ordering on the A and B surfaces

Following the analysis of Sec. 5.2.2, we run similar simulations of the liquid-solid interface, but with the addition of an As atom in the liquid. We run different REMD[129] trajectories for each interface at various temperatures, in which the atom is positioned on top of the surface, but is not constrained to the interface alone. We still expect that it will spend the majority of the time trapped at the surface, given the large minima in the FES observed in the last section.

In Fig. 5.7 we provide the same qualitative analysis of the preferential position of As and Ga during the simulation. Here, we highlight with the lime colour the position of the extra As at the interface.

For the A polar surface the As atom sits on top of the last Ga layer, as expected, and the liquid Ga begins to show an ordering around it that resembles the ZB crystal structure. On the B polar surface, the As atom prefers the ZB-crystal structure to the WZ one, which is also visited by the atom. An analysis of the free energy of the As atom at the interface would allow us to quantify the preference of one structure over the other, although the analysis would be limited

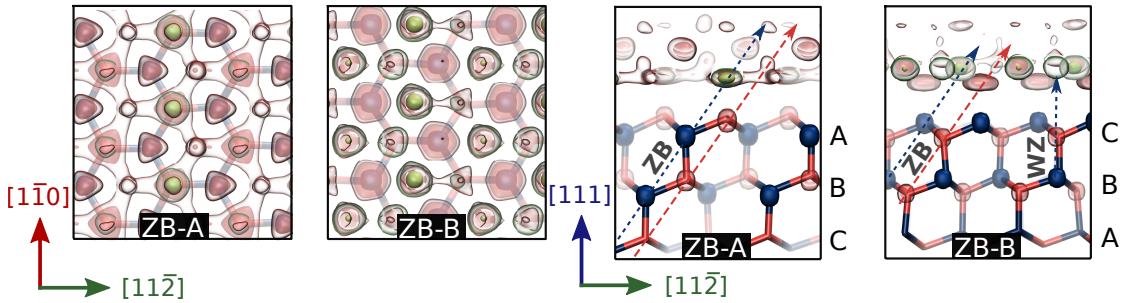

Figure 5.7 – Simulations of the Ga(l)-GaAs(s) interface at 300K. The same isocontours as Fig. 5.5 are used, with the addition of the lime-coloured surface corresponding to the extra As atom in the liquid. The isocontour for Ga is also made transparent to ease the view of the extra As. Objects further from the viewer are represented with less contrast as a depth cue. We provide lines to guide the eye towards the expected positions for the ZB and WZ crystal structures.

to the bulk.

Future work in this direction would involve the addition of multiple As atoms in the liquid Ga, to study the early phases of the precipitation and the preferential surface reconstruction in these conditions. Moreover, we would like to repeat these analyses at the triple-phase point, to include the known macroscopic parameters that influence the growth, such as the contact angle.

6 Conclusions

Machine learning potentials have changed the long-standing paradigm of having to choose between computationally cheap –but with limited applicability– empirical forcefields and accurate –but limited mostly to small systems– *ab initio* methods. By computing only a handful of structures at the desired level of theory, we are able to interpolate the known data to compute the energies and forces for larger systems and longer timescales, thus allowing to investigate complex materials' properties.

However, the construction of a machine learning potential is often tied to the system of interest and the user has to make a number of choices that can impact the success of the learning task. Therefore, there are great opportunities available to simplify the construction of the potentials, optimize the representations used for the systems, compare different learning strategies, and automatize the construction of the training set.

In this thesis we tried to tackle a few of what we considered to be important unresolved issues in the training and use of the potentials, then demonstrating their effectiveness on a potential for the full binary phase diagram of gallium arsenide, which has then been used to investigate currently standing problems on the GaAs nanowires' growth.

We first focused on the representations. It is natural to question the quality and the efficiency of the plethora of representations to learn a given dataset. Thus, we compared some commonly used frameworks on a dataset of water dimers and trimers, finding that they perform very similarly for the given task. In particular, we observed that the regions of the phase space that are better predicted are also the regions that are sampled better, showing the importance of generating a training set that is compatible with the learning task. For example, we obtained better results by training models on a uniformly sampled selection of points compared to a random selection on the water dimers dataset.

Then, we studied how to select an optimal set of features from a larger number of candidates. We demonstrated that simple heuristic methods such as CUR selection and farthest point sampling can greatly reduce the number of features needed to learn the energy of some

selected systems, at almost no loss of accuracy. This is particularly relevant for the case of symmetry functions, because it allows to quickly select a suitable set of symmetry functions for any new system. We first tested it on a dataset of bulk water structures, where our selection achieved the same accuracy as a published model of carefully selected functions, at a lower computational cost. Then, we selected symmetry functions to train a potential for the Al-Mg-Si alloy and, in the second part of the thesis, to train a potential for GaAs, with equally satisfactory results. Similarly, we applied these feature selection methods to the power spectrum of the SOAP representation, finding that a reduced power spectrum can perform as well as the full one in the low data regime. This can be helpful to reduce the computational load for calculations where several species are involved.

A second topic that we covered is the ability to quantify the uncertainty arising from the use of machine learning potentials. Training a potential on a set of reference data introduces an error that is often ignored, even though it can affect the interpretation of the results obtained with the potentials. Therefore, we introduced a quick and reliable method to measure the uncertainty of thermodynamical properties computed with the aid of machine learning potentials. We tested it on numerous systems and quantities, to demonstrate its straightforward application when a committee of potentials is used. Moreover, we generalized our formulas to any case where we the uncertainty of the variable and the potential is known. Similarly, this method can also be applied to recognize when the potential is unable to confidently predict the forces in a system and use a fallback potential to avoid fully unphysical behaviour and safely explore the relevant phase space.

In the rest of the thesis we used the methods that we introduced, together with other state of the art techniques, to train a potential for the Ga_xAs_{1-x} system, spanning all the liquid and solid phases, which encompass both metallic and semiconducting structures. This complex system is an excellent testing ground to understand the abilities, as well as the limits, of the potentials. We first generated a suitable training set for this system. We began with a limited set of structures restricted to a specific region of the phase space, to which we added bulk configurations to be able to correctly compute static properties. Finally, we added new configurations to predict correctly the behaviour over the non-sampled regions of the phase space. In this last task, in particular, we made extensive use of the uncertainty estimation and the farthest point sampling selection, which allowed to quickly find the structures that needed to be recomputed.

To test the potential, we ran simulations to compute various solid and liquid properties, most of which are out of reach to *ab initio* methods, such as surface tension and the diffusion in large cells. Furthermore, we explored the reliability of the potential over the whole phase space, computing the binary phase diagram for the potential and comparing to the experimental one. Overall, we find a qualitative agreement, although we underestimate the melting point of GaAs and Ga, likely due to the short-comings of the level of theory that we use (DFT with pseudopotentials at the GGA level).

Finally, we provided an example of simulations of systems of experimental interest that can be run and analyzed only thanks to the use of machine learning potentials. We simulated the liquid-solid interface of GaAs nanowires, in the bulk, which is considered to be of importance during the growth of the nanowires. Our simulations were compared to novel experimental images of the interface, finding evidence that the pre-ordering of liquid Ga depends on the polarity of the nanowire, in agreement with the experimental results. Our simulations further allowed to investigate the three-dimensional nature of the ordering, clearly distinguishing the different behaviours. Furthermore, we predicted quantities, such as the free energy of As atoms in the liquid on the z-axis and the xy plane on the surface, that can be useful to pinpoint the best conditions for the growth of these nanowires.

To summarise, machine learning potentials have had an enormous impact on our field, and we must work in order to lessen the burden on the user to train a new model. Selecting efficient representations and providing methods to ascertain the confidence of the predictions are necessary steps in this direction. With these methods available, we can move to simulate larger systems, to study complex phenomena and predict with accurate methods a number of properties that have been out of our reach until now.

Bibliography

[1] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*. Vol. 1. Elsevier, 2001.

[2] Huziel E Sauceda et al. "Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces". In: *The Journal of chemical physics* 150.11 (2019), p. 114102.

[3] David M Wilkins et al. "Accurate molecular polarizabilities with coupled cluster theory and machine learning". In: *Proceedings of the National Academy of Sciences* 116.9 (2019), pp. 3401–3406.

[4] Andreas Mardt et al. "VAMPnets for deep learning of molecular kinetics". In: *Nature communications* 9.1 (2018), pp. 1–11.

[5] Bhadeshia Hkdh. "Neural networks in materials science". In: *ISIJ international* 39.10 (1999), pp. 966–979.

[6] Thomas B Blank et al. "Neural network models of potential energy surfaces". In: *The Journal of chemical physics* 103.10 (1995), pp. 4129–4137.

[7] Steven Hobday, Roger Smith, and Joe Belbruno. "Applications of neural networks to fitting interatomic potential functions". In: *Modelling and Simulation in Materials Science and Engineering* 7.3 (1999), p. 397.

[8] Helmut Gassner et al. "Representation of intermolecular potential functions by neural networks". In: *The Journal of Physical Chemistry A* 102.24 (1998), pp. 4596–4605.

[9] Ryosuke Jinnouchi et al. "On-the-fly active learning of interatomic potentials for large-scale atomistic simulations". In: *The Journal of Physical Chemistry Letters* 11.17 (2020), pp. 6946–6955.

[10] Jonathan Vandermause et al. "On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events". In: *npj Computational Materials* 6.1 (2020), pp. 1–11.

[11] Volker L Deringer, Chris J Pickard, and Gábor Csányi. "Data-driven learning of total and local energies in elemental boron". In: *Physical review letters* 120.15 (2018), p. 156001.

[12] Mitchell A Wood et al. "Data-driven material models for atomistic simulation". In: *Physical Review B* 99.18 (2019), p. 184305.

[13] Daniel Schwalbe-Koda, Aik Rui Tan, and Rafael Gómez-Bombarelli. "Adversarial Attacks on Uncertainty Enable Active Learning for Neural Network Potentials". In: *arXiv preprint arXiv:2101.11588* (2021).

[14] Felix Musil et al. "Physics-inspired structural representations for molecules and materials". In: *arXiv preprint arXiv:2101.04673* (2021).

[15] Jörg Behler and Michele Parrinello. "Generalized neural-network representation of high-dimensional potential-energy surfaces". In: *Physical review letters* 98.14 (2007), p. 146401.

[16] Albert P Bartók, Risi Kondor, and Gábor Csányi. "On representing chemical environments". In: *Physical Review B* 87.18 (2013), p. 184115.

[17] Andrea Grisafi et al. "Symmetry-adapted machine learning for tensorial properties of atomistic systems". In: *Physical review letters* 120.3 (2018), p. 036002.

[18] Jörg Behler. "Atom-centered symmetry functions for constructing high-dimensional neural network potentials". In: *The Journal of chemical physics* 134.7 (2011), p. 074106.

[19] Tobias Morawietz et al. "How van der Waals interactions determine the unique properties of water". In: *Proceedings of the National Academy of Sciences* 113.30 (2016), pp. 8368–8373. DOI: 10.1073/pnas.1602375113.

[20] Hagai Eshet et al. "Ab initio quality neural-network potential for sodium". In: *Physical Review B* 81.18 (2010), p. 184107.

[21] Nongnuch Artrith, Tobias Morawietz, and Jörg Behler. "High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide". In: *Physical Review B* 83.15 (2011), p. 153101.

[22] J Behler et al. "RuNNer: A Neural Network Code for High-Dimensional Potential-Energy Surfaces". In: *Universität Göttingen* (2018).

[23] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. "ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost". In: *Chemical science* 8.4 (2017), pp. 3192–3203.

[24] Albert P Bartók et al. "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons". In: *Physical review letters* 104.13 (2010), p. 136403.

[25] Daniel Marchand et al. "Machine learning for metallurgy I. A neural-network potential for Al-Cu". In: *Physical Review Materials* 4.10 (2020), p. 103601.

[26] Ryo Kobayashi et al. "Neural network potential for Al-Mg-Si alloys". In: *Physical Review Materials* 1.5 (2017), p. 053604.

[27] Daniele Dragoni et al. "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.

[28] Yury Lysogorskiy et al. "Performant implementation of the atomic cluster expansion (PACE): Application to copper and silicon". In: *arXiv preprint arXiv:2103.00814* (2021).

[29] Albert P Bartók et al. “Machine learning a general-purpose interatomic potential for silicon”. In: *Physical Review X* 8.4 (2018), p. 041048.

[30] Gabriele C Sosso et al. “Neural network interatomic potential for the phase change material GeTe”. In: *Physical Review B* 85.17 (2012), p. 174103.

[31] Claudio Zeni et al. “Building machine learning force fields for nanoclusters”. In: *The Journal of Chemical Physics* 148.24 (2018), p. 241739.

[32] Piero Gasparotto, Robert Horst Meißner, and Michele Ceriotti. “Recognizing local and global structural motifs at the atomic scale”. In: *Journal of chemical theory and computation* 14.2 (2018), pp. 486–498.

[33] Frank Noé and Cecilia Clementi. “Kinetic distance and kinetic maps from molecular dynamics simulation”. In: *Journal of Chemical Theory and Computation* 11.10 (2015), pp. 5002–5011.

[34] Leonid Kahle et al. “Unsupervised landmark analysis for jump detection in molecular dynamics simulations”. In: *Physical Review Materials* 3.5 (2019), p. 055404.

[35] Andrea Anelli et al. “Generalized convex hull construction for materials discovery”. In: *Physical Review Materials* 2.10 (2018), p. 103804.

[36] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. “Inverse molecular design using machine learning: Generative models for matter engineering”. In: *Science* 361.6400 (2018), pp. 360–365.

[37] Fritz Körmann et al. “Temperature dependent magnon-phonon coupling in bcc Fe from theory and experiment”. In: *Physical review letters* 113.16 (2014), p. 165503.

[38] Tilmann Hickel et al. “Advancing density functional theory to finite temperatures: methods and applications in steel design”. In: *Journal of Physics: condensed matter* 24.5 (2011), p. 053202.

[39] Thomas E Markland and Michele Ceriotti. “Nuclear quantum effects enter the mainstream”. In: *Nature Reviews Chemistry* 2.3 (2018), pp. 1–14.

[40] Edgar A Engel, Bartomeu Monserrat, and Richard J Needs. “Anharmonic nuclear motion and the relative stability of hexagonal and cubic ice”. In: *Physical Review X* 5.2 (2015), p. 021033.

[41] Venkat Kapil et al. “Assessment of Approximate Methods for Anharmonic Free Energies”. In: *Journal of chemical theory and computation* 15.11 (2019), pp. 5845–5857.

[42] Venkat Kapil and Edgar A Engel. “A complete description of thermodynamic stabilities of molecular crystals”. In: *arXiv preprint arXiv:2102.13598* (2021).

[43] Michael Gastegger, Jörg Behler, and Philipp Marquetand. “Machine learning molecular dynamics for the simulation of infrared spectra”. In: *Chemical science* 8.10 (2017), pp. 6924–6935.

[44] Tobias Morawietz et al. “The interplay of structure and dynamics in the Raman spectrum of liquid water over the full frequency and temperature range”. In: *The journal of physical chemistry letters* 9.4 (2018), pp. 851–857.

[45] Nathaniel Raimbault et al. “Using Gaussian process regression to simulate the vibrational Raman spectra of molecular crystals”. In: *New Journal of Physics* 21.10 (2019), p. 105001.

[46] Federico M Paruzzo et al. “Chemical shifts in molecular solids by machine learning”. In: *Nature communications* 9.1 (2018), pp. 1–10.

[47] Michael W Mahoney and William L Jorgensen. “A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions”. In: *The Journal of chemical physics* 112.20 (2000), pp. 8910–8922.

[48] Max Veit et al. “Equation of state of fluid methane from first principles with machine learning potentials”. In: *Journal of chemical theory and computation* 15.4 (2019), pp. 2574–2586.

[49] Huziel E Sauceda et al. “Dynamical strengthening of covalent and non-covalent molecular interactions by nuclear quantum effects at finite temperature”. In: *Nature Communications* 12.1 (2021), pp. 1–10.

[50] Kristof T Schütt et al. “SchNet–A deep learning architecture for molecules and materials”. In: *The Journal of Chemical Physics* 148.24 (2018), p. 241722.

[51] Sebastian Wille et al. “An experimentally validated neural-network potential energy surface for H-atom on free-standing graphene in full dimensionality”. In: *Physical Chemistry Chemical Physics* 22.45 (2020), pp. 26113–26120.

[52] Volker L Deringer et al. “Origins of structural and electronic transitions in disordered silicon”. In: *Nature* 589.7840 (2021), pp. 59–64.

[53] Nongnuch Artrith, Björn Hiller, and Jörg Behler. “Neural network potentials for metals and oxides–First applications to copper clusters at zinc oxide”. In: *physica status solidi (b)* 250.6 (2013), pp. 1191–1203.

[54] Nongnuch Artrith and Alexie M Kolpak. “Understanding the composition and activity of electrocatalytic nanoalloys in aqueous solvents: A combination of DFT and accurate neural network potentials”. In: *Nano letters* 14.5 (2014), pp. 2670–2676.

[55] A Hamedani et al. “Insights into the primary radiation damage of silicon by a machine learning interatomic potential”. In: *Materials Research Letters* 8.10 (2020), pp. 364–372.

[56] Thuong T Nguyen et al. “Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions”. In: *The Journal of chemical physics* 148.24 (2018), p. 241725.

[57] Giulio Imbalzano et al. “Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials”. In: *The Journal of Chemical Physics* 148.24 (2018), p. 241730.

[58] 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.

[59] Ralf Drautz. "Atomic cluster expansion for accurate and transferable interatomic potentials". In: *Physical Review B* 99.1 (2019), p. 014104.

[60] Sergey N Pozdnyakov et al. "Incompleteness of atomic structure representations". In: *Physical Review Letters* 125.16 (2020), p. 166001.

[61] Markus Bachmayr et al. "Atomic cluster expansion: Completeness, efficiency and stability". In: *arXiv preprint arXiv:1911.03550* (2019).

[62] Berk Onat, Christoph Ortner, and James R Kermode. "Sensitivity and dimensionality of atomic environment representations used for machine learning interatomic potentials". In: *The Journal of Chemical Physics* 153.14 (2020), p. 144106.

[63] Miguel A Caro. "Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials". In: *Physical Review B* 100.2 (2019), p. 024112.

[64] Volodymyr Babin, Claude Leforestier, and Francesco Paesani. "Development of a "first principles" water potential with flexible monomers: Dimer potential energy surface, VRT spectrum, and second virial coefficient". In: *Journal of chemical theory and computation* 9.12 (2013), pp. 5395–5403.

[65] Jörg Behler. "Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations". In: *Physical Chemistry Chemical Physics* 13.40 (2011), pp. 17930–17955.

[66] KV Jovan Jose, Nongnuch Artrith, and Jörg Behler. "Construction of high-dimensional neural network potentials using environment-dependent atom pairs". In: *The Journal of chemical physics* 136.19 (2012), p. 194111.

[67] Jörg Behler. "First principles neural network potentials for reactive simulations of large molecular and condensed systems". In: *Angewandte Chemie International Edition* 56.42 (2017), pp. 12828–12840.

[68] Jörg Behler. "Constructing high-dimensional neural network potentials: A tutorial review". In: *International Journal of Quantum Chemistry* 115.16 (2015), pp. 1032–1050.

[69] Bastiaan J Braams and Joel M Bowman. "Permutationally invariant potential energy surfaces in high dimensionality". In: *International Reviews in Physical Chemistry* 28.4 (2009), pp. 577–606.

[70] Jun Li, Bin Jiang, and Hua Guo. "Permutation invariant polynomial neural network approach to fitting potential energy surfaces. II. Four-atom systems". In: *The Journal of chemical physics* 139.20 (2013), p. 204103.

[71] Bin Jiang and Hua Guo. "Permutation invariant polynomial neural network approach to fitting potential energy surfaces". In: *The Journal of chemical physics* 139.5 (2013), p. 054112.

[72] Bin Jiang and Hua Guo. “Permutation invariant polynomial neural network approach to fitting potential energy surfaces. III. Molecule-surface interactions”. In: *The Journal of chemical physics* 141.3 (2014), p. 034109.

[73] Cas van der Oord et al. “Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials”. In: *Machine Learning: Science and Technology* 1.1 (2020), p. 015004.

[74] Alexander V Shapeev. “Moment tensor potentials: A class of systematically improvable interatomic potentials”. In: *Multiscale Modeling & Simulation* 14.3 (2016), pp. 1153–1173.

[75] Han Wang et al. “DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics”. In: *Computer Physics Communications* 228 (2018), pp. 178–184.

[76] Felix A Faber et al. “Alchemical and structural distribution based representation for universal quantum machine learning”. In: *The Journal of chemical physics* 148.24 (2018), p. 241717.

[77] Michael Gastegger et al. “wACSF—Weighted atom-centered symmetry functions as descriptors in machine learning potentials”. In: *The Journal of chemical physics* 148.24 (2018), p. 241709.

[78] Nicholas J Browning et al. “Genetic optimization of training sets for improved machine learning models of molecular properties”. In: *The journal of physical chemistry letters* 8.7 (2017), pp. 1351–1359.

[79] 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.

[80] Albert P Bartók and Gábor Csányi. “Gaussian approximation potentials: A brief tutorial introduction”. In: *International Journal of Quantum Chemistry* 115.16 (2015), pp. 1051–1057.

[81] Rose K Cersonsky et al. “Improving sample and feature selection with principal covariates regression”. In: *Machine Learning: Science and Technology* (2021).

[82] Daniel J Rosenkrantz, Richard E Stearns, and Philip M Lewis II. “An analysis of several heuristics for the traveling salesman problem”. In: *SIAM journal on computing* 6.3 (1977), pp. 563–581.

[83] 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.

[84] Sandip De et al. “Comparing molecules and solids across structural and alchemical space”. In: *Physical Chemistry Chemical Physics* 18.20 (2016), pp. 13754–13769.

[85] Albert P Bartók et al. “Machine learning unifies the modeling of materials and molecules”. In: *Science advances* 3.12 (2017), e1701816.

[86] Felix A Faber et al. “Prediction errors of molecular machine learning models lower than hybrid DFT error”. In: *Journal of chemical theory and computation* 13.11 (2017), pp. 5255–5264.

[87] CJ Burnham et al. “The vibrational proton potential in bulk liquid water and ice”. In: *The Journal of chemical physics* 128.15 (2008), p. 154519.

[88] Sandip De et al. “Mapping and classifying molecules from a high-throughput structural database”. In: *Journal of cheminformatics* 9.1 (2017), pp. 1–14.

[89] Gareth A Tribello, Michele Ceriotti, and Michele Parrinello. “Using sketch-map coordinates to analyze and bias molecular dynamics simulations”. In: *Proceedings of the National Academy of Sciences* 109.14 (2012), pp. 5196–5201.

[90] Yunxing Zuo et al. “Performance and Cost Assessment of Machine Learning Interatomic Potentials”. In: *The Journal of Physical Chemistry A* 124.4 (2020), pp. 731–745.

[91] Alexander Goscinski et al. “The role of feature space in atomistic learning”. In: *Machine Learning: Science and Technology* (2021).

[92] Venkat Kapil, Jörg Behler, and Michele Ceriotti. “High order path integrals made easy”. In: *The Journal of chemical physics* 145.23 (2016), p. 234103.

[93] Bingqing Cheng, Jörg Behler, and Michele Ceriotti. “Nuclear quantum effects in water at the triple point: Using theory as a link between experiments”. In: *The journal of physical chemistry letters* 7.12 (2016), pp. 2210–2215.

[94] Matti Hellström and Jörg Behler. “Concentration-dependent proton transfer mechanisms in aqueous NaOH solutions: From acceptor-driven to donor-driven and back”. In: *The journal of physical chemistry letters* 7.17 (2016), pp. 3302–3306.

[95] Suresh Kondati Natarajan, Tobias Morawietz, and Jörg Behler. “Representing the potential-energy surface of protonated water clusters by high-dimensional neural network potentials”. In: *Physical Chemistry Chemical Physics* 17.13 (2015), pp. 8356–8371.

[96] Vanessa Quaranta, Matti Hellström, and Jörg Behler. “Proton-transfer mechanisms at the water–ZnO interface: The role of presolvation”. In: *The journal of physical chemistry letters* 8.7 (2017), pp. 1476–1483.

[97] Steve Plimpton. “Fast parallel algorithms for short-range molecular dynamics”. In: *Journal of computational physics* 117.1 (1995), pp. 1–19.

[98] Daniele Giofré et al. “Ab initio modelling of the early stages of precipitation in Al-6000 alloys”. In: *Acta Materialia* 140 (2017), pp. 240–249.

[99] Graeme Henkelman and Hannes Jónsson. “A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives”. In: *The Journal of chemical physics* 111.15 (1999), pp. 7010–7022.

[100] Graeme Henkelman and Hannes Jónsson. “Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points”. In: *The Journal of chemical physics* 113.22 (2000), pp. 9978–9985.

[101] Paolo Giannozzi 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.

[102] M Mantina et al. “First principles impurity diffusion coefficients”. In: *Acta Materialia* 57.14 (2009), pp. 4102–4108.

[103] Grégoire Montavon et al. “Machine learning of molecular electronic properties in chemical compound space”. In: *New Journal of Physics* 15.9 (2013), p. 095003.

[104] Félix Musil et al. “Machine learning for the structure–energy–property landscapes of molecular crystals”. In: *Chemical science* 9.5 (2018), pp. 1289–1300.

[105] Giulio Imbalzano et al. “Uncertainty estimation by committee models for molecular dynamics and thermodynamic averages”. In: *arXiv preprint arXiv:2011.08828* (2020).

[106] Kevin Tran et al. “Methods for comparing uncertainty quantifications for material property predictions”. In: *Machine Learning: Science and Technology* 1.2 (2020), p. 025006.

[107] Carl Edward Rasmussen and Christopher K I Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[108] Andrew A Peterson, Rune Christensen, and Alireza Khorshidi. “Addressing uncertainty in atomistic machine learning”. In: *Physical Chemistry Chemical Physics* 19.18 (2017), pp. 10978–10985.

[109] Alexander Shapeev et al. “Active learning and uncertainty estimation”. In: *Machine Learning Meets Quantum Physics* (2020), pp. 309–329.

[110] Dimitris N Politis and Joseph P Romano. “Large sample confidence regions based on subsamples under minimal assumptions”. In: *The Annals of Statistics* (1994), pp. 2031–2050.

[111] Bradley Efron. “Bootstrap methods: another look at the jackknife”. In: *Breakthroughs in statistics*. Springer, 1992, pp. 569–593.

[112] Muhammed Shuaibi et al. “Enabling robust offline active learning for machine learning potentials using simple physics-based priors”. In: *Machine Learning: Science and Technology* (2020).

[113] Kevin Rossi et al. “Simulating solvation and acidity in complex mixtures with first-principles accuracy: the case of CH3SO3H and H2O2 in phenol”. In: *Journal of Chemical Theory and Computation* (2020).

[114] Christoph Schran, Krystof Brezina, and Ondrej Marsalek. “Committee neural network potentials control generalization errors and enable active learning”. In: *arXiv preprint arXiv:2006.01541* (2020).

[115] Ryosuke Jinnouchi et al. “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.

[116] Felix Musil et al. “Fast and accurate uncertainty estimation in chemical machine learning”. In: *Journal of chemical theory and computation* 15.2 (2019), pp. 906–915.

[117] MBBJM Tuckerman, Bruce J Berne, and Glenn J Martyna. “Reversible multiple time scale molecular dynamics”. In: *The Journal of chemical physics* 97.3 (1992), pp. 1990–2001.

[118] Thomas E Markland and David E Manolopoulos. “A refined ring polymer contraction scheme for systems with electrostatic interactions”. In: *Chemical Physics Letters* 464.4-6 (2008), pp. 256–261.

[119] Raghunathan Ramakrishnan et al. “Big data meets quantum chemistry approximations: the Δ-machine learning approach”. In: *Journal of chemical theory and computation* 11.5 (2015), pp. 2087–2096.

[120] Zhenwei Li, James R Kermode, and Alessandro De Vita. “Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces”. In: *Physical review letters* 114.9 (2015), p. 096405.

[121] Justin S Smith et al. “Less is more: Sampling chemical space with active learning”. In: *The Journal of chemical physics* 148.24 (2018), p. 241733.

[122] Jon Paul Janet et al. “A quantitative uncertainty metric controls error in neural network-driven chemical discovery”. In: *Chemical science* 10.34 (2019), pp. 7913–7922.

[123] Chiheb Ben Mahmoud et al. “Learning the electronic density of states in condensed matter”. In: *Physical Review B* 102.23 (2020), p. 235130.

[124] Glenn M Torrie and John P Valleau. “Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling”. In: *Journal of Computational Physics* 23.2 (1977), pp. 187–199.

[125] Gabor Csányi et al. ““Learn on the fly”: A hybrid classical and quantum-mechanical molecular dynamics simulation”. In: *Physical review letters* 93.17 (2004), p. 175503.

[126] Michele Ceriotti et al. “The inefficiency of re-weighted sampling and the curse of system size in high-order path integration”. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 468.2137 (2012), pp. 2–17.

[127] Venkat Kapil et al. “i-PI 2.0: A universal force engine for advanced molecular simulations”. In: *Computer Physics Communications* 236 (2019), pp. 214–223.

[128] Andreas Singraber, Jörg Behler, and Christoph Dellago. “Library-Based LAMMPS Implementation of High-Dimensional Neural Network Potentials”. In: *Journal of chemical theory and computation* 15.3 (2019), pp. 1827–1840.

[129] Riccardo Petraglia et al. “Beyond static structures: Putting forth REMD as a tool to solve problems in computational organic chemistry”. In: *Journal of computational chemistry* 37.1 (2016), pp. 83–92.

[130] Balint Aradi, Ben Hourahine, and Th Frauenheim. “DFTB+, a sparse matrix-based implementation of the DFTB method”. In: *The Journal of Physical Chemistry A* 111.26 (2007), pp. 5678–5684.

[131] Michael Gaus, Albrecht Goez, and Marcus Elstner. “Parametrization and benchmark of DFTB3 for organic molecules”. In: *Journal of Chemical Theory and Computation* 9.1 (2013), pp. 338–354.

[132] Michael Gaus et al. “Parameterization of DFTB3/3OB for sulfur and phosphorus for chemical and biological applications”. In: *Journal of chemical theory and computation* 10.4 (2014), pp. 1518–1537.

[133] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. “Effect of the damping function in dispersion corrected density functional theory”. In: *Journal of computational chemistry* 32.7 (2011), pp. 1456–1465.

[134] Michael W Schmidt et al. “General atomic and molecular electronic structure system”. In: *Journal of computational chemistry* 14.11 (1993), pp. 1347–1363.

[135] Mark S Gordon and Michael W Schmidt. “Advances in electronic structure theory: GAMESS a decade later”. In: *Theory and applications of computational chemistry*. Elsevier, 2005, pp. 1167–1189.

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

[137] Stephan N Steinmann and Clemence Corminboeuf. “A system-dependent density-based dispersion correction”. In: *Journal of chemical theory and computation* 6.7 (2010), pp. 1990–2001.

[138] Stephan N Steinmann and Clemence Corminboeuf. “Comprehensive benchmarking of a density-dependent dispersion correction”. In: *Journal of chemical theory and computation* 7.11 (2011), pp. 3567–3577.

[139] Stephan N Steinmann and Clemence Corminboeuf. “A generalized-gradient approximation exchange hole model for dispersion coefficients”. In: *The Journal of chemical physics* 134.4 (2011), p. 044117.

[140] Ansgar Schäfer, Hans Horn, and Reinhart Ahlrichs. “Fully optimized contracted Gaussian basis sets for atoms Li to Kr”. In: *The Journal of Chemical Physics* 97.4 (1992), pp. 2571–2577.

[141] Lori A Burns et al. “The BioFragment Database (BFDb): An open-data platform for computational chemistry analysis of noncovalent interactions”. In: *The Journal of chemical physics* 147.16 (2017), p. 161727.

[142] Andreas Singraber et al. “Parallel Multistream Training of High-Dimensional Neural Network Potentials”. In: *Journal of chemical theory and computation* 15.5 (2019), pp. 3075–3092.

[143] Bingqing Cheng et al. *Dataset: Ab Initio Thermodynamics of Liquid and Solid Water*. 2018. DOI: 10.24435/materialscloud:2018.0020/v1.

[144] Bingqing Cheng et al. “Ab initio thermodynamics of liquid and solid water”. In: *Proceedings of the National Academy of Sciences* 116.4 (2019), pp. 1110–1115.

[145] Kevin Rossi et al. *Dataset: Simulating Solvation and Acidity in Complex Mixtures with First-Principles Accuracy: The Case of CH_3SO_3H and H_2O_2 in Phenol*. 2020. DOI: 10.24435/MATERIALSCLOUD:Z9-ZR.

[146] Ulf R Pedersen et al. "Computing Gibbs free energy differences by interface pinning". In: *Physical Review B* 88.9 (2013), p. 094101.

[147] Wolfgang Lechner and Christoph Dellago. "Accurate determination of crystal structures based on averaged local bond order parameters". In: *The Journal of chemical physics* 129.11 (2008), p. 114707.

[148] Paul J Steinhardt, David R Nelson, and Marco Ronchetti. "Bond-orientational order in liquids and glasses". In: *Physical Review B* 28.2 (1983), p. 784.

[149] Gareth A Tribello et al. "PLUMED 2: New feathers for an old bird". In: *Computer Physics Communications* 185.2 (2014), pp. 604–613.

[150] Federico Giberti et al. "Iterative unbiasing of quasi-equilibrium sampling". In: *Journal of chemical theory and computation* 16.1 (2019), pp. 100–107.

[151] Mahdi Zamani et al. "3D Ordering at the Liquid–Solid Polar Interface of Nanowires". In: *Advanced Materials* (2020), p. 2001030.

[152] Sergey Pozdnyakov, Michael Willatt, and Michele Ceriotti. *Dataset: Randomly-Displaced Methane Configurations*. 2020. DOI: 10.24435/MATERIALSCLOUD:QY-DP.

[153] Michele Ceriotti, Giovanni Bussi, and Michele Parrinello. "Colored-noise thermostats à la carte". In: *Journal of Chemical Theory and Computation* 6.4 (2010), pp. 1170–1180.

[154] Giovanni Bussi, Davide Donadio, and Michele Parrinello. "Canonical sampling through velocity rescaling". In: *The Journal of chemical physics* 126.1 (2007), p. 014101.

[155] Paolo Giannozzi et al. "Advanced capabilities for materials modelling with Quantum ESPRESSO". In: *Journal of Physics: Condensed Matter* 29.46 (2017), p. 465901.

[156] Giulio Imbalzano and Michele Ceriotti. "Properties of a III-V semiconductor across temperatures and compositions from first principles". In: *arXiv preprint arXiv:2103.09041* (2021).

[157] Jesper Wallentin et al. "Hard X-ray detection using a single 100 nm diameter nanowire". In: *Nano letters* 14.12 (2014), pp. 7071–7076.

[158] JP Boulanger et al. "Characterization of a Ga-assisted GaAs nanowire array solar cell on Si substrate". In: *IEEE Journal of Photovoltaics* 6.3 (2016), pp. 661–667.

[159] Xue Chen et al. "Analysis of the influence and mechanism of sulfur passivation on the dark current of a single GaAs nanowire photodetector". In: *Nanotechnology* 29.9 (2018), p. 095201.

[160] Hyunseok Kim et al. "Monolithic InGaAs nanowire array lasers on silicon-on-insulator operating at room temperature". In: *Nano letters* 17.6 (2017), pp. 3465–3470.

[161] Martin Friedl et al. “Template-assisted scalable nanowire networks”. In: *Nano letters* 18.4 (2018), pp. 2666–2671.

[162] Martin Friedl et al. “Remote doping of scalable nanowire branches”. In: *Nano Letters* (2020).

[163] Enrique Barrigón et al. “Synthesis and applications of III–V nanowires”. In: *Chemical reviews* 119.15 (2019), pp. 9170–9220.

[164] Xiangfeng Duan et al. “High-performance thin-film transistors using semiconductor nanowires and nanoribbons”. In: *Nature* 425.6955 (2003), pp. 274–278.

[165] PC McIntyre and A Fontcuberta i Morral. “Semiconductor nanowires: to grow or not to grow?” In: *Materials Today Nano* 9 (2020), p. 100058.

[166] Karsten Albe et al. “Modeling of compound semiconductors: Analytical bond-order potential for Ga, As, and GaAs”. In: *Physical Review B* 66.3 (2002), p. 035205.

[167] DA Murdick et al. “Analytic bond-order potential for the gallium arsenide system”. In: *Physical Review B* 73.4 (2006), p. 045206.

[168] Claudia Mangold et al. “Transferability of neural network potentials for varying stoichiometry: Phonons and thermal conductivity of Mn x Ge y compounds”. In: *Journal of Applied Physics* 127.24 (2020), p. 244901.

[169] David Vanderbilt. “Soft self-consistent pseudopotentials in a generalized eigenvalue formalism”. In: *Physical review B* 41.11 (1990), p. 7892.

[170] Gianluca Prandini et al. “Precision and efficiency in solid-state pseudopotential calculations”. In: *npj Computational Materials* 4.1 (2018), pp. 1–13.

[171] Michele Ceriotti et al. “Nuclear quantum effects in water and aqueous systems: Experiment, theory, and current challenges”. In: *Chemical reviews* 116.13 (2016), pp. 7529–7550.

[172] Giovanni Bussi, Tatyana Zykova-Timan, and Michele Parrinello. “Isothermal-isobaric molecular dynamics using stochastic velocity rescaling”. In: *The Journal of chemical physics* 130.7 (2009), p. 074101.

[173] Benjamin A Helfrecht et al. “Structure-property maps with Kernel principal covariates regression”. In: *Machine Learning: Science and Technology* 1.4 (2020), p. 045021.

[174] Ask Hjorth Larsen et al. “The atomic simulation environment—a Python library for working with atoms”. In: *Journal of Physics: Condensed Matter* 29.27 (2017), p. 273002.

[175] Anubhav Jain et al. “The Materials Project: A materials genome approach to accelerating materials innovation”. In: *APL Materials* 1.1 (2013), p. 011002. DOI: 10.1063/1.4812323.

[176] SY Chiang and GL Pearson. “Properties of vacancy defects in GaAs single crystals”. In: *Journal of Applied Physics* 46.7 (1975), pp. 2986–2991.

[177] X Liu et al. “Native point defects in low-temperature-grown GaAs”. In: *Applied Physics Letters* 67.2 (1995), pp. 279–281.

[178] Peter A Schultz and O Anatole Von Lilienfeld. “Simple intrinsic defects in gallium arsenide”. In: *Modelling and Simulation in Materials Science and Engineering* 17.8 (2009), p. 084007.

[179] Marc-André Malouin, Fedwa El-Mellouhi, and Normand Mousseau. “Gallium self-interstitial relaxation in GaAs: An ab initio characterization”. In: *Physical Review B* 76.4 (2007), p. 045211.

[180] G Zollo and RM Nieminen. “Small self-interstitial clusters in GaAs”. In: *Journal of Physics: Condensed Matter* 15.6 (2003), p. 843.

[181] Akihiro Ohtake. “Surface reconstructions on GaAs (001)”. In: *Surface Science Reports* 63.7 (2008), pp. 295–327.

[182] Nikolaj Moll et al. “GaAs equilibrium crystal shape from first principles”. In: *Physical Review B* 54.12 (1996), p. 8844.

[183] JM Gibson, ML McDonald, and FC Unterwald. “Direct imaging of a novel silicon surface reconstruction”. In: *Physical review letters* 55.17 (1985), p. 1765.

[184] Shin-ichi Kitamura, Tomoshige Sato, and Masashi Iwatsuki. “Observation of surface reconstruction on silicon above 800 C using the STM”. In: *Nature* 351.6323 (1991), pp. 215–217.

[185] Walter A Harrison. “Surface reconstruction on semiconductors”. In: *Surface Science* 55.1 (1976), pp. 1–19.

[186] DJ Chadi. “Atomic structure of GaAs (100)-(2× 1) and (2× 4) reconstructed surfaces”. In: *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films* 5.4 (1987), pp. 834–837.

[187] DK Biegelsen et al. “Surface reconstructions of GaAs (100) observed by scanning tunneling microscopy”. In: *Physical Review B* 41.9 (1990), p. 5701.

[188] MD Pashley et al. “Structure of GaAs (001)(2× 4)- c (2× 8) determined by scanning tunneling microscopy”. In: *Physical review letters* 60.21 (1988), p. 2176.

[189] Binglun Yin, Zhaoxuan Wu, and WA Curtin. “Comprehensive first-principles study of stable stacking faults in hcp metals”. In: *Acta Materialia* 123 (2017), pp. 223–234.

[190] https://www.azom.com/properties.aspx?ArticleID=1132.

[191] VA Ananichev, AE Voronova, and LN Blinov. “Thermal expansion of arsenic”. In: *Russian journal of applied chemistry* 75.10 (2002), pp. 1709–1710.

[192] R Feder and T Light. “Precision Thermal Expansion Measurements of Semi-insulating GaAs”. In: *Journal of Applied Physics* 39.10 (1968), pp. 4870–4871.

[193] VM Glazov and AS Pashinkin. “Thermal expansion and heat capacity of GaAs and InAs”. In: *Inorganic materials* 36.3 (2000), pp. 225–231.

[194] T Soma, Y Saito, and H Matsuo. “Phonon Dispersion Curves near the Covalent-Metallic Transition Pressure of GaAs, GaSb, and InSb”. In: *physica status solidi (b)* 103.2 (1981), K173–K177.

[195] SI Novikova. “Investigation of thermal expansion of GaAs and ZnSe”. In: *Soviet Physics-Solid State* 3.1 (1961), pp. 129–130.

[196] TF Smith and GK White. “The low-temperature thermal expansion and Gruneisen parameters of some tetrahedrally bonded solids”. In: *Journal of Physics C: Solid State Physics* 8.13 (1975), p. 2031.

[197] PW Sparks and CA Swenson. “Thermal expansions from 2 to 40 K of Ge, Si, and four III-V compounds”. In: *Physical Review* 163.3 (1967), p. 779.

[198] George B Adams Jr, Herrick L Johnston, and Eugene C Kerr. “The heat capacity of gallium from 15 to 320 K. The heat of fusion at the melting point”. In: *Journal of the American Chemical Society* 74.19 (1952), pp. 4784–4787.

[199] C Travis Anderson. “THE HEAT CAPACITIES OF ARSENIC, ARSENIC TRIOXIDE AND ARSENIC PENTOXIDE AT LOW TEMPERATURES1”. In: *Journal of the American Chemical Society* 52.6 (1930), pp. 2296–2300.

[200] NA Gokcen. “The As (arsenic) system”. In: *Bulletin of Alloy Phase Diagrams* 10.1 (1989), pp. 11–22.

[201] JS Blakemore. “Semiconducting and other major properties of gallium arsenide”. In: *Journal of Applied Physics* 53.10 (1982), R123–R181.

[202] Ulrich Piesbergen. “Die durchschnittlichen Atomwärmen der AIII Bv-Halbleiter AlSb, GaAs, GaSb, InP, InAs, lnSb und die Atomwärme des Elements Germanium zwischen 12 und 273° K”. In: *Zeitschrift für Naturforschung A* 18.2 (1963), pp. 141–147.

[203] BD Lichter and P Sommelet. “THERMAL PROPERTIES OF A 3-B 5 COMPOUNDS. PT. 2. HIGH- TEMPERATURE HEAT CONTENTS AND HEATS OF FUSION OF INAS AND GAAS”. In: *TRANS MET SOC AIME* 245.5 (1969), pp. 1021–1027.

[204] TC Cetas, CR Tilford, and CA Swenson. “Specific heats of Cu, GaAs, GaSb, InAs, and InSb from 1 to 30 K”. In: *Physical Review* 174.3 (1968), p. 835.

[205] Atsushi Togo and Isao Tanaka. “First principles phonon calculations in materials science”. In: *Scripta Materialia* 108 (2015), pp. 1–5.

[206] Slawomir Biernacki and Matthias Scheffler. “Negative thermal expansion of diamond and zinc-blende semiconductors”. In: *Physical review letters* 63.3 (1989), p. 290.

[207] Nataliya Lopanitsyna, Chiheb Ben Mahmoud, and Michele Ceriotti. “Finite-temperature materials modeling from the quantum nuclei to the hot electrons regime”. In: *arXiv preprint arXiv:2011.03874* (2020).

[208] MC Bellissent-Funel et al. “Structure factor and effective two-body potential for liquid gallium”. In: *Physical Review A* 39.12 (1989), p. 6310.

[209] James WE Drewitt et al. “Structural Ordering in Liquid Gallium under Extreme Conditions”. In: *Physical Review Letters* 124.14 (2020), p. 145501.

[210] Pier Luigi Silvestrelli et al. “Atomic Structure of Glassy GeTe4 as a Playground to Assess the Performances of Density Functional Schemes Accounting for Dispersion Forces”. In: *The Journal of Physical Chemistry B* 124.49 (2020), pp. 11273–11279.

[211] R Bellissent et al. “Structure of liquid As: A Peierls distortion in a liquid”. In: *Physical review letters* 59.6 (1987), p. 661.

[212] Claire Bergman et al. “ON THE ATOMIC STRUCTURE OF LIQUID GaAs”. In: *Le Journal de Physique Colloques* 46.C8 (1985), pp. C8–97.

[213] LH Xiong et al. “Temperature-dependent structure evolution in liquid gallium”. In: *Acta Materialia* 128 (2017), pp. 304–312.

[214] Haiyang Niu et al. “Ab initio phase diagram and nucleation of gallium”. In: *Nature Communications* 11.1 (2020), pp. 1–9.

[215] Q-M Zhang et al. “Atomic structure and bonding in liquid GaAs from Iab-initioP molecular dynamics”. In: *Physical Review B* 42.8 (1990), p. 5071.

[216] C Molteni, L Colombo, and L Miglio. “Structure and properties of amorphous gallium arsenide by tight-binding molecular dynamics”. In: *Physical Review B* 50.7 (1994), p. 4371.

[217] Vitaliy V Godlevsky, Jeffrey J Derby, and James R Chelikowsky. “Ab initio molecular dynamics simulation of liquid CdTe and GaAs: semiconducting versus metallic behavior”. In: *Physical review letters* 81.22 (1998), p. 4959.

[218] A Gheorghiu et al. “COMPARATIVE STUDY OF THE STRUCTURE OF AMORPHOUS Ge AND AMORPHOUS III-V COMPOUNDS”. In: *Le Journal de Physique Colloques* 46.C8 (1985), pp. C8–545.

[219] Nigel J Shevchik and William Paul. “The structure of tetrahedrally coordinated amorphous semiconductors”. In: *Journal of Non-Crystalline Solids* 13.1 (1973), pp. 1–12.

[220] SC Hardy. “The surface tension of liquid gallium”. In: *Journal of Crystal Growth* 71.3 (1985), pp. 602–606.

[221] Rajaram Shetty, Raghuraman Balasubramanian, and William R Wilcox. “Surface tension and contact angle of molten semiconductor compounds: I. Cadmium telluride”. In: *Journal of crystal growth* 100.1-2 (1990), pp. 51–57.

[222] Burkhard Dünweg and Kurt Kremer. “Molecular dynamics simulation of a polymer chain in solution”. In: *The Journal of chemical physics* 99.9 (1993), pp. 6983–6997.

[223] 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 (2004), pp. 15873–15879.

[224] Othonas A Moultos et al. “System-size corrections for self-diffusion coefficients calculated from molecular dynamics simulations: The case of CO2, n-alkanes, and poly (ethylene glycol) dimethyl ethers”. In: *The Journal of Chemical Physics* 145.7 (2016), p. 074109.

[225] Seyed Hossein Jamali et al. “Finite-size effects of binary mutual diffusion coefficients from molecular dynamics”. In: *Journal of chemical theory and computation* 14.5 (2018), pp. 2667–2677.

[226] Marc J Assael et al. “Reference data for the density and viscosity of liquid cadmium, cobalt, gallium, indium, mercury, silicon, thallium, and zinc”. In: *Journal of Physical and Chemical Reference Data* 41.3 (2012), p. 033101.

[227] Nikolay Blagoveshchenskii et al. “Self-diffusion in liquid gallium and hard sphere model”. In: *EPJ Web of Conferences*. Vol. 83. EDP Sciences. 2015, p. 02018.

[228] Koichi Kakimoto and Taketoshi Hibiya. “Temperature dependence of viscosity of molten GaAs by an oscillating cup method”. In: *Applied physics letters* 50.18 (1987), pp. 1249–1250.

[229] Massimiliano Bonomi et al. “Promoting transparency and reproducibility in enhanced molecular simulations”. In: *Nature methods* 16.8 (2019), pp. 670–673.

[230] 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 (2013), p. 174502.

[231] Edoardo Baldi. *Atomistic modeling of the solid-liquid interface of metals and alloys*. Tech. rep. EPFL, 2020.

[232] Félix Musil et al. “Efficient implementation of atom-density representations”. In: *arxiv:2101.08814* (2021).

[233] Francesco Maresca et al. “Screw dislocation structure and mobility in body centered cubic Fe predicted by a Gaussian Approximation Potential”. In: *npj Computational Materials* 4.1 (2018), pp. 1–7.

[234] M Cobelli et al. “Metal-semiconductor transition in the supercooled liquid phase of the Ge 2 Sb 2 Te 5 and GeTe compounds”. In: *Physical Review Materials* 5.4 (2021), p. 045004.

[235] Yu-Xing Zhou et al. “Structure and Dynamics of Supercooled Liquid Ge2Sb2Te5 from Machine-Learning-Driven Simulations”. In: *physica status solidi (RRL)–Rapid Research Letters* 15.3 (2021), p. 2000403.

[236] a RS Wagner and s WC Ellis. “Vapor-liquid-solid mechanism of single crystal growth”. In: *Applied physics letters* 4.5 (1964), pp. 89–90.

[237] Peter Krogstrup et al. “Single-nanowire solar cells beyond the Shockley–Queisser limit”. In: *Nature Photonics* 7.4 (2013), pp. 306–310.

[238] Nikolay Panev et al. “Sharp exciton emission from single InAs quantum dots in GaAs nanowires”. In: *Applied Physics Letters* 83.11 (2003), pp. 2238–2240.

[239] Xin Yan et al. “Formation mechanism and optical properties of InAs quantum dots on the surface of GaAs nanowires”. In: *Nano letters* 12.4 (2012), pp. 1851–1856.

[240] Frank Glas, Jean-Christophe Harmand, and Gilles Patriarche. “Why does wurtzite form in nanowires of III-V zinc blende semiconductors?” In: *Physical review letters* 99.14 (2007), p. 146101.

[241] Philippe Caroff et al. “Controlled polytypic and twin-plane superlattices in III–V nanowires”. In: *Nature nanotechnology* 4.1 (2009), pp. 50–55.

[242] Eleonora Russo-Averchi et al. “High yield of GaAs nanowire arrays on Si mediated by the pinning and contact angle of Ga”. In: *Nano letters* 15.5 (2015), pp. 2869–2874.

[243] Federico Panciera et al. “Phase selection in self-catalyzed GaAs nanowires”. In: *Nano letters* 20.3 (2020), pp. 1669–1675.

[244] VG Dubrovskii. “Development of growth theory for vapor–liquid–solid nanowires: contact angle, truncated facets, and crystal phase”. In: *Crystal Growth & Design* 17.5 (2017), pp. 2544–2548.

[245] HEA Huitema, MJ Vlot, and JP Van Der Eerden. “Simulations of crystal growth from Lennard-Jones melt: Detailed measurements of the interface structure”. In: *The Journal of chemical physics* 111.10 (1999), pp. 4714–4723.

[246] So/ren Toxvaerd. “The structure and thermodynamics of a solid–fluid interface”. In: *The Journal of Chemical Physics* 74.3 (1981), pp. 1998–2005.

[247] Jeremy Q Broughton and George H Gilmer. “Molecular dynamics investigation of the crystal–fluid interface. I. Bulk properties”. In: *The Journal of chemical physics* 79.10 (1983), pp. 5095–5104.

[248] Ming Tang, W Craig Carter, and Rowland M Cannon. “Diffuse interface model for structural transitions of grain boundaries”. In: *Physical Review B* 73.2 (2006), p. 024102.

[249] Shen J Dillon et al. “Complexion: a new concept for kinetic engineering in materials science”. In: *Acta Materialia* 55.18 (2007), pp. 6208–6218.

[250] Wayne D Kaplan et al. “A review of wetting versus adsorption, complexions, and related phenomena: the rosetta stone of wetting”. In: *Journal of Materials Science* 48.17 (2013), pp. 5681–5717.

[251] Paolo Raiteri, Julian D Gale, and Giovanni Bussi. “Reactive force field simulation of proton diffusion in BaZrO3 using an empirical valence bond approach”. In: *Journal of Physics: Condensed Matter* 23.33 (2011), p. 334213.

[252] Rienk E Algra et al. “Formation of wurtzite InP nanowires explained by liquid-ordering”. In: *Nano letters* 11.1 (2011), pp. 44–48.

[253] Xiaoming Yuan et al. “Tunable polarity in a III–V nanowire by droplet wetting and surface energy engineering”. In: *Advanced Materials* 27.40 (2015), pp. 6096–6103.

[254] Mahdi Zamani et al. “Optimizing the yield of A-polar GaAs nanowires to achieve defect-free zinc blende structure and enhanced optical functionality”. In: *Nanoscale* 10.36 (2018), pp. 17080–17091.

[255] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. “Metadynamics”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.5 (2011), pp. 826–843.

[256] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. “Well-tempered metadynamics: a smoothly converging and tunable free-energy method”. In: *Physical review letters* 100.2 (2008), p. 020603.

Giulio IMBALZANO

Personal Data

BIRTH PLACE AND DATE:	Milan, Italy | 11 October 1990
ADDRESS:	Rue de Monthoux 27, 1201 Geneva, Switzerland
PHONE:	+41 78 648 4306
EMAIL:	giulio.imbalzano@gmail.com
GITHUB:	https://github.com/giulioi

Work Experience

MAY-NOV 2016 — Tutor for students with learning impairments

I have worked with students with learning impairments (e.g., dyslexia, dyscalculia), teaching them strategies to overcome their difficulties, making use of the learning aids they were provided.

Education

Current – DEC 2016 — PhD student in MATERIALS SCIENCE

École Polytechnique Fédérale de Lausanne, EPFL - Switzerland

Thesis: Transferable machine-learning models of complex materials: the case of GaAs.

Supervisor: Michele Ceriotti

Topics: Machine learning potentials, uncertainty estimation, feature selection, GaAs nanowires

DEC 2015 – SEP 2012 — MSc in NUCLEAR ENGINEERING

Politecnico di Milano - Italy, 110/110

Thesis: First principle calculations of the residual resistivity of defects in metals.

Supervisors: Pär Olsson, Carlo Bottani

Topics: Radiation protection, reactor physics, solid state physics.

MAY 2015 – SEP 2013 — MSc in NUCLEAR ENGINEERING (double degree)

KTH Royal Institute of Technology - Sweden

Topics: Radiation damage in materials, nuclear fusion, plasma physics.

Scholarships and Awards

2016 **Ermenegildo Zegna Founder's Scholarship**
50k€ for promising Italian students who pursue higher education abroad.

2015 **Sigvard Eklund prize**
35k SEK for best Swedish Master thesis in Nuclear Engineering

2013 **Scholarship from Politecnico di Milano**
for students studying abroad

Publications

2021 The role of feature space in atomistic learning
Goscinski, A., Fraux, G., **Imbalzano, G.**, and Ceriotti, M. (2021). Machine Learning: Science and Technology, 2(2), 025028.

2021 Properties of a III-V semiconductor across temperatures and compositions from first principles
Imbalzano, G., Ceriotti, M. arXiv preprint arXiv:2103.09041.

2021 Uncertainty estimation for molecular dynamics and sampling
Imbalzano, G., Zhuang, Y., Kapil, V., Rossi, K., Engel, E. A., Grasselli, F., and Ceriotti, M. (2021). The Journal of Chemical Physics, 154(7), 074102.

2020 3D Ordering at the Liquid–Solid Polar Interface of Nanowires
Zamani, M., **Imbalzano, G.**, Tappy, N., Alexander, D. T., Martí-Sánchez, S., ... Ceriotti, M., and Fontcuberta i Morral, A. (2020). Advanced Materials, 32(38), 2001030.

2018 Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials
Imbalzano, G., Anelli, A., Giofré, D., Klees, S., Behler, J., Ceriotti, M. The Journal of chemical physics, 148(24), 241730.

2018 Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions
Nguyen, T.T., Székely, E., **Imbalzano, G.**, Behler, J., Csányi, G., Ceriotti, M., Götz, A.W., Paesani, F. The Journal of chemical physics, 148(24), 241725

Software and Programming Languages

Linux, Python, C, C++, LAMMPS, VASP, QUANTUM ESPRESSO, Git, slurm, scikit-learn, pyTorch, TensorFlow,MATLAB, Mathematica, R, SERPENT, Inkscape, Blender, LaTeX, SOLIDWORKS, Windows, Microsoft suite

Languages

ITALIAN:	Native	FRENCH:	Fluent
ENGLISH:	Fluent	SWEDISH:	Good

Interests and Activities

Technology, open-source, programming
Running, martial arts, cooking, travelling

Other information

30, Married, driving license, B Swiss permit, Italian citizenship (no Military obligations)