EPFL

Thèse n° 10 071

# Machine-learning the electronic density of states: electronic properties without quantum mechanics

Présentée le 20 mars 2023

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

pour l'obtention du grade de Docteur ès Sciences

par

**Chiheb BEN MAHMOUD**

Acceptée sur proposition du jury

Prof. A. Mortensen, président du jury
Prof. M. Ceriotti, directeur de thèse
Prof. J. Kermode, rapporteur
Prof. R. Meißner, rapporteur
Prof. N. Marzari, rapporteur

École polytechnique fédérale de Lausanne

2023

# Acknowledgements

I want to take this opportunity to acknowledge the many people involved in making this PhD possible. I cannot begin to express my thanks to my supervisor Prof. Michele Ceriotti for supporting me up to the completion of this thesis. His guidance and his ability to keep attention to the details while maintaining a global vision have been an inspiration. Thanks to him, I had the opportunity to contribute to an exciting field and participate in several collaborations. I am forever grateful for his enthusiasm and his open-door (and Slack) policy that helped keep the discussion going.

I would like to extend my sincere gratitude to Dr. Andrea Anelli and Dr. Federico Grasselli. I was lucky to collaborate with them on exciting projects and learn from their experiences. They extended me a great amount of assistance in expanding my skill set. Their enthusiasm, productivity, and detail-oriented approach to research have helped me transform objectives into achievements.

I have been lucky to call COSMO my academic home. I am thankful to every current and past group member for insightful and thought-provoking discussions. Their diverse (scientific) background has been essential in expanding my knowledge beyond the scope of my project. Many interactions in COSMO occur around coffee and/or food. In particular, I would like to thank Andrea, Félix, Alexander, Natasha, Federico Gr., Venkat, Guillaume, Max, Jigyasa, Federico Gi., Rose, Raymond, Sergey, Lorenzo, and Davide for all the discussions, support and advice. Special thanks to Anne for being a source of positivity, even during the darkest (academic) days.

I cannot leave Lausanne without thanking all the non-COSMO friends I made along the way. Special thanks to Norma, Sarra, Francesco, Farzad, and Simran for being a part of this journey since its beginning.

My appreciation and gratitude go towards my wife, Cyrine, for coping with the highs and the many lows of a PhD life. I cannot wait to finally start our life together and make up for our lost time. Last but not least, I want to express my gratitude to my parents, Mongi and Naziha, and my sister, Hadil, for being a constant source of support and inspiration. I am forever indebted!

*Lausanne, 16 February 2023* C.B.M

# Abstract

The electronic density of states (DOS) quantifies the distribution of the energy levels that can be occupied by electrons in a quasiparticle picture and is central to modern electronic structure theory. It also underpins the computation and interpretation of experimentally observable material properties such as optical absorption, electrical conductivity, and heat capacity. It can be accurately computed through expensive first-principle calculations, limiting the size of the problems that can be simulated easily to a few thousand atoms. Machine-learning (ML) techniques are a promising alternative to these calculations, as they were successfully applied to study many atomic-scale problems by generalising information from small configurations to large and complex structures. However, most efforts focused on learning the ground-state Born-Oppenheimer energies and the atomic forces, which are scalar quantities, unlike the DOS, which is a multivariate function of the energy.

In this thesis, we discuss the inherent challenges in constructing an ML framework that predicts the DOS as a combination of local contributions that depend, in turn, on the geometric configuration of neighbours around each atom. We compare different approaches to represent the DOS as a learning target and the accuracy of predicting quantities such as the Fermi level, the electron density at the Fermi level, or the band energy, either directly or as a side product of the evaluation of the DOS.

As a first benchmark, we evaluate our model on a challenging case study that includes configurations of silicon spanning a broad set of thermodynamic conditions, ranging from bulk structures to clusters and from semiconducting to metallic behaviour.

Then, we leverage the atom-centredness of the model to compute the DOS of large amorphous silicon samples, for which it would be prohibitively expensive to compute the DOS by direct electronic structure calculations. Besides the size transferability, we show that this decomposition of the DOS can extract physical insights into the connections between structural and electronic features to describe their transitions in disordered silicon phases.

Finally, we explore two approaches to using the DOS in integrated ML frameworks to model the properties of materials, where the DOS is used to incorporate the effect of thermal excitations

of electrons. We propose to combine simulations from well-established ML interatomic potentials with band energy calculations extracted from DOS predictions on the already-produced trajectories. This procedure successfully describes the heat capacity of molten nickel and is in agreement with the experiments. However, we show that this method is only valid when the dynamics of the ions are, to a large extent, not affected by the electronic excitations, and it would fail in conditions with higher temperatures, such as those found in astrophysical settings. Therefore, we introduce an integrated ML framework that includes these thermal effects in constructing the interatomic potential. The novelty of this method is that the electronic temperature is an external parameter of the simulation because one only needs access to ground-state energies, forces and DOS. We successfully apply our model to study metallic hydrogen in the conditions of a young Jupiter core. We reconstruct its equation of state and its heat capacity and find that they are compatible with their first-principle-derived counterparts. The work of this thesis demonstrates the impact of a physics-inspired universal model describing structural and electronic properties inexpensively and its ability to enable more accurate and predictive materials modelling and design.

Keywords: machine-learning, multivariate, electronic structure, density of states, finite temperature

# Résumé

La densité d'états électroniques (DOS) quantifie la distribution des niveaux énergétiques qui peuvent être occupés par les électrons dans une image de quasi-particule et est au cœur de la théorie moderne de la structure électronique. Elle sous-tend également le calcul et l'interprétation des propriétés matérielles observables expérimentalement, telles que l'absorption optique, la conductivité électrique et la capacité thermique. Elle peut être calculée avec précision par des calculs ab initio coûteux, ce qui limite à quelques milliers d'atomes la taille des problèmes qui peuvent être simulés facilement. Les techniques d'apprentissage automatique (ML) constituent une alternative prometteuse à ces calculs, car elles ont été appliquées avec succès à l'étude de nombreux problèmes à l'échelle atomique en généralisant les informations provenant de petites configurations à des structures larges et complexes. Cependant, la plupart des efforts se sont concentrés sur l'apprentissage des énergies de Born-Oppenheimer à l'état fondamental et des forces atomiques, qui sont des quantités scalaires, contrairement à la DOS, qui est une fonction multidimensionnelle en l'énergie.

Dans cette thèse, nous discutons les défis inhérents à la construction d'un cadre ML qui prédit la DOS comme une combinaison de contributions locales qui dépendent, à leur tour, de la configuration géométrique des voisins de chaque centre atomique. Nous comparons différentes approches pour représenter la DOS comme cible d'apprentissage et la précision de la prédiction de quantités telles que le niveau de Fermi, la densité d'électrons au niveau de Fermi ou l'énergie de bande, soit directement, soit comme produit secondaire de l'évaluation de la DOS.

Comme premier point de référence, nous évaluons notre modèle sur une étude de cas difficile qui comprend des configurations de silicium couvrant un large éventail de conditions thermodynamiques, allant du comportement semi-conducteur au comportement métallique.

Ensuite, nous tirons parti de la centralité atomique du modèle pour calculer la DOS de grands échantillons de silicium amorphe, pour lesquels il serait excessivement coûteux de calculer la DOS par des calculs directs de structure électronique. Outre la transférabilité de la taille, nous montrons que cette décomposition de la DOS permet d'extraire des informations physiques

sur les connexions entre les caractéristiques structurelles et électroniques afin de décrire les transitions dans les phases désordonnées du silicium.
Enfin, nous explorons deux approches de l'utilisation de la DOS dans des cadres ML intégrés pour modéliser les propriétés des matériaux, où la DOS est utilisée pour incorporer l'effet des excitations thermiques des électrons. Nous proposons de combiner des simulations à partir de potentiels interatomiques ML bien établis avec des calculs d'énergie de bande extraits des prédictions de la DOS sur les trajectoires déjà produites. Cette procédure décrit avec succès la capacité thermique du nickel fondu et est en accord avec les expériences. Cependant, nous montrons que cette méthode n'est valable que lorsque la dynamique des ions n'est, dans une large mesure, pas affectée par les excitations électroniques, et qu'elle échouerait dans des conditions de températures plus élevées, telles que celles rencontrées dans les milieux astrophysiques. Par conséquent, nous introduisons un cadre intégré ML qui inclut ces effets thermiques dans la construction du potentiel interatomique. La nouveauté de cette méthode est que la température électronique est un paramètre externe de la simulation car il suffit d'avoir accès aux énergies, forces et DOS de l'état fondamental. Nous appliquons avec succès notre modèle pour étudier l'hydrogène métallique dans les conditions d'un jeune noyau de Jupiter. Nous reconstruisons son équation d'état et sa capacité thermique et constatons qu'elles sont compatibles avec leurs équivalents dérivés des premiers principes. Le travail de cette thèse démontre l'impact d'un modèle universel inspiré de la physique décrivant les propriétés structurelles et électroniques de manière peu coûteuse et sa capacité à permettre une modélisation et une conception des matériaux plus précises et prédictives.

Mots clefs : apprentissage automatique, fonction multivariée, structure électronique, densité des états, temperature finie

# Contents

# List of Figures

# List of Tables

# 1 Introduction

Computational materials science is an interdisciplinary field that combines physics, chemistry, computer science and data analysis to study and design materials using modelling, simulation and theory. It aims to acquire complete insight into materials' microscopic and macroscopic behaviour, leading to a better understanding of the processing challenges, experimental feasibility, and conditions necessary to optimise specific properties. In general, computer simulations both provide explanations for experimentally observed phenomena [1], and can make experimentally-verifiable discoveries like the effect of dopants on hydrogen adsorption on $CeO_2$ surfaces [2], the stability and functional properties of inorganic materials like TaCoSn and $TaCo_2Sn$ [3] or proton exchange on water membrane catalysis [4].

Computational approaches to materials design avoid problems associated with experimental methods, such as the high setup cost or handling of poisonous substances. These techniques also aim to guide manufacturing processes by providing insight into how a material would behave in certain situations or by determining the optimal conditions to enhance a particular property. Depending on the demands and design of our system, available computational power, and the length and time scales of the studied system, there are many appropriate options for computational theory and software. As a rule of thumb, the larger the system is, the more approximations we need to make in order to simulate a material due to the lack of computational resources, even despite the extraordinary development of computing platforms and hardware. The different needs of simulating different length scales led to the development of different categories of methods. If a method leverages the atomic behaviour of a material in order to infer its properties, it falls under the so-called atomistic methods. Examples of atomistic techniques include molecular dynamics (MD), Monte Carlo (MC) and kinetic Monte Carlo. If the method is interested in systems of the size of real-world devices, they fall under the continuum methods, such as the finite-element solvers. A schematic representation including

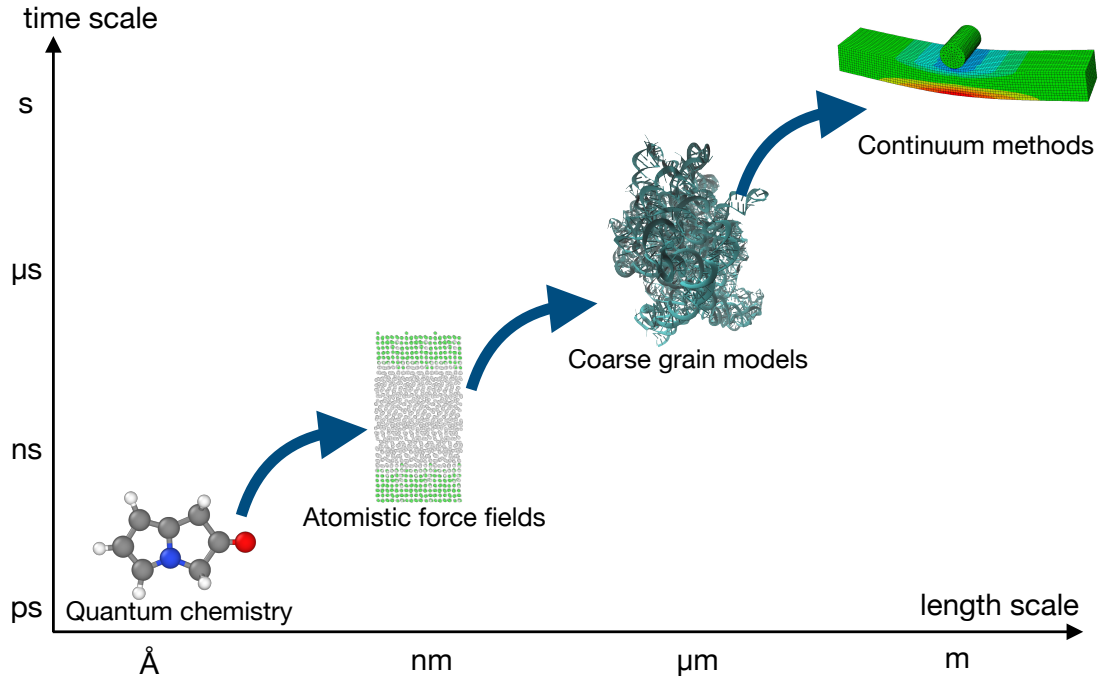


Figure 1.1: An overview of the length scale and time scale of some of the techniques used in computational materials science ranging from the study of the electronic structure to the mechanical properties and reactions at a human time scale. The continuum figure (top right corner) is adapted from Ref. [5]. Original figure published under the CC-BY 4.0 license (https://creativecommons.org/licenses/by/4.0/)

some of these methods, the length/time scales and examples of possible applications is given in Fig. 1.1.The combination of atomistic and continuum methods can lead to the development of multi-scale modelling, where the properties and behaviour of materials at the atomic scale are used to inform and improve the predictions made at the macroscopic scale. This allows for a more accurate and comprehensive understanding of the behaviour of materials across multiple scales.

This thesis focuses on the first set of methods: the atomistic techniques. Their goal is to understand and model the motion of individual atoms, as their collective behaviour can describe a material's macroscopic response to deformation and phase transition, for example. They are usually paired with an interatomic potential that can describe the nature of the interactions between the individual atoms with different levels of accuracy. Historically, interatomic potentials have a functional form with a restricted number of parameters that can be tuned to reproduce an experimental observable like the shear viscosity of the pair correlation function. These parameters usually have physical interpretations, such as the case

of the bond equilibrium distance in Lennard-Jones or Morse potentials. They can also be tuned to reproduce quantities derived from a quantum mechanical (QM) calculation, like the lattice constant or the cohesive energy. More complicated and non-parametric, i.e. potentials with hundreds of parameters or more, are tuned nowadays using machine learning algorithms that we introduce later in this chapter. One objective of these potentials is to be transferable in the broad sense of the term. The transferability of an interatomic potential can be expressed in terms of system size, system chemical composition, system thermodynamic conditions and chemical reactions as well. Some of these problems can be circumvented if one leverages QM calculations to compute the interatomic forces instead of fitting a few parameters, performing *ab initio* or first-principle simulations.

Ab initio methods solve the Schrödinger equation with various levels of approximations, leading to a wide range in approaches, ranging from a simple parameterisation of the electronic interactions like the local combination of atomic orbitals (LCAO) and tight binding [6] methods to the post-Hartree-Fock methods like the coupled cluster approach [7].
The most famous of such techniques is density functional theory (DFT), an elegant solution to the ground-state many-body time-independent Schrödinger problem. DFT is relatively cheap computationally, compared to higher-level-theory QM calculations such as Quantum Monte Carlo, and has been optimised for modern high-performance computing and with scalable implementations. DFT is a ground-state formulation of the many-body problem, which leverages the Born-Oppenheimer approximation that decouples the ionic and electronic degrees of freedom, based on the existence of a mapping between the atomic coordinates of a system and its electronic structure. With DFT, we can perform many sophisticated applications and compute numerous properties of materials ranging from vibrational [8, 9], thermodynamic [10, 11], electrical [12] and mechanical [13, 14] properties. The conditions of these numerical experiments cover a wide range of pressures and temperatures ranging from those similar to the surface of the earth up to the core of exoplanets, where the temperature and pressure are several orders of magnitude higher than we live in. However, we are limited in the system sizes and time scales we can probe using DFT (a few thousand atoms and tens of picoseconds of trajectories), as most current implementations of DFT scale cubically with the number of particles in the system.
The basics of DFT rely on the Hohenberg-Kohn [15] and Kohn-Sham [16] theorems that reduce the many-body problem into a system of non-interacting quasi-particles "swimming" in an average field of interactions from the other particles. The parameterisation of DFT often occurs in the choice of the exchange-correlation functional. These functionals are either fitted to certain empirical data or are constructed to comply with theoretical constraints, such as reproducing properties predicted from Quantum Monte Carlo. One of the first attempts to

solve the exchange-correlation problem is the local density approximation (LDA) [16], where a system of homogenous gas approximates the electron cloud. Later, different techniques emerged, such as the Perdew–Burke–Ernzerhof [17] (PBE) functional, which is a general-purpose functional because it is a non-empirical functional with reasonable accuracy over a wide range of systems [18]. There is also a range of hybrid functionals targeted to specific electronic properties, such as the Heyd-Scuseria-Ernzerhof [19, 20] (HSE06) functional, which usually predicts the energy band gap accurately [21]. These functionals are usually derived to satisfy physical constraints or to reproduce properties computed from a higher-level theory calculation. The development of this myriad of exchange-correlation functionals raised the question of the conformity of the calculations and the subsequent property predictions. This uncertainty and these discrepancies led to the development of the ensemble Bayesian error estimation functional (BEEF) [22] as a method to make the best compromise between the target properties of the included functionals.

In the Kohn-Sham DFT framework, it is possible to consider the total energy of an atomic system as a functional of two primary quantities of the electronic structure theory, the charge density ($\rho$) and the electronic density of states (DOS). $\rho$ is an experimental observable and is defined as the modulus of the wavefunction. It describes the probability of finding an electron at a specific coordinate in space. Experimental techniques such as X-ray diffraction or transmission electron microscopy (TEM) [23] are used to measure this electron density. These measurements locate the atomic positions, identify chemical bonds or indicate molecular size and shape [24].
The DOS is a measure of the number of electronic states available for occupation by the electrons within a given energy range. Beyond its utility in computing the total energy, the DOS allows us to determine the material's metallicity and electrical conductivity at the ground state. When the transition rates are known, it is possible to determine the optical electron transitions via the Fermi golden rule. The DOS also allows one to evaluate the contribution of thermally activated electronic states by examining its value near the Fermi energy, which can then be used to quantify the contribution of electrons to the specific heat capacity of a system. In a textbook example, this can be directly seen from the Sommerfeld approximation, where the heat capacity, entropy, and band energy are a function of fundamental constants and the value of the DOS at the Fermi level.

As mentioned earlier, DFT suffers from cubic scaling, making its use limited to a few thousand atoms. One possible approach to avoid performing expensive calculations is to build surrogate models. A popular choice to perform this task is Machine learning (ML). It can be defined as the set of statistical methods employed to extract trends and laws from a set of data in order to make automated inferences about new observations. ML techniques can use structural

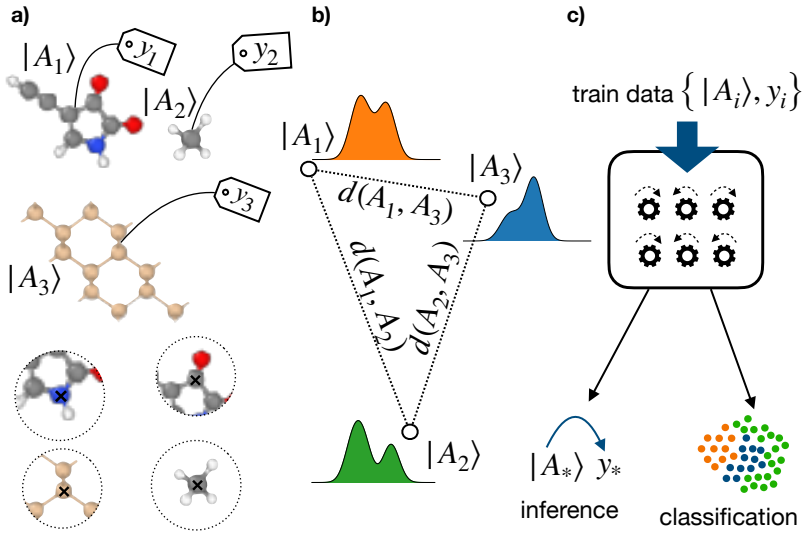


Figure 1.2: A representation of the three main ingredients in an ML workflow for atomistic modelling and the construction of MLIPs: (a) the label input structures or molecules (b) the structural representation, and (c) the learning algorithm.

properties of materials, like chemical composition, atomic positions, or volume, to build and train models able to reproduce target properties like the ionisation energy, chemical shifts or bulk modulus. The ML models can also accelerate atomistic simulations when trained on QM data like DFT energies and/or forces by fitting machine-learning interatomic potentials (MLIP)s. It becomes possible to perform MD or MC simulations with quantum-mechanical accuracy at a fraction of the cost [25, 26, 27] or design molecules and materials with specific properties using generative algorithms [28, 29]. This is possible due to the linear scaling of MLIPs with system size, expressed by the additivity of locally-defined quantities and based on the assumption of nearsightedness of matter [30] in contrast with the cubic scaling of QM methods, and one usually expects three to four orders of magnitude speed-up per calculation step. These gains opened the door to performing more efficient high-throughput screening of materials with the desired thermal or dynamical properties [31, 32, 33]. Nowadays, it is becoming common to read about works involving millions of atoms simulating nano-scale devices [34, 35, 36] or studying phenomena requiring a large length scale and hence a high number of atoms, e.g. ice nucleation [37] and amorphous silicon compression [38].

The other family of ML models is known as the unsupervised techniques, which are helpful in identifying trends, studying similarities and evaluating patterns and clustering in the data. We also see newly applied methods that combine supervised and unsupervised approaches to guide the construction of maps that are aware of target properties and the geometrical correlations or bias target properties with the geometrical features of the training data [39].

There are three key ingredients one should keep in mind when constructing MLIPs for atomistic modelling and ML models in general, as illustrated in Fig. 1.2. The first significant ingredient is the training data. The goal of building an MLIP is to sample a potential energy surface (PES), which implies the need to reproduce not only the energy values of a structure but also the atomic forces and the stress tensor since they are gradients of the energy. Depending on the chemical nature of the system and the target property, there are several datasets and repositories organised by class of materials and data sources, including the QM9 [40] and the Cambridge Structural Database [41]. However, they are not always sufficient to build an accurate ML model, e.g. they may lack the necessary information to capture the effect of thermal fluctuations. A typical approach is to add structures from an MD trajectory driven targeting several conditions. In recent years, different systematic strategies have been proposed to reduce human intervention to tackle the augmentation issue. Some of them rely on uncertainty quantification by training MLIPs iteratively or by biasing the models towards some areas of the phase space [33].

The second essential ingredient is the encoding of the atomic configuration, also known as features or descriptors. These descriptors transform a structure into a vector in a feature space and then be used as inputs for the MLIPs. While many end-to-end algorithms take only the atomic Cartesian coordinates as inputs, this is inefficient as the prediction of energies and forces will depend on the Cartesian reference and the order of the atoms. However, this does not mean that these features cannot be derived from any physical or structural considerations. In fact, it is important to select the descriptors that provide a comprehensive description of the atomic arrangements and correlate well with the target properties. For MLIPs, it is important to choose a descriptor that respects the smoothness and the symmetries of the potential energy surface. Also, it is essential to provide a unique description to every structure to ensure the reliability of the predictions made by the MLIPs. Density-based descriptors are a promising class of descriptors that are able to encode the rotational and translational symmetries of a structure by acting on the Cartesian coordinates of atoms. Examples of this class of features include radial distribution function, the smooth overlap of atomic positions (SOAP) [42], the Behler-Parrinello symmetry functions [43] and the atomic cluster expansion (ACE) [44]. One common aspect between these atomic descriptors is that they allow for atom-centred representation of a system, which results in a natural size extension formulation.

The last ingredient of the ML workflow is the learning algorithm itself. These algorithms try to establish the mapping between the inputs, like the atomic descriptors, and the labels, like the DFT energies or the DOS, in the case of supervised learning or identify patterns in unlabeled data in the case of unsupervised learning. Regardless of the task at hand, they usually have a set of hyperparameters that need to be optimised in order to guarantee the

main objective of ML, which is the generalisation of the rules identified during the training phase to new cases (or atomic structures). Depending on the task at hand, some approaches are more suitable than others. Examples of supervised ML models include artificial neural networks [45], graph neural networks [46], Gaussian Processes [47] and decision trees. Among unsupervised learning, we can mention support-vector machines and principal component analysis.

The recent developments in the use of ML methods to circumvent the expensive costs of QM calculations focus on modelling the potential energy surface and its gradients, i.e. the atomic forces and the stress tensor, since they are necessary to run MD simulations. These surrogate ML models successfully overcome the drawbacks of both classical interatomic potentials (low accuracy) and quantum mechanical calculations (high cost). The use of ML in materials modelling goes beyond the total energies and forces to other electronic structure properties such as the positions of the Wannier centres [48], the exchange-correlation energies since they are responsible for the accuracy of the DFT approach [49], or the kinetic energy functional [50, 51], in an orbital-free DFT approach. We also see efforts tackling surrogate models for primary outputs of a DFT calculation, such as the local density of states resolved in energy and space [52, 53], the electron density $\rho$ [54, 55, 56, 57], and the electronic DOS [58, 59, 60, 61, 58, 62, 63]. Learning these quantities yields predictions that can also be used to calculate some quantities of interest indirectly, in a physics-inspired approach, like the band energy or the exchange-correlation energy, sometimes providing better accuracy than models targeting these quantities directly [64]. Most of these surrogate models rely on additivity from atom-centred terms, which has implications for the interpretability of the correlations between structural features and target properties. In particular, there have been works that define surrogate models for DOS based on additivity from individual atoms and use these fingerprints to characterise atomic environments further, contributing to more comprehensive structure-property relations [65, 66, 38]. Finally, it is worth mentioning a new trend in atomistic ML techniques, which focuses on modelling the Hamiltonian [67, 68]. The latter holds the necessary information to compute the different electronic structure properties, like the DOS, hence reducing the number of surrogate ML models while maintaining the highest levels of fidelity to the QM description of the material.

## 1.1 Summary

The development and use of universal models describing structural and electronic properties inexpensively set the foundations for more accurate and predictive materials modelling and design and removing barriers between physics and data-driven modelling. This thesis aims to

bridge a gap in the modelling of the fundamental electronic structure properties using ML techniques. In particular, we focus on modelling the electronic density of states and explore how it can be used within ML workflows to assist with their predictions.

This thesis is organised as follows: Chapter 2 introduces the general machine learning framework that allows us to model the electronic density of states as a vector-valued function of the energy, utilising the similarity measure between atomic environments introduced by the SOAP representation. Chapter 3 describes the numerical experiments used to validate the ML model for the electronic DOS in a challenging data set of silicon structures describing several thermodynamic conditions and structures ranging from distorted bulk snapshots to clusters. It also explores the use of the DOS as a fingerprint to identify the possible geometrical features of local environments to describe structural and electronic transitions in disordered phases. Chapter 4 introduces a theoretical framework to perform finite temperature modelling of condensed matter using the DOS without recomputing data sets with expensive QM methods and based only on a single temperature calculation, usually the ground state. It also discusses different levels of possible approximations and an application of this model to study hydrogen in the high-pressure/high-temperature regime. Conclusions are drawn in Chapter 5.

## List of publications

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

1. **Ben Mahmoud, C.**, Anelli, A., Csányi, G., Ceriotti, M., 2020. Learning the electronic density of states in condensed matter. Phys. Rev. B 102, 235130.
2. Deringer, V.L., Bernstein, N., Csányi, G., **Ben Mahmoud, C.**, Ceriotti, M., Wilson, M., Drabold, D.A., Elliott, S.R., 2021. Origins of structural and electronic transitions in disordered silicon. Nature 589, 59–64.
3. Lopanitsyna, N., **Ben Mahmoud, C.**, Ceriotti, M., 2021. Finite-temperature materials modelling from the quantum nuclei to the hot electrons regime. Phys. Rev. Materials 5, 043802
4. **Ben Mahmoud, C.**, Grasselli, F., Ceriotti, M., 2022. Predicting hot-electron free energies from ground-state data. Phys. Rev. B 106, L121116

# 2 Methods

## 2.1 Introduction

Machine-learning (ML) is the field of study interested in building methods to infer knowledge from data. This means that ML focuses on constructing statistical methods and algorithms to analyse datasets and potentially identify patterns and trends. This acquired knowledge is then used to predict the behaviour of future (test) cases, even those that have not been considered during the learning or training phase. This aspect of generalisation, without explicit knowledge about the analytical relations and correlations between the data, makes ML techniques intriguing and compelling. They require minor human intervention and constraints in the training phase. The generalisation is also an important metric to evaluate the quality of the training data and the ML algorithm.

Generally, the more data we add, the more accurate the ML model is and the more generalisable it becomes. However, the raw input data, e.g. atomic coordinates in atomistic systems, is rarely a suitable candidate for building the training datasets. Let us consider a molecule in a vacuum. Its cohesive energy should remain constant under any rotation or translation, although these operations will affect the atomic coordinates. This thought experiment justifies the need for proper descriptors, also known as fingerprints or representations, that encode the symmetries of the problem and correlate with the modelled observation. Needless to say that the representation must be unique for every point in the data set because, otherwise, the ML models would assign the same target property like the cohesive energy for two different configurations, which makes the ML models unreliable.

Depending on the available data, there are two main families of ML techniques: unsupervised and supervised learning. The significant difference between the two families is the labelling of the data. The first set focuses on unlabelled data by exploring hidden correlations between

input points and finding rules to cluster them. The second set deals with labelled data by training models that classify the observations or predict the outcomes accurately.

This chapter focuses on two fundamental blocks in the ML workflow in materials science and atomistic simulations. First, we are interested in structure representation as a method to increase the robustness of the model to symmetry operations. In particular, we go over the basics of density-based descriptors. Next, we dive deeply into one of supervised ML's widely-used algorithms, Gaussian Process regression (GPR). We focus on its kernel formulation and its extension to vector-valued functions. The latter is considered the backbone of modelling and learning the electronic density of states (DOS).

## 2.2 Density based descriptors

We begin by providing a brief overview of some of the strategies to build atomic descriptors and representations. The field of geometrical descriptions of atomic arrangements has experienced very rapid growth in the past decade. Even though dozens of alternative descriptors have been proposed, most of the atomic descriptors ranging from Behler-Parrinello symmetry functions [43], first introduced in 2007, till the latest descriptors such as the Atom Cluster Expansion [44] (ACE) or NICE [69], can be derived from the same considerations [70].

Taking one step back, atomic fingerprints are not a new idea that appeared with the emergence of machine-learning techniques for atomistic modelling. In particular, empirical potentials included some sort of atomic descriptors in their functional forms to describe the nature of interactions between the different atomic species present in a system, including the Lennard-Jones potential or the Gupta empirical potential [71] that use the pairwise distances as descriptors. The most basic atomic descriptor is the set of Cartesian coordinates of atoms in space. This approach requires a feature space of $3N$ dimensions, with $N$ the number of atoms in a particular frame, to describe the atoms' arrangement. Despite being simple, this descriptor is used as a fingerprint for several applications, including the challenging characterisation of structural motifs in biomolecular complexes [72]. However, it presents multiple disadvantages. First, it is linked to a fixed Cartesian reference. So it would produce different fingerprints for the same atomic arrangement if we rotate the reference or translate ins origin. For instance, if one considers two water molecules where one is obtained by rotation of the other around its principal axis, we will get two different sets of atomic positions. Also, we obtain different fingerprints if we permute the labels of the two hydrogen atoms in the water molecule. Second, it does not provide a systematic way to compare structures with different types and numbers of atoms. Finally, depending on the nature of the studied problem, some properties might not be invariant, i.e. their value depends on the orientation of the molecule. Suppose one is

interested in modelling the dipole moment of the water molecule. The descriptor should be able to encode how a given quantity changes when a rigid transformation is applied to the global system. In that case, one does not need to explore all the possible molecule's rotations to find the corresponding dipole moment. In fact, they should have the same magnitude but differ only in direction.
This simple example highlights certain essential aspects needed for designing structural descriptors. They must respect the internal symmetries of a structure, including translational, rotational and permutational symmetries. While doing so, it must also be ensured that the representation is smooth with respect to the atomic displacements since most of the physical observables are also smooth with respect to these perturbations. Some of these issues have already been solved in the empirical potential community by moving to internal coordinates, like pairwise distances and dihedral angles.

They should also provide a natural system size extension, which is usually achieved by assuming the additivity of local or atomic fingerprints. The locality assumption is also justified by the nearsightedness of matter [30], a key assumption in several branches of physical sciences, including the theory behind the linear scaling density functional theory [73], and also in machine-learning approaches for atomistic modelling. It is this underlying assumption that we will use throughout this manuscript. In practice, the locality assumption can be tested for atomic properties like the forces by considering an atomic environment defined by a centre and a radial cutoff. The procedure consists in measuring the variance of the atomic property computed for the central atom by randomly displacing atoms outside the environment as a function of the radial cutoff. Finally, one other important property should be completeness, i.e. an atomic fingerprint must be unique for every structure, and two different arrangements of atoms must yield different fingerprints by the descriptor. This assumption turns out to be the most difficult to achieve. In atom-density descriptors, one must have access to high body-order correlations to ensure the uniqueness of fingerprints [74].

In general, the atomic descriptors used in the context of atomistic modelling can be classified into two main groups: global and local. Global descriptors include the Coulomb matrices [75], bag of bonds [76], many body tensor representation [77] and the mixture of local environments [78]. It has been shown in Ref. [70] that they are projections on different basis sets of the same algebraic object, i.e. the atomic density field $\rho$. We use the notations first introduced in Ref. [70], mimicking the Dirac notations (bra-ket) used in quantum mechanics. The reasoning behind this choice is that they highlight the basis-independence of the atom-density representations.
We define $|A\rangle$ as the object holding all the information about a structure $A$ and the atoms it contains, including the chemical composition and the spatial arrangement of the atoms. It

is possible to complement the indication in $|A\rangle$ with the nature of the representation, and we write the abstract object $|A;\text{rep.}\rangle$, where "rep." is a shorthand that describes the kind of correlations that underlies the featurisation. For example, we use the $|A;\rho\rangle$ notation to emphasise the nature of the atom density field used to describe the structure $A$. Examples of this class of descriptors are the smooth overlap of atomic position (SOAP) [42], the features of the moment tensor potential [79], or the features of the SNAP [80] architecture. We proceed by decomposing the global representation of the structure $A$ into a sum of its atomic representations:

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

where $|\mathbf{r}_i; g\rangle$ is the representation of the $i$th atomic centre placed at $\mathbf{r}_i$, and $g$ is usually a smooth localised function at the atomic centres, e.g. a Gaussian. The general expression of Eq. (2.1) highlights these structural representations should be regarded as formal/abstract mathematical objects independent of the basis, e.g. real space or plane waves, or the smooth function $g$. In the case of a projection of the ket $|A;\rho\rangle$ on real space represented by the bra $\langle\mathbf{x}|$, we obtain the following:

$$\langle\mathbf{x}|A;\rho\rangle = \sum_{i\in A} \langle\mathbf{x}|\mathbf{r}_i; g\rangle \equiv \sum_{i\in A} g(\mathbf{x}-\mathbf{r}_i). \tag{2.2}$$

By taking the limit of the smooth function $g$ to a delta Dirac function $\delta$ at every atom, we recover the Cartesian coordinates of the atoms within the structure $A$. In general, the change of basis is obtained following the same formulas introduced in most quantum mechanics courses, also known as the completeness relation:

$$\langle T|A\rangle = \int \mathrm{d}Q \langle T|Q\rangle\langle Q|A\rangle, \tag{2.3}$$

where $|T\rangle$ is a basis of projections and the set of smooth functions $|Q\rangle$ forms a complete basis. The equality holds for sums over countable discrete indices if the basis is not continuous.

Eq. (2.2) is not translationally invariant as it relies explicitly on the atomic positions. The first step towards translation symmetry is to integrate the density over all the possible $\mathbb{R}^3$ translations $\hat{t}$ in our bra-ket notations $\int \mathrm{d}\hat{t}\langle\mathbf{x}|\hat{t}|A;\rho\rangle$. This operation is known as the Haar integration, and its objective is to symmetrise an operator with respect to elements in a particular group. One problem that arises from performing the symmetrisation on the global field $|A;\rho\rangle$ is the integration of all the spatial degrees of freedom, and the final result is a constant proportional to the total number of particles in the simulation box. A similar problem would occur if we consider the Haar integration over the $SO(3)$ group of elemental rotations and lose all the angular information.

One possible method to mitigate these issues is to consider the tensor product of density fields instead of the density. The $\hat{t}$-symmetrised two-point correlation function is

$$\langle a_1\mathbf{x}_1; a_2\mathbf{x}_2|\langle\rho\otimes\rho\rangle_{\mathbb{R}^3}\rangle = \int \mathrm{d}\hat{t}\langle a_1\mathbf{x}_1|\hat{t}|\rho\rangle\langle a_2\mathbf{x}_2|\hat{t}|\rho\rangle \propto \sum_{ij}\delta_{a_1a_j}\delta_{a_2a_i}\langle(\mathbf{x}_1-\mathbf{x}_2)|(\mathbf{r}_j-\mathbf{r}_i);\tilde{g}\rangle, \quad (2.4)$$

where we write $|A;\rho\rangle \equiv |\rho\rangle$ to simplify the notations a bit, $\tilde{g}$ is the cross-correlation of two localised density functions, $a_1$ and $a_2$ are the chemical species of the atomic centres, $\mathbf{x}_1$and $\mathbf{x}_2$ are the position operators of the same centres, and $\delta$ is the Kronecker symbol and vanishes if the chemical elements in the subscript are different. We can formally define $\overline{\rho^{\otimes 2}} = \langle\rho\otimes\rho\rangle$, the averaged tensor product of the density. This expression of Eq. (2.4) allows us to define translationally invariant features, and as a bonus, these features are local. In fact, Eq. (2.4) can be re-written as:

$$\langle a_1\mathbf{x}_1; a_2\mathbf{x}_2|\overline{\rho^{\otimes 2}}\rangle = \sum_{i\in A}\delta_{a_2a_i}\langle a_1(\mathbf{x}_1-\mathbf{x}_2)|A;\rho_i\rangle, \quad (2.5)$$

where $|A,\rho_i\rangle$ is an abstract object describing the local density field, or environment of the $i$th atom:

$$\langle a\mathbf{x}|A,\rho_i\rangle = \sum_{j\in A}\delta_{aa_j}f_{\mathrm{cut}}(|\mathbf{r}_j-\mathbf{r}_i|)\langle\mathbf{x}|\mathbf{r}_j-\mathbf{r}_i;g\rangle \quad (2.6)$$

In this expression, we introduce a (radial) cutoff function $f_{\mathrm{cut}}(r)$ that depends on the distance between the $i$th and $j$th atoms to emphasise further the local aspect of translationally invariant $|A,\rho_i\rangle$. Alternatively, we can hide the dependence on the cutoff function in the definition of the local environments $A_i$ and we write:

$$\langle a\mathbf{x}|A,\rho_i\rangle = \sum_{j\in A_i}\delta_{aa_j}\langle\mathbf{x}|\mathbf{r}_j-\mathbf{r}_i;g\rangle, \quad (2.7)$$

where the sum is performed over all atoms $j$ defined by the cutoff function $f_{\mathrm{cut}}(r)$. One should note that this construction of the translationally invariant $\langle a\mathbf{x}|A,\rho_i\rangle$ is equivalent to using displacement vectors as done in the standard SOAP implementation [42], for example.

The next step in the feature construction is to force rotational invariance. As mentioned earlier, we are looking into performing Haar integration over the $O(3)$ symmetry group. However, this time we perform the symmetrisation on the local density field $|A;\rho_i\rangle$. To make notations less heavy, we use this simplified version $|\rho_i\rangle \equiv |A;\rho_i\rangle$. We may perform the tensor product operation $(\nu)$ times, and we define $|\overline{\rho_i^{\otimes\nu}}\rangle = |\langle\underbrace{\rho_i\otimes\cdots\otimes\rho_i}_{\nu\text{ times}}\rangle_{O(3)}\rangle$ as the $(\nu+1)$-body-order symmetrised field. And then we symmetrise this new averaged tensor product by using the expression in

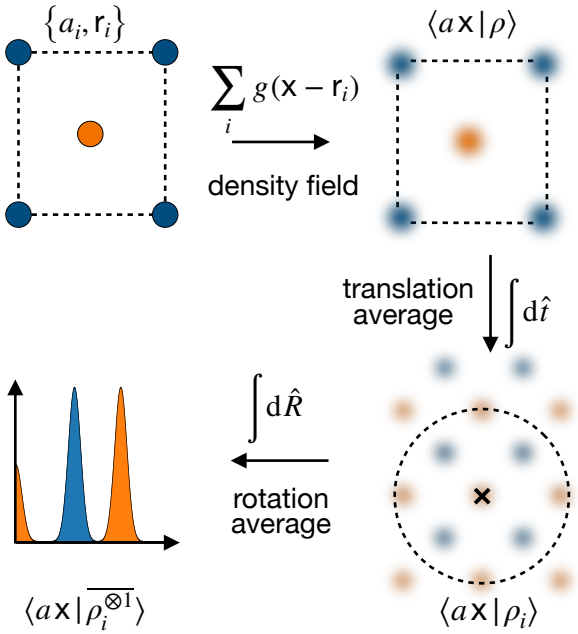


Figure 2.1: Summary of the steps in symmetrised field construction. This figure is reproduced from Ref. [81]. Original figure published under the CC-BY 4.0 license (https://creativecommons.org/licenses/by/4.0/)

Eq. (2.2), allowing us to obtain the following:

$$\langle a_1\mathbf{x}_1 \cdots a_\nu\mathbf{x}_\nu|\overline{\rho_i^{\otimes\nu}}\rangle = \sum_{k=0,1}\int_{SO(3)} \mathrm{d}\hat{R}\, \underbrace{\langle a_1\mathbf{x}_1|\hat{i}^k\hat{R}|\rho_i\rangle \cdots \langle a_\nu\mathbf{x}_\nu|\hat{i}^k\hat{R}|\rho_i\rangle}_{\nu \text{ times}}, \tag{2.8}$$

where the sum over $k$ describes the inversion symmetry. Fig. 2.1 is an illustration of the symmetrised local field for the two-body order case ($\nu = 1$). Since we are working with rotation operators, one easy (and fast) method to evaluate these integrals is to involve spherical harmonics. We can project Eq. (2.6) on an orthonormal basis of radial functions $R_n(x) \equiv \langle x|n\rangle$ and a basis of spherical harmonics $Y_m^l(\hat{\mathbf{x}}) \equiv \langle\hat{\mathbf{x}}|lm\rangle$, making sure to use the completeness relation of Eq. (2.3). The expansion coefficients of the localised density field on the basis of orthonormal radial functions and the spherical harmonics are:

$$\begin{aligned}\langle anlm|A;\rho_i\rangle &= \sum_{j\in A_i}\delta_{aa_j}\langle nlm|\mathbf{r}_j - \mathbf{r}_i; g\rangle \\ &= \int \mathrm{d}\mathbf{x}\langle n|x\rangle\langle lm|\hat{\mathbf{x}}\rangle\langle a\mathbf{x}|A;\rho_i\rangle \\ &= \int \mathrm{d}\mathbf{x} R_n(x)\, Y_m^l(\hat{\mathbf{x}}) \sum_{j\in A_i} g(\mathbf{x} - (\mathbf{r}_j - \mathbf{r}_i))\end{aligned} \tag{2.9}$$

where $\langle nlm|\mathbf{r}_j - \mathbf{r}_i; g\rangle$ is the expansion coefficient of a Gaussian centred on the interaction

between the atoms $i$ and $j$. The use of spherical harmonics $|lm\rangle$ makes the evaluation of the integrals involving the rotation operator $\hat{R}$ easy, like in Eq. (2.8), using Wigner-D matrices:

$$\langle lm|\hat{R}|l'm'\rangle = \delta_{ll'} D^{l}_{mm'}(\hat{R}). \tag{2.10}$$

This is because the Wigner-D matrices are an irreducible representation of the group of elementary rotations $SO(3)$. We obtain explicit expressions for the symmetrised fields for different values of $\nu$ or body-orders. In particular, the SOAP power spectrum representation [42], which we use as the main atomic representation in the later chapter of this manuscript, is obtained for $\nu = 2$ as we can write from Eq. (2.8):

$$\langle a_1 n_1 l_1 m_1; a_2 n_2 l_2 m_2|\overline{\rho_i^{\otimes 2}}\rangle = \delta_{l_1 l_2}\delta_{m_1 m_2}\frac{8\pi^2}{2l_1+1}\sum_s (-1)^{s-m_1}\langle a_1 n_1 l_1 s|\rho_i\rangle\langle a_1 n_2 l_2(-s)|\rho_i\rangle, \tag{2.11}$$

which can be rearranged as follows to obtain the final form of the power spectrum:

$$\begin{aligned}\langle a_1 n_1; a_2 n_2; l|\overline{\rho_i^{\otimes 2}}\rangle &= \frac{(-1)^l}{\sqrt{2l+1}}\sum_m (-1)^m \langle a_1 n_1 lm|\rho_i\rangle\langle a_2 n_2 l(-m)|\rho_i\rangle \\ &\propto \frac{1}{\sqrt{2l+1}}\sum_m c^{i,a_1}_{n_1 lm}(c^{i,a_2}_{n_2 lm})^*\end{aligned} \tag{2.12}$$

where $c^{i,a}_{nlm} = \langle anlm|A;\rho_i\rangle$. From this construction, we can prove that the power spectrum is a 3-body-order representation of atomic environments. At this level, one could affirm that the SOAP power spectrum presents several hyperparameters that need optimising depending on the studied problem: the number of radial channels $n_{\max}$, the maximum of the angular channels $l_{\max}$ and $m_{\max}$, the radial cutoff of the local density field, and any other parameters linked to the Gaussian density smearing function $|g\rangle$. In particular, one could be inspired by the locality of most target quantities from a simulation and the contribution of neighbours to the structure-property relation and introduce a radial scaling of the SOAP (or any other representation) features that is implemented as an additional weighting to the radial cutoff function $f_{\text{cut}}$ of the contributions from the neighbours:

$$u(r_{ij}) = \frac{c}{c + (r_{ij}/r_0)^m} \tag{2.13}$$

where $c$, $m$ and $r_0$ are parameters to be optimised with respect to the target property of the learning scheme. An optimised radial scaling can substantially improve the performance of a model, similar to what can be obtained by the use of multiple kernels with different length scales [82].

A different, but equivalent formulation of the density representation, especially in the limit of $|\rho\rangle \to |\delta\rangle$, reveals that this class of representations can be seen through a set of all pairwise

distances and angles between the different atomic species in the structure. We can thus formally prove that many atomic descriptors used in the field of machine-learning atomistic modelling, like the DeepMD kit [83] or the Behler-Parrinello symmetry functions [43], are equivalent to the SOAP power spectrum, as all of these representations arise from $\nu = 2$. Using similar arguments, considering the case of $\nu = 3$ would lead to constructing the SOAP bispectrum, and we can confirm it arises from 4-body correlations arguments. Needless to mention that projections over the radial and spherical harmonics basis sets are not the only way to exploit these local features. One could, as mentioned earlier, utilise a basis set of plane waves $|\mathbf{k}\rangle$.

Despite being successful in providing robust fingerprints for atomistic modelling applications and structure stability, these representations do not meet the completeness requirement, including the SOAP bispectrum [74]. In fact, the formulation of these descriptors in terms of internal coordinates, i.e. pairwise distances and dihedral angles, reveals that it is possible to construct two different atomic arrangements with the same features. This becomes problematic when the two configurations have different energies, creating "confusion" for the machine-learning algorithm and hindering its accuracy in the small dataset regime. Also, this behaviour affects generative models, where the goal is to find the atomic configurations from their geometrical fingerprints. Possible solutions might include going higher in the body-order correlations. Notable approaches include the NICE [69] construction and the ACE representation [44], which is known to be complete for high body-orders as $\nu \to +\infty$. Determining the optimal body-order $\nu$ is still an open question at the moment of writing this manuscript. The uniqueness problem also affects different architectures, deemed complete, like graph neural networks [84]. Although the picture seems a bit gloomy, this particular problem appears to have limited impact in most real-world applications, where usually labelling the atomic centres with their chemical species and the possibility of using representations centred on multiple atoms in the structures usually helps (but might not be enough) to the lift of the degeneracy of the atomic descriptors. These options should be explored further in order to assess their viability for machine-learning workflows for atomistic modelling applications.

## 2.3 GPR for scalar-valued functions

One of the goals behind the use of machine-learning techniques in atomistic modelling is to establish a mapping between the configurational phase space of molecules and materials, in the broader sense, and their physical observables, e.g. cohesive energy of a molecular compound or the electronic bandgap of a semiconductor, or to construct ML interatomic potentials by learning their energies and atomic forces. Formally, this problem belongs to

the class of supervised machine-learning methods, and in particular of regression. It can be formulated as follows: we want to find the functional relationship $f$ between an input space $\mathscr{X}$, that could be $\mathbb{R}^N$ or of infinite dimensions, and an output space that we consider to be $\mathbb{R}$. In this section, we present the basics of a well-known method to perform regressions, i.e. Gaussian process regression, applied to learning scalar (or single output) properties. We also discuss the different ways to interpret Gaussian processes.

### 2.3.1 Function space point of view

A Gaussian Process (GP) is a stochastic process that is a collection of random variables. The particularity of a GP is that the distribution of any finite collection of its random variables follows a multivariate Gaussian distribution. It is fully characterised by a mean function $m : \mathscr{X} \to \mathbb{R}$ and a covariance function $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ that is symmetric and positive definite, and we usually note:

$$f \sim \mathscr{GP}(m, k), \tag{2.14}$$

where for any $X \in \mathscr{X}$ we have $m(X) = \mathbb{E}[f(X)]$ and for any $X, X' \in \mathscr{X}$ we have $k(X, X') = \mathbb{E}[(f(X) - m(X))(f(X') - m(X'))]$. Note, that the covariance between the outputs, despite describing similarities between the random processes $f(X)$, is usually written as a function of the inputs [47]: $\mathrm{cov}\big(f(X), f(X')\big) = k(X, X')$. The covariance function $k$ may be referred to as the kernel function. For atomistic modelling applications, $k$ could be a measurement of the similarity between two data points $X$ and $X'$. Without any loss of generality, we can assume that $m(X) \equiv 0$ since we can subtract the mean function, which does not affect the covariance. This procedure is common for machine-learning applications. In what follows, we assume we have a finite dataset $\mathscr{D}\left\{\left(X_i, y_i\right)_{i \in N}\right\} \equiv \mathscr{D}\{(\mathbf{X}, \mathbf{Y})\}$ collecting $N$ input-observation pairs, where the $X_i \in \mathscr{X}$ and the $y_i \in \mathbb{R}$.

The GP, as defined in Eq. (2.14), can be used as a prior distribution within the Bayesian inference framework. This leads to performing function regression, thus defining the Gaussian process regression (GPR). Since the mean function and the covariance completely specify the GP, the collection of random variables $\{f(X_i)\}_{X_i \in \mathscr{X}}$ follows a Gaussian distribution and we write the prior distribution as follows:

$$f(\mathbf{X}) \sim \mathscr{N}(m(\mathbf{X}), \mathbf{K}), \tag{2.15}$$

where $f(\mathbf{X}) = [f(X_1), .., f(X_N)]$ is a vector containing the GP prior for every observation in the dataset $\mathscr{D}$, $m(\mathbf{X}) = [m(X_1), .., m(X_N)]$ a vector containing the mean predictive function of every observation and $\mathbf{K}$ is the kernel matrix of size $N \times N$, and whose elements are $[\mathbf{K}]_{ij} = k(X_i, X_j)$.

The notation of Eqs. (2.14) and 2.15 is a shorthand for the multivariate Gaussian probability of $f(\mathbf{X})$:

$$p(f) = \frac{1}{\sqrt{(2\pi)^N \det(K)}} \exp\left(-\frac{1}{2}\left(f(\mathbf{X}) - m(\mathbf{X})\right)^\intercal \mathbf{K}^{-1}\left(f(\mathbf{X}) - m(\mathbf{X})\right)\right). \tag{2.16}$$

The GP formally specifies the prior belief about the properties of the function $f$. These beliefs are updated in the presence of the training data $\mathscr{D}$. This is achieved via a likelihood function relating the prior assumptions and the true observations. The result of this approach is a posterior distribution that we use to infer on, potentially, unseen test cases. To illustrate this idea, we assume that our observations $y_i$ are noisy. This assumption usually helps with the robustness of the model. In addition, noise-free modelling is not interesting in itself because we would only get observations drawn from a known distribution.The incorporation of the noise in the modelling of the observations is typical for modelling using "real life" data. It can be explained physically by uncertainties in the measuring methods, for example. This noisy model is formulated as follows:

$$y_i = f(X_i) + \varepsilon_i\,, \text{ with } \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2), \tag{2.17}$$

where the noise components $\varepsilon_i$ are independent and identically distributed (i.i.d) with variance $\sigma_i^2$ and they can be assigned to each observation individually. However, in practice, it is possible to assign the same value for all the $\sigma_i^2 = \sigma^2$. Examining Eq. (2.17) reveals an interesting result. The observations $y_i$ form themselves a GP with the same mean function $m(\mathbf{X})$ of $f$ and with a different kernel matrix $\mathbf{K}_y = \mathbf{K} + \sigma_i^2 \mathbb{I}_N$, where $\mathbb{I}_N$ is the identity matrix of size $N \times N$. This can be deduced thanks to the additivity property of the Gaussian distributions of $f$ and the noise. The Gaussian likelihood function of the single noisy observations reads:

$$p(y_i | f, X_i, \sigma_i^2) = \mathcal{N}(f(X_i), \sigma_i^2). \tag{2.18}$$

Let us introduce a new input-observation pair $(X_{N+1},\, y_{N+1})$ and propose to find the model linking them according to the dataset $\mathscr{D}$. The joint distribution for the test case and the training data $\mathbf{Y}$ is such that:

$$\begin{bmatrix} \mathbf{Y} \\ \\ f(X_{N+1}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{X}) \\ \\ m(X_{N+1}) \end{bmatrix}, \begin{bmatrix} \mathbf{K} + \sigma^2 \mathbb{I}_N & \mathbf{k}(X_{N+1}, \mathscr{D}) \\ \\ \mathbf{k}(X_{N+1}, \mathscr{D})^\intercal & k(X_{N+1}, X_{N+1}) \end{bmatrix}\right), \tag{2.19}$$

where $\mathbf{k}(X_{N+1}, \mathscr{D})$ is a vector of size $N$ containing the kernel function value between the new data point $X_{N+1}$ and the data points in the training set and its elements are $[\mathbf{k}(X_{N+1}, \mathscr{D})]_i =$

$k(X_{N+1}, X_i)$. Utilising the law of conditional probability:

$$p(f(X_{N+1})|\mathbf{Y}) = \frac{p(f(X_{N+1}), \mathbf{Y})}{\int p(f, \mathbf{Y}) \mathrm{d}f}, \tag{2.20}$$

the joint distribution gives access to the conditional probability of the observation of the test point $X_{N+1}$, and we obtain the updated posterior that is also a Gaussian:

$$p(f(X_{N+1})|\mathscr{D}, X_{N+1}, \sigma_i^2) = \mathscr{N}(f(X_{N+1}), \mathrm{var}(f(X_{N+1}))), \tag{2.21}$$

where

$$f(X_{N+1}) = \mathbf{k}^\intercal(X_{N+1}, \mathscr{D}) \cdot (\mathbf{K} + \sigma^2 \mathbb{I}_N)^{-1} \mathbf{Y} \tag{2.22}$$

and

$$\mathrm{var}(f(X_{N+1})) = k(X_{N+1}, X_{N+1}) - \mathbf{k}^\intercal(X_{N+1}, \mathscr{D}) \cdot (\mathbf{K} + \sigma^2 \mathbb{I}_N)^{-1} \cdot \mathbf{k}(X_{N+1}, \mathscr{D}). \tag{2.23}$$

Based on this last definition, $k(X_{N+1}, X_{N+1})$ is simply the variance associated with new data point $X_{N+1}$ as implied from the kernel function $k$ and we can write $k(X_{N+1}, X_{N+1}) = \sigma_{N+1}^2$.

A closer inspection of Eq. 2.22 shows that the mean predictive function of a GP is nothing more than a linear expansion over the covariance function of the inputs in the training set $\mathscr{D}$ and the new input:

$$f(X_{N+1}) = \sum_{i=1}^{N} \alpha_i k(X_{N+1}, X_i), \tag{2.24}$$

where the linear expansion coefficients are

$$\boldsymbol{\alpha} = (\mathbf{K} + \sigma^2 \mathbb{I}_N)^{-1} \mathbf{Y}. \tag{2.25}$$

Eq. 2.23 reveals one crucial aspect of GPs: they incorporate information about the uncertainty of their modelling capabilities since they are a distribution over functions. The variance of a GP does not depend on the observations but only on the input training data $X_i$ and the new point $X_{N+1}$. However, it is possible to influence the uncertainty by choosing kernel functions that correlate strongly with the outputs. This shifts the burden of designing a solid ML framework to constructing suitable covariance functions for the input data, whose parameters (i.e. the covariance function) can be regarded as hyperparameters of the GPR model. The community developed several kernel functions to provide initial starting points for GPR. Notable general-purpose examples include:

- the Gaussian kernel, or the *squared exponential* (to avoid confusion with Gaussian

Process), also known as the radial basis function (RBF) kernel:

$$k(X, X') = s^2 \exp\left(\frac{||X - X'||^2}{2\sigma_l^2}\right),$$

where $s^2$ and $\sigma_l^2$ are hyperparameters of the model describing the magnitude and the variation scale desired between the features and should be tuned according to the problem. This kernel function is used by default by many ML and GPR users.

- the linear kernel or the dot product kernel:

$$k(X, X') = X \cdot X',$$

and by substituting the linear kernel in Eqs. (2.25) or (2.29), we should recover the well-known expression of linear ridge regression.

- the polynomial kernel:

$$k(X, X') = (X \cdot X')^{\zeta}, \text{where } \zeta > 1,$$

and from this family, one can expand towards piecewise polynomial kernels.

Other interesting kernels could be formulated when one realises that the sum or the product of two kernels is also a kernel. In atomistic modelling, this observation resulted in interesting constructions, like the multi-scale kernels [85] where one builds a different kernel for different ranges of separation between the atomic centres.
In addition to the hyperparameters introduced by the kernels, we may also consider the variance of noise components for the training data $\sigma_i^2$ as hyperparameters of the GPR models. All of the free parameters could be tuned by maximising the log marginal likelihood $\log\left(p(\mathbf{Y}|\mathscr{D}, \sigma_i^2)\right)$ given the training set or by N-fold cross-validation to test the generalisation of the model, for example. These techniques are representatives of the field of model selection, whose aim is to make sure that the ML models are practical tools in real-world applications, in particular, making sure that the models are able to generalise to new unseen test cases and avoid *overfitting* on the training data. Our objective is to infer, based on the training set, the parameters of the covariance function and compare models based on clear performance metrics.

This approach, demonstrated by Eq. (2.25), shows the importance of the Gaussian noise $\varepsilon_i$ as it assures that the kernel matrix is not singular and we can compute its inverse, in case of degeneracy in the training data. It also highlights the data-driven nature of GPs as a linear combination of input information that entirely determines the posterior distribution. In

the world of ML, this point of view of interpreting GPR is called the *function space* point of view. The linear combination has as many elements as the dataset $\mathscr{D}$ and scales with the size of the training data, earning this approach the name of *full GPR*. Also, this highlights the *non-parametric* aspect of GPR, as they leverage *all* the data provided to the model. We calculate the covariance of the entire dataset with itself. One can immediately see that this approach is expensive as it requires inverting the matrix $(\mathbf{K}+\sigma^2\mathbb{I}_N)$ of size $N\times N$. The inversion of this matrix scales as $\mathscr{O}(N^3)$ and also poses some limitations on the computer memory. In order to mitigate against this problem, we discuss a widely-used approximation to the full GPR framework, and we refer to it the *sparse GPR*.

The sparse GPR formalism is represented by several approximations, such as the Nyström approximation or the projected process (PP) approximation. They all share the idea of projecting the entire function space on a latent space generated by a subset of fixed functions. The PP approximation has the particularity of making use of the whole dataset as it is based on minimising the likelihood over all the data points. This is equivalent to choosing $M \ll N$ representatives to form a subset $\mathscr{M}$, also called the sparse points or the *active set*, from the training set $\mathscr{D}$and expressing the predictive mean of the GP in their basis:

$$f(X_{N+1}) = \sum_{m=1}^{M} \alpha_m k(X_{N+1}, X_i). \tag{2.26}$$

In this approximation, one no longer needs to compare all the data points with themselves but with only a sub-selection. Hence, it significantly reduces the computational cost of GPs. However, as the training set size increases, $M$ should also increase in order to maintain the accuracy of the model. This is because a larger training set contains more information about the underlying distribution of the data, and a larger number of sparse points is needed to capture this information accurately. The predictive mean and the variance are:

$$f(X_{N+1}) = \mathbf{k}^\intercal(X_{N+1}, M)\cdot\tilde{\mathbf{K}}^{-1}\mathbf{K}_{NM}^\intercal\mathbf{Y} \tag{2.27}$$

and

$$\mathrm{var}(f(X_{N+1})) = k(X_{N+1}, X_{N+1}) - Q_{X_{N+1},X_{N+1}} + \mathbf{k}^\intercal(X_{N+1}, M)\cdot\tilde{\mathbf{K}}^{-1}\mathbf{k}(X_{N+1}, M), \tag{2.28}$$

where $\mathbf{K}_{MN}$ is a matrix of size $(M\times N)$ whose elements are the kernel function values between the training data and the active set, $\mathbf{k}(X_{N+1}, M)$ is the kernel vector between the new data point and the active set, $\tilde{\mathbf{K}} = (\mathbf{K}_{NM}^\intercal\mathbf{K}_{NM} + \sigma^2\mathbf{K}_{MM})$, $\mathbf{K}_{MM}$ is a kernel matrix of the active set elements and $Q_{X_{N+1},X_{N+1}} = \mathbf{k}^\intercal(X_{N+1}, M)\mathbf{K}_{MM}^{-1}\mathbf{k}(X_{N+1}, M)$. We notice that all the matrices within the sparse GPR framework are smaller than the ones in full GPR. We only need to invert matrices

of size $M \times M$, in contrast to $N \times N$ in the previous case, reducing the computational cost significantly. Within these notations, we can write the linear expansion coefficients $\alpha_m$ under the form:

$$\boldsymbol{\alpha} = \tilde{\mathbf{K}}^{-1}\mathbf{K}_{NM}^{\intercal}\mathbf{Y} = (\mathbf{K}_{NM}^{\intercal}\mathbf{K}_{NM} + \sigma^2\mathbf{K}_{MM})^{-1}\mathbf{K}_{NM}^{\intercal}\mathbf{Y}. \tag{2.29}$$

Considering the limit of $M \rightarrow N$, we recover the expression of the linear expansion for the full GPR of Eq. (2.25) by exploiting the fact that $\mathbf{K}_{NN} = \mathbf{K}$ and $\mathbf{K}^{\intercal} = \mathbf{K}$.

The selection of the sparse points is always a debate, and a quick check of the literature reveals many methods discussing this particular problem [47]. In a materials science context, greedy and stochastic methods are the norm. We can cite the furthest point sampling (FPS) [86] and the CUR decomposition [87, 88]. These methods can also be used to reduce the size of the feature vectors of the density-based descriptors discussed in Section 2.2. Hence, the optimisation is achieved at two levels of the ML workflow: the feature generation and the ML algorithm.

### 2.3.2 Weight space point of view and links to RKHS

The function space interpretation of GPs is not the only way to derive an expression for the GPR predictions. We can also rely on a definition closer to another popular method within the ML community, the kernel ridge regression (KRR). We express the predictive mean of the GP as a linear expansion on the basis set of the kernel function $k(X_{N+1}, X_i)$ evaluated between the new point $X_{N+1}$ and a subset of the $X_i$ in the training set $\mathscr{D}$ with size $U \leq N$ :

$$f(X_{N+1}) = \sum_{i=1}^{U} c_i k(X_{N+1}, X_i). \tag{2.30}$$

The coefficients $c_i$ are obtained from the minimisation of a loss function

$$\mathscr{L} = \frac{1}{\sigma^2}||\mathbf{f}(X) - \mathbf{Y}||^2 + R, \tag{2.31}$$

in which $R = \sum_{i,j}^{U} c_i k(X_i, X_j) c_j$ is the Tikhonov regularisation term. It can be interpreted as the norm of the coefficients vector $\mathbf{c}$ in the space of the kernel functions of the environments $X_i$. It ensures that the model does not overfit. The functional form of the predictive mean of GP is potentially very flexible. In fact, if we ignore the regularisation and assume a full-rank kernel, the minimised loss will find the coefficients $c_i$ constructing a function that passes by all the training observations $y_i$. This usually leads to poor performances when tested on new input data, which defeats the purpose of building an ML model.

One should notice that the linear expansion coefficients, despite not arising from the same considerations, lead to equivalent interpretations of a GP. This observation also applies to the noise $\epsilon_i$ and their variance $\sigma_i^2$. One should note that the loss in the regularisation/weights point of view represents the cost we pay when predicting "wrong" outputs. In the Bayesian/function-space point of view, on the other hand, the loss does not affect the inference of the posterior distribution and weighs the incorrect decisions given their uncertainties. Also, the loss only appears at a later step in the workflow.

One last thing we should mention when presenting (sparse) GPR is the design of the kernel or the covariance $k(X, X')$ and how it is linked to reproducing kernel Hilbert space (RKHS). An RKHS with a reproducing kernel $k$ is a Hilbert space $\mathscr{H}_k$ of functions on a non-empty set $\mathscr{X}$ equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}_k}$ if the following conditions are satisfied:

1. for every $X \in \mathscr{X}$ we have $k(\cdot, X) \in \mathscr{H}_k$
2. the *reproducing property*: for every $X \in \mathscr{X}$ and $f \in \mathscr{H}_k$ we have $f(X) = \langle f, k(\cdot, X) \rangle_{\mathscr{H}_k}$

As a reminder, $k(X, X')$ is a function that measures the similarity between two reference data points $X$ and $X'$. We call it a kernel because it satisfies the conditions of a kernel, i.e. $k(X, X')$ is symmetric and positive definite. The Moore-Aronszajn theorem [89] states that every symmetric and positive-definite kernel defines a unique RKHS, i.e. the kernel function defines an RKHS.
From the definition of the RKHS, in particular the reproducing property, we can write the kernel of two data points $X$ and $X'$ as:

$$k(X, X') \equiv \mathrm{cov}\big(f(X), f(X')\big) = \langle k(\cdot, X'), k(\cdot, X) \rangle_{\mathscr{H}_k}. \tag{2.32}$$

We immediately notice that the covariance is defined by a dot product between two functions in the Hilbert space $\mathscr{H}_k$. In ML terminology, the $\phi_X : X' \to k(X', X)$ is known as the *feature map* of $X$ and has potentially infinite dimensions. Without any loss of generality, we can write the kernel function as a function of the feature maps:

$$k(X, X') = \langle \phi_X, \phi_{X'} \rangle_{\mathscr{H}_k} \tag{2.33}$$

The consequences of such a result are crucial: we can embed infinite features and transform them into similarity measures with values in $\mathbb{R}$ as long as we are able to find the appropriate dot product and Hilbert space. This operation is known as the *kernel trick*. The structural representations discussed in Section 2.2 are an example of these feature maps.
These observations hint at a strong link between (sparse)GPR and regression in the Hilbert

space $\mathscr{H}_k$. In fact, the minimisation solution of the loss, as defined in Eq. (2.31), can be obtained through the representer theorem and allow us to write the predictive mean of GPR as a linear expansion of functions of the form $(X \rightarrow k(\cdot, X))$:

$$f(X) = \sum_{i=1}^{N} \alpha_i k(X, X_i), \tag{2.34}$$

where the sum runs over all the training data points $X_i$. We obtain the same solution as the weight space interpretation of GPs of Eq. (2.30).

### 2.3.3 GPR in atomistic modelling

The GPR framework applies naturally to several problems in the materials science community, including establishing mappings between materials classes and some physical observables like the prediction of NMR chemical shifts [90] and the molecular orbital energy [91].
However, it is mainly used in one of the fundamental problems in atomistic modelling: constructing surrogate models for the potential energy surface (PES), i.e. the energies of configurations and its gradients, i.e. atomic forces and global stress tensors, creating machine-learning interatomic potentials (MLIPs). This is manifested by the popularity of the Gaussian Approximation Potential (GAP) [92] approach in studying multiple systems [26, 93, 94, 95, 96], especially when paired with kernels constructed using atom-centred features like the SOAP representations but not only [97]. The GAP approach is based on decomposing the total energy of an atomic system $E_{\text{tot}}(A)$ into contributions from its local atomic energies $E(A_i)$:

$$E_{\text{tot}}(A) = \sum_{i \in A} E(A_i).$$

We can re-write this equation using the bra-ket notations from Section 2.2, by considering the energy as an operator $|E\rangle$ on which we project the object holding the structural information $|A\rangle$:

$$E_{\text{tot}}(A) \equiv \langle E|A\rangle = \sum_{i \in A} E(A_i) \equiv \sum_{i \in A} \langle E|A_i\rangle. \tag{2.35}$$

One should notice that the concept of local energy is ill-defined, despite its popularity in the field of classical interatomic potentials (or force fields), and that it does not translate to a physical observable. However, there are a few efforts to use ML local energies (and other ill-defined local quantities) as fingerprints to characterise local environments [39] or to optimise some hyperparameters of structural representations like the radial cutoff [93].

Sometimes, one does not model the total energy from a quantum mechanical calculation, such as the different flavours of density functional theory or post-Hartree-Fock methods, but

usually the difference between the energy computed from a "cheaper" calculation and the quantum mechanical one. This approach can be referred to by *delta learning* or *baselining*. The baseline potential provides numerical stability to the machine-learned PES, and it can be computed from empirical potentials, other machine-learning models, or another quantum mechanical calculation. The use of a many-body descriptor, like the SOAP representation, is found to fail to describe the repulsive interaction between dimers [93]; hence, the ML potential fails to describe the short-range energies and forces and leads to unphysical structures. The difference between the baseline potential calculated on a training set and the reference (quantum mechanical) potential becomes the target of the machine-learning workflow, and that is the $E_{\text{tot}}$ in Eq. (2.35).

GAP can be built using the full GPR, but this presents a significant computational problem: the training of the model might scale unfavourably with the training set size (we mentioned earlier that the cost of inverting the kernel matrix scales as $\mathcal{O}(N^3)$). To mitigate against this issue, we may recourse to the sparse GPR framework and select a subset $\mathscr{M}$ of atomic (and local) environments constituting the sparse points, of size $M \ll N$, and we write the energy of a single environment $A_i$ treating the chemical species $a_i$ in separate channels:

$$\langle E|A_i\rangle = \sum_{I\in\mathscr{M}} \delta_{a_i a_I} \langle E; a_I|M_I\rangle k(A_i, M_I), \tag{2.36}$$

where $M_I$ is the $I$th sparse point and $k(A_i, M_I)$ is the kernel evaluated between the representation of the local environment $A_i$ and the sparse point $M_I$. Please note that, formally, there is no restriction on how the sparse points are chosen or whether they should be drawn from the training set, but usually, they are selected from the training environments. As we stated earlier, the sparse GPR scales more favourably with training set size when training the model (as $\mathcal{O}(M^3)$). Reducing the number of parameters might hinder the quality of the model and increase prediction errors, but in practice, this rarely occurs. The choice of a minimal sparse point basis results in an insignificant decrease in the model's performance.

The formulation of Eq. (2.36) allows us to easily access, at least formally, the gradients of the PES, i.e. the atomic forces $\mathbf{F}_j$ by taking the gradient with respect to atomic displacements of the $j$th atom

$$\mathbf{F}_j = -\nabla_j \langle E|A\rangle = -\sum_{I\in\mathscr{M}} \langle E; a_I|M_I\rangle \sum_{i\in A} \delta_{a_i a_I} \nabla_j k(A_i, M_I) \tag{2.37}$$

and the virial stress tensor by computing the derivative with respect to a deformation $\boldsymbol{\eta}$ of the simulation cell

$$\frac{\partial}{\partial \boldsymbol{\eta}} \langle E|A\rangle = \sum_{I\in\mathscr{M}} \langle E; a_I|M_I\rangle \sum_{i\in A} \delta_{a_i a_I} \sum_{j\in A} \mathbf{r}_{ji} \otimes \nabla_j k(A_i, M_I), \tag{2.38}$$

where $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ the direction vector between the $i$th and the $j$th atoms. Both expressions depend on the gradients of the kernel elements with respect to the neighbours of every atomic centre, making the evaluation of these gradients computationally expensive. However, the final expressions are simple because of their independence of the structural representation, the kernel function, and the ML algorithm used to perform the regression. This observation makes kernel methods and the GAP formulation an elegant approach to creating physically-motivated ML models for the PES as we would see in Section 4.2. Finally, it is worth mentioning that it is possible to construct GAP by learning exclusively from the derivatives of the total energy, i.e. atomic forces and stress tensor. This is possible due to the kernel formulation of GAP, and the fact that taking derivatives is a linear operation on the total energy and, therefore, on the kernels. Details of this construction can be found in Refs. [92, 98].

## 2.4 Extension to vector-valued functions

Scalar properties are not the only physical invariant observables we are interested in modelling. This extends to *vector-valued* properties, i.e. the properties that depend on an external variable, usually independent from the input space of configurations. Some of these properties, that do not depend on the orientation of the piece of a material, are the photo-emission spectra of molecules and the electronic density of states of semiconductors and metals. We call them vector-valued functions to make the distinctions with "true" vectors that obey strict mathematical formalism and transformation rules. The modelling of vector-valued functions has been treated extensively, especially in the context of geostatistics, where GPR models are referred to by kriging [99]. In a deeper dive into the general machine-learning literature, one notices the interchangeable use of terms like multi-output and multitask learning. However, they should not be equivalent. Multi-output learning refers to the general class of modelling vector-valued functions, while multitask learning deals with vector-valued functions where each entry has its own (distinct) input space. We can also classify these models depending on how they are trained: isotopic learning is used when the outputs share the same inputs and heterotopic learning is when some inputs have their own input space, similar to multitask learning.

We should state at this level that there are different approaches applied within the machine-learning community to tackle the problem of multivariate learning, e.g. adding the dimension of the output space to the training data [100] or decomposing the modelling of the $D$ outputs into independent single-output problems and proceeding with the GPR for scalar-valued functions as in Section 2.3 with or without independent training sets for every output channel. The latter approach disregards potential correlations between the dimensions but could

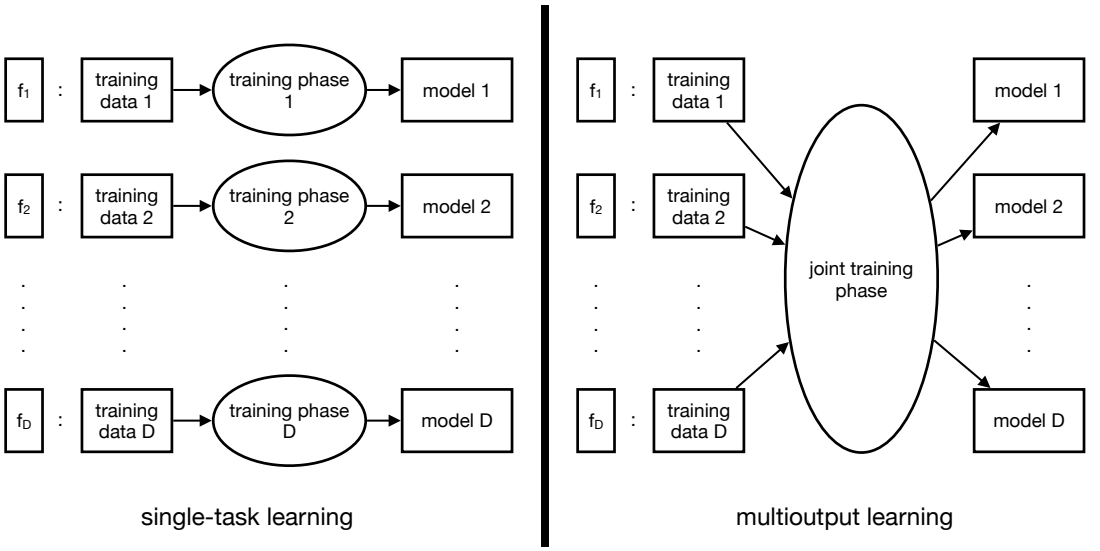


Figure 2.2: A schematic representation of the difference between the single task and multitask approaches to tackle the multivariate learning problem.

outperform models with knowledge transfer between the output processes [101]. Fig. 2.2 illustrates the general problem of multitask learning.

In this section, we present an overview of GPR applied to learning vector-valued functions and how they are linked to the case of scalar properties. Formally, the problem we are interested in is learning an unknown functional relationship $\mathbf{f}$ between an input space $\mathscr{X}$, similar to the one treated in Section 2.3, and an output space $\{1,\dots,D\}$ of multi-output functions that we can consider to be $\mathbb{R}^D$, with potentially $D \gg 1$ where $D$ is the number of processes.

### 2.4.1 Function space and RKHS

GPs for vector-valued functions follow the same approach as in the case of scalar targets of Section 2.3, with a slight difference in the definitions of the mean function of the GP and the kernel function. The former becomes vector-valued, with the same dimension as the output space, and the latter becomes kernel-valued, i.e. the covariance between two realisations $\mathbf{f}(X)$ and $\mathbf{f}(X')$ of the GP is a kernel matrix of size $D \times D$. In this subsection, we re-visit the main steps for the derivation of the full and sparse GPR predictor of vector-valued functions. To avoid unnecessary repetition, we just provide the function-space point of view and the construction of the RKHS.

### Function space

The vector-valued functions can be viewed as a collection of random processes $\{f_d\}_{d\in 1,D}$ organised in a vector. A vector-valued function $\mathbf{f}$ follows a GP if there exists a mean function $\mathbf{m}: \mathscr{X} \to \mathbb{R}^D$ and reproducing kernel that is matrix-valued: $\mathbf{K}: \mathscr{X} \times \mathscr{X} \to \mathbb{R}^{D\times D}$, and we can use similar notations to Section 2.3:

$$\mathbf{f} \sim \mathscr{GP}(\mathbf{m}, \mathbf{K}), \tag{2.39}$$

where $\mathbf{m} \in \mathbb{R}^D$ is a vector containing the mean function of every output process and $\mathbf{K}$ is a covariance matrix describing the similarity between two outputs associated with two different processes $d$ and $d'$. The elements $[\mathbf{K}(X, X')]_{d,d'}$ of the covariance are matrices, in contrast with the GPs for scalar properties where they were just scalars, and they describe the similarity between the processes $f_d(X)$ and $f_{d'}(X')$. The assumption of zero mean $\mathbf{m}$ can also be applied. Then, we follow the same procedure as the GP for scalar-valued functions. We assume a finite dataset $\mathscr{D}\left\{\left(X_i, \mathbf{y}_i\right)_{i\in N}\right\} \equiv \mathscr{D}\{(\mathbf{X}, \mathbf{Y})\}$ collecting collecting $N$ input-observation pairs where the $X_i \in \mathscr{X}$ and the $\mathbf{y}_i \in \mathbb{R}^D$.

The GP defines a Gaussian prior distribution over the random variables $\{\mathbf{f}(X_i)\}_{X_i\in\mathscr{X}}$ as we note:

$$\mathbf{f}(\mathbf{X}) \sim \mathcal{N}(\mathbf{m}, \mathbf{K}(\mathbf{X}, \mathbf{X})), \tag{2.40}$$

where $\mathbf{f}(\mathbf{X}) = [\mathbf{f}(X_1), \ldots, \mathbf{f}(X_N)]$ a matrix containing the GP prior for every observation in the dataset $\mathscr{D}$, $\mathbf{m}(\mathbf{X}) = [\mathbf{m}(X_1), \ldots, \mathbf{m}(X_N)]$ a matrix containing the mean predictive function of every observation, and $\mathbf{K}$ is block partitioned kernel matrix of size $(ND \times ND)$, and it is written as :

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \begin{bmatrix} (\mathbf{K}(X_1, X_1))_{1,1} & \ldots & \mathbf{K}(X_1, X_D))_{1,D} \\ (\mathbf{K}(X_2, X_1))_{2,1} & \ldots & \mathbf{K}(X_2, X_D))_{2,D} \\ \vdots & \ldots & \vdots \\ (\mathbf{K}(X_D, X_1))_{D,1} & \ldots & \mathbf{K}(X_D, X_D))_{D,D} \end{bmatrix}, \tag{2.41}$$

where every block $(\mathbf{K}(X_i, X_j))_{d,d'}$ is a matrix of size $N \times N$. In this expression, we assume that every dimension $d$ has its own training set.

We assume that our observations $\mathbf{y}_i$ are corrupted, and we model this behaviour by an independent and identically distributed Gaussian noise $\{\varepsilon_d\}_{d=1,D}$ and we write the model as follows:

$$y_{i,d} = f_d(X_i) + \varepsilon_{i,d}, \text{ with } \varepsilon_{i,d} \sim \mathcal{N}(0, \sigma^2_{i,d}). \tag{2.42}$$

In order to perform a regression, we write the Gaussian likelihood function for a single observation $\mathbf{y}_i$ as follows:

$$p(\mathbf{y}_i|\mathbf{f}, X_i, \Sigma) = \mathcal{N}(\mathbf{f}(X_i), \Sigma), \tag{2.43}$$

where $\Sigma$ is a $D \times D$ diagonal matrix, whose elements are the $\sigma_d^2$. We can obtain the predictive mean of the GP and its variance by introducing a new data point made of a pair of input-observation $(X_{N+1}, \mathbf{y}_{N+1})$. Following the same reasoning in Section 2.3.1, we find the predictive mean of the multioutput regression:

$$\mathbf{f}(X_{N+1}) = \sum_{i=1}^{N} \boldsymbol{\alpha}_i \mathbf{K}(X_{N+1}, X_i), \tag{2.44}$$

where the linear expansion coefficients are matrices and have the same formal form as Eq. (2.25):

$$\boldsymbol{\alpha} = (\mathbf{K}(\mathbf{X},\mathbf{X}) + \Sigma)^{-1}\mathbf{Y}. \tag{2.45}$$

The sparsification of this *full GPR* approach follows the same logic as the case of the scalar-valued functions. In particular, we can apply the same methodology of the projected process approximation and project the processes on a subset of *active set* points and end up with another linear expansion expression for the predictive mean.

#### RKHS for vector-valued functions

Similar to the case of scalar-valued functions, the RKHS of vector-valued functions is characterised by its matrix-valued kernel, and we note the reproducibility condition for any function $\mathbf{f}$ in the RKHS $\mathscr{H}_{\mathbf{K}}$ as:

$$\langle \mathbf{f}, \mathbf{K}(\cdot, X) \cdot \mathbf{c} \rangle_{\mathscr{H}_{\mathbf{K}}} = \mathbf{f}(X)^{\intercal} \cdot \mathbf{c}, \tag{2.46}$$

where the notation $\langle \cdot, \cdot \rangle_{\mathscr{H}_{\mathbf{K}}}$ refers to the inner dot product of the RKHS. The kernel trick can also be applied in this case, and one can extend the definition of features maps $\phi$ to vector-valued functions. In particular, and depending on the nature of the processes $f_d$, one can assign different maps $\phi_d$ to different processes.

If we follow the regularisation route of the scalar-valued functions as shown in Eq. (2.31), the regression problem can be formulated as a minimisation of a loss function:

$$\mathscr{L} = \frac{1}{N} ||\mathbf{f}(X_i) - \mathbf{y}_i||^2 + \lambda ||\mathbf{f}||^2_{\mathscr{H}_{\mathbf{K}}}. \tag{2.47}$$

The solution to this problem can be obtained thanks to the representer theorem [102]:

$$\mathbf{f}(X) = \sum_{i=1}^{N} \mathbf{K}(X, X_i) \cdot \mathbf{c}_i, \tag{2.48}$$

where $\mathbf{c} = (\mathbf{K}(\mathbf{X},\mathbf{X}) + \lambda N \mathbb{I}_{ND})^{-1}\mathbf{Y}$ is the vector holding the linear expansion coefficients of

size $ND$. It is obtained by concatenating the expansion coefficients vectors of size $D$, $\mathbf{c}_i$ as $\mathbf{c} = \begin{pmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_N \end{pmatrix}$. If we dispose of a new test point $X_{N+1}$, the corresponding expression for the regressor is given by:

$$\mathbf{f}(X_{N+1}) = \mathbf{K}(X_{N+1}, \mathscr{D})^\intercal \cdot \mathbf{c}, \tag{2.49}$$

where $\mathbf{K}(X_{N+1}, \mathscr{D})$ is a matrix of dimensions $ND \times D$, and whose entries are $(\mathbf{K}(X_{N+1}, X_i))_{d,d'}$.

### 2.4.2 Kernel design for vector-valued functions

Designing kernels for vector-valued functions is delicate and is heavily influenced by the nature of the processes and their intrinsic correlations. It is possible only to consider the most general case and build a different GPR model for every dimension by ignoring possible correlations and potentially requiring an increasing number of training sets. However, it is advantageous to model connections between the different processes. Let us consider the case of learning a continuum spectrum like the electronic density of states or the photoemission spectrum. These spectra are usually constructed by convoluting the discrete spectrum with a smooth function and stored as a collection of discrete values. The resulting "vector", or more appropriately, array, contains several adjacent points that are highly correlated. It would be ideal if one could establish a strategy to leverage and model these "hidden" correlations.

We showcase an elegant approach to designing kernels for vector-valued functions based on decoupling the information from the input space $\mathscr{X}$ and the output space $\{1, \ldots, D\}$, known as *separable kernels*. The starting point is writing the matrix-valued kernel function of the multi-output GP as a product between a kernel function of the input space and another one for the output space. The general form of an element of this matrix within this class of kernels is written as follows:

$$(\mathbf{K}(X, X'))_{d,d'} = k(X, X') k_T(d, d'), \tag{2.50}$$

where $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ and $k_T : \{1, \ldots, D\} \times \{1, \ldots, D\} \to \mathbb{R}$ are the two scalar kernel functions. Equivalently, we can consider the matrix representation of these kernels, exploiting the fact the output space is fixed for a given problem:

$$\mathbf{K}(X, X') = k(X, X')\mathbf{B}, \tag{2.51}$$

where $\mathbf{B}$, also known as coregionalisation matrix, is a symmetric semidefinite matrix, with $D \times D$ dimension, and it is the matrix representation of $k_T$. The link between the two expressions can be visualised if we consider the training set $\mathscr{D}$and inject Eq. (2.50) into Eq. (2.41) to obtain:

$$\begin{aligned}\mathbf{K}(\mathbf{X},\mathbf{X}) &= \mathbf{k}(\mathbf{X},\mathbf{X}) \otimes \mathbf{B} \\ &= \begin{pmatrix} k(X_1,X_1)\mathbf{B} & \cdots & k(X_1,X_N)\mathbf{B} \\ \vdots & & \vdots \\ k(X_N,X_1)\mathbf{B} & \cdots & k(X_N,X_N)\mathbf{B} \end{pmatrix}\end{aligned} \tag{2.52}$$

where $\mathbf{k}(\mathbf{X},\mathbf{X})$ is the kernel matrix of input space associated with the dataset $\mathscr{D}$.

We can recover the general case of independent output values by setting the covariance matrix $\mathbf{B}$ to the identity $\mathbb{I}_D$ or equivalently have the kernel function of the form: $k_T(d,d') = \delta_{d,d'}$, where $\delta_{d,d'}$ is the Kronecker symbol. The kernel matrix associated with $\mathscr{D}$ becomes equivalent to:

$$\mathbf{K}(\mathbf{X},\mathbf{X}) \equiv \begin{pmatrix} k(X_1,X_1)\mathbb{I}_D & & (0) \\ & \ddots & \\ (0) & & k(X_N,X_N)\mathbb{I}_D \end{pmatrix}, \tag{2.53}$$

which is the same as the kernel matrix from the scalar-valued functions. Intuitively, we can deduce that the off-diagonal terms of the matrix $\mathbf{B}$ encode the dependencies among the output processes, and ensure the knowledge transfer between the outputs. However, even within this simple approach, the correlation between the outputs still exists, implicitly because we set the same hyperparameters for all the outputs, including those of the kernel matrix, from all the data provided in the training set $\mathscr{D}$.

The key question that needs to be addressed is how one should design the matrix $\mathbf{B}$. As a reminder, for atomistic modelling problems, we can use kernels based on the atomic representations discussed in Section 2.2. The most straightforward approach can be found in standard geostatistics literature [99], in particular, by adopting the weight-space point of view. The method consists of evaluating, in a separate inference step, the matrix $\mathbf{B}$ from the covariance matrix of the outputs of the training data $\mathbf{y}_i$. Alternatively, we can achieve the same interpretation of the covariance matrix within an intrinsic regionalisation model (ICM) [103] point of view, one of the wide-spread formulations for multivariate learning with applications in geology [104, 105] or environmental studies [106, 107]. Its main idea consists in drawing $R$ independent realisations $u^r(X)$ of the same process $\mathbf{u}(X)$ following a GP: $\mathbf{u}(X) \sim \mathscr{GP}(\mathbf{m},k)$, and expressing every process $f_d$ as a linear combination of the drawn samples as:

$$f_d(X) = \sum_{r=1}^{R} a_d^r u^r(X), \tag{2.54}$$

where the $a_d^r$ are the linear expansion coefficients. Here we emphasise that all the independent

processes $u^r(X)$ are drawn using the same kernel function $k$. The covariance of the function $\mathbf{f}$ is:

$$\text{cov}(\mathbf{f}(X), \mathbf{f}(X')) = \mathbf{A}\mathbf{A}^\top k(X, X') \equiv \mathbf{B}k(X, X'), \tag{2.55}$$

where $\mathbf{A}$ is a matrix holding the linear expansion coefficients with elements $[\mathbf{A}]_{r,d} = a_d^r$. And we obtain the same formulation as the separable kernels in Eq. (2.51). The covariance matrix $\mathbf{B}$ has a rank equal to $R$ because the processes $u^r(X)$ are independent. This formulation highlights that we can express the outputs $f_d$ as a linear expansion of independent and orthogonal latent functions, the $u^r$. In particular, it is possible to choose a representative set of $R < D$ latent functions in order to reduce the computational costs of the inference step. Doing so requires extra inductive bias to the model, which could be regarded as a method to make the model more data efficient by imposing a certain "structure" on the data. As mentioned earlier, the most straightforward option is to use the covariance matrix of the outputs $\mathbf{y}_i$ of the training set $\mathscr{D}$. The latent functions, in this case, can be chosen to be the eigenvectors of the $D \times D$ covariance matrix, and then a low-rank approximation can then be employed. For example, we visit the case of projecting $\mathbf{B}$ on a subset of its eigenvectors, which is equivalent to defining a new coordinate system in which $\mathbf{B}$ becomes diagonal. We assume the mean value of every output is 0; otherwise, it may be necessary to remove them. We write the eigenvalue decomposition of $\mathbf{B}$ as:

$$\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top, \tag{2.56}$$

where $\mathbf{\Lambda}$ is a diagonal matrix holding the (ordered from highest to lowest) eigenvalues $\Lambda_k$ of the matrix $\mathbf{B}$, and $\mathbf{U}$ is the matrix holding the eigenvectors, also known as principal components. They coincide with the vector form $\mathbf{u}_d$ of the latent functions $u^r$ of Eq. (2.54). At this level, we can apply a low-rank approximation, which is a projection on the $R$ non-degenerate principal components. We denote by $\mathbf{C}$ the matrix holding the coefficients $\mathbf{c}_i$ of Eq. (2.48): $\mathbf{C} \equiv (\mathbf{c}_1, \ldots, \mathbf{c}_N)$. We define new expansion coefficients $\tilde{\mathbf{c}}^d = (\mathbf{c}_1 \cdot \mathbf{u}_d, \ldots, \mathbf{c}_N \cdot \mathbf{u}_d)$ and write $\mathbf{C} = \sum_{d=1}^{D} \tilde{\mathbf{c}}^d \otimes \mathbf{u}_d$. Similarly, we re-write the target outputs as $\tilde{\mathbf{Y}} = \sum_{d=1}^{D} \tilde{\mathbf{y}}^d \otimes \mathbf{u}_d$ with $\tilde{\mathbf{y}}^d = (\mathbf{y}_1 \cdot \mathbf{y}_d, \ldots, \mathbf{y}_N \cdot \mathbf{u}_d)$. These transformations can be simply seen as rotations in the output space. From these (re)definitions, we can write the linear system to solve for the linear coefficients as:

$$\begin{aligned} \mathbf{C} = (\mathbf{K}(\mathbf{X},\mathbf{X}) + \lambda N \mathbb{I}_N)^{-1}\mathbf{Y} &= \sum_{d=1}^{D} (\mathbf{B} \otimes k(\mathbf{X},\mathbf{X}) + \lambda N \mathbb{I}_N)^{-1} \tilde{\mathbf{y}}^d \otimes \mathbf{u}_d \\ &= \sum_{d=1}^{D} (\Lambda_d k(\mathbf{X},\mathbf{X}) + \lambda N \mathbb{I}_N)^{-1} \tilde{\mathbf{y}}^d \otimes \mathbf{u}_d. \end{aligned} \tag{2.57}$$

The (eigen)vectors $\mathbf{u}_d$ are orthogonal, and this allows us to solve this problem as independent

$R$ problems:

$$\tilde{\mathbf{c}}^d = (k(\mathbf{X},\mathbf{X}) + \frac{\lambda}{\Lambda_d}\mathbb{I}_N)^{-1}\frac{\tilde{\mathbf{y}}^d}{\Lambda_d}, d = 1,\ldots,R. \tag{2.58}$$

Fig. 2.3 summarises the training of the model with this approach using the matrix representation of the problem. This derivation shows that, within the separable kernels approximation, one could reduce the dimensionality of the (potentially correlated) outputs by projecting them on a basis set of orthogonal functions, e.g. the principal components of the outputs of the training set. The projection coefficients become the targets of our learning scheme. This means that the problem becomes that of learning independent-by-construction tasks; hence, we can utilise all the tricks from the scalar-valued GPR framework. The independent basis functions, on which we project the outputs, hold information about all the outputs, therefore allowing for an efficient transfer of "knowledge" between them and significantly reducing the computation cost. In Section 3.1, we encounter the same result of decomposing the outputs on a basis set of principal components, but from considerations coming from the correlated nature of the outputs, and not from kernel design considerations as we encountered in this section.

To wrap up this discussion, it is possible to extend the ICM to cases where the different latent functions are drawn from distinct kernel functions. This class of models is known as the linear regionalisation models (LCM) [103, 108], and we write the covariance of $\mathbf{f}$ as a linear combination of coregionalisation matrices $\mathbf{B}_q$ and kernel functions $k_q$:

$$\text{cov}(\mathbf{f}(X),\mathbf{f}(X')) = \sum_{q=1}^{Q}\mathbf{B}_q k_q(X,X'). \tag{2.59}$$

The LMC amounts to represent the outputs variables as linear combinations of sets of uncorrelated latent variables. This approach has the advantage of making use of several input spaces and potentially offering a better knowledge transfer between the outputs. In an atomistic modelling context, one could imagine applying LMC within a multi-scale kernels approach (as discussed in Section 2.2), where each kernel pair $(k_q,\mathbf{B}_q)$ is built to target a portion of the outputs.

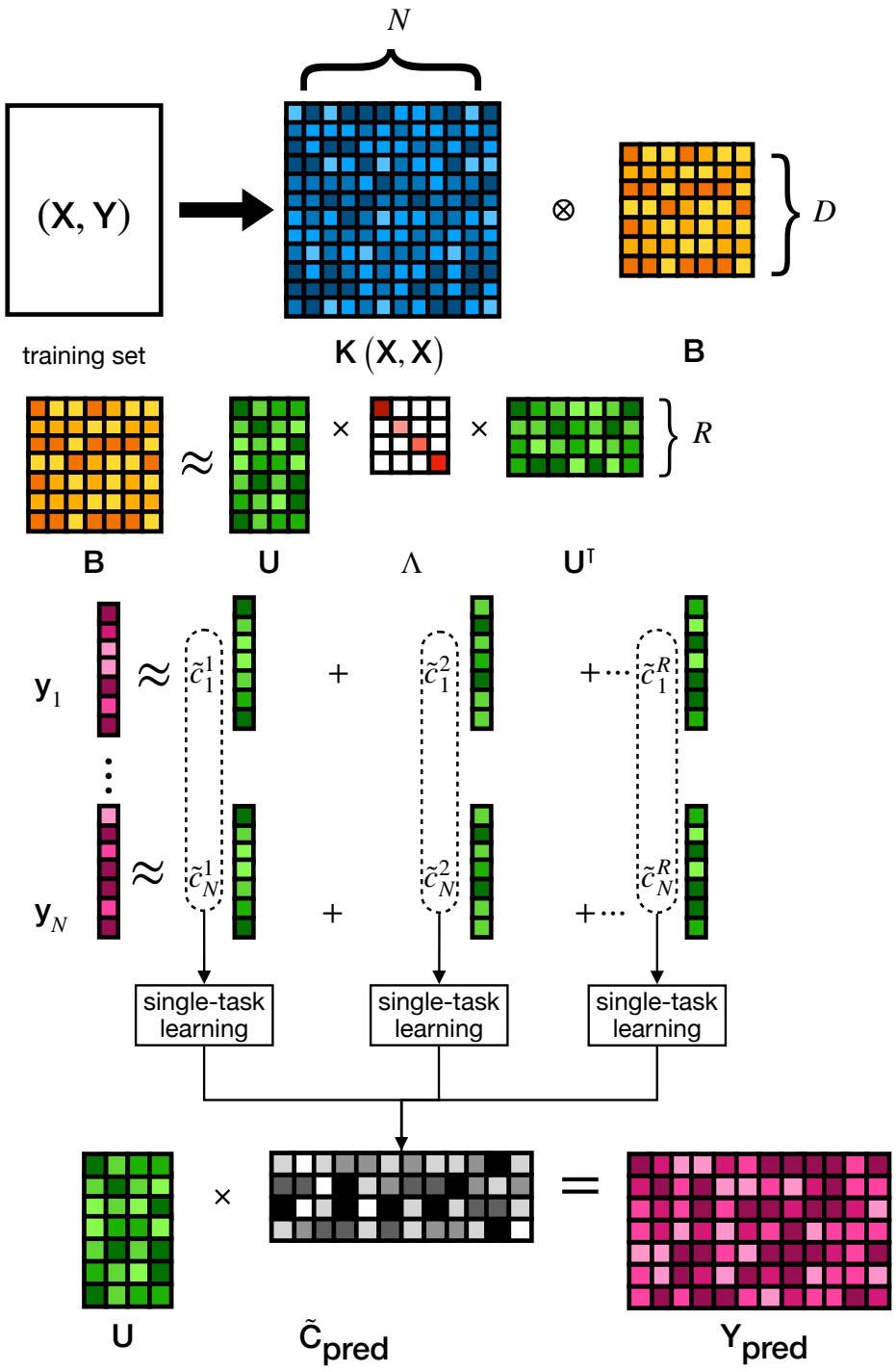


Figure 2.3: A matrix representation of the separable kernels approach with a low-rank approximation as described in Eq. (2.54)

# 3 Learning the electronic density of states[1]

The electronic density of states (DOS) is a fundamental quantity in the electronic structure theory. It describes the distribution of the energy levels that can be occupied by electrons in a quasiparticle picture. It can be used to calculate the electronic contribution to the heat capacity in metals, the density of free charge carriers in semiconductors, and is an indirect proxy for properties such as the energy band gap, the band energy, and also the optical absorption spectrum. All of these "derived" properties justify the efforts for building surrogate models to a quantum-mechanically computed DOS.

In this chapter, we present a machine-learning (ML)-based atom-centred model for the electronic DOS using the geometrical descriptors and the multivariate fitting algorithms presented in Chapter 2. We also assess the relative performance of models that directly predict electronic properties, such as the band energy and the optical absorption spectrum, with those of models that use the DOS as an intermediate quantity. We use as a benchmark a challenging data set of silicon structures that includes different solid phases, liquid and amorphous configurations, and gas phase clusters, spanning a wide range of behaviours from metallic to semiconducting. Finally, we demonstrate the transferability of the model to predict the DOS of large amorphous configurations, exploiting the local description to link atomic environments to their contributions to the overall density of energy levels and describe the electronic transitions in disordered phases of silicon.

[1] This chapter is an adaptation of my contributions to Refs. [109, 38]. It also includes some unpublished results and analysis. Sections 3.1 and 3.2 are adapted from Ref. [109]. I contributed to running the reference DFT calculations, building the ML models, and performing the analysis. Section 3.3 is an adaptation of the contribution to Ref. [38], where I trained the DOS models, including performing the hybrid-DFT calculations and performed the analysis leading to the construction of the features/properties maps.

## 3.1 Atom-centred model for the DOS

We define the DOS as a sum of Dirac distributions centred around the eigenvalues of the single-particle Hamiltonian, $\epsilon_n(\mathbf{k})$ describing the energy levels that the electrons can occupy at each point $\mathbf{k}$ of the electronic Brillouin zone (BZ):

$$\mathrm{DOS}(\epsilon) = \frac{2}{N_k} \sum_{n}^{\text{bands}} \sum_{\mathbf{k}} \delta(\epsilon - \epsilon_n(\mathbf{k})),$$

where "2" accounts for the spin degeneracy of the electrons in case of the absence of an external magnetic field and $N_k$ is the number of $\mathbf{k}$ points used to sample the BZ. This definition approximates the integral over the BZ and extends readily to a-periodic systems by removing the summation over the $\mathbf{k}$ points. In order to obtain a continuous distribution, the Dirac distributions are broadened with a Gaussian broadening $g_\mathrm{b}$: $\delta(\epsilon - \epsilon_n(\mathbf{k})) \rightarrow (g_\mathrm{b}\sqrt{2\pi})^{-1} \exp[-(\epsilon - \epsilon_n(\mathbf{k}))^2/(2g_\mathrm{b}^2)]$. The choice of $g_\mathrm{b}$ should reflect the level of the fine structure of the DOS curve that one wants to keep: a small value of $g_\mathrm{b}$ would lead to more sharp peaks and describe more accurately the conductivity/metallicity or the band gap of a structure, while a large value of $g_\mathrm{b}$ would lead to a "featureless" DOS curve and underestimate the band gap by accumulating more states near the Fermi energy. For ML applications, it is possible to optimise the value of $g_\mathrm{b}$ using Bayesian optimisation by maximising a log-likelihood, for example. The Gaussian broadening is a popular option to construct the continuous DOS spectrum, but it is not the only option. Other approaches could use Fermi-Dirac broadening or a combination of polynomials and Lorentzian functions as presented in Refs. [110, 111].

The smooth DOS gives us access to several "derived" quantities. In particular, we will be interested in these quantities:

- The Fermi energy ($\varepsilon_F$)

$$\varepsilon_F : \int \mathrm{d}\epsilon\, \mathrm{DOS}(\epsilon) f_\mathrm{FD}(\epsilon - \varepsilon_F, T = 0) = N_\mathrm{val}, \tag{3.1}$$

  where $f_\mathrm{FD}(\epsilon - \varepsilon_F, T = 0)$ is the occupation of the energy level according to Fermi-Dirac statistics at $T = 0$, which only describe the occupied states, and $N_\mathrm{val}$ is the number of valence electrons.

- The density of states at the Fermi energy ($\mathrm{DOS}(\varepsilon_F)$)

- The band energy

$$\epsilon_\mathrm{band} = \int \mathrm{d}\epsilon\, \mathrm{DOS}(\epsilon)\epsilon\ f_\mathrm{FD}(\epsilon - \varepsilon_F, T = 0).$$

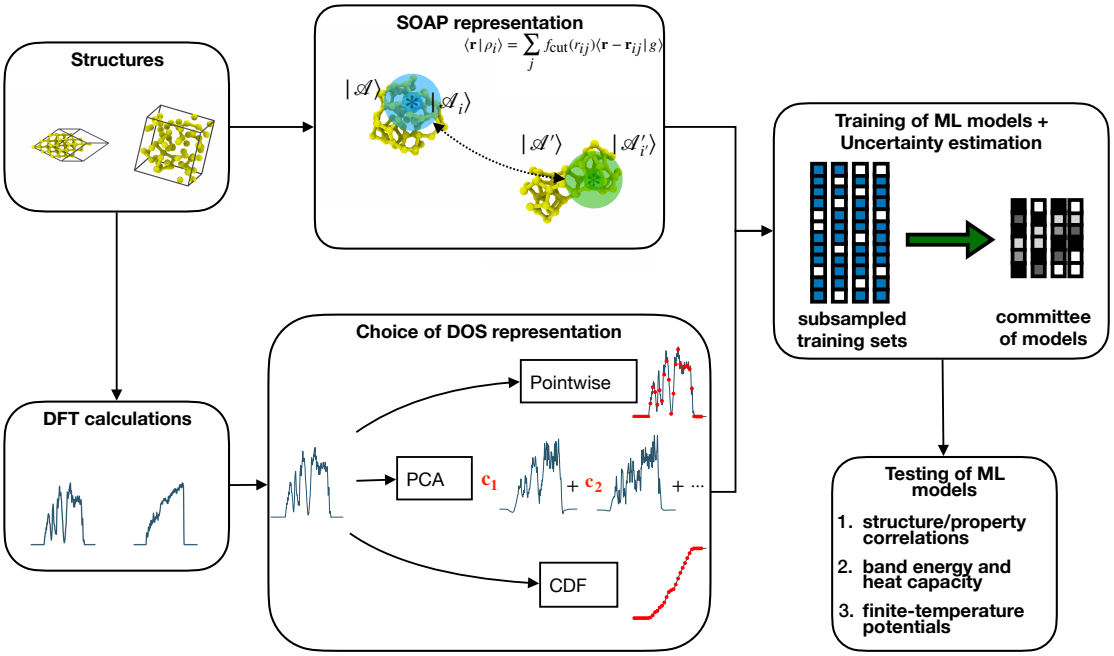


Figure 3.1: Schematic representation of the workflow of the ML DOS model. $A$ indicates full atomic configurations, $A_i$ represents the atomic environments in these structures, the blue curves are the DOS from DFT calculations and the red dots are the targets of the ML models.

- The distribution of excitations

$$A(\Delta) = \int\int \mathrm{d}\epsilon\, \mathrm{d}\epsilon'\, \mathrm{DOS}(\epsilon)\, f_{\mathrm{FD}}(\epsilon - \varepsilon_F, T=0)$$
$$\mathrm{DOS}(\epsilon')\, (1 - f_{\mathrm{FD}}(\epsilon - \varepsilon_F, T=0))\delta(\epsilon - \epsilon' - \Delta).$$

  $A(\Delta)$ mimics the adsorption spectrum, where $\Delta$ corresponds to the absorbed photon's energy and we ignore the amplitude of the transition. The shape of $A(\Delta)$ for small excitation energies reveals the presence and the magnitude of a band gap.

These quantities demonstrate the usefulness of modelling the DOS, because a single inference step leads to estimating several important properties of an atomic configuration. We encounter in Chapter 4 other use cases for DOS-based surrogate models within a physics-driven modelling approach of materials.

In the following paragraphs, we introduce the different components of our strategy to predict the DOS based on a regression model trained on a small number of reference configurations. Fig. 3.1 is a summary of our strategy and the workflow used to build and evaluate the ML DOS model.

We aim to build a model of the DOS for a structure $A$ by decomposing the total DOS into a sum of local contributions from each of its atomic environments $A_i$, i.e.,

$$\mathrm{DOS}(A,\epsilon) = \sum_{i\in A} \mathrm{LDOS}(A_i,\epsilon). \tag{3.2}$$

Implementing such a model requires the definition of a framework to parameterise the shape of the LDOS, and a framework to represent the structure and the composition of the environment surrounding each atom. Given these, one can determine the parameters $\mathbf{x}$ of the LDOS model by minimising a loss function of the form:

$$\mathscr{L}^2(A,\mathbf{x}) = \int \mathrm{d}\epsilon \left| \mathrm{DOS}(A,\epsilon) - \sum_{i\in A} \mathrm{LDOS}_{\mathbf{x}}(A_i,\epsilon) \right|^2. \tag{3.3}$$

The model can then be used to make predictions for new, possibly more complex, structures. We model $\mathrm{LDOS}_{\mathbf{x}}(A_i,\epsilon)$ using the Projected Process (PP) approximation of the Gaussian Process Regression (GPR) [47], presented in Section 2.3. We use the most diverse $M$ atomic environments, selected according to a Farthest Point Sampling (FPS) scheme [88] among the training structures, as the *active set* that defines the basis on which the target is expanded in local contributions:

$$\mathrm{LDOS}_{\mathbf{x}}(A_i,\epsilon) = \sum_{j\in M} x_j(\epsilon) k(A_i,M_j), \tag{3.4}$$

where $k(A_i,M_j)$ is a positive-definite kernel basis function that expresses the similarity between the environment $A_i$ and an environment from the active set $M_j$. We use a polynomial kernel constructed from the power spectrum of the smooth overlap of atomic positions (SOAP) representation. Given the additive nature of the DOS model, as introduced in Eq. (3.2), we define the kernel between entire structures as the sum of the kernels between the atomic environments that constitute the structures, $k(A,A') = \sum_{i\in A, i'\in A'} k(A_i,A'_{i'})$. The linear expansion coefficients $x_j(\epsilon)$ depend on the energy and should be discretised in a way that reflects the representation of the DOS, which is discussed in the next paragraph. We optimise the coefficients $\mathbf{x}_M(\epsilon)$ by minimising the following empirical loss function, in the same spirit of ridge regression models, based on knowledge of the targeted DOS for the training structures:

$$\mathscr{L}^2_\lambda(\mathbf{x}_M) = \sum_{A\in\mathrm{train}} \mathscr{L}^2(A,\mathbf{x}_M) + \lambda^2 \mathbf{x}_M^\intercal K_{MM} \mathbf{x}_M. \tag{3.5}$$

Here, $\lambda$ is the regularisation parameter and $K_{MM}$ is the kernel matrix, whose entries are the kernel functions between the active-set environments. The optimal solution to this problem is obtained as a function of the kernel matrix of the active set $K_{MM}$ and the kernel matrix of

the training structures and the active set $K_{NM}$:

$$\mathbf{x}_M(\epsilon) = (\lambda^2 K_{MM} + K_{MN}K_{MN}^{\intercal})^{-1} K_{MN}\mathbf{y}_N(\epsilon), \tag{3.6}$$

where $\mathbf{y}_N$ is a vector containing the values of $\mathrm{DOS}(A,\epsilon)$ for the $N$ training structures. Once we find the optimal solution for our problem, usually using a k-fold cross-validation scheme, the DOS of a new structure $A_*$ can be obtained as a simple dot product:

$$\mathrm{DOS}(A_*,\epsilon) = \mathbf{k}_{A_*M}^{\intercal} \cdot \mathbf{x}_M(\epsilon), \tag{3.7}$$

where $\mathbf{k}_{A_*M}$ is the vector that contains the kernels between the structure $A_*$ and the $M$ active set environments.

### Uncertainty estimation

GPR models have a built-in variance estimator that makes it possible to assess the statistical uncertainty – and hence the reliability – of the prediction for a specific structure, as discussed in Section 2.3. For computational efficiency and to simplify the propagation of uncertainty from the atom-centred decomposition to the full density of states of a structure, we build instead a committee of $N_{RS}$ GPR models of size $n < N$, as discussed in Ref. [112]. If the models are built based on the PP approximation, keeping a fixed active set for all the models, each model corresponds to a different weight vector $\mathbf{x}_M^{(s)}(\epsilon)$, and predictions can be obtained inexpensively as the kernel vector in Eq. (3.7) must be computed only once for each new structure. The average of the predictions $\mathrm{DOS}^{(s)}(A_*,\epsilon)$ made by the models in the committee is taken as the best estimate:

$$\mathrm{DOS}_{RS}(A_*,\epsilon) = \frac{1}{N_{RS}} \sum_s \mathrm{DOS}^{(s)}(A_*,\epsilon),$$

while their variance is taken as a measure of the uncertainty

$$\sigma_{RS}^2(A_*,\epsilon) = \frac{\alpha_{RS}(\epsilon)}{N_{RS}-1} \sum_s \left(\mathrm{DOS}^{(s)}(A_*,\epsilon) - \mathrm{DOS}_{RS}(A_*,\epsilon)\right)^2. \tag{3.8}$$

The factor $\alpha_{RS}(\epsilon)$ serves to compensate for the correlations that are present between the training points and between the different resampled models; $\alpha_{RS}(\epsilon)$ can be determined by calibrating the uncertainty estimate with a likelihood maximisation criterion, using a validation set or an internal reference [112]. We obtain an unbiased estimator [113] for the rescaling

factor $\alpha_{RS}(\epsilon)$:

$$\alpha_{RS}(\epsilon) = -\frac{1}{N_{RS}} + \frac{N_{RS}-3}{N_{RS}-1} \sum_{A \in \text{val}} \frac{\left|\text{DOS}^{(s)}(A,\epsilon) - \text{DOS}_{RS}(A,\epsilon)\right|^2}{\sigma^2(A,\epsilon)}, \tag{3.9}$$

where $\sigma^2$(A, $\epsilon$) is the variance of the prediction of the DOS(A, $\epsilon$) and can be obtained from variance of the predictions of the committee of models or from the built-in variance of the GPR as in Eq. (2.23). Note that it is possible to realise this calibration process by rescaling the predictions around the mean, i.e.

$$\begin{aligned} \text{DOS}^{(s)}(A_*,\epsilon) \leftarrow & \text{DOS}_{RS}(A_*,\epsilon) + \\ & \sqrt{\alpha_{RS}(\epsilon)}\left[\text{DOS}^{(s)}(A_*,\epsilon) - \text{DOS}_{RS}(A_*,\epsilon)\right]. \end{aligned} \tag{3.10}$$

We can use this calibrated ensemble of predictions to perform post-processing tasks, such as the assembly of atom-centred predictions, in a way that automatically incorporates correlations between the predictions of different models.

### DOS representation

As discussed earlier in Section 2.4, we have already encountered several strategies to build kernel functions for multivariate GPR. We concluded that, for some of them, designing a kernel function is equivalent to projecting the target property on a different basis.
In this paragraph, we explore some approaches to represent the space of the (L)DOS. In order to build a practical model of the DOS, one needs to represent the (L)DOS as a set of discrete values, $y_j(A)$, that can then be used to construct a multivariate regression model. A pointwise trivial approach is to discretise the energy axis over a finite range $[\epsilon_0, \epsilon_0 + (N_{\text{DOS}}-1)\delta\epsilon]$, and take the smooth DOS computed at each energy point as a regression target,

$$y_j^{\text{PW}}(A) = \text{DOS}(A, \epsilon_0 + j\delta\epsilon). \tag{3.11}$$

In the Gaussian process regression model we use, each regression task is independent of the others, which means that the number of these descriptors can become arbitrarily high depending on the level of complexity (defined by the Gaussian broadening) and the density of the considered energy points (linked to the discretisation width $\delta\epsilon$), which results in an increase in the number of prediction models.
This pointwise representation is not necessarily the most efficient: it potentially requires training and evaluating hundreds of ML models. It ignores the fact that variations in the DOS between different structures and different energy levels are correlated, both because of physical reasons and because of the Gaussian broadening. We should note here that in a

regression scheme, it is possible to ensure the smoothness of the predicted DOS by utilising a smooth regularising function of the energy channels $\lambda(\epsilon)$ in Eq. (3.5), in contrast with the approach of optimising $\lambda(\epsilon)$ for every energy channel $\epsilon$ independently.

It is possible to reduce the degrees of freedom of the problem by projecting the DOS on an orthogonal basis set of latent functions and learning the expansion coefficients, as presented in Section 2.4.2 when discussing the implications of the separable kernel approach. In order to capture the correlations between the variations of the DOS at different points, we construct a data-adapted basis by evaluating the principal components (PC) of the DOS within the training set. In practice and given that we aim to build a predictor for the LDOS, but we only have information on the total DOS of a structure, we normalise the DOS of each structure by the number of electronic states and construct the matrix

$$\tilde{Y}_{Aj} = \frac{y_j(A)}{N_A} - \frac{1}{N_{\text{train}}} \sum_{A'} \frac{y_j(A')}{N_{A'}}. \tag{3.12}$$

We then compute the eigendecomposition of the covariance matrix

$$\mathbf{Y}^\intercal \mathbf{Y} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\intercal. \tag{3.13}$$

The columns of the unitary matrix $\mathbf{U}$ that are associated with the largest (non-zero) eigenvalues $\Lambda_k$ describe uncorrelated modes of variation of the (L)DOS. The truncation of the expansion to a small number of PCs determines the error one makes in approximating the DOS and corresponds effectively to an additional smoothing of the DOS. Building a model in the PC representation amounts to computing the projection of the DOS on the basis functions

$$\tilde{y}_k^{\text{PC}}(A) = \sum_j y_j(A)\, U_{jk}, \tag{3.14}$$

training a regression model on each of the $\tilde{y}_k^{\text{PC}}$ coefficients and then reconstructing the prediction in terms of the principal vectors,

$$y_j(A) \approx \sum_k \tilde{y}_k^{\text{PC}}(A)\, U_{jk}, \tag{3.15}$$

A third approach to represent the DOS can be derived to address the fact that a loss of the form Eq. (3.3) cannot discriminate between distributions that differ by the position of peaks that have negligible overlap – a problem that is frequently encountered when comparing spectral functions. The Wasserstein distance (also known as earth mover's distance) is a metric to compare distributions designed to address this problem, and that can be easily computed as the norm of the pointwise difference between the inverse cumulative distribution

functions [114]. Inspired by the Wasserstein metric, we propose to represent the DOS in terms of the associated cumulative distribution function (CDF), which can be computed as partial sums over the pointwise representation, which approximates the integral over the energy:

$$y_k^{\mathrm{CDF}}(A) = \sum_{j=0}^{k} y_j(A). \tag{3.16}$$

Even though a Euclidean norm that uses this vector is *not* a precise implementation of the earth mover's distance (that is based on the Euclidean distance between *inverse* CDFs), it is still sensitive to shifts in peak position. It also preserves the additive construction of the total DOS based on atom-centred contributions – a physical constraint that would be lost by using a metric based on the inverse CDF.

## 3.2 Benchmarks on silicon data set

### Data set

We use, as a training and validation data set, a challenging combination of 1039 silicon structures containing configurations that correspond to elastically and thermally distorted bulk diamond and $\beta$-tin structures, snapshots from molten silicon simulations, amorphous configurations obtained at different quenching rates, as well as some cluster configurations. We extract these structures from the training data set used to build an ML interatomic potential for silicon [26]. Fig. 3.2 demonstrates the heterogeneity of the data set, showing a projection on the two largest principal components of the average SOAP power spectrum vectors of the different configurations. The parameters of the SOAP representation are the same as those used for the regression models discussed below. The map reflects the presence of several distinct groups of structures that have been obtained with simulations performed at different temperatures and pressures. In what follows, we randomly selected 800 structures that we used to train the different models and used the remainder of the data set for testing. The random selection ensures approximately uniform sampling of the different portions of the configuration space. Unless otherwise specified in Sections 3.2 and 3.2, test errors are averaged over 16 random splittings of the overall data set. We compute the single-particle energy levels for this system by running density functional theory (DFT) calculations using the FHI-aims all-electrons code [115]. We use the "tight" settings, and the Perdew–Burke-Ernzerhof (PBE) [17] exchange-correlation functional. We keep a constant k-point spacing of $0.01\text{Å}^{-1}$ for the periodic structures. The energy levels are aligned by zeroing the vacuum level of the Hartree potential for isolated structures, and its constant, $G = 0$ component for periodic structures. As discussed below, we compute the density of states by summing over

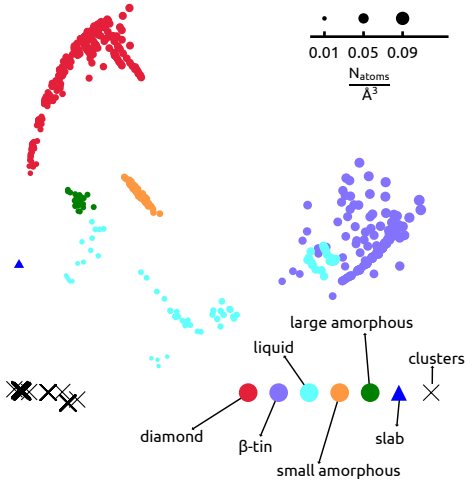


Figure 3.2: Clustering of the structures in the silicon data set based on the first 2 kernel principal components of every configuration. These components hold ≈ 94% of the variance in features space. The different subsets are well separated in features space, except for a few liquid structures computed at high pressure, that partly overlap with $\beta$-tin phase configurations. We also plot, on the same axes, the position of carefully-equilibrated, large amorphous silicon supercells (discussed in Section 3.3) and of one Si(100) slab (discussed in Fig. 3.8), used to assess the extrapolative capabilities of the model.

the single-particle eigenvalues, using different levels of Gaussian broadening. In appendix A.1, we discuss a few strategies for the energy reference for the DOS representations and how they affect the ML model's performance.

### Model hyperparameters and training

As introduced in Section 2.2, the SOAP representation has several hyperparameters that need to be tuned depending on the training data and the target property. Given that, as discussed above, we aim to compare the performance of the model with a different representation of the DOS and different target properties, one would need to perform hundreds of separate optimisation procedures for these hyperparameters. Instead, we performed a single optimisation using the cohesive energy as the target property; this avoids biasing explicitly our comparison towards one of the DOS learning protocols and is representative of a scenario in which one wants to re-use the features that underlie an ML interatomic potential to estimate additional electronic-structure properties.

We use the metric induced by a preliminary set of SOAP features to select, by farthest point sampling (FPS) [88], 3000 environments out of ≈22000 training environments to use as an active set for the PP approximation. As shown in Fig. 3.3, increasing the active set size further leads to a negligible prediction error reduction when using the PW representation of the DOS

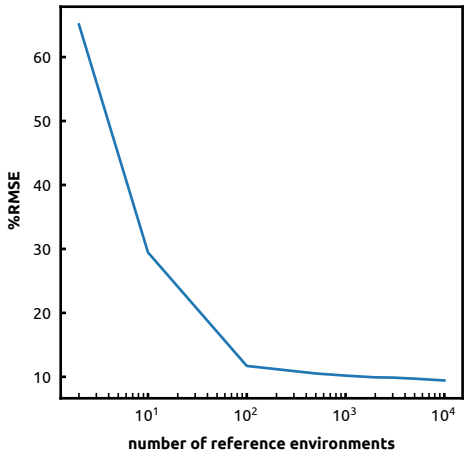


Figure 3.3: Evolution of the errors in the pointwise DOS prediction as a function of the active set size. The reference DOS is constructed using $g_b = 0.3$eV.

and a Gaussian broadening of $g_b = 0.3$eV of the target DOS. In particular, we notice that the reduction of prediction errors between an active set of size 3000 and an active set of size 10000 is smaller than 0.1%. A discussion of the effect of $g_b$ on the performance of the ML models will follow in this Section and is summarised in Fig. 3.9. We then select the best hyperparameters using a 5-fold cross-validation regression scheme and a grid search, using a single random ordering of the full data set. We obtain the smallest prediction errors for the cohesive energy for the following set of parameters in the notations of *librascal* [116] for the SOAP representation: *interaction_cutoff*=6Å, *max_radial*=12, *max_angular*=9, and *gaussian_sigma_constant*=0.45, and for the radial scaling: *rate*=1Å, *exponent*=5, and *scale*=3.0Å. This choice of parameters leads only to a marginal degradation of performance in comparison to a model that has been specifically optimised to reproduce the PW representation of the DOS, as shown in Fig. 3.4). For consistency, given that the change of hyperparameters modifies the kernel, and the kernel-induced distance, we then re-select 3000 active environments using these optimal values. It should be noted, however, that selecting new active points leads to negligible improvement in the accuracy of the model.

We build eight models using the same active set but different 50% subsampling of the 800 train structures. We report the mean of the model as our best prediction (which has an accuracy comparable to that of a single model trained on the full 800 structures), and rescale the spread of the models around the mean, as discussed in Section 3.1, to obtain an ensemble of predictions based on which we can easily propagate our uncertainty quantification. In order to investigate the impact of the representation of the DOS on the performance of the model, we consider three values of the Gaussian broadening $g_b$: 0.1eV, 0.3eV and 0.5eV. We discretise

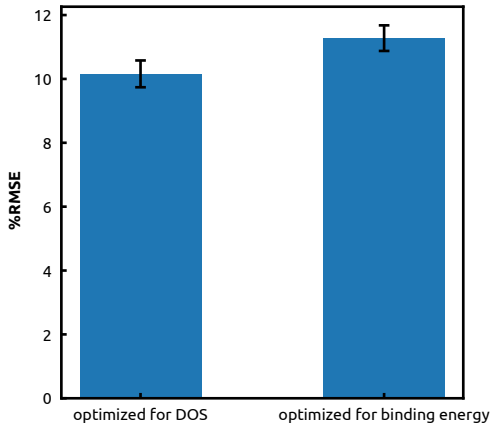


Figure 3.4: Comparison between the average errors in the DOS prediction over 8 train/test split using the hyperparameters optimized for the DOS prediction and the hyperparameters optimized for the binding energy prediction. The error bars represent the standard error of the mean. The Gaussian broadening of the reference DOS is $g_b = 0.3$eV.

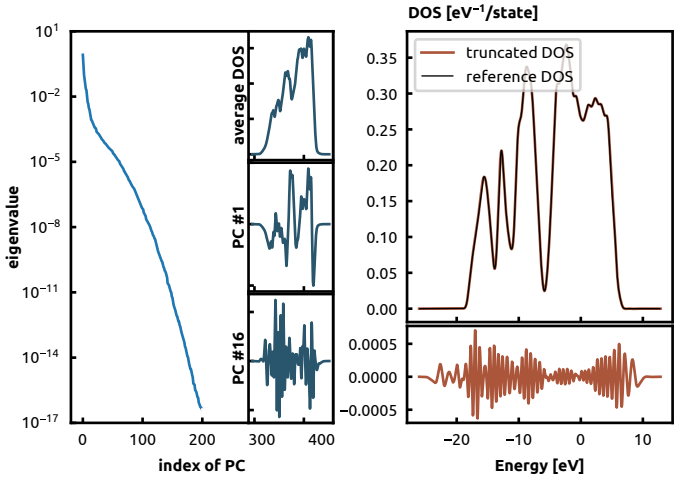


Figure 3.5: (Left) Distribution of the first 200 eigenvalues of the covariance matrix of the DOS at a $g_b = 0.3$ eV. The 3 small panels on the right represent the shape of the average DOS in the data set, the 1$^{st}$ principal component and the 16$^{th}$ principal component. (Right) The reference DOS curve of a silicon diamond structure at 0.3eV broadening and its reconstruction from the first 80 PCs. The lower panel shows the errors at every energy level. The total error for this structure is $\approx 5.11 \times 10^{-3}$ eV$^{-1}$/atom.

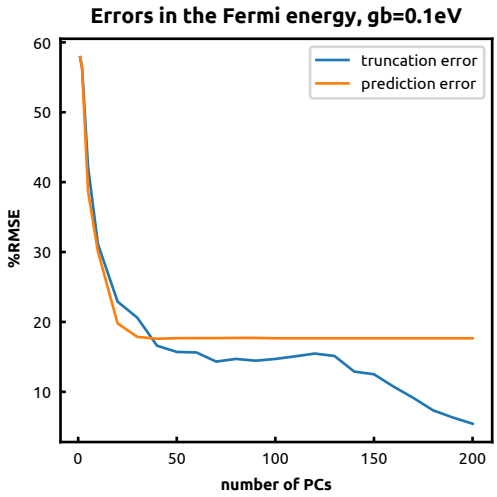


Figure 3.6: Evolution of the systematic errors in the Fermi energy $\varepsilon_F$ due to the truncation of the DOS (in blue) and the ML errors of $\varepsilon_F$ predicted from truncated DOS (in orange). The prediction error saturates starting from 38 PCs used to build the DOS. The reference DOS is built using $g_b = 0.1$eV.

the DOS on a grid where the points are spaced by $\delta\epsilon = 0.05$eV, which ensures that we are able to sample the fine structure of DOS($\epsilon$) when using a $g_b = 0.1$eV Gaussian broadening. We use this representation as the pointwise representation from Eq. (3.11). We use the same grid to compute the cumulative integral of the DOS and obtain the CDF representation of Eq. (3.16). We recover the DOS from the CDF representation by taking the derivative of the CDF with respect to the energy axis. In practice, we use a central derivative approach to determine DOS($\epsilon$):

$$\text{DOS}(\epsilon) = \frac{\text{CDF}(\epsilon + \delta\epsilon) - \text{CDF}(\epsilon - \delta\epsilon)}{2\delta\epsilon}, \tag{3.17}$$

where $\delta\epsilon$ is the energy discretisation step.

For the PC representation, Eq. (3.14), we select the principal eigenvectors of the covariance matrix computed for the 800 training structures. The left panel in Fig. 3.5 demonstrates the rapid decay of the eigenvalues of the covariance for $g_b = 0.3$eV, and the shape of the average DOS of the data set, the $1^{st}$ and the $16^{th}$ eigenvectors. One notices that the principal components corresponding to the high eigenvalues contribute to the general structure of the DOS curve, while the ones corresponding to lower eigenvalues describe its fine structure. We choose the number of PCs to retain 99.999% of the variance, which corresponds to 200, 80 and 35 PCs for $g_b = 0.1, 0.3, 0.5$eV, respectively. Even though the error resulting from this approximation is visually very small, as shown in the right panel of Fig. 3.5, it leads to non-negligible errors when using the DOS to compute the Fermi energy, as shown in Fig. 3.6.

## Benchmarking the models

Having discussed the details of the ML DOS model, we now turn to assess its performance on the database of silicon structures. We begin by comparing the accuracy as a function of the broadening of the density of states, and its representation, and proceed to determine how the error in the prediction of DOS($\epsilon$) translates to the error in quantities that can be obtained from it, such as the band energy or the Fermi level.

To facilitate the comparison of the model performance between different properties and scenarios, we normalise the root mean squares error (RMSE) by the standard deviation of the target property, expressed as a percentage. This is particularly important because reducing the Gaussian broadening substantially increases the complexity of the DOS, measured in terms of the variance over the full data set (c.f. Table3.1). For scalar properties, the expression reads:

$$\mathrm{RMSE}_{\mathrm{scalar}} = 100 * \frac{\sqrt{\frac{1}{N}\sum_i (y^i_{\mathrm{pred}} - y^i)^2}}{\sqrt{\frac{1}{N}\sum_i (y^i - \bar{y})^2}} \tag{3.18}$$

This expression is easily extended to cover the properties that have an energy dependence (such as the DOS and the distribution of excitations), that require the simultaneous regression of multiple coefficients, by comparing the $L^2$ distance to the deviation from the average vector representing the target property of the training set:

$$\mathrm{RMSE}_{\mathrm{vec}} = 100 * \frac{\sqrt{\frac{1}{N}\sum_i \int (\mathbf{y}^i_{\mathrm{pred}} - \mathbf{y}^i)^2}}{\sqrt{\frac{1}{N}\sum_i \int (\mathbf{y}^i - \bar{\mathbf{y}})^2}}, \tag{3.19}$$

where $\bar{\mathbf{y}}$ is the vector containing the average of each coefficient in the $\mathbf{y}^i$. We will use either of the definitions as appropriate and indicate the error simply as %RMSE.

### Comparison of the DOS representations

We begin by showing, for the pointwise representation of the DOS and $g_{\mathrm{b}} = 0.3$eV, a plot of the model DOS($\epsilon$) for the diamond and liquid structures with the lowest, median and highest predicted uncertainty (Figure 3.7). The figure demonstrates that a single model is capable of predicting the behaviour of Si across the semiconductor-to-metal transition and that the uncertainty quantification correctly identifies the most problematic test structures. As an example of the stability of the model when performing extrapolative predictions, we estimate the DOS of a 96-atom Si(100) slab, with one surface truncated to the bulk geometry, and the other reconstructed with a $c4 \times 2$ geometry (Fig. 3.8, geometry from Ref. [117]), using the

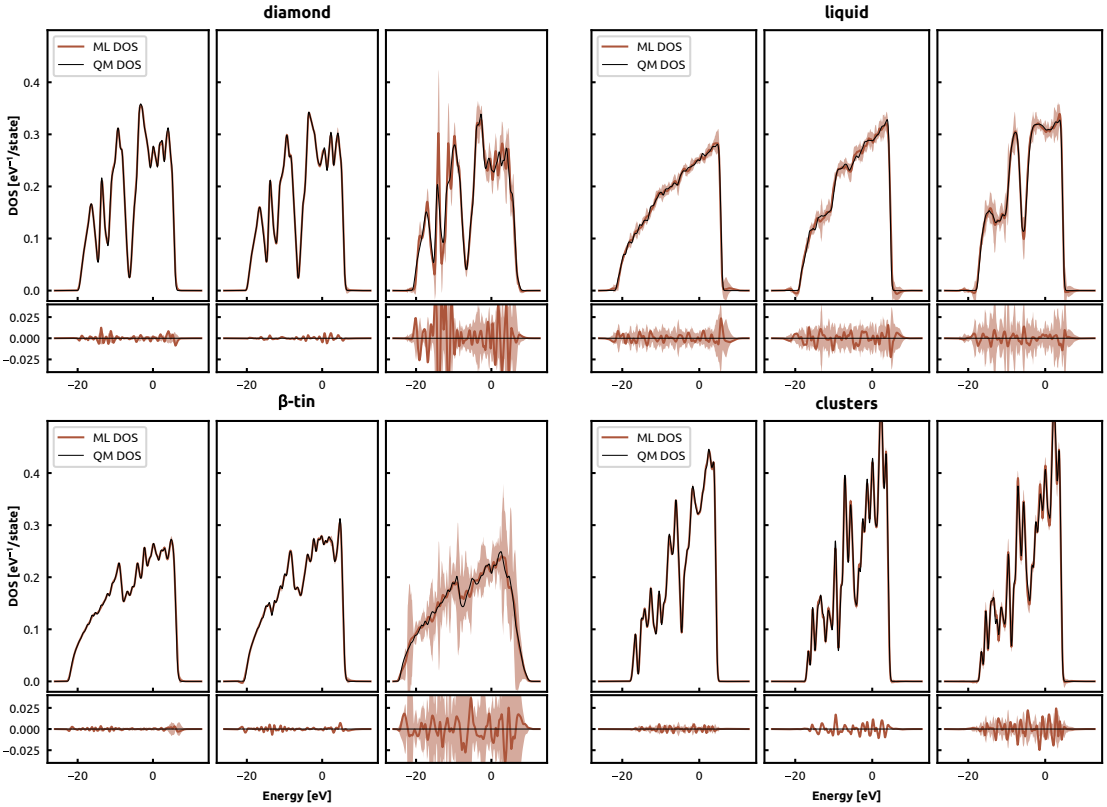


Figure 3.7: Representative examples of DOS predictions for the silicon data set. The four panels represent one of the subsets in the data set: diamond, liquid, $\beta$-tin structures and clusters. Every panel shows three cases corresponding to the best, median and worst predicted uncertainty in the test set, compared to the reference DFT DOS (left to right). The reference DOS is constructed using $g_b = 0.3$eV. The shaded areas indicate the uncertainty estimates of the ML model at every energy level based on a committee of 8 GPR models. The lower section of each plot depicts the residuals, colour-coded in the same way as the plot of the prediction.

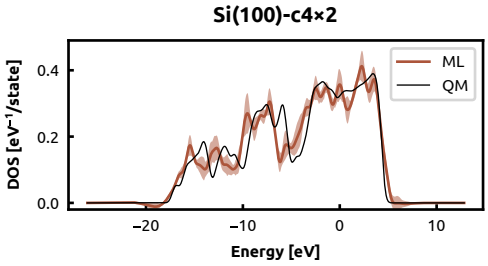


Figure 3.8: Comparison of DFT and ML DOS of 96-atom slab model of the Si(100) − $c4\times2$ surface reconstruction. The reference DOS is constructed using $g_b = 0.3$eV. The ML model uses the pointwise representation of the DOS. The shaded area represents the uncertainty of the ML model at each energy level.

| broadening (eV) | 0.1 | 0.3 | 0.5 |
|---|---|---|---|
| DOS (states/atom) | 0.435 | 0.335 | 0.297 |
| $\epsilon_F$ (eV) | 0.849 | 0.834 | 0.814 |
| DOS($\epsilon_F$) (eV$^{-1}$/atom) | 0.064 | 0.072 | 0.082 |
| $\epsilon_{\text{band}}$ (eV) | 2.483 | 2.477 | 2.474 |
| A($\Delta$) | 0.267 | 0.272 | 0.278 |

Table 3.1: Standard deviation of the density of states curve, the Fermi energy, the DOS value at the Fermi energy, the band energy and the excitation distribution over the entire silicon data set.

pointwise representation and a target Gaussian broadening of $g_b = 0.3$eV. As shown in Fig. 3.2, this structure is isolated in phase space, which results in both the predicted uncertainty and the actual error being large, comparable to the worse-case scenarios in Fig. 3.7. Even in this challenging example, however, the qualitative features of the DOS are correctly reproduced, and the large uncertainty could be used in an active learning setting to signal the need to refine the training set. This prediction could be useful to investigate, visually, the main features of the DOS of this 2D structure, but it may not be sufficient to extract derived quantities from the DOS like the DOS($\epsilon_F$) or the band energy because of the high predicted uncertainty. This result shows the extrapolative capabilities of the ML DOS model trained on 3D structures to 2D structures. Also, it highlights the need to tune the ML models with respect to the phase space region of interest. We investigate this aspect further in Section 3.3.

In order to assess more quantitatively the accuracy of the model for different values of broadening $g_b$ and different representations of the DOS, we then compute the %RMSE of predictions (Eq. (3.19)) using the full training set of 800 structures, shown in Fig. 3.9. To account for the dependency of the accuracy on the test/train split, we repeat the regression and testing on 16 random splits and report the mean and standard error of the mean over the different splits.

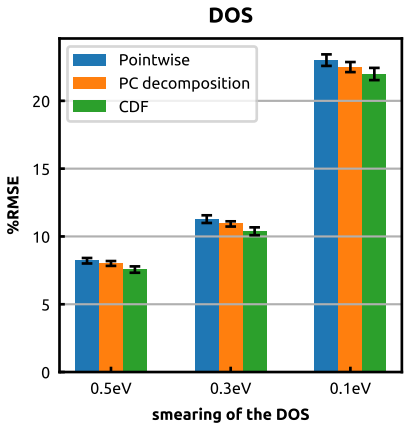


Figure 3.9: Average errors in the DOS over 16 train/test splits in the Si data set using 3 different representations of the DOS curves: the pointwise approach, the decomposition on a basis set of principal components, and the DOS derived from the learnt CDF, and for 3 different Gaussian broadening values: 0.5eV, 0.3.eV and 0.1eV. Different representations lead to comparable errors, that grow systematically with decreasing $g_b$. The error bars represent the standard error of the mean. Errors for the PC decomposition due to the truncation of the PC basis are negligible.

Even though we have renormalised the error on the intrinsic variance of the data, which is larger for the smaller values of the Gaussian broadening (c.f. Tab. 3.1), the error in the predicted DOS is clearly much larger for the finer $g_b$. The errors jump from roughly 8% for the 0.5eV broadening and 11% for the 0.3eV broadening to 22% for the 0.1eV broadening. The representation of the DOS has a small impact on the accuracy of the model, with the CDF showing a slight but systematic advantage over the pointwise and the PC representations. The projection errors for the latter representation (i.e. the PC representation) are too small to affect the errors of the DOS predictions, unlike what we will encounter later when we discuss the derived quantities in Section 3.2.

We also investigate the relevance of the error estimation method we utilise in this work. In Figure 3.10, we report the parity plots between the integrated uncertainty based on 8 committee models and the RMSE of the pointwise DOS for different values of broadening $g_b$ = 0.5eV, 0.3eV and 0.1eV. We define the integrated uncertainty of a DOS curve of a structure $A$ as follows:

$$\sigma(A) = \sqrt{\int \sigma^2_{RS}(A,\epsilon)\mathrm{d}\epsilon},$$

where $\sigma^2_{RS}(\epsilon)$ is the uncertainty of DOS$(A,\epsilon)$. We notice that, on average, the uncertainty increases when we decrease $g_b$. We also notice that higher uncertainty correlates well with higher prediction error of the pointwise DOS.

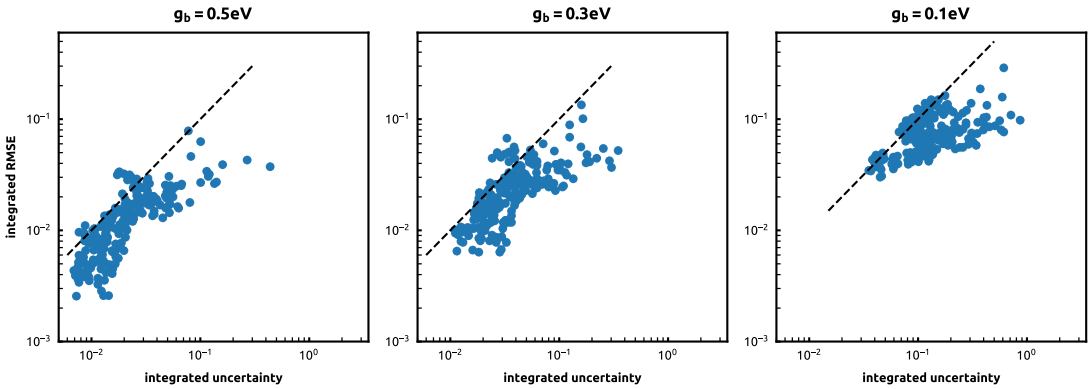


Figure 3.10: Parity plot in log-log scale of the integrated uncertainty vs the integrated RMSE of the pointwise ML DOS of single structures in the data set. The reference DOS are constructed using $g_b = 0.5$eV, 0.3eV and 0.1eV, respectively. The dashed black line represents the parity line $y = x$.

In order to understand the limitations of our data set and to validate the training set size we use in this work, we examine the learning curves (LC) for the different representations of the DOS, as shown in the left panel of Fig. 3.11 for the most challenging case of $g_b = 0.1$ eV. For training sizes less than $\approx 100$ structures, we notice that all the LCs are decreasing algebraically, and for larger sizes, the rate decreases, and we can see that the LCs start to saturate, which indicates that adding more data to train the ML models does not result in a significant improvement of the models' performance. Despite small differences at the smaller train set sizes, the three representations show similar convergence behaviour. The same observation holds even when looking at the LCs of the pointwise representation, in the central panel of Fig. 3.11 for different Gaussian broadening values $g_b = 0.5,\ 0.3,$ and 0.1eV. However, the LC associated with the largest broadening of $g_b = 0.5$eV seems to saturate for a larger train set size compared to the smaller broadening values. The PC representation can provide some insight into the origin of the plateau. The right panel of Fig. 3.11 shows the LCs of the first, third and sixteenth projections of the DOS on the PC basis, with the fractional error referring to the variance of each component. The first two elements are well-learnt: errors are around 10% for 100 structures in the training set, and the corresponding LCs saturate at low validation errors, 2% and 8%, respectively. In contrast with the first few principal components, the learning is slower and more difficult for lower-variance components. The error is below 60% for the $16^{\text{th}}$ component only for the full training data set. In general, we observe that the convergence of these individual errors gets slower as we consider higher PCs of the DOS. As shown in Fig. 3.5, these smaller-variance PCs are associated with high-frequency, "noisy" modes that are necessary to describe the fine structure of the DOS. For many applications, a large Gaussian broadening does not hinder using the density of states – and indeed, previous attempts at

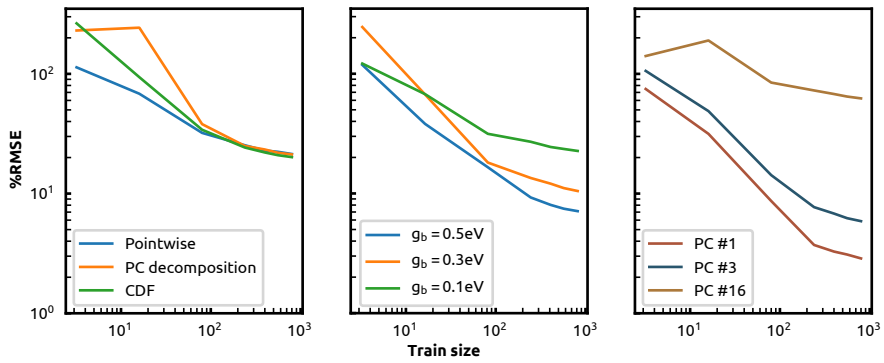


Figure 3.11: Learning curves for (Left) the 3 different representations of the DOS curve where the reference DOS is computed with $g_b = 0.1$ eV, (Centre) for the pointwise representation the reference DOS is computed with $g_b = 0.5,\ 0.3,$ and 0.1eV, and (Right) the projection on the $1^{st}$, $3^{rd}$ and $16^{th}$ PC as a function of the training structures. All the errors are normalised with respect to the standard deviation of the training set.

using ML to predict the DOS used large broadening values (e.g. 0.2 eV in Ref. [52]). Whenever a higher resolution is needed (e.g. to identify the position of individual defect states or to determine precisely the band gap), a higher density of data, possibly in combination with more complex models, is needed. In Section 3.3, we show that - when focusing on a more restricted set of configurations - it is possible to achieve quantitative accuracy with a fine DOS broadening by a relatively minor tuning and an extension of the training set.

### Learning from the DOS

Besides its intrinsic interest as a description of the single-particle energy levels in a condensed-phase system, the DOS($\epsilon$) can be used as the starting point to derive other quantities that relate to experimental observables. We consider four quantities: the Fermi energy ($\varepsilon_F$), the density of states at the Fermi energy (DOS($\varepsilon_F$)), the band energy ($\epsilon_{\text{band}}$), and the distribution of excitations ($A(\Delta)$) because they are easily extracted from a smooth DOS curve. The definition of these quantities was given at the beginning of Section 3.1. For each of these properties, we build an ML model using the same kernel parameters and train set, minimising a loss function analogous to Eq. (3.3). We compare these direct predictions to *indirect* models built by first predicting DOS($\epsilon$) and then using it to compute $\varepsilon_F$, DOS($\varepsilon_F$), $\epsilon_{\text{band}}$, $A(\Delta)$. Whenever an expression depends on $\varepsilon_F$, we use the value computed consistently from the predicted DOS. We build different models with various values of $g_b$ and representation of the DOS, as discussed in the previous paragraph.

The average prediction errors for the four properties and for the different models are illustrated

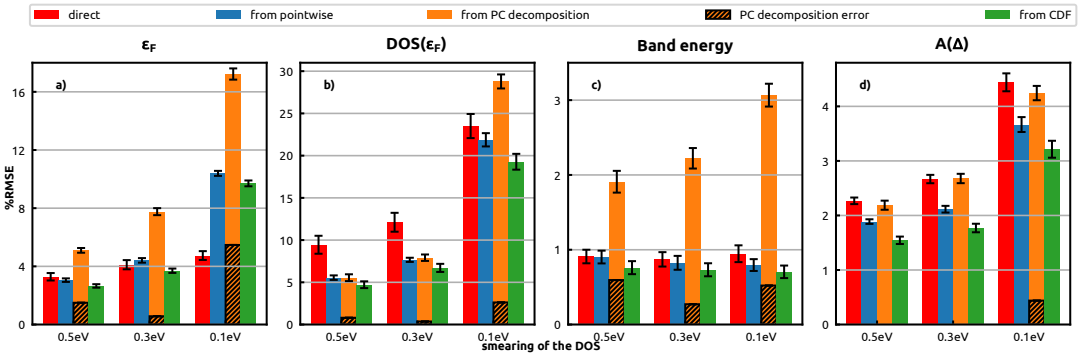


Figure 3.12: Average errors in the derived quantities over 16 train/test splits in the Si data set using three different representations of the DOS curves: the pointwise approach, the decomposition on a basis set of principal components and derived from the learnt CDF, and for three different $g_b$ values: 0.5 eV, 0.3 eV and 0.1 eV. The error bars represent the standard error of the mean, and hatched areas represent the average systematic errors due to the projection on the basis of PCs. The test errors grow systematically when we decrease the value of $g_b$.

in Fig. 3.12. Similar to what we observe for the DOS learning, the prediction errors increase as we decrease the value of $g_b$, including the direct method. The truncation errors in the PC decomposition representation contribute significantly to the overall errors.

Let us focus on each individual quantity, starting with the Fermi energy, whose accurate determination is particularly important, as it enters the definition of the other three quantities. The prediction errors for the direct models are low, between 3.5% and 4.5%. They increase, although less dramatically than in the case of the DOS, for decreasing values of the Gaussian broadening $g_b$. Errors for the indirect predictions are comparable at $g_b = 0.3$ eV and $g_b = 0.5$ eV, except for the PC decomposition, for which the errors are almost twice as large. The truncation of the DOS contributes to the larger error. As shown in Fig. 3.6, a large number of PC components is necessary to obtain an accurate estimate for $\epsilon_F$. In combination with difficulty in learning the fine-grained components, this explains the poor performance of the PC approach. Similarly to what was observed for the DOS modes, the error increases substantially for the smallest broadening value, and the direct prediction outperforms, by nearly a factor of 2, all the indirect predictions.

The prediction errors of the DOS($\epsilon_F$) of the direct model follow the same trend as the DOS, where they grow from 9.5% for $g_b = 0.5$eV to 23% for $g_b = 0.1$eV. In contrast to the case of the Fermi energy, here the errors of the indirect models are significantly lower for $g_b = 0.5$eV ($\approx 5\%$ error) and $g_b = 0.3$eV ($\approx 7\%$ error) and comparable between the three approaches, with a minor advantage for the CDF scheme. For $g_b = 0.1$eV, errors increase substantially, but

the indirect models still outperform a direct prediction, except for the PC decomposition, where the errors are close to 29%. The DOS truncation error contributes partly to the poor performance of the PC scheme, similar to what was observed for the Fermi energy – whose internally-consistent prediction is used as the point at which the DOS is computed.

The prediction errors of the direct model of the band energy are low compared to the intrinsic variability, below 1% and largely independent of the value of $g_\mathrm{b}$. The fact that the error in the predicted band energy is largely independent of the broadening suggests that the averaging procedure that is associated with the evaluation of $\epsilon_\mathrm{band}$ reduces the sensitivity to the fine details of the DOS and that the large error that is observed on the prediction of DOS($\epsilon$) for $g_\mathrm{b} = 0.1$eV is not reflected in coarser-grained features of the distribution of energy levels. The prediction errors of the indirect models are slightly lower than those of the direct model, once again with the exception of the PC decomposition, where the errors jump to 2% for $g_\mathrm{b} = 0.5$eV to 3% for $g_\mathrm{b} = 0.1$eV. Even though the use of PCs does help improve the accuracy of the predicted DOS marginally, this is clearly not reflected in the accuracy of derived quantities, because of the presence of high-frequency components in the DOS that contribute to the value of $\varepsilon_F$, $\epsilon_\mathrm{band}$ and DOS($\varepsilon_F$) and are either discarded or very difficult to learn. Finally, the direct prediction errors of $A(\Delta)$ are usually low (between 2% and 4.5%), with errors that grow gently as $g_\mathrm{b}$ is reduced. The errors of the indirect models are systematically lower than the direct model, with the CDF model consistently showing the best performance.

Overall, these examples show that using a model of the DOS as an intermediate step in the calculation of electronic-structure properties can outperform, marginally or substantially, a direct prediction. The improvement is most noticeable when learning properties such as the excitation density $A(\Delta)$ or the DOS at the Fermi level that clearly depend in a non-trivial way on non-locality – i.e. the presence of a localised defect can change in a non-additive manner the value of the property for the entire system. Another advantage of an indirect model is that, based on a single prediction, one can compute a multitude of physical observables – in addition to those we mention here, the electronic contributions to a material's heat capacity or Gibbs free energy, the band gap, etc. – and that these predictions are consistent with each other, rather than affected by independent model errors. Contrary to what we observed when building a model of DOS($\epsilon$), the strategy used to represent the target function has an important effect on the accuracy of the indirect predictions. In particular, learning the separate principal components in the data set leads consistently to degraded performance, in part because of the error incurred by truncating the PC expansion, but also in part because higher-order components have very poor learning rates, at least when using a single set of features for all the components. The CDF-based model is consistently the best of the indirect models, suggesting that a Wasserstein-type metric is the most relevant way to assess the quality of a predicted DOS.

| | variability/eV$^{-1/2}$ $\sqrt{\frac{1}{N}\sum_i \int (\mathbf{y}^i - \bar{\mathbf{y}})^2}$ | | | RMSE/eV$^{-1/2}$ $\sqrt{\frac{1}{N}\sum_i \int (\mathbf{y}^i_{\text{pred}} - \mathbf{y}^i)^2}$ | | | |
|---|---|---|---|---|---|---|---|
| $g_b$/eV | 0.5 | 0.3 | 0.1 | 0.5 | 0.3 | 0.1 | $n_{\text{test}}$ |
| diamond | 0.118 | 0.149 | 0.22 | 0.02 | 0.036 | 0.103 | 79 |
| $\beta$-tin | 0.092 | 0.105 | 0.13 | 0.017 | 0.025 | 0.057 | 61 |
| liquid | 0.122 | 0.139 | 0.163 | 0.021 | 0.032 | 0.074 | 16 |
| clusters | 0.095 | 0.181 | 0.418 | 0.010 | 0.020 | 0.087 | 60 |
| amorphous | 0.02 | 0.029 | 0.063 | 0.018 | 0.028 | 0.07 | 23 |
| TOTAL | 0.254 | 0.295 | 0.393 | 0.017 | 0.029 | 0.084 | 239 |

Table 3.2: Overview of the test set intrinsic variability and RMSE for the pointwise prediction of the DOS, for one of the 16 random train/test splits and for different values of $g_b$ = 0.5eV, 0.3eV and 0.1eV.

It is worth noting that the three approaches are simple linear transformations of the same data, which is, in a multivariate kernel regression framework, equivalent to the choice of a non-diagonal regularisation of the regression weights that couples different target properties. One could envisage explicitly optimising the regularisation to improve the accuracy in the desired derived quantities, by choosing the regularisation value that reduces the (cross-)validation error on the derived quantity of interest instead of minimising the validation error of the DOS.

## 3.3 Electronic fingerprints in amorphous silicon

We use carefully-equilibrated large-scale configurations of amorphous silicon [118] to demonstrate two advantageous features of a local machine-learning model, such as the one we use here. On the one hand, it allows predicting properties of large structures with a cost that scales linearly with system size; on the other, it provides a data-driven decomposition of the DOS in local contributions that can be used to analyse structure-property relationships.

### Large-scale evaluation of the DOS

We consider a series of larger amorphous silicon structures, with a size ranging between 216 and 4096 atoms, that were obtained by slowly annealing a molten Si configuration using an ML interatomic potential, following the protocol described in Ref. [118]. For all but the largest size, we compute DFT reference values following the same scheme we used for the train set. As can be seen in Fig. 3.2, the larger size and careful equilibration lead to structures that are quite different from the 64-atom amorphous silicon, but very similar to each other, concentrated in a narrow region close to that occupied by liquid configurations. As shown in Fig. 3.13, the general-purpose model we benchmark in Section 3.2 achieves an accuracy

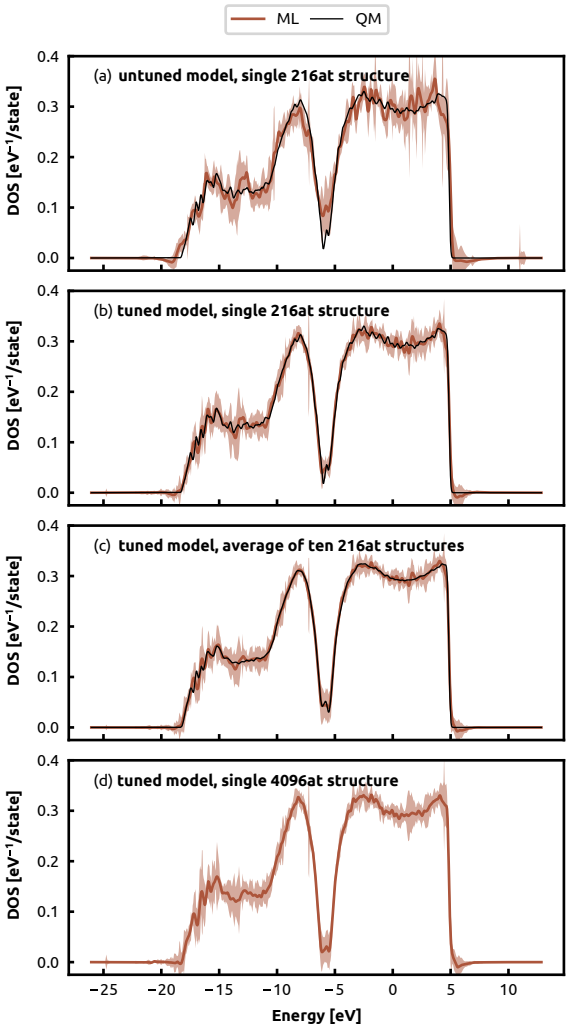


Figure 3.13: Comparison of DFT and ML electronic DOS of (a) a 216-atom structure where the ML DOS is based on the general-purpose ML model discussed in Section 3.2, (b) a 216-atom structure where the ML DOS is based on an ML model that has been further tuned for this class of structures (c) an average of ten 216-atom structures with the tuned ML model (d) a 4096-atom structure with the tuned ML model. For all panels, the reference DOS is constructed using $g_b = 0.1$eV, and ML models use the CDF representation. The shaded area indicates the uncertainty in the ML prediction.

comparable to that observed for liquid configurations – exhibiting clearly the qualitative features of the reference DOS. However, as noted earlier, a more finely-tuned training set is needed to achieve quantitative prediction accuracy with a high-resolution, $g_b = 0.1$eV DOS reference. To demonstrate that this fine-tuning can be easily achieved when focusing on a targeted application, we modify the training set by eliminating the cluster configurations that occupy a completely disconnected portion of phase space, exhibit very large variance, and are built using a different band alignment reference with respect to bulk structures (see Fig. 3.2 and Tab. 3.2). We also add 10 amorphous structures of 128 atoms each, generated by the same potential as the other large structures, to ensure that the train set contains disordered configurations that are more representative of the large, slowly quenched configurations. We represent the DOS using the CDF approach and with a Gaussian broadening $g_b = 0.1$eV. We also use the same SOAP parameters we adopted for the benchmarking of the ML DOS models from Section 3.2.

When computing the properties of materials in realistic thermodynamic conditions, it is necessary to average over multiple configurations, to compute a mean value that is consistent with a (quasi)-equilibrium probability distribution. We observe that this ensemble average smooths the DOS, and that, as a result, the agreement between ML predictions and DFT reference values improves substantially (Fig. 3.13). The features of the DOS are well reproduced, including the presence of a small peak around the Fermi energy (i.e near $\approx -5.7$eV). The cost of an ML prediction of the DOS is several orders of magnitude smaller than that of a DFT calculation, even for the smaller system sizes. Furthermore, the cost of a multivariate GPR prediction scales linearly with system size as opposed to the cubic scaling of DFT. We report in Fig. 3.14 the representative timings for calculating the DOS of amorphous silicon samples of increasing size. The ML model allows computing inexpensively the DOS of the largest structure of 4096 atoms (bottom right panel of Fig. 3.13), for which an explicit DFT calculation would require the application of a linear-scaling approach at still substantial cost, as a cubic-scaling DFT simulation using the same setup we applied to smaller structures would require more the 1 million CPU hours and 20 TB of RAM. The DOS predicted for this structure is consistent with the average DOS of the smaller structures – which indicates that this larger sample contains motifs that are found, with similar probability, in smaller simulations.

### Interpreting the local machine-learning DOS

Having shown that the ML model accurately reproduces the total DOS of the sample, we assess whether the local contributions can be given a meaningful interpretation. The atom-centred decomposition of the DOS that underlies our model can yield LDOS contributions

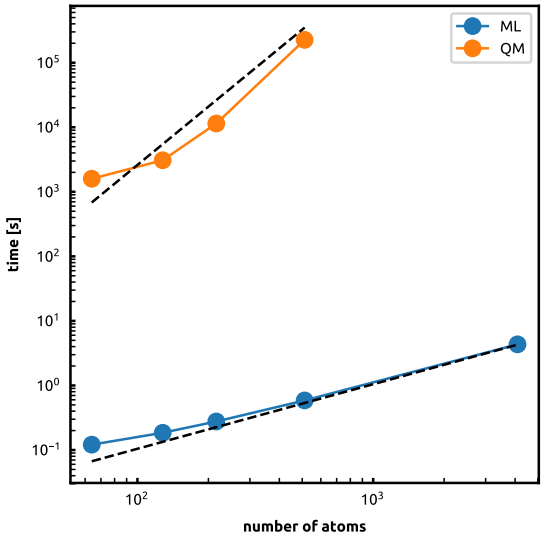


Figure 3.14: CPU time needed to produce the DOS for amorphous silicon structures of different sizes ranging from 64 atoms to 512 atoms when performing a DFT calculation and from 64 atoms to 4096 atoms when using the tuned ML model from Section V.A. In the latter case, the time of the SOAP representation is included. The dashed line near the blue curve represents the linear scaling with the number of atoms in the structure. The dashed line near the orange curve represents the cubic scaling with the number of atoms in the structure.

that are negative over some energy ranges. These negative contributions, which might appear unphysical, are a consequence of the fact that each atom-centred term reflects information from all of the atoms within the cutoff distance, so that atoms with large positive and negative LDOS($\varepsilon_F$) combine to yield the observed total DOS, which is the only physical observable given as target. As discussed in Refs. [119, 120], in the absence of an explicit, physics-based local learning target, atomic ML predictions reflect the interplay between structures and properties *mediated by the choice of representation.* This is particularly relevant in light of the recent observation that 3-body correlation features are incomplete [74], and that learning of global properties relies on neighbouring atoms to disambiguate pairs of environments that have different structures but the same features (and hence the same ML-predicted local properties).

To investigate how the ML predictions of the LDOS can be used to analyse the structural motifs found in the 4096-atom amorphous silicon structure, we use the recently-developed kernel principal covariates regression technique (KPCovR) [39] to obtain a visually-interpretable description of the structure-property relations. KPCovR finds a low-dimensional projection of the kernel-induced features that correlate linearly with a set of target properties – in this case, the energy-resolved LDOS. It can be seen as a modified kernel principal-component analysis

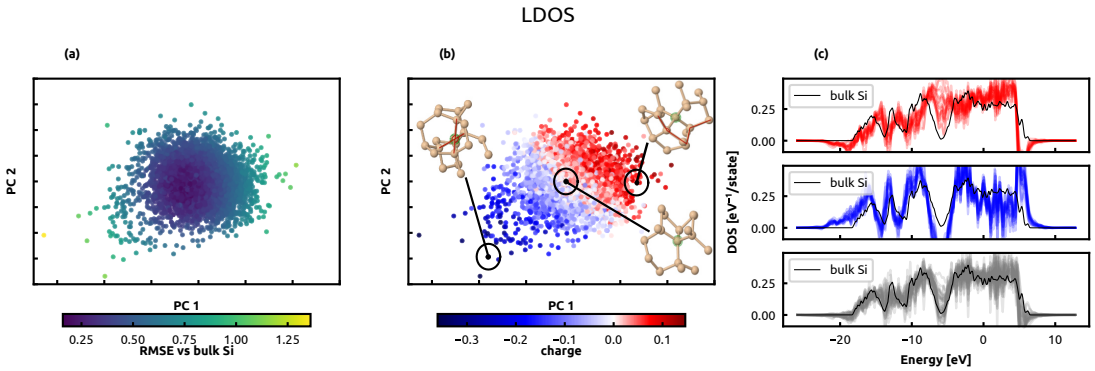


Figure 3.15: KPCovR map [39] of the Si environments in the 4096-atom amorphous configuration, built based on a combination of SOAP kernels and LDOS predictions, with a mixing parameter $\alpha = 0.05$. Points are coloured according to (a) the RMSE of the LDOS and (b) the local charge computed based on the LDOS. Snapshots of selected environments are also shown, with highly-distorted Si-Si-Si angles highlighted in dark red. (c) Comparison between the LDOS of selected configurations compared to the DOS of bulk Si. From top to bottom, the panels correspond to P, N, and O type environments (respectively, the 10 environments with the lowest, highest and median local charge).

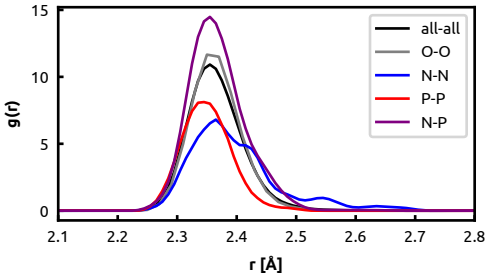


Figure 3.16: Pair correlation functions between Si atoms, resolved according to the classification between N (negatively-charged, $Q < -0.05$), P (positively charged, $Q > 0.05$), O (neutral, $-0.05 \leq Q \leq 0.05$) atoms, introduced in the text.

in which one uses a modified kernel with a scaling parameter $\alpha$ combining the structural information encoded in $\mathbf{K}$ with the target properties (more precisely, their best GP estimate) $\hat{\mathbf{Y}}$:

$$\tilde{\mathbf{K}} = \alpha\mathbf{K} + (1-\alpha)\hat{\mathbf{Y}}\hat{\mathbf{Y}}^{\mathrm{T}} \tag{3.20}$$

As shown in Fig. 3.15a, KPCovR constructs a purely structure-based latent space that correlates well with the ML LDOS, and identifies environments that have a DOS similar to bulk silicon, and defective environments with substantially different (local) electronic properties. We can further analyse the link between geometric and electronic structure by computing a "local

charge" indicator, defined as

$$Q(A_i) = N_{\text{val}} - \int \mathrm{d}\epsilon \, \text{LDOS}(A_i, \epsilon) f_{FD}(\epsilon - \varepsilon_F)_{|T=0}, \tag{3.21}$$

where $N_{\text{val}}$ is the number of valence electrons. This local charge correlates strongly with the KPCovR map (see Fig. 3.15b). Type N environments (to the left of the plot) are negatively charged, type P environments (to the right of the plot) have a net positive local charge, and type O environments (in the middle, having a DOS most similar to bulk Si structures) are approximately neutral. By visualising the environments with the interactive structure analyser Chemiscope [121] (input available in the Supplemental Material of Ref. [109]), we can recognise the structural features associated with the principal axes of the KPCovR latent space. Type N environments have a distorted structure, with tetrahedral angles that approach 180 degrees. Type P environments have a relatively regular distribution of nearest neighbours, but *their neighbours* have a highly distorted configuration, similar to that observed for environments at the right end of the map. Environments at the centre of the plot (type O) have both the central atom and its neighbours in a regular tetrahedral structure, and often with the same relative orientation, as one would find in crystalline Si structures. Looking more closely at the LDOS associated with these structures (Fig. 3.15c), one sees that, as anticipated above, the atom-centred ML predictions exhibit strongly unphysical features, with large negative values of the LDOS being associated with P environments. The fact that physical values of the total DOS can arise by combining unphysical predictions is also apparent in the fact that the positions of N and P environments are strongly correlated. As shown in Fig. 3.16, N and P atoms are less often found as first neighbours of an environment of the same type, while N–P pairs are encountered more often than one would expect from a random distribution. One possibility to recover a more physically-interpretable prediction is to compute *local averages* of the ML LDOS. In other terms, since each atom appears in multiple environments, it should get a share of the prediction for each of the environments it contributes to. This suggests the following expression for a locally-averaged DOS (LADOS) leveraging the radial scaling function $u(r)$ from Eq. (2.13) of the SOAP representation:

$$\text{LADOS}(A_i, \epsilon) = \sum_{j \in A} \frac{f_{\text{cut}}(r_{ij}) u(r_{ij}) \text{LDOS}(\epsilon, A_j)}{\sum_{k \in A} f_{\text{cut}}(r_{jk}) u(r_{jk})}, \tag{3.22}$$

in which we use a weighting of the contributions that corresponds to that used to construct the local density features. This formulation also ensures that the same prediction for the global DOS of the structure can be obtained by summing over the LADOS values, as well as over the raw LDOS (i.e. pre local averaging). As shown in Fig. 3.17, local averaging reduces the variability in the atom-centred predictions and leads to largely positive-(semi)-definite

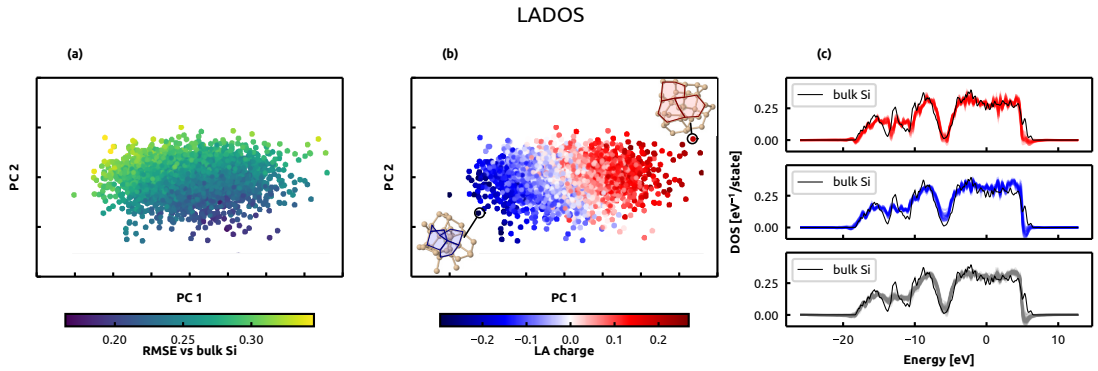


Figure 3.17: (left) KPCovR map [39] of the Si environments in the 4096-atom amorphous configuration, built based on a combination of SOAP kernels and LADOS predictions, with a mixing parameter $\alpha = 0.05$. Points are coloured according to (a) the RMSE of the LADOS and (b) the local charge computed based on the LADOS. Snapshots of selected environments are also shown, with 5-rings highlighted in blue, and 7-rings highlighted in red. (c) Comparison between the LDOS of selected configurations compared to the DOS of bulk Si. From top to bottom, the panels correspond to P, N and O type environments (respectively, the 10 environments with the lowest, highest and median local charge).

values of the local density of states – which is consistent with the fact that averages over large portions of a structure tend to a well-defined total DOS. The principal axis of the KPCovR map is still largely correlated to a local charge value $Q(A_i)$ computed based on the LADOS, but the structural features that are associated with the local charge are less apparent and more delocalised than those we found for the ML LDOS. We observe that structures with a large positive local charge tend to be associated with rings of 7 Si atoms, while structures with a large negative charge tend to be close to many 5-membered rings. Furthermore, the LDOS and LADOS-based values of $Q(A_i)$ correlate poorly with each other and with commonly-used structural descriptors, such as the tetrahedrality index [122, 123] (for an interactive view, see the Chemiscope input in the Supplemental Material of Ref. [109]), and do not show strong signals for the over- and under-coordinated structures that are found to exhibit a localised band gap state in physics-based local DOS calculations [124].

Another direction to interpret the L(A)DOS-based values of $Q(A_i)$ is to compare them to physics-based atomic charges. We were provided with the atom-projected DOS as calculated by the Local Orbital Basis Suite Towards Electronic-Structure Reconstruction (LOBSTER) [125] software of a 512-atom amorphous silicon structure. These atom-projected DOS can be interpreted as the LDOS within our atom-centred approach to model the DOS (Eq. (3.2)). We compare these LDOS to the ones obtained by our ML approach. In Fig. 3.18, we show examples of the LDOS defined by the LOBSTER analysis and our ML approach for three atomic environments. We notice that our ML LDOS is able to recreate the physical LDOS in the lower-

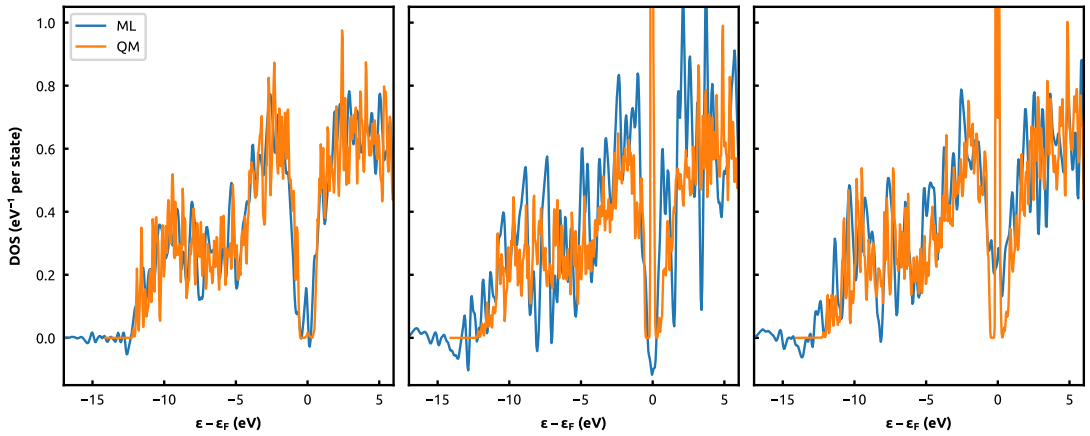


Figure 3.18: Examples of ML LDOS predictions compared to their LOBSTER counterpart for three environments in a 512-atom amorphous silicon structure. The Fermi energy $\varepsilon_F$ is set as the energy zero. The reference ML DOS is constructed using $g_\mathrm{b} = 0.1$eV. The reference LOBSTER LDOS is constructed using $g_\mathrm{b} = 0.05$eV.

energy regions but usually fails to describe the behaviour near the global Fermi energy of the structure. In particular, our ML LDOS model cannot predict the localised highly occupied state in the band gap. This can be explained by the fact that the ML model is only trained on the global DOS, which does not present these trends. In our tests, the ML LADOS suffers from the same issues as the ML LDOS. We can use the LDOS of LOBSTER to define local-atomic-based charges $Q(A_i)$, and we call them the LOBSTER charges. We compare these values of the LOBSTER charges to the values of LADOS-based values of $Q(A_i)$ computed from three cutoff radii: 0.0Å corresponding to no local averaging, 3.0Å corresponding to averaging over the first neighbours as described by the pair distribution function, and 6.0Å corresponding to the SOAP representation cutoff and to what we used for Fig. 3.17. We report the result of this comparison in Fig. 3.19. In particular, the "best" match between the two sets of locally defined charges occurs for an average over the first shell of neighbours. However, the local charge predictions for the under-coordinated (3-fold) and over-coordinated (5-fold) atoms are not in agreement with their reference values from the LOBSTER analysis. This behaviour can be explained by the fact the locally averaged ML LDOS are not able to recreate the localised state at the band gap, a characteristic of these environments according to the LOBSTER analysis, because the ML DOS model is not explicitly trained to reproduce these electronic features. Further investigation of the effect of the DOS representation on the local atomic charges, without local averaging, can be found in Appendix A.3.

An analysis of the atom-centred ML DOS, in combination with a hybrid supervised/unsupervised learning method such as KPCovR, facilitates the identification of the impact of structural

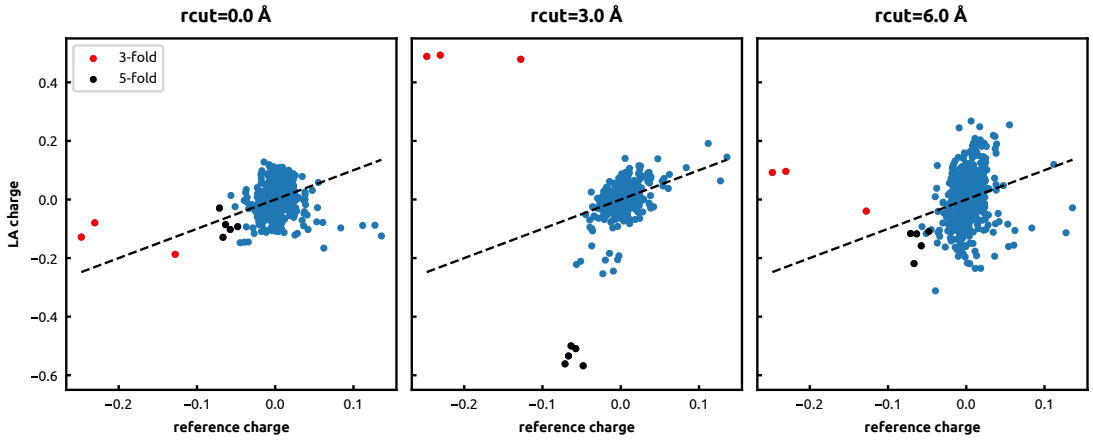


Figure 3.19: Parity plots of the atomic charge as computed from a LOBSTER [125] analysis of a 512-atom amorphous silicon structure and the locally averaged charge with different averaging cutoffs in Eq. (3.22) where the cutoff radius is (Left) 0.0Å corresponding to no local averaging, (Centre) 3.0Å corresponding to averaging over the first neighbours as described by the radial distribution function, and (Right) 6.0Å. The red and black dots correspond to the under-coordinated and over-coordinated atoms within the structure, respectively.

heterogeneity on the electronic structure of disordered materials, but one should not over-interpret an analysis that is also influenced by the details of the ML representation and the regression scheme.

## 3.4 Application: Origins of electronic transitions in disordered silicon

The study of disordered phases of materials is challenging for computer simulations at a quantum mechanical level because it requires access to large system sizes and long simulation times. ML interatomic potentials are a powerful tool to address these challenges because of their transferability, as one could train a model on small configurations and let the model generalise the mapping to much larger systems. In particular, we collaborated with a group of researchers on a study of the structural and electronic transitions while compressing a large box of 100,000 atoms of amorphous silicon with ML models, and it resulted in Ref. [38].

The main focus of this publication is to describe the mechanisms of structural transitions of amorphous silicon under high pressure. Diamond anvil cell experiments have indicated an amorphous–amorphous transition upon compressing amorphous silicon to several gigapascals, evidenced by the sudden disappearance of high-frequency Raman fingerprints and by a concomitant sharp increase of the electrical conductivity (a semiconductor–metal

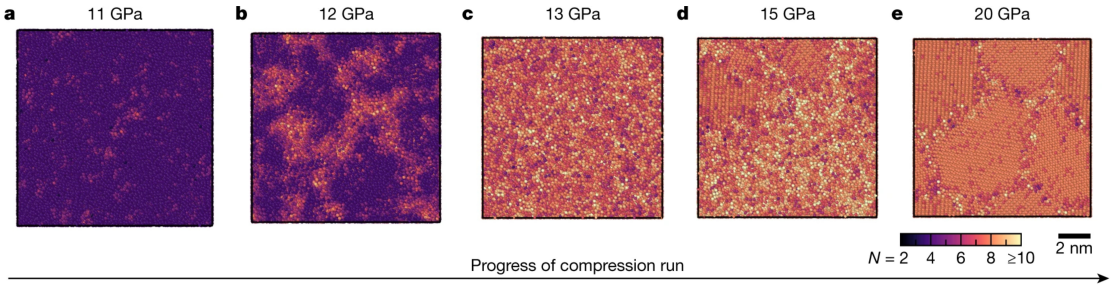


Figure 3.20: Amorphous silicon at high and very-high pressure. Colour coding indicates coordination numbers, N (spatial cut-off = 2.85Å) (a-e) Structural snapshots during isothermal compression at 500K using the GAP model, showing the coexistence of LDA-like ($N = 4$) and HDA-like ($N > 4$) regions up to 11GPa, the collapse into a transient VHDA phase ($N \gg 4$) at 12-13 GPa, and finally the formation of sh crystallites.

transition), both indicative of a major change in atomistic structure [126, 127, 128]. Increasing the pressure even further, to about 14 GPa, was seen to induce crystallisation of the simple hexagonal (sh) phase of silicon (thereby demarcating the existing limit of dense disordered phases) [129, 130]. Although experiments made it possible to identify the transition in the first place, they can provide relatively little insight into the atomistic structure of the amorphous high-density phase(s). Until this work, there were no atomistic simulations that successfully reproduced the pressure-induced crystallisation of amorphous silicon. Our collaborators carried out GAP-driven [92] simulations of the 100,000-atom amorphous silicon system under isothermal compression. Hydrostatic pressure was applied at a constant rate of 0.1 GPa $ps^{-1}$ while the temperature was held at 500 K, which is high enough to overcome local energy barriers but below the melting line. A description of the behaviour of amorphous silicon when compressed is summarised by Fig. 3.20. Up to 11GPa (c.f. Fig. 3.20a), most atoms remained in fourfold-coordinated environments, similar to the low-density amorphous (LDA) phase. A striking result is the coexistence of LDA and high-density amorphous (HDA)-like regions at the same temperature and pressure; that is, the simulations indicate the presence of polyamorphism over a range of several GPa, rather than an abrupt transition to an almost completely fivefold-coordinated single HDA phase. Upon further compression, beginning at around 12GPa (c.f. Fig. 3.20b), regions with much higher coordination ($\geq 7$) suddenly emerged. These highly coordinated regions rapidly coalesced into a dense form that is distinct from both LDA and HDA (c.f. Fig. 3.20c). We refer to this phase as very-high-density amorphous (VHDA). The rapid structural collapse during VHDA formation reduced the volume from around 18 to around 14$Å^3$ per atom. Importantly, this VHDA phase was transient, and crystalline regions rapidly nucleated (c.f. Fig. 3.20d). The key finding of the present work is not just the formation of sh silicon at high pressure, but the observation of a multistep crystallisation process that proceeds through an entirely distinct VHDA precursor contrasting with the assumptions in

previous works of direct HDA → crystalline transitions [129, 130]. Having reached 20GPa, the system had fully transformed into a polycrystalline ("pc") phase exhibiting hexagonally packed layers, stacked to form an sh structure (c.f. Fig. 3.20e).

Among the experimental indicators for the amorphous–amorphous transition in silicon is a sudden increase in the electrical conductivity [127]. We studied the electronic structure of our 100,000-atom systems by leveraging the ML DOS approach of this Chapter. We developed a regression model for the DOS in disordered silicon, requiring only atomic coordinates as input. The new parameterisation is fitted to hybrid-DFT data for representative structural models of all relevant polyamorphs, including VHDA, as well as the pertinent crystalline phases. In particular, we use SOAP features with radial scaling and sparse Gaussian processes to build an ML model for the total DOS of a given atomistic structure. We represent the DOS as a target of the machine learning models by its cumulative distribution function (CDF). As a reminder, this approach yielded systematically lower prediction errors than models using the DOS curve directly for the silicon data set, because it is sensitive to shifts in peak positions. The SOAP cut-off radius was 6.0 Å; the smoothness parameter was set to 0.5 Å. The radial scaling parameters used for the SOAP representation of Eq. (2.13) are $c = 1$, $r_0 = 3.0$Å and the exponent $m = 5$. We selected 3000 reference atomic environments by farthest point sampling (FPS) for the sparse GPR. The training data consisted of 658 structures supplemented by 100 small amorphous silicon snapshots (64 atoms per cell) at 0GPa from the data set investigated in this Section 3.2, and 30 small dense disordered silicon structural models (64 atoms per cell) that were drawn from the new reference data set used to fit an ML interatomic potential used to validate the findings of this study, over a range of pressures between 11 and 20 GPa. The latter part serves to represent the high-density phases and their electronic DOS properly. Electronic structure calculations to extract the DOS for labelling the input data were performed using the FHI-aims package [115] with the intermediate convergence settings. The HSE06 hybrid functional [19, 20], which is known to usually provide reliable estimates of the band structure of systems with small band gaps [21], was used to determine the self-consistent Kohn–Sham eigenvalues. The latter were then used to compute the reference DOS. The **k**-point spacing was 0.01Å. We also performed uncertainty quantification of the ML DOS model by building a committee of 8 models, each containing a subset of 394 structures randomly selected from the training set. The average prediction of the DOS from the committee of models was taken as the final prediction, and their variance as the uncertainty estimate. The models of the committee are correlated, and so we rescaled the variance around the mean, determining the calibration coefficient with a likelihood-maximisation criterion.

With this model in hand, we are able to make hybrid-DFT-quality predictions for the electronic DOS of large simulation cells within minutes alongside the ML model's uncertainty estimate as

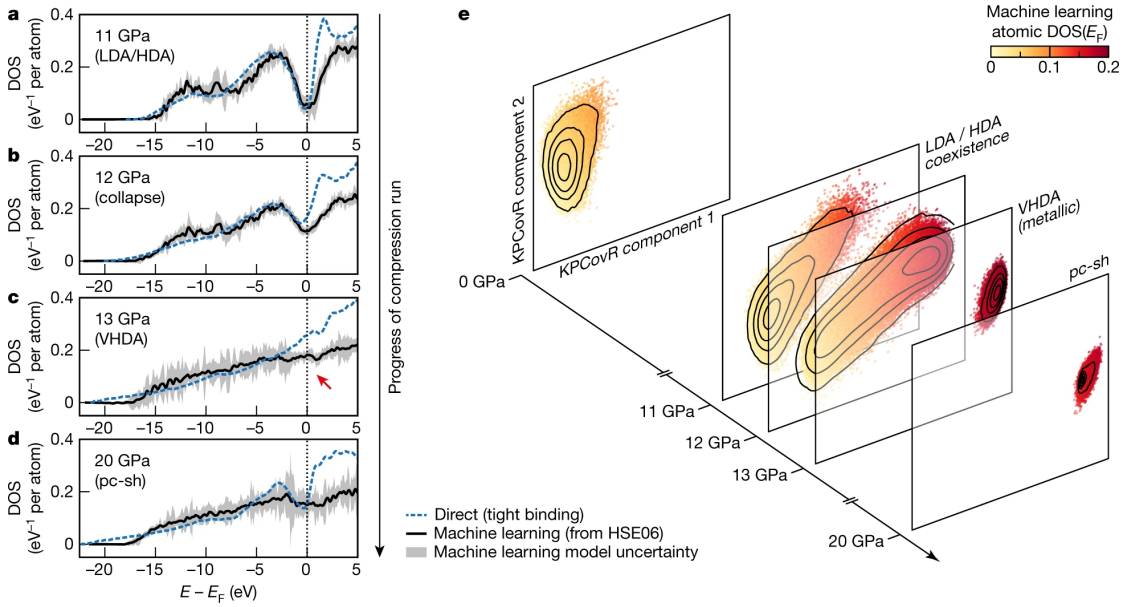


Figure 3.21: Electronic fingerprints of structural transitions. (a-d), Electronic densities of states (DOS) at various stages of the compression run (compare with Fig. 3.20a-e). Black lines indicate the result of a machine learning model for hybrid DFT data using the HSE06 functional; grey shading indicates the associated uncertainty quantification. Blue dashed lines show the result of direct tight-binding computations for the 100,000-atom systems. We note that the tight-binding basis set is minimal (one s and three p valence orbitals per atom), and therefore states above the Fermi level, $\varepsilon_F$, are less well represented because of incompleteness effects. Details about the tight-binding calculations can be found in the methods section of Ref. [38]. A red arrow marks the filling-in of the pseudo gap upon VHDA formation, as discussed in the text. In all plots, $\varepsilon_F$ is set as the energy zero. (e), Evolution of the atomic environments during our compression simulation, visualised using KPCovR. The axes (components) provide the two-dimensional projection of the SOAP kernel features that give the best balance between discriminating the structural diversity of the environments, and linearly predicting the locally averaged machine learning DOS($\varepsilon_F$). The latter quantity, as a fingerprint of electronic structure and metallisation, is used to colour-code the points associated with individual atomic environments. Contour lines indicate the distribution of atomic environments in the KPCovR space and emphasise the structural and electronic transition upon VHDA formation.

shown in Fig. 3.21; direct electronic-structure computation at this high level would have been restricted to system sizes of a few hundred atoms at most. The value of the DOS at the Fermi level, DOS($\varepsilon_F$), is a primary signature of electrical conductivity [131], and its increase during compression (Fig. 3.21a-c) indicates metallisation in the transient VHDA phase, qualitatively consistent with the rapid conductivity increase between 10-12GPa that is observed in diamond anvil cell experiments [127]. At 13GPa-when the VHDA formation was complete in our simulation-the pseudo gap was entirely filled in (marked by an arrow in Fig. 3.21c). The prediction of this distinct electronic feature might be tested by ultrafast spectroscopy techniques, which have been previously applied to the liquid-liquid phase transition in silicon [132] and can access timescales that indeed correspond to those in our simulations. Machine learning models for the DOS, as shown in Fig. 3.21, might have a key role in this regard, by giving access to experimentally relevant system sizes (unlike DFT). Another implication of the onset of metallicity is a possible link to superconductivity, analogous to what has been observed for the metallic high-pressure form of the heavier congener, amorphous germanium [133], and indeed for crystalline sh silicon (with a critical temperature of about 8K at 14.8GPa) [134]. This question, however, requires further experimental study.

Finally, by combining the structural information from the SOAP representation and the machine-learned local electronic fingerprints LDOS, we may construct structure-property maps for atomic environments using kernel principal covariates regression (KPCovR) [39]. In particular, we use the locally averaged DOS (LADOS) as defined in Eq. (3.22) and restricted to the $[-4,4]$eV energy interval to highlight the correlation between the local environments and their corresponding (LA)DOS in the vicinity of the Fermi energy $\varepsilon_F$. These LADOS values are used, together with the same kernel used to regress the DOS, to build a map of the KPCovR map (represented in Fig. 3.21e) that reflects both structural diversity (dissimilarity) and the correlations between structure and the LADOS. This approach yields two-dimensional slices that map out the atomic environments arranged so as to reflect structural diversity and also the relationship between structure and metallicity, for which the LADOS($\varepsilon_F$) is used as a proxy. The two principal components used to draw the maps in Fig. 3.21e were determined by training the KPCovR model on 164,000 environments selected by FPS from 41 structures at pressures ranging from 0 to 20 GPa. All remaining atomic environments were then projected on these two coordinates and used for further analysis. We then arranged the slices in three dimensions to study their evolution through the transitions, with pressure as the third coordinate (Fig. 3.21e). We observed a unimodal distribution of data points in LDA silicon at 0GPa, reflecting the coexistence of locally ordered semiconducting environments and highly defective environments that contribute to the DOS in the electronic band gap. The distribution gradually shifted and broadened towards environments with higher local DOS($\varepsilon_F$)

as polyamorphic HDA regions developed up to 11GPa. The structural collapse at 12GPa led to a new maximum in the map: this indicates a transition between two distinct phases, also seen in Fig. 3.20b. The VHDA phase was localised in a very different region of the map than the LDA/HDA environments, consistent with the marked increase in coordination numbers (Fig. 3.20c) and local DOS($\varepsilon_F$) contributions. For the sh crystallites (at 20GPa), the data points remained in an overall similar region of the map but became more sharply focused compared to VHDA silicon and shifted slightly to a region of lower DOS($\varepsilon_F$), indicative of the formation of a small pseudo gap (also seen in Fig. 3.21d). We expect that such maps, in both two and three dimensions, will become useful tools for studying structural and electronic transitions in diverse phases of matter.

## 3.5 Conclusion

In this Chapter, we presented an ML framework based on sparse Gaussian process regression, a SOAP-based representation of local environments, and an additive decomposition of the electronic density of states to learn and predict the DFT-computed DOS for a diverse data set of silicon structures, covering a broad range of thermodynamic conditions and different phases. We discussed the effect of the Gaussian broadening values usually used to smooth the DOS curves on the prediction process. We also compare 3 different methods to represent the DOS that can be linked to the use of different metrics to assess the error in the predictions: the pointwise discretisation approach, a decomposition on the basis of selected principal components, and the description of the DOS as a derivative of its associated cumulative distribution function. We find that the different representations are fairly compatible, with a slight advantage to the latter independently from the broadening values.

We also investigated the accuracy of derived properties, that can be computed from the predicted DOS, against a direct ML model. In particular, we considered the Fermi energy, the DOS value at the Fermi energy, the band energy and the excitation spectrum. In general, we find that the indirect models lead to small but consistent improvements over direct predictions. This improvement is remarkable because a direct model has the possibility to focus on the structure-property relations that are more relevant to the target. The fact that going through the DOS improves predictions indicates that the density of states is more amenable to an additive, local decomposition with respect to properties like $\varepsilon_F$ that depend on the global imposition of charge neutrality.

We demonstrated an application of our ML model to the prediction of the DOS of some amorphous silicon configurations, including one containing 4096 atoms for which a brute-force DFT calculation would be prohibitively expensive. We observe excellent accuracy in

the ML predictions and that the averaging over multiple configurations – which is necessary to obtain predictions consistent with experimental observations – considerably reduces the discrepancy between the ML model and the DFT reference. A data-driven analysis of the local density of states reveals the interplay between structural motifs and electronic structure in amorphous silicon, but a physical interpretation of the local DOS contributions computed by an ML model should not disregard the role played by choice of structural features and regression scheme, which affects the atom-centred predictions in the absence of explicit reference values for the LDOS.

We showed that the ML model for the DOS was a powerful tool for studying electronic transitions in disordered phases by leveraging its generalisation capabilities. We trained a model on a higher theory level of DFT and on small snapshots and applied it on the ultra-scale necessary to describe the transition mechanisms when compressing amorphous silicon. This model provided consistent observations with experimental observables while giving insights into the structural/electronic fingerprints interplay.

Our ML framework makes it possible to estimate, based exclusively on atomic configurations, one of the most essential descriptors of the electronic structure. Combining it with one of the well-established potential energy models makes it possible to compute the electronic contributions to macroscopic properties, such as the heat capacity of metals, to perform simulations that take into account finite-electronic-temperature effects [135], and provides another brick in the construction of a full surrogate ML model of the properties of molecules and materials. The possibility of computing atomic charges by enforcing global charge neutrality and then using local DOS to determine charge partitioning provides an interesting line of investigation to realise a “grand-canonical ML” framework that combines a local model with a physics-based charge equilibration scheme.

# 4 Thermal excitations[1]

In this chapter, we explore two methods to leverage the electronic density of states to account for the electronic contribution to physical observables. First, we use the Born-Oppenheimer approximation to decouple the electronic and ionic degrees of freedom and compute the electronic contribution as an *a posteriori* correction to the observable derived from the dynamics of the ions. Second, we derive a framework where the thermal excitations of the electrons are taken into consideration when constructing the finite-temperature potential energy surface and its gradients. In both cases, we utilise the atom-centred machine-learning (ML) model for the electronic density of states from Section 3.1, trained on ground-state data. These results show the power of physics-inspired modelling, where combining different ML models for structural and electronic properties of materials leads to more accurate modelling of the properties of materials across a wide range of conditions.

## 4.1 Electronic thermal excitations as an a posteriori correction

As a first step to address the question of modelling condensed matter at high temperatures, we propose to study a nickel over a wide range of temperatures up to its melting temperature, as reported in Ref. [136]. This work is a collaboration, where the author's contribution consisted in building the ML model for the density of states and using the model to compute thermodynamic properties. The main objective of this study is to build an accurate machine-learning interatomic potential to study the structural, mechanical and thermodynamic properties of nickel. This is achieved by using sampling techniques at finite temperatures to compute bulk and interfacial properties of nickel from cryogenic temperatures up to above the melting point. The ion-based predictions are complemented by taking into account the effect of the electronic excitations without performing extra electronic structure calculations with respect

[1] This chapter is an adaptation of my contributions to Refs. [136, 137]

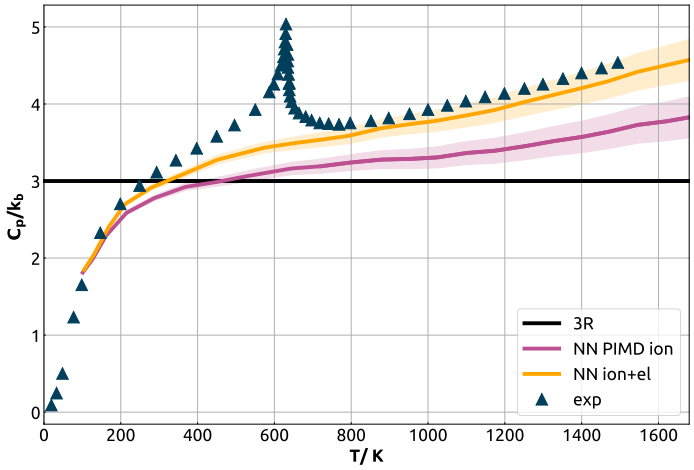


Figure 4.1: Constant pressure heat capacity $C_p$ as a function of temperature. Triangles indicate experimental observations. The solid magenta line represents the heat capacity computed in this work with path integral molecular dynamics. The dark yellow solid line represents the heat capacity computed, including electronic corrections based on an ML model of the DOS. The solid black line describes the classical prediction for heat capacity.

to the training phase, as described below.

We propose to incorporate the electronic degrees of freedom, i.e. the thermal excitation effects, as a posteriori correction to the thermodynamic averages, heat capacity in this case, obtained from the dynamics of ions. In particular, we can use the DOS model of Section 3.1 to compute the electronic contributions to several thermodynamic properties, such as the Helmholtz free energy at finite temperature

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

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

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

and an entropy term

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

and the electronic contribution to the high-temperature heat capacity

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

These expressions are written in a "non-self-consistent" approximation, where we consider the density of states to be fixed to that computed from the Kohn-Sham eigenvalues obtained self-consistently at $T^{\mathrm{el}} = 0$, and the temperature dependence is due to the occupation of the energy levels, which is given by a Fermi-Dirac distribution $f_{\mathrm{FD}}(\varepsilon - \varepsilon_F, T)$, and by the Fermi energy $\varepsilon_F(T)$ which can be computed for the DOS at each temperature $T$ by enforcing charge neutrality as described in Eq. (3.1). Given a molecular dynamics trajectory, one could predict the electronic DOS for every frame, then use them to estimate the electronic energy $U^{el}$. The electronic contribution to heat capacity $C_p$ can be obtained by finite differences, for example.

In this work, the thermodynamic averages of the ions are deduced from molecular dynamics simulations driven by an interatomic potential, in this case, a neural network potential (NNP) using the Behler-Parrinello symmetry functions [43] and trained on several classes of structures covering bulk configurations, point defects and interfaces. The NNP was then validated on static lattice properties by computing the stability of the *fcc*, *hcp* and *bcc* phases of bulk nickel. It showed good agreement with the DFT results in calculating mechanical properties like elastic constants, the bulk modulus and the formation energy of point defects, which requires performing simulations on large systems. Through finite temperature simulations, we computed several finite temperature properties from thermodynamic averages. In low-temperature simulations, it is important to account for quantum nuclear effects. While in the high-temperature regime, it is necessary to account for magnetic and electronic excitations, making a direct comparison to DFT results difficult, and this explains the choice of comparing the NNP results to existing classical force fields [138] and experiments when possible. Despite its success in describing several properties of nickel, it fails to describe the heat capacity of nickel near the melting temperature, as shown in Fig. 4.1, despite providing satisfactory results and in agreement with experiments, below the Curie temperature. This approach underestimates the heat capacity by ≈20% compared to the experimental observations at high temperatures. The discrepancy (which is also observed in explicit first-principles molecular dynamics [139] and in simulations that use the reference classical force field) is due to electronic contributions.

To train a model of the DOS, we use a subset containing 1069 structures of the data set used to train the NNP, while discarding those corresponding to solid-liquid, liquid-vacuum and solid-vacuum interfaces, and we complement it with 123 independent structures extracted from liquid and solid trajectories at the melting temperature. We use the radial cutoff $r_0 = 6.$Å

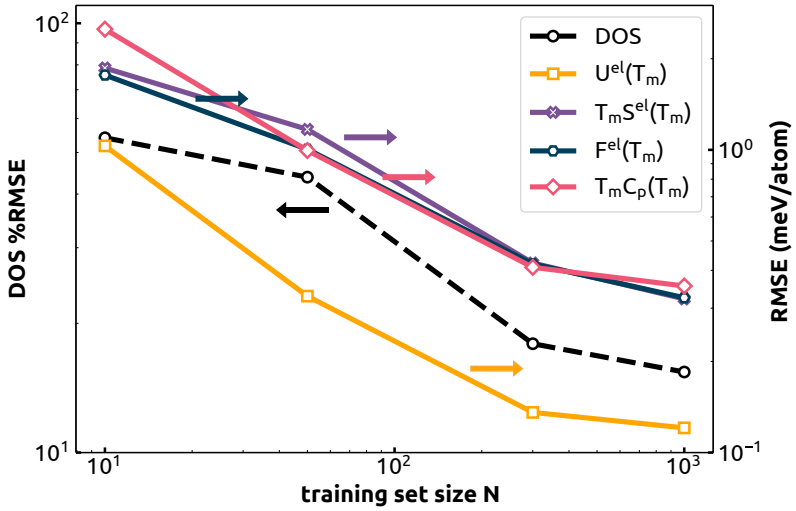


Figure 4.2: Evolution of the prediction errors in the validation set as a function of the training set size for the pointwise representation of the ML DOS (in black), as well as for quantities derived from the DOS prediction for thermal excitations computed at $T_m$ = 1700K (namely, the band energy $U^{\mathrm{el}}(T_m)$, the electronic entropy term $T_m S^{\mathrm{el}}(T_m)$, the free energy $F^{\mathrm{el}}(T_m)$, and the heat capacity $T_m C_p(T_m)$ written in energy units).The reference DOS is generated with a Gaussian broadening of 0.1eV. The arrows point to the axis on which the errors can be read.

and an atomic density smoothing 0.45 for the SOAP features. The active set contains 15000 environments selected by FPS out of the ≈ 127000 that are present in the training set. We determine the regression weights $\mathbf{x}_M$ using a regularisation parameter that is optimised by a 10-fold cross-validation scheme in order to ensure the model is not in the over-fitting regime.

The learning curves, computed by reporting errors on a fixed test set of the predictions of models trained on an increasing fraction of the remaining 1000 structures, are shown in Fig. 4.2. The figure shows both the error on the DOS, computed as the integrated root mean square error (RMSE) of the ML DOS and DFT DOS normalised by the integrated standard deviation of the reference DFT DOS, as well as errors for the quantities in Eqs. (4.1-4.4), computed on the predicted DOS and checked against those obtained from the reference DFT curve. Learning curves are not saturating, indicating that a more accurate model could be obtained, if needed, by increasing further the train set size. In practice, this model is sufficiently accurate: even though the normalised error on the DOS is large (%RMSE=14.71% for the largest train set size), this translates into sub-meV errors for the key properties at the melting temperature $T_m$ = 1700K. For the band energy $U^{\mathrm{el}}(T_m)$: %RMSE=3.30% and RMSE=0.12meV/atom; for the entropy $T_m S^{\mathrm{el}}(T_m)$: %RMSE=5.81% and RMSE=0.32meV/atom; for the free energy $F^{\mathrm{el}}(T_m)$: %RMSE=9.04% and RMSE=0.32meV/atom and for the heat capacity $T_m C_p^{\mathrm{el}}(T_m)$: %RMSE=4.25% and RMSE=0.36 meV/atom.

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

Table 4.1: Average band energy, entropy contribution and free energy of solid and liquid phases at the melting temperature of Nickel $T_m = 1700K$, together with their difference. The values are computed from the ML DOS estimated for $\approx$ 15000 snapshots extracted from an NNP simulation of the liquid and solid phase at $T_m$. The uncertainties are derived by separately computing each quantity using a separate prediction of the calibrated DOS model, and computing the standard deviation of the end results.

The dark yellow line of Fig. 4.1 confirms our hypothesis that the discrepancy between the computational and experimental values of $C_p$ is due to the "hot" electrons. This cheap ML model is also in agreement with much more elaborated and accurate simulations using density functional theory and quasi-harmonic simulations [139]. We should acknowledge that, despite these remarkable achievements, our integrated modelling approach has poor results in the region around the Curie temperature, where magnetic excitations become important. Even though we do not include them in this model, adding a description of magnetism constitutes an interesting direction for future studies. This could be achieved by incorporating the magnetic moments into the structural fingerprints or building ML models for up-spin and down-spin atoms. As for the DOS, one could also build two independent models for the spin-up and spin-down electronic systems obtained from spin-polarised calculations.

We can also use the ML model of the DOS to compute the contributions to the free energy associated with electronic excitations, Eq. (4.1), averaged over trajectories of the bulk solid and liquid phases at temperatures around the melting temperature $T_m$. The melting temperature can be calculated from the dynamics of the ions by using the interface pinning technique [140], which works by applying a harmonic bias potential to a two-phase system, which couples to an order-parameter $\Phi$ that discriminates between the two phases of interest. The Gibbs free energy difference between the phases is determined by the average force that the pinning potential exerts on the system and allows us to extract a difference in the chemical potentials of the phases $\Delta\mu_{\text{sl}}$. By performing multiple simulations at different temperatures, one can identify the dependence of $\Delta\mu_{\text{sl}}$ on $T$. The temperature at which $\Delta\mu_{\text{sl}} = 0$ identifies the melting point $T_m$, and the slope is equal to the entropy of melting.

Applying the adiabatic approximation, one could argue that the difference $\Delta F^{\text{el}}(T) = F_{\text{l}}^{\text{el}}(T) - F_{\text{s}}^{\text{el}}(T)$ between the electronic free energies of the liquid phase $F_{\text{l}}^{\text{el}}(T)$ and the solid phase $F_{\text{s}}^{\text{el}}(T)$ could contribute to shifting the chemical potential, and hence leading to a change in the predicted $T_m$. As shown in Table 4.1, even though the electronic excitations give a

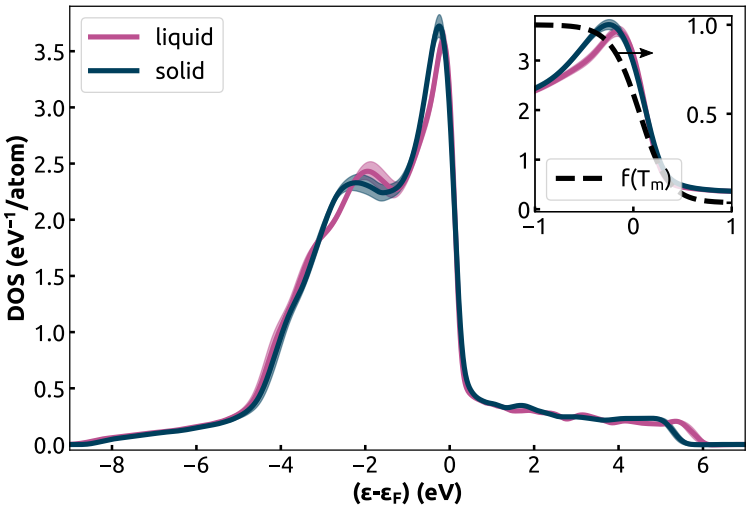


Figure 4.3: Average predicted DOS curve for the solid and liquid trajectories at the melting temperature $T_m = 1700K$. The shaded area represents the standard deviation of DOS($E$) over the considered trajectories, and the inset shows a close-up of the region around the Fermi energy. The dashed curve represents the Fermi-Dirac function $f_{\mathrm{FD}}(\epsilon - \epsilon_F, T_m)$

very substantial contribution to the free energy of Ni around $T_m$, the contributions from the solid and the molten phases cancel out almost perfectly, so that the impact on the melting temperature is less than 10K – in agreement with the observations made in Ref. [141]. It should also be noted that converging these quantities to the level required to resolve the small difference between solid and liquid phases is far from trivial – both in terms of the statistical error over an MD trajectory and in terms of the ML error computed following Ref. [112], by first generating a committee of predictions for the DOS, and then using each curve to obtain a separate estimation of $\Delta F^{\mathrm{el}}(T)$.

The averaged DOS over the solid and liquid phases at $T_m$, shown in Fig. 4.3, demonstrate that the cancellation between $F_{\mathrm{l}}^{\mathrm{el}}(T)$ and $F_{\mathrm{s}}^{\mathrm{el}}(T)$ is to be expected, given the minor differences, particularly in the vicinity of the Fermi level. More significant effects could appear in systems that, upon melting, undergo a substantial change in electronic properties, e.g. from semiconducting to metallic like in Section 3.4.

Here, we have discussed our attempt to incorporate electronic excitations directly into ML simulations. It consisted of predicting the single-particle density of states and using it to evaluate a-posteriori corrections to the thermodynamic quantities, e.g. heat capacity or melting temperature, extracted from the MD of ions whose machine-learning interatomic potential is trained on ground-state data. However, we show in the next sections that this approach is limited to condensed matter well below the Fermi temperature, where atomic

forces are almost unaffected by the electronic excitations.

## 4.2 Approximating the finite temperature free energy

Having demonstrated the impact of using an ML model for the density of states to compute better estimates for thermal properties of nickel, we move to discuss the challenges of building a "universal" model that incorporates the thermal excitations of electrons in the machine-learning interatomic potential (MLIP). Current ML strategies are usually designed to reproduce the ground state Born-Oppenheimer (BO) potential energy surface, and do not account for the temperature-dependent electronic excitations which play a significant role in metallic matter at planetary conditions, like warm dense matter (WDM) [142, 143, 144, 145, 146]. These fluctuations introduce subtle but important corrections in the thermophysical properties of ordinary metals [147, 148]. The most common strategy to treat finite electron temperature is to replace the BO potential with a temperature-dependent electronic free energy $A(T^{\mathrm{el}})$. In traditional ML interatomic potentials (MLIPs) frameworks that rely exclusively on nuclear coordinates as inputs, switching from the BO potential to $A(T^{\mathrm{el}})$ would require training a separate model for every target electronic temperature $T^{\mathrm{el}}$, recomputing also the training set – although the temperature can be included as an input of the model, which yields MLIPs that are explicitly temperature-dependent, and interpolate between training data at different electron temperature [149].

Our objective is to construct a framework that allows us to construct temperature-dependent MLIPs, beyond the ground-state BO approximation. In this section, we prove that within a density functional theory framework, it is possible to rigorously approximate the total free energy, atomic forces and the stress tensor of an atomic system, the three ingredients needed to reconstruct the finite temperature free energy surface, as the sum of a $T^{\mathrm{el}} = 0\mathrm{K}$ contribution and a finite-$T^{\mathrm{el}}$ correction depending exclusively on the *ground-state* electronic density of states (DOS). This general result underpins a framework that relies only on ground-state calculations to learn $A(T^{\mathrm{el}})$ and its derivatives in the presence of thermally-excited electrons. Thus, a consistent ground-state training set and model can be generated and used to sample the finite-electron-temperature distributions, using $T^{\mathrm{el}}$ as an external parameter.

### Framework derivation

Let us start by considering the standard representation of the DFT energy:

$$E = E_{\mathrm{band}} - E_{\mathrm{dc}} + E_{\mathrm{ion}} \tag{4.5}$$

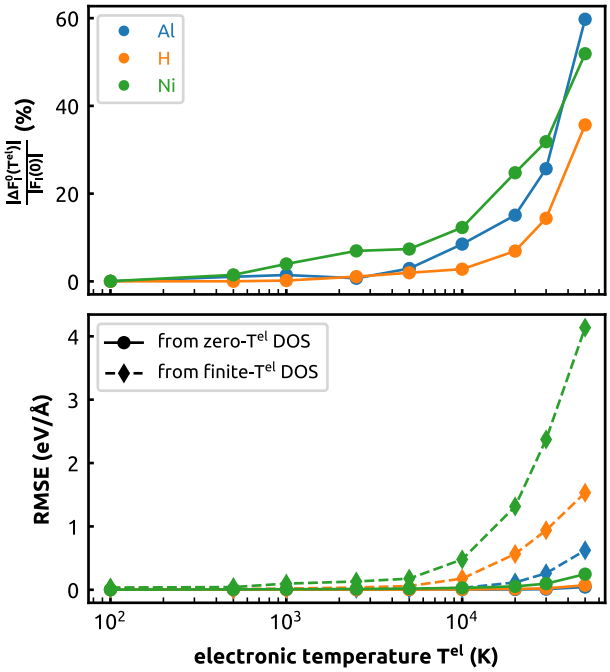


Figure 4.4: (Upper panel) Relative deviation of Hellmann-Feynman atomic force versus the electronic temperature with respect to the ground state force for a given ion and a Cartesian direction. (Lower panel) Root mean square errors (RMSE) of 10 force components computed with Eq. (4.32) compared to their Hellmann-Feynman counterparts. Solid lines: using the DOS from $T^{\mathrm{el}} = 0$K calculations. Dashed lines: using the DOS from finite-$T^{\mathrm{el}}$ calculations. Blue: aluminium; orange: hydrogen; green: nickel.

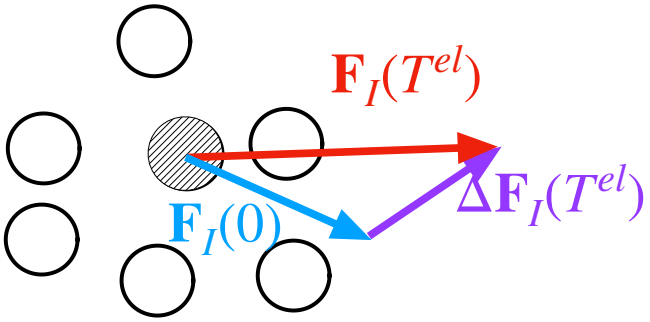


Figure 4.5: The sketch represents the decomposition of the finite-$T^{\mathrm{el}}$ atomic force component within our framework

as a sum of the electrostatic interactions between the ions $E_{\text{ion}}$, the band energy $E_{\text{band}} = \sum_i f_i \epsilon_i$, expressed in terms of the Kohn-Sham (KS) eigenvalues $\epsilon_i$ and level occupations $f_i$, and the "double-counting term"

$$E_{\text{dc}} = \frac{1}{2} \iint \frac{\rho(\mathbf{r}')\rho(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r} d\mathbf{r}' - E_{\text{xc}}[\rho] + \int V_{\text{xc}}[\rho](\mathbf{r})\rho(\mathbf{r}) d\mathbf{r}. \tag{4.6}$$

Here $E_{\text{xc}}$ is the exchange-correlation (XC) functional, $V_{\text{xc}}[\rho](\mathbf{r}) = \delta E_{\text{xc}}/\delta\rho(\mathbf{r})$ is the XC potential, and $\rho(\mathbf{r}) = \sum_i f_i |\phi_i(\mathbf{r})|^2$ is the DFT density, expressed in terms of the KS eigenfunctions $\phi_i(\mathbf{r})$ and occupations $f_i$.

Whenever an electronic temperature $T^{\text{el}}$ is introduced, the $f_i$ become fractional, and the correct energy functional becomes the Helmholtz free energy [150, 135, 111]

$$A(T^{\text{el}}) = E(0) + \Delta E(T^{\text{el}}) - T^{\text{el}} S(T^{\text{el}}), \tag{4.7}$$

where $E(0)$ is the ground-state energy, $\Delta E(T^{\text{el}})$ is the finite-$T^{\text{el}}$ contribution to the energy, and $S(T^{\text{el}})$ is the KS electronic entropy. From Eq. (4.7) one can obtain the finite-$T^{\text{el}}$ Hellmann-Feynman forces [151], whose relative deviation with respect to $T^{\text{el}}$ = 0K-forces becomes significant at large $T^{\text{el}}$, as reported in the upper panel of Fig. 4.4. In principle, a $T^{\text{el}}$-dependent XC functional should be employed [152]. However, it is often possible to rely on the Zero Temperature Approximation (ZTA), where the XC functional depends on $T^{\text{el}}$ only through the $T^{\text{el}}$-dependence of the density: $E_{\text{xc}}[\rho(T^{\text{el}})]$. The ZTA performs well at both low and high $T^{\text{el}}$ and also satisfies exact conditions as discussed in Ref. [153], and we adopt it as the basis of our framework.

A change in the occupation of the levels, e.g. as a consequence of thermal excitations, determines a change in the density, and thus, self-consistently, in the KS eigenenergies and eigenfunctions. For instance, the functional derivative of $E_{\text{band}}$ with respect to $f_i$ is

$$\frac{\delta E_{\text{band}}}{\delta f_i} = \epsilon_i + \sum_j f_j \frac{\delta \epsilon_j}{\delta f_i} \tag{4.8}$$

Nonetheless, it can be proved, following reasoning similar to that used in Ref. [154], and in Ref. [155] for the energy variation due to infinitesimal atomic displacements, that the second term in Eq. (4.8) *cancels exactly* with the variation of the double-counting term, $\delta E_{\text{dc}}/\delta f_i$. Therefore, the change in $E$ due to a finite change in the occupations can be approximated by

$$\Delta E \approx \Delta E^0_{\text{band}} \equiv \sum_i \epsilon^0_i \Delta f_i, \tag{4.9}$$

where $\epsilon_i^0 \equiv \epsilon_i(\{\Delta f_k = 0\})$ are the unperturbed Kohn-Sham eigenenergies computed at vanishing variation on all the $f_k$. The "0" superscript labels quantities obtained from unperturbed eigenenergies computed a reference electronic temperature, which we consider for the remainder of this discussion to be $T^{\mathrm{el}} = 0\mathrm{K}$.

### Mathematical proof of Eq. (4.9)

In order to prove the equality of Eq. (4.9), we start with the variation of the band energy due to a change in the occupation of the $j$-th level:

$$\frac{\delta E_{\mathrm{band}}}{\delta f_j} = \epsilon_j + \sum_i f_i \frac{\delta \epsilon_i}{\delta f_j} \tag{4.10}$$

In first-order perturbation theory, we have

$$\left.\frac{\partial \rho(\mathbf{r})}{\partial f_j}\right|_{\Delta \mathbf{f}=0} = |\phi_j^0(\mathbf{r})|^2 =: \rho_j^0(\mathbf{r}), \tag{4.11}$$

with no change in the eigenfunctions due to further self-consistent cycles. The eigenenergy of the $i$-th KS state is

$$\epsilon_i = \langle \phi_i | \hat{T} + \hat{V}_{\mathrm{ion}} + \hat{V}_{\mathrm{scf}} | \phi_i \rangle, \tag{4.12}$$

where $\hat{V}_{\mathrm{ion}}$ is the ion potential and

$$V_{\mathrm{scf}}(\mathbf{r}) = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \mathrm{d}\mathbf{r}' + V_{\mathrm{xc}}[\rho](\mathbf{r}) \tag{4.13}$$

Therefore, the variation of $\epsilon_i$ under a change in the occupation of the $j$-th KS state is

$$\frac{\delta \epsilon_i}{\delta f_j} = \langle \phi_i | \frac{\delta \hat{V}_{\mathrm{scf}}}{\delta f_j} | \phi_i \rangle = \iint \rho_i^0(\mathbf{r}) \left( \frac{1}{|\mathbf{r}-\mathbf{r}'|} + \frac{\delta V_{\mathrm{xc}}[\rho]}{\delta \rho(\mathbf{r}')} \right) \rho_j^0(\mathbf{r}') \mathrm{d}\mathbf{r}' \mathrm{d}\mathbf{r} \tag{4.14}$$

since both $\frac{\delta \hat{V}_{\mathrm{ion}}}{\delta f_j}$ and $\frac{\delta \hat{T}}{\delta f_j}$ vanish [2], and

$$\frac{\delta V_{\mathrm{scf}}(\mathbf{r})}{\delta f_j} = \int \frac{\rho_j^0(\mathbf{r}') \mathrm{d}\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|} + \int \frac{\delta V_{\mathrm{xc}}[\rho]}{\delta \rho(\mathbf{r}')} \rho_j^0(\mathbf{r}') \mathrm{d}\mathbf{r}'. \tag{4.15}$$

[2] The first step in Eq. (4.14) is justified by the Hellmann-Feynman theorem even without the approximation of Eq. (4.11). The ZTA enters Eq. (4.14), where we neglect the *explicit* dependency of $V_{\mathrm{xc}}$ upon the electronic temperature.

In conclusion, the variation of the band energy due to a change in $f_j$ is

$$\frac{\delta E_{\text{band}}}{\delta f_j} = \epsilon_j + \sum_i f_i \left[ \iint \frac{\rho_i^0(\mathbf{r}')\rho_j^0(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' + \iint \frac{\delta V_{\text{xc}}}{\delta \rho(\mathbf{r}')} \rho_i^0(\mathbf{r})\rho_j^0(\mathbf{r}')\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \right]. \tag{4.16}$$

For the double-counting term, we have

$$-\frac{\delta E_{\text{dc}}}{\delta f_j} = -\frac{\delta E_{\text{H}}}{\delta f_j} + \int \underbrace{\frac{\delta E_{\text{xc}}}{\delta \rho(\mathbf{r})}}_{V_{\text{xc}}[\rho](\mathbf{r})} \underbrace{\frac{\partial \rho(\mathbf{r})}{\partial f_j}}_{\rho_j^0(\mathbf{r})} \mathrm{d}\mathbf{r} - \frac{\delta}{\delta f_j} \int V_{\text{xc}}[\rho](\mathbf{r}) \sum_i f_i \rho_i^0(\mathbf{r}) \mathrm{d}\mathbf{r} \tag{4.17}$$

where

$$E_{\text{H}} = \frac{1}{2} \iint \frac{\rho(\mathbf{r}')\rho(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \tag{4.18}$$

is the Hartree energy. Therefore,

$$-\frac{\delta E_{\text{H}}}{\delta f_j} = -\frac{\partial}{\partial f_j} \frac{1}{2} \sum_i \sum_l f_i f_l \iint \frac{\rho_i^0(\mathbf{r}')\rho_l^0(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' = -\sum_i f_i \iint \frac{\rho_i^0(\mathbf{r}')\rho_j^0(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}', \tag{4.19}$$

while

$$-\frac{\partial}{\partial f_j} \int V_{\text{xc}}[\rho](\mathbf{r}) \sum_i f_i \rho_i^0(\mathbf{r}) \mathrm{d}\mathbf{r} = -\int V_{\text{xc}}[\rho](\mathbf{r}) \rho_j^0(\mathbf{r}) \mathrm{d}\mathbf{r} - \iint \frac{\delta V_{\text{xc}}[\rho]}{\delta \rho(\mathbf{r}')} \underbrace{\frac{\partial \rho(\mathbf{r}')}{\partial f_j}}_{\rho_j^0(\mathbf{r}')} \sum_i f_i \rho_i^0(\mathbf{r}) \mathrm{d}\mathbf{r}' \mathrm{d}\mathbf{r}. \tag{4.20}$$

Therefore, the first-order derivative of the double-counting term is

$$-\frac{\delta E_{\text{dc}}}{\delta f_j} = -\sum_i f_i \left[ \iint \frac{\rho_i^0(\mathbf{r}')\rho_j^0(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' + \iint \frac{\delta V_{\text{xc}}[\rho]}{\delta \rho(\mathbf{r}')} \rho_i^0(\mathbf{r})\rho_j^0(\mathbf{r}')\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \right], \tag{4.21}$$

which exactly cancels $\sum_i f_i \frac{\delta \epsilon_i}{\delta f_j}$ when Eq. (4.14) is used, thus proving Eq. (4.9). This derivation is based on variations of $f_i$ with respect to given reference values, which need not necessarily be those at $T = 0$, which provides a way to extend further the range of applicability of the approximation at very high $T^{\text{el}}$ by computing data at a few self-consistent reference temperatures.

### Specific case of Fermi-Dirac occupation

We now focus on the specific case where the set of $f_i$ are Fermi-Dirac distributed, $f_i = f_{\text{FD}}(\epsilon_i - \mu(T^{\text{el}}), T^{\text{el}})$, $\mu(T^{\text{el}})$ being the chemical potential of the electron system at a temperature $T^{\text{el}}$

and $k_B$ the Boltzmann constant. Here, we define the Fermi-Dirac function as:

$$f_{\mathrm{FD}}(x,T) = \frac{1}{1+e^{\frac{x}{k_B T}}}$$

From Eq. (4.9), the finite-$T^{\mathrm{el}}$ correction to the DFT energy is

$$\Delta E^0_{\mathrm{band}}(T^{\mathrm{el}}) = \int_{-\infty}^{+\infty} \epsilon \mathrm{DOS}^0(\epsilon) \left[ f_{\mathrm{FD}}\left(\epsilon - \mu(T^{\mathrm{el}}), T^{\mathrm{el}}\right) - f_{\mathrm{FD}}\left(\epsilon - \mu(0), T = 0^+\right)\right] \mathrm{d}\epsilon, \tag{4.22}$$

where $\mathrm{DOS}^0(\epsilon) = \sum_i \delta(\epsilon - \epsilon^0_i)$ is the electronic DOS. In order to finish building the approximation of Eq. (4.7), one needs to find the expressions of the chemical potential of the electron system $\mu(T^{\mathrm{el}})$ and the entropy contribution $S(T)$. In particular, should we use the zero-$T^{\mathrm{el}}$ or the finite-$T^{\mathrm{el}}$ values to compute such approximations?
Notice that since the energy eigenstates do not enter directly (but only through the occupations), the order is higher than in the case of the internal energy. What we aim at is a good approximation of a possible correction

$$-T\delta^T S \equiv -T[S^T(T) - S^0(T)] \tag{4.23}$$

where

$$S^\tau(T) \equiv -k_B \sum_i f^\tau_i(T) \ln[f^\tau_i(T)] + [1 - f^\tau_i(T)] \ln[1 - f^\tau_i(T)] \tag{4.24}$$

and

$$f^\tau_i(T) \equiv f_{\mathrm{FD}}\left(\epsilon^\tau_i - \mu^\tau(T), T\right) \tag{4.25}$$

Here $\epsilon^\tau_i$ is the $i$-th eigenenergy computed at temperature $\tau$, while $\mu^\tau(T)$ is the chemical potential obtained from the normalisation relation when the employed states are $\epsilon^\tau_i$ but the temperature used to populate the states is $T$. The symbol $\delta^\tau$ denotes a change, at fixed population temperature $T$, due to a variation $\tau$ of the temperature used in the calculation of the eigenstates:

$$\delta^\tau f_i(T) \equiv f_{\mathrm{FD}}\left(\epsilon^\tau_i - \mu^\tau(T), T\right) - f_{\mathrm{FD}}\left(\epsilon^0_i - \mu^0(T), T\right). \tag{4.26}$$

At the first order in $\delta^\tau f(T)$ we obtain

$$\begin{aligned} -T\delta^\tau S &\approx -T \sum_i \left.\frac{\partial S}{\partial f_i}\right|_{f^0_i(T)} \delta^\tau f_i(T) \\ &= -k_B T \sum_i \ln\left(\frac{1 - f^0_i(T)}{f^0_i(T)}\right) \delta^\tau f_i(T) \\ &= -\sum_i (\epsilon^0_i - \mu^0(T)) \delta^\tau f_i(T) \end{aligned} \tag{4.27}$$

Taking $\tau = T$, the term $-\sum_i \epsilon_i^0 [f_i^T(T) - f_i^0(T)]$ *cancels* with an analogous term arising in $\Delta E$, as it is evident when we add and subtract $f_i^0(T)$ in $\Delta f_i = f_i^T(T) - f_i^0(0)$ used in Eq. (4.9), while $\sum_i \mu^0(T)\delta^\tau f_i(T) = 0$ because the total number of electrons is fixed.

This derivation proves that it is sufficient to use the zero-$T^{\rm el}$ levels, i.e. the zero-$T^{\rm el}$ DOS, and hence justify the use of $\mathrm{DOS}^0$ in the charge conservation relation to determine $\mu(T^{\rm el})$:

$$N = \int_{-\infty}^{+\infty} \mathrm{DOS}^0(\epsilon)\, f_{\rm FD}\left(\epsilon - \mu(T^{\rm el}), T^{\rm el}\right) \mathrm{d}\epsilon, \tag{4.28}$$

and the electronic entropy $S(T^{\rm el})$:

$$S(T^{\rm el}) \approx S^0(T^{\rm el}) \equiv \int_{-\infty}^{+\infty} \mathrm{DOS}^0(\epsilon)\, s\left(\epsilon - \mu(T^{\rm el}), T^{\rm el}\right) \mathrm{d}\epsilon, \tag{4.29}$$

where $N$ is the number of valence electrons and $s(x,T) = f_{\rm FD}\ln f_{\rm FD} + (1-f_{\rm FD})\ln(1-f_{\rm FD})$. Therefore, our approximation for the free energy yields a Mermin-like functional:

$$A(T^{\rm el}) \approx E(0) + \Delta E^0_{\rm band}(T^{\rm el}) - T^{\rm el} S^0(T^{\rm el}). \tag{4.30}$$

Our derivation directly translates to the calculation of derivatives of the free energy, i.e. the atomic forces and the stress tensor. For instance, according to the Born-Oppenheimer approximation, the force acting on the $I-$th nucleus in the DFT ensemble is

$$\mathbf{F}_I(T^{\rm el}) = -\nabla_I A(T^{\rm el}) \approx \mathbf{F}_I(0) + \Delta\mathbf{F}^0_I(T^{\rm el}), \tag{4.31}$$

where

$$\begin{aligned} \mathbf{F}_I(0) &\equiv -\nabla_I E(0) \\ \Delta\mathbf{F}^0_I(T^{\rm el}) &\equiv -\nabla_I [\Delta E^0_{\rm band}(T^{\rm el}) - T^{\rm el} S^0(T^{\rm el})]. \end{aligned} \tag{4.32}$$

In this decomposition, the electronic temperature $T^{\rm el}$ enters as an *external parameter*. Fig. 4.5 is a schematic representation of this framework, using the atomic force as a representative.

These equations would be of limited practical value if the end goal were to compute $A(T^{\rm el})$ for a given structure and temperature by means of a self-consistent electronic structure calculation. However, they become very useful in the context of data-driven modelling, as they provide a rigorous basis for the development of a ML framework to learn finite-$T^{\rm el}$ interatomic forcefields without the need to train on finite-$T^{\rm el}$ calculations. The $T^{\rm el} = 0$K quantities, $E(0)$ and $\mathbf{F}_I(0)$, can be modelled by any of the widely used MLIPs [92, 43, 83, 46, 156]. The hot-electron correction, Eq. (4.32), can be accessed by training an ML model for the

DOS.

Our derivation justifies other approximations made in the literature, such as the fixed-DOS approximation of Refs. [157], which assumes that the electronic DOS is approximately independent of $T^{\mathrm{el}}$. In fact, the cancellations ensure the validity of Eq. (4.9), even if the self-consistent energy levels (and thus the DOS itself) changed substantially by changing $T^{\mathrm{el}}$.

### What about using the finite-$T^{\mathrm{el}}$ DOS?

If one wanted to go beyond this ground-state approximation, it would not be sufficient to obtain the finite-$T^{\mathrm{el}}$ DOS and to use it in expressions similar to Eqs. (4.22), (4.28) and (4.29). Without access to the self-consistent finite-temperature $E_{\mathrm{dc}}$, doing so would lead to *worse* results, as shown in the lower panel in Fig. 4.4. Also, any mixture of the finite-$T^{\mathrm{el}}$ and zero-$T^{\mathrm{el}}$ DOS in the terms of Eq. (4.30) would lead to worse predictions of the free energy and the atomic forces as well as shown in Fig. 4.6 in an aluminium supercell, a liquid hydrogen structure and a liquid nickel structure. Figure 4.7 shows, in particular, the use of the DOS computed at $T^{\mathrm{el}} = 50,000\mathrm{K}$ and trying to recover the free energy and force of lower temperature calculations compared to the finite temperature DFT. The relative errors are larger at low temperatures, which indicates that the approximation of this work is better used when the reference calculation is done at low temperatures. This behaviour occurs despite a minimal change in the DFT DOS computed at $T^{\mathrm{el}} = 0\mathrm{K}$ and $T^{\mathrm{el}} = 50,000\mathrm{K}$. Fig. 4.8 shows the DOS of a liquid hydrogen structure computed at the mentioned temperatures.

This observation opens the door to a general approach. Suppose one was prepared to perform self-consistent calculations at multiple temperatures. In that case, our perturbative expressions could also be applied to a reference temperature different from $T^{\mathrm{el}} = 0$ and serve as the basis of more accurate temperature-interpolation schemes. We consider five different cases for a single force component of a liquid hydrogen snapshot, which is the same one used to illustrate the approximation developed in this work in Figs. 4.6 and 4.7:

- the force extrapolation from this work's approximation, with reference calculation done at the ground-state $T_0 = 0\mathrm{K}$, $F^{T_0 \to T_{\mathrm{max}}}(T^{\mathrm{el}})$
- the force extrapolation from this work's approximation, with reference calculation done at $T_{\mathrm{max}} = 50,000\mathrm{K}$, $F^{T_{\mathrm{max}} \to T_0}(T^{\mathrm{el}})$
- a linear fit between the ground-state force and the finite-$T^{\mathrm{el}}$ force computed at $T^{\mathrm{el}} = 50,000\mathrm{K}$
- a linear combination of the previous two predictions in a way to approximate the finite-

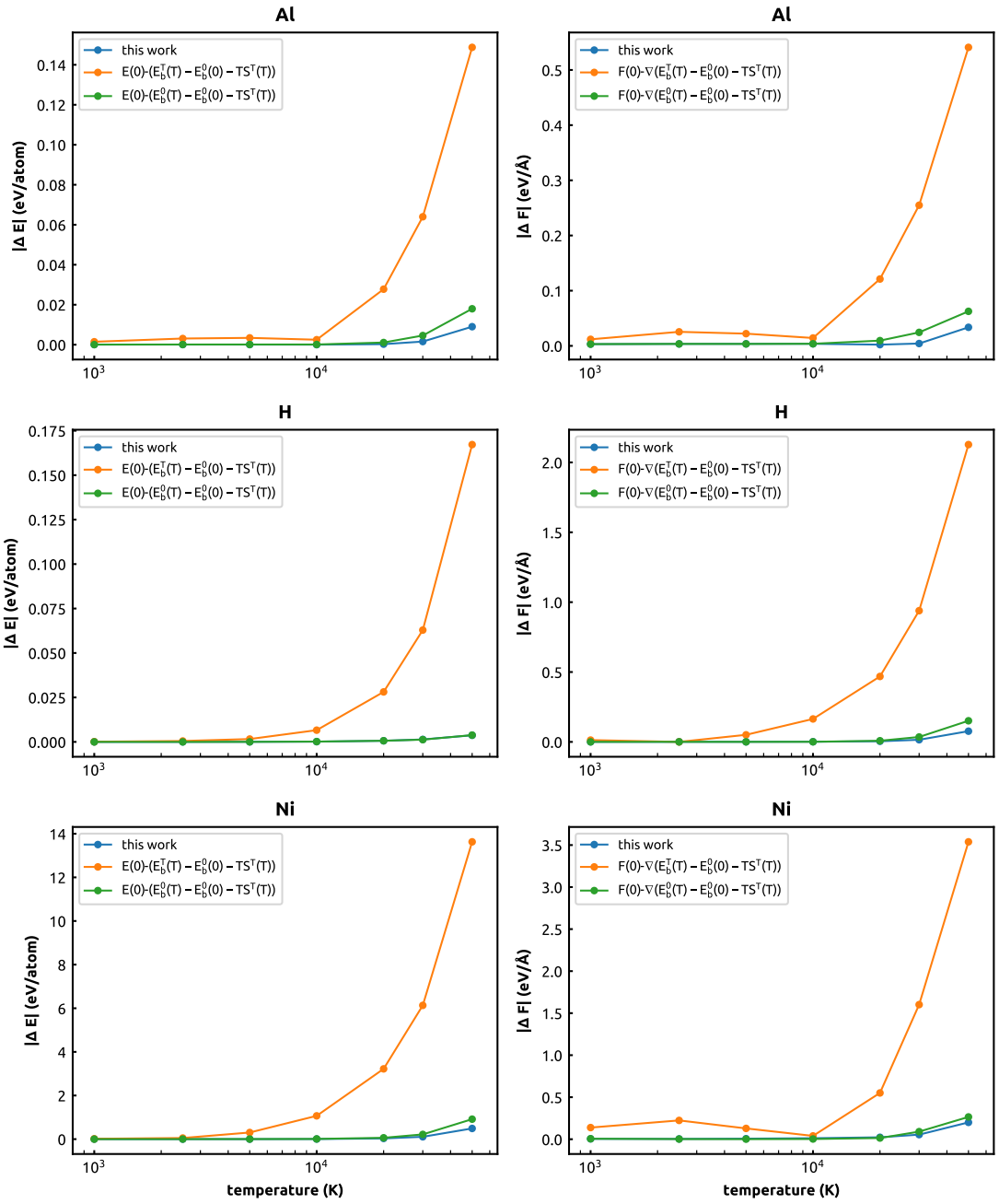


Figure 4.6: Errors, with respect to the finite-$T^{\text{el}}$ results, of different methods to compute a finite-$T^{\text{el}}$ correction to the total energy and the atomic force, computed for a single force component in an aluminium supercell, a liquid hydrogen structure and a liquid nickel structure as a function of the electronic temperature.

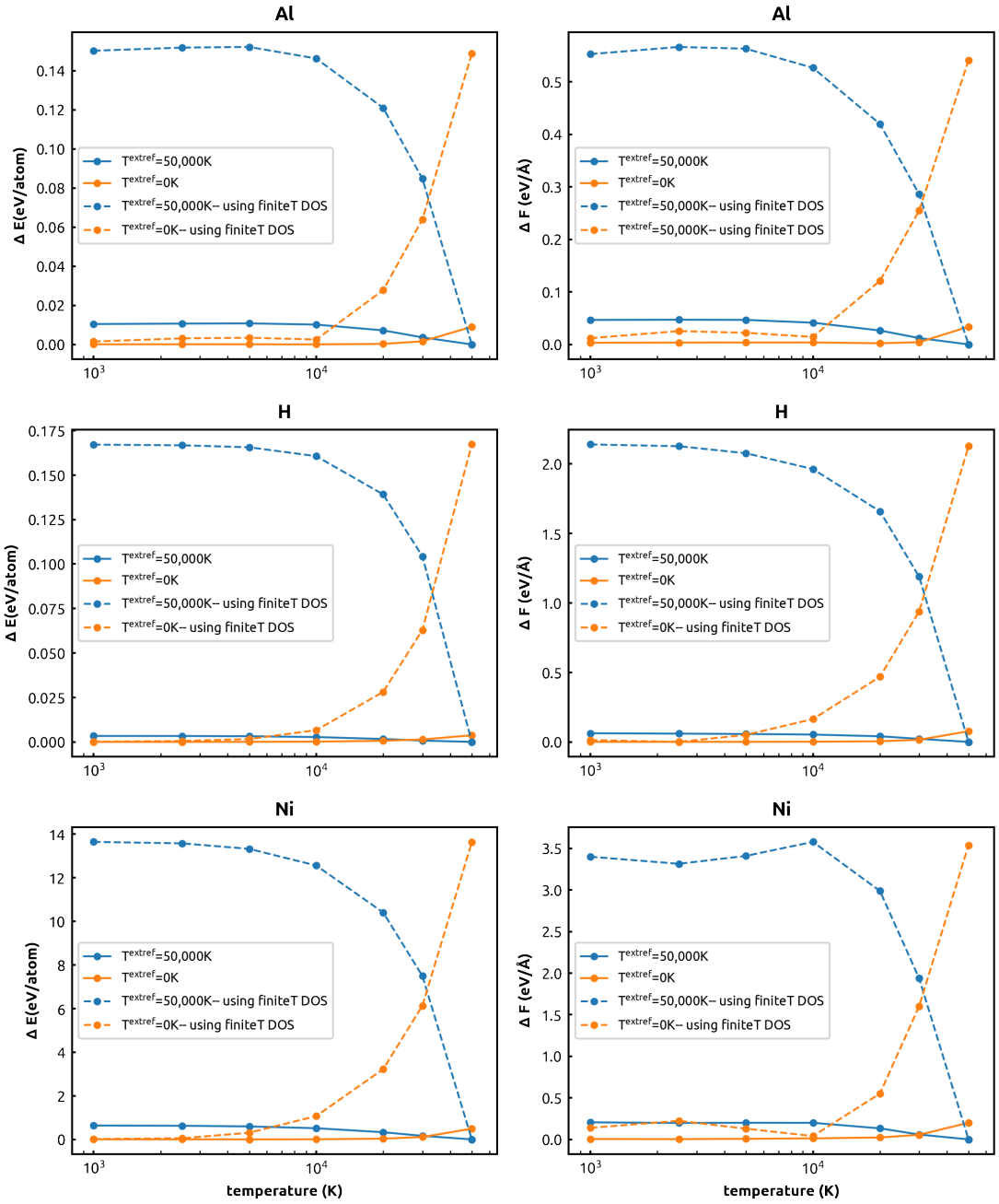


Figure 4.7: Errors, with respect to the finite-$T^{\mathrm{el}}$ results, of different methods to compute a finite-$T^{\mathrm{el}}$ and using two reference calculations done at $T^{\mathrm{el}} = 50,000\mathrm{K}$ and $T^{\mathrm{el}} = 0\mathrm{K}$, to compute a correction to the total energy and the atomic force, computed for a single force component in an aluminium supercell, a liquid hydrogen structure and a liquid nickel structure as a function of the electronic temperature.

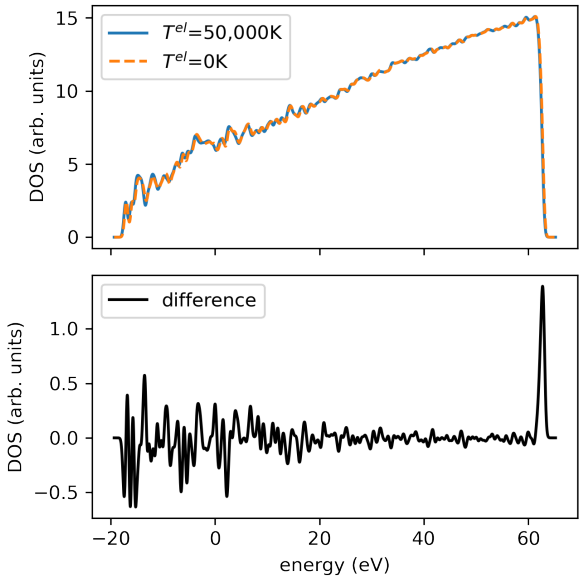


Figure 4.8: (Upper panel) The DOS of a liquid hydrogen structure using two electronic temperatures. Blue: ground-state, orange: $T^{\mathrm{el}}$=50,000K. The Gaussian broadening used to construct the DOS is 0.3eV. (Lower panel) The residual between the two DOS computed at $T^{\mathrm{el}} = 0$K and $T^{\mathrm{el}} = 50,000$K.

$T^{\mathrm{el}}$ force with the better approximation according to the following mixing:

$$F^{\mathrm{two-point}}(T^{\mathrm{el}}) = \frac{T^{\mathrm{el}} - T_0}{T_{\mathrm{max}} - T_0} F^{T_0 \to T_{\mathrm{max}}}(T^{\mathrm{el}}) + \frac{T_{\mathrm{max}} - T^{\mathrm{el}}}{T_{\mathrm{max}} - T_0} F^{T_{\mathrm{max}} \to T_0}(T^{\mathrm{el}}),$$

- a cubic fitting using the two previous predictions $F^{T_0 \to T_{\mathrm{max}}}(T^{\mathrm{el}})$ and $F^{T_{\mathrm{max}} \to T_0}(T^{\mathrm{el}})$ using our approximation

$$F^{\mathrm{two-point,cubic}}(T^{\mathrm{el}}) = \left[1 - \lambda\left(\frac{T^{\mathrm{el}} - T_0}{T_{\mathrm{max}} - T_0}\right)\right] F^{T_0 \to T_{\mathrm{max}}}(T^{\mathrm{el}}) + \lambda\left(\frac{T^{\mathrm{el}} - T_0}{T_{\mathrm{max}} - T_0}\right) F^{T_{\mathrm{max}} \to T_0}(T^{\mathrm{el}}),$$

where $\lambda(x)$ is the cubic polynomial that satisfies $\lambda(0) = 0$, $\lambda(1) = 1$, $\lambda'(0) = \lambda'(1) = 0$

We report our findings about the errors of these methods compared to finite-$T^{\mathrm{el}}$ computed forces in Fig. 4.9. The simple linear fit yields worse results than all the other methods, except at the extremes (0K and 50,000K). Our Mermin-like functional used as a single-point extrapolation seems to be a good approximation on its own when using a round-state reference, while the high-$T$ reference leads to high errors at low temperatures. A linear mixing of the high and low-temperature extrapolations improves the accuracy in the high-temperature limit and can be further improved by considering that in our Mermin-functional expansion, the

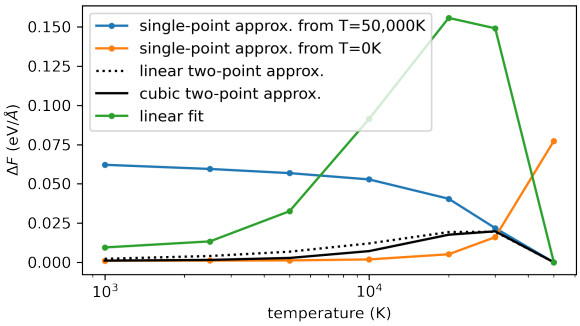


Figure 4.9: Calculation errors compared to the finite-$T^{\mathrm{el}}$ force in a liquid hydrogen structure.

error has zero first derivative close to the edges and that this property can be preserved by using a mixing function that has zero derivatives at the edges.

### Derivation of finite-$T^{\mathrm{el}}$ gradients

As mentioned earlier, our derivation for the Mermin-like functional is compatible with most MLIPs and ML DOS models presented in the scientific literature. We show a general derivation of the different finite-$T^{\mathrm{el}}$ force components in Eq. (4.32): $\Delta\mathbf{F}_I^0(T^{\mathrm{el}}) \equiv -\nabla_I[\Delta E^0_{\mathrm{band}}(T^{\mathrm{el}}) - T^{\mathrm{el}}S^0(T^{\mathrm{el}})]$. Also, we demonstrate that if one opts for a kernel model for the DOS, like the one presented in Chapter 3, it is possible to derive simple expressions for the force components in order to make the implementation simpler, regardless of the representation of the DOS.

From the definitions of the band energy, in Eq. (4.22), and the entropy in Eq. (4.29), we notice that taking the gradient of these components requires the determination of the gradient of the Fermi level $\mu(T)$ with respect to the atomic positions because of the chain rule. One easy way to determine this gradient is by taking the derivative of Eq. (4.28) at a fixed total number of electrons is unaffected by perturbing the system and we obtain:

$$\nabla_I\mu(T) = -\frac{1}{\int \mathrm{d}\epsilon\ \mathrm{DOS}^0(\epsilon)\frac{\partial f}{\partial \mu}}\int_{-\infty}^{+\infty}\mathrm{d}\epsilon\ f_{\mathrm{FD}}(\epsilon-\mu(T),T)\nabla_I\mathrm{DOS}^0(\epsilon). \tag{4.33}$$

In the limit of $T = 0$K, the Fermi level becomes the Fermi energy $\mu(0) = \varepsilon_F$ and we obtain:

$$\nabla_I\varepsilon_F = -\frac{1}{\mathrm{DOS}^0(\varepsilon_F)}\int_{-\infty}^{\varepsilon_F}\mathrm{d}\epsilon\ \nabla_I\mathrm{DOS}^0(\epsilon).$$

The thermal band energy of Eq. (4.22) is the difference between two similar terms, which is why we only focus on the first term depending on $T > 0$, $E_b(T) = \int \mathrm{d}\epsilon\ \epsilon\mathrm{DOS}^0(\epsilon) f_{\mathrm{FD}}(\epsilon-\mu(T^{\mathrm{el}}),T^{\mathrm{el}})$, since it is the general case of the $T = 0$K term. In order to make notations easier to follow, we

write $f = f_{\mathrm{FD}}(\epsilon - \mu(T), T)$. We start by writing the gradient of $E_b(T)$:

$$\nabla_I E_b(T) = \int \mathrm{d}\epsilon\, \epsilon f \nabla_I \mathrm{DOS}^0(\epsilon) + \int \mathrm{d}\epsilon\, \epsilon \mathrm{DOS}^0(\epsilon) \frac{\partial f}{\partial \mu} \nabla_I \mu. \tag{4.34}$$

By plugging Eq. (4.33) in the previous equation and rearranging terms, we obtain the following expression for the gradients of the band energy as a function of the gradient of the electronic density of states:

$$\nabla_I E_b(T) = \int \mathrm{d}\epsilon\, (\epsilon - \Sigma) f \nabla_I \mathrm{DOS}^0(\epsilon), \tag{4.35}$$

where

$$\Sigma = \frac{\int \mathrm{d}\epsilon\, \epsilon \mathrm{DOS}^0(\epsilon) \frac{\partial f}{\partial \mu}}{\int \mathrm{d}\epsilon\, \mathrm{DOS}^0(\epsilon) \frac{\partial f}{\partial \mu}}$$

is an average shift term appearing due to the conservation of the Fermi level. In the $T = 0$ limit, this shift is the Fermi energy of the electron system: $\Sigma(T = 0) = \varepsilon_F$.

We follow the same logic in determining the gradient of the entropy $S^0(T)$:

$$\begin{aligned}\nabla_I S^0(T) = -k_B \int \mathrm{d}\epsilon\, [f\log(f) + (1-f)\log(1-f)] \nabla_I \mathrm{DOS}^0(\epsilon) \\ - k_B \int \mathrm{d}\epsilon\, \frac{\partial [f\log(f) + (1-f)\log(1-f)]}{\partial \mu} \nabla_I \mu.\end{aligned} \tag{4.36}$$

In this expression, we need to simplify the derivative with respect to $\mu$ in the second integral and we obtain:

$$\frac{\partial [f\log(f) + (1-f)\log(1-f)]}{\partial \mu} = \log\left(\frac{f}{1-f}\right) \frac{\partial f}{\partial \mu} = -\beta(\epsilon - \mu) \frac{\partial f}{\partial \mu}. \tag{4.37}$$

We plug this expression alongside Eq. (4.33) in Eq. (4.36), rearrange terms and obtain the following expression for the gradient of the entropy:

$$\nabla_I S^0(T) = -k_B \int \mathrm{d}\epsilon\, [f\log(f) + (1-f)\log(1-f)] \nabla_I \mathrm{DOS}^0(\epsilon) + \frac{1}{T}(\mu - \Sigma) \int \mathrm{d}\epsilon\, f \nabla_I \mathrm{DOS}^0(\epsilon), \tag{4.38}$$

where we notice that the shift term $\Sigma$ appears again, acting on the Fermi level of the electron system. This derivation proves that we can write the hot electron forces just in terms of the gradients of the DOS, which can be computed easily within the kernel-based ML model, e.g. Eq. (3.7). In fact, if one considers the pointwise representation of the $\mathrm{DOS}(A, E) = \mathbf{k}_{AM}^{\intercal} \cdot \mathbf{x}_M(E)$, the weights matrix $\mathbf{x}_M$ only depends on the energy grid and is independent of the atomic positions. Hence, we can take the kernel vector out of the integrals in Eq. (4.35) and Eq. (4.38). For the remainder of this short discussion, we use Eq. (4.35) as a demonstration and the same

result that we obtain from the derivation also holds for the entropy in Eq. (4.38). The gradients of the band energy of Eq. (4.35) become:

$$\nabla_I E_b(T) = \nabla_I(\mathbf{k}_{AM}^{\mathsf{T}}) \cdot \left( \int \mathrm{d}\epsilon \; (\epsilon - \Sigma) f \mathbf{x}_M(\epsilon) \right). \tag{4.39}$$

The practical implementation of the CDF representation, as shown in Eq. (3.17), makes the implementation of the gradients of the DOS also straightforward. One only needs to replace, in Eq. (4.39), the weights matrix at $E$ by the average of the matrix values at $(E-\delta E)$ and $(E+\delta E)$: $\mathbf{x}_M(E) \leftarrow \frac{\mathbf{x}_M(E+\delta E)-\mathbf{x}_M(E-\delta E)}{2\delta E}$.
The implementation of the gradients in the PC decomposition approach is slightly more delicate than the previous two representations. The DOS of a structure $A$ can be decomposed as $\mathrm{DOS}(A,E) = \sum_k c_k(A)\mathbf{U}_k(E)$, where the $c_k(A)$ are the linear expansion coefficients and $\mathbf{U}_k(E)$ are the values of the principal components (or latent functions) at the energy level $E$. Each coefficient $c_k$ is a target for a regression model and we write $c_k(A) = \mathbf{k}_{AM}^{\mathsf{T}} \cdot \mathbf{w}_k$, where the $\mathbf{w}_k$ are the regression weights. In this case, the DOS model becomes

$$\mathrm{DOS}(A,E) = \sum_k (\mathbf{k}_{AM}^{\mathsf{T}} \cdot \mathbf{w}_k)\mathbf{U}_k(E),$$

where only the $\mathbf{U}_k$ hold the dependence on the energy levels. Therefore, Eq. (4.35) can be implemented as:

$$\nabla_I E_b(T) = \sum_k (\nabla_I(\mathbf{k}_{AM}^{\mathsf{T}}) \cdot \mathbf{w}_k) \cdot \left( \int \mathrm{d}\epsilon \; (\epsilon - \Sigma) f \mathbf{U}_k(\epsilon) \right). \tag{4.40}$$

One should notice that these derivations can be extended to the case of the gradients of the free energy $A(T^{\mathrm{el}})$ with respect to the cell vectors in order to obtain the stress virial. One only needs to replace the gradients of the DOS with respect to the atomic positions with the gradients of the DOS with respect to the variations of the cell vectors.

## 4.3 Hydrogen in planetary conditions

We demonstrate the practicality of our theoretical framework in ML workflows incorporating the electronic finite temperature effects in atomistic simulations by constructing the EOS of metallic liquid hydrogen at conditions similar to those found in the core of a young Jupiter [158], and we compare our ML approach to explicit first-principles molecular dynamics (FPMD) simulations results at finite-$T^{\mathrm{el}}$. We build a training set made of $\sim 28,000$ structures, each containing 128 atoms, and densities ranging between 0.6g cm$^{-3}$ and 1.77g cm$^{-3}$. It consists of configurations from Ref. [159], complemented by snapshots obtained from MD

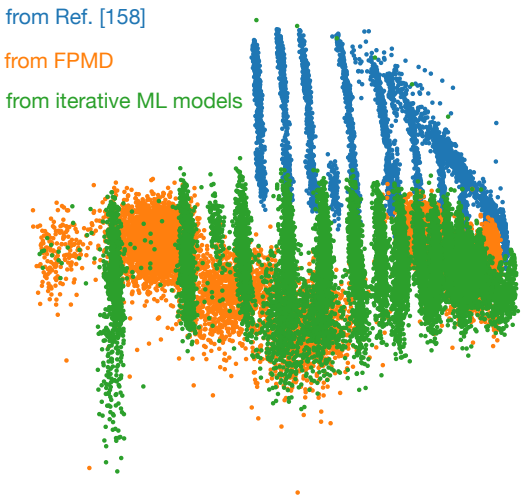


Figure 4.10: Clustering of the structures in the hydrogen data set based on the first 2 principal components of the SOAP representation of every configuration.

simulations performed with preliminary versions of the MLIP. Fig. 4.10 is a map of the first two principal components of the SOAP features of the structures in the dataset showing its diversity. We employ QUANTUM ESPRESSO [160, 161, 162] (QE) for DFT calculations of the data set, using the Optimized Norm-Conserving Vanderbilt pseudopotential [163] version 1.2, which is shown to perform well even at ~TPa pressures [164]. Dispersion interactions are included via a van der Waals density functional [165, 166, 167, 168]. We use a plane-wave energy cutoff on wave functions of 100Ry. We use the Marzari-Vanderbilt cold smearing [169] of 0.01Ry, which yielded similar energies and forces to calculations done at higher k-point density and with a Fermi-Dirac smearing of 10K. The self-consistency accuracy in the electron density is $1 \times 10^{-12}$Ry. We use dense k-point mesh targeting at least $0.01\text{Å}^{-1}$ spacing to ensure the convergence of the DOS and the stress tensor. We perform these convergence tests on a hydrogen structure containing 64 atoms and of density 0.733g cm$^{-3}$. The results of the convergence tests with respect to the k-point grid for the DOS (to the left) and the $S_{xx}$ component of the stress tensor (to the right) are shown in Figure 4.11. In our calculations, we make sure to compute a sufficient number of bands to accommodate the tail of the distribution, which reaches high energies at large $T^{\text{el}}$. In practice, we choose to compute 3 bands per atom, which results in the computation of 6 electronic states per atom. This choice ensures that the Fermi-Dirac occupation of the highest energy level is below $1 \times 10^{-5}$ when $T^{\text{el}} = 50,000$K. We find that this value is sufficiently low not to affect the values of the atomic forces.

We train a GAP model, to learn and predict the $T^{\text{el}} = 0$ contribution to the total free energy on the total DFT energies and a single Hellmann-Feynman force component per structure. We

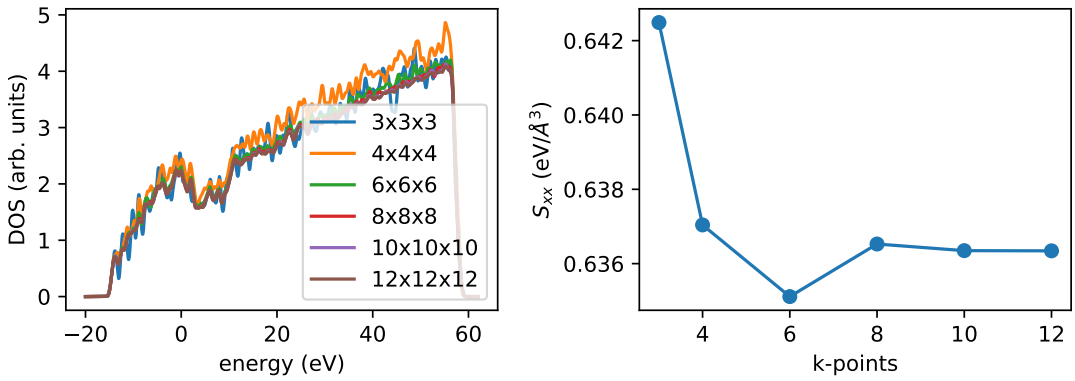


Figure 4.11: Convergence of some key quantities with respect to the k point grid for a 64-atom liquid hydrogen structure. Left: electronic density of states constructed using a Gaussian broadening value of 0.3eV. Right: the $S_{xx}$ component of the Hellmann-Feynman stress tensor. In both cases, $8 \times 8 \times 8$ corresponds to the targeted k-point spacing of 0.01Å$^{-1}$.

use a two-body baseline for the GAP, as discussed in Section 2.3.3. As a reminder, we write:

$$E = E_{2B} + E_{MB}$$

where $E_{2B}$ is a baseline two-body term and $E_{MB}$ is a many-body term. The baseline potential is fitted on dissociating hydrogen dimers placed in a large box of length 15Å. The distances range from 0.075Å to 0.73Å, which corresponds to the repulsive part of the pair interaction. This 2-body term ensures the stability of our molecular dynamics simulations at high temperatures (>20,000K). In a preliminary study using only the many-body term of the GAP model, we found that the MLIP could not recreate the binding curve of a hydrogen dimer and favoured configurations where the two hydrogen atoms were superposed. This behaviour of the many-body term of GAP is also observed in other studies, such as in Ref. [93]. In practice, we tabulate the repulsive potential values and use LAMMPS [170] to handle the calculations of the atomic energies and forces of the two-body interactions. The many-body term is obtained from the Smooth Overlap of Atomic Positions (SOAP) [42] representation with radial scaling [70]. The SOAP representation requires the optimisation of several hyperparameters that we achieve using a grid search. We select the parameters that minimise the prediction error on the total energy in a subset of 5000 structures using a 2-fold cross-validation regression scheme. The best SOAP parameters are (in the notation of *librascal* [116]): *max_radial*=8, *max_angular*=6, *interaction_cutoff*=2Å, *gaussian_sigma_constant*=0.1 and the best radial scaling parameters are: *rate*=1.0, *scale*=2.0 and *exponent*=4. It is worth noting that in a preliminary study, we tried to use 2-body MLIP, with a large radial cutoff (>7Å), to model the pair interaction of the hydrogen dimer. We found that this operation resulted in an increase in the variance of the

| | RMSE |
|---|---|
| $E(0)$ | 11.05meV/atom |
| $\Delta E^0_{\text{band}}(T^{\text{el}}) - T^{\text{el}}S^0(T^{\text{el}})$ | 13.43meV/atom |
| $A(T^{\text{el}})$ | 12.22meV/atom |
| $\mathbf{F}_I(0)$ | 0.87eV/Å |
| $\Delta\mathbf{F}_I(T^{\text{el}})$ | 0.66eV/Å |
| $\mathbf{F}_I(T^{\text{el}})$ | 0.81eV/Å |

Table 4.2: Table of the validation root mean square errors (RMSE) of the ML models on the energies and forces compared to the reference DFT data, at the same level of theory introduced in Eqs. (4.30) and (4.32). The electronic temperature is $T^{\text{el}} = 35,000$K. The training set consists of 28,000 structures and the errors are reported for a validation set of 2,500 configurations.

total energies learned by the many-body SOAP MLIP; hence, the performance of the final model was poor.

We follow a similar approach to the many-body term to construct an atom-centred model for the electronic density of states (DOS), as explained in Chapter 3. We build the DOS from the Kohn-Sham eigenenergies of the same calculations for the total energies using a Gaussian broadening $g_{\text{b}}$ of 0.5eV. We also build a single DOS-gradient component, computed by finite displacement of a single atom by $1 \times 10^{-3}$Å. The model also relies on the SOAP representation. We use a grid-search on the same subset to find the optimal hyperparameters that minimise the prediction errors on the DOS: *max_radial*=12, *max_angular*=8, *interaction_cutoff*=4Å, *gaussian_sigma_constant*=0.1 and the best radial scaling parameters are *rate*=1.0, *scale*=1.0 and *exponent*=2.

Before moving towards the validation of the ML workflow with the finite-$T^{\text{el}}$ approximation, it is important to address one last issue of learning the DOS. The alignment of the energy bands is, in principle, irrelevant to the definition of the finite-$T^{\text{el}}$ correction, according to Eq. (4.22). While the gradients of the DOS with respect to the atomic positions or the cell vectors may depend on the chosen alignment, the finite-$T^{\text{el}}$ correction to the forces and stress involving them do not. We only need to make sure that the alignment is done consistently when computing the energy terms and the DOS gradients. While there are several strategies to align the DOS, like aligning with respect to the deep core levels, the valence band minimum (VBM) or the Fermi energy, the choice of the alignment should only depend on the quality of the ML prediction. In this case, we find that aligning the DOS with respect to the VBM yields the lowest prediction errors because it uniformises the energy range on which the DOS is defined. This strategy helps create uniform targets. In Appendix A.2, we discuss further the challenges of building an ML DOS for this data set containing structures spanning a wide density interval, and the need to numerically fit states in the conduction band.

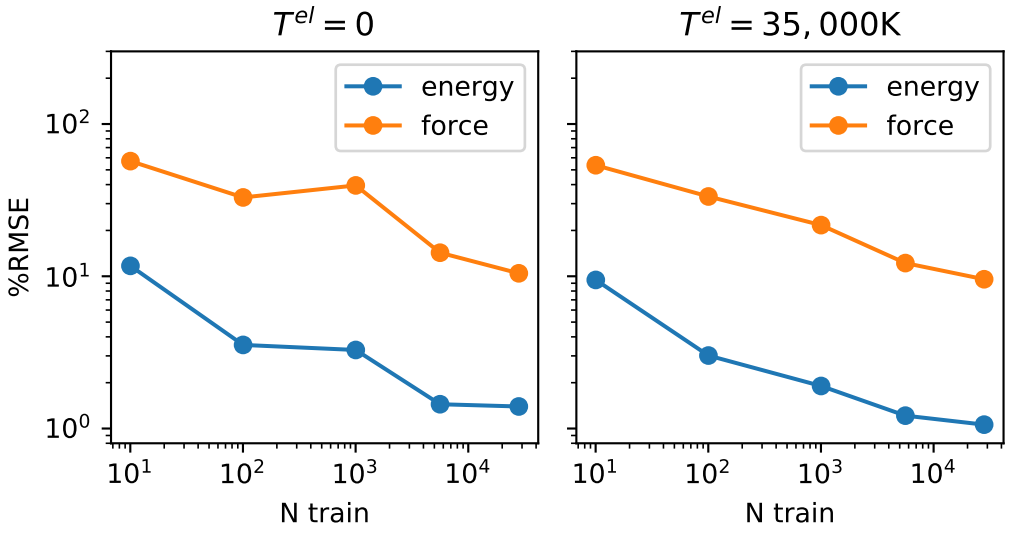


Figure 4.12: Root mean square error (RMSE) as a percentage of the total variance of the energies and forces as a function of the size of the training set. Left: from the GAP model; right: from the GAP and the finite-$T^{\mathrm{el}}$ correction at $T^{\mathrm{el}} = 35,000$K.

We use the Projected Process approximation of the Gaussian process regression framework to train the two ML models. The idea is to select a subselection of the training environments and use them as a basis to expand the target quantities (energy or DOS). We select 7000 environments for the GAP model and 5000 environments for the DOS model. These environments are chosen by a greedy algorithm, the furthest point sampling [88]. We validate our two ML models on a validation set containing 2500 structures. Figure 4.12 shows the learning curves (LCs) of the energies and forces obtained from the GAP model (left) and the full framework, i.e. GAP and the thermal electronic correction (right) at $T^{\mathrm{el}} = 35,000$K. The LCs are still linear for all the considered quantities in the log-log plots and suggest that the accuracy of the models can be enhanced by training on more configurations. Table. 4.2 shows the root mean square error (RMSE) of the different (free) energies and forces for $T^{\mathrm{el}} = 0$K and at $T^{\mathrm{el}} = 35,000$K. The ML models are in good agreement with the corresponding DFT calculations and the RMSE of the total free energy is well below the typical thermal energy at the temperatures we consider in this study, and comparable to the values observed in previous simulations of liquid systems at high ionic temperatures [93].

In order to gauge the importance of finite-$T^{\mathrm{el}}$ effects, and to obtain accurate reference calculations consistent with our computational setup, we run two sets of FPMD trajectories targeting the pressures 400GPa, 800GPa, 1,200GPa, and 1,600GPa for each of the ionic temperatures $T^i$ = 10,000K, 20,000K, 35,000K, and 50,000K. The electronic temperature of the first set is

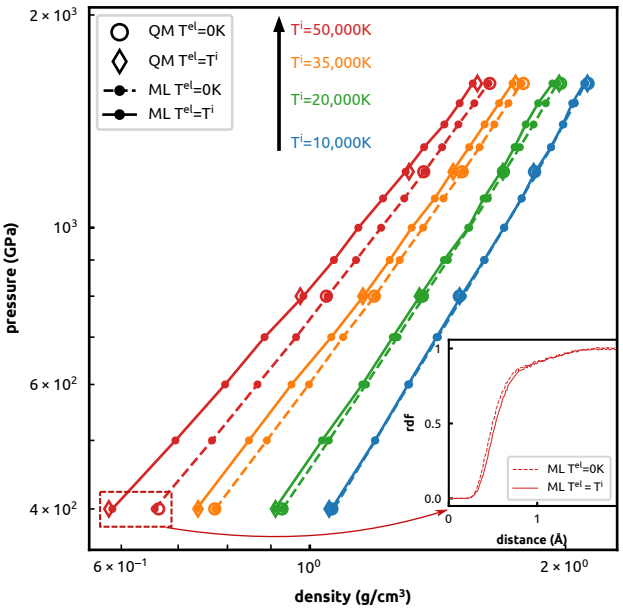


Figure 4.13: Hydrogen isotherms of different equations of state (EOS). The empty circles correspond to the EOS computed with "cold"-electron-Γ-point DFT. The empty diamonds correspond to the EOS computed with finite-electron-temperature-Γ-point DFT. The dashed lines correspond to the EOS computed with ML trained on $T^{\mathrm{el}} = 0$K data. The solid lines correspond to the EOS computed with finite temperature ML framework, where $T^{\mathrm{el}} = T^i$. The temperatures range from 10,000K to 50,000K are denoted by the different colours as shown in the legend. The statistical error bars computed by block averages are smaller than the size of the markers. The small inset represents the radial distribution functions of hydrogen at $P = 400$GPa and $T^i = 50{,}000$K computed from the ML trajectories. The dashed line corresponds to $T^{\mathrm{el}} = 0$K and the solid line to $T^{\mathrm{el}} = T^i$.

$T^{\mathrm{el}} = 0$K, while $T^{\mathrm{el}} = T^i$ is in the second set. The DFT calculations are performed with QE and $\Gamma$-point sampling. We evolve the ion dynamics with i-PI for at least 8ps, after an equilibration phase of 1ps, with a time step of 0.1fs. $T^i$ is controlled by stochastic velocity rescaling [171] with a time constant $\tau = 5$fs, and an isotropic barostat [172] with a time constant $\tau = 20$fs, thermalised with an optimal-sampling generalised Langevin thermostat [173]. Due to the high temperature and the fast intrinsic time scale of hydrogen, such relatively short simulations are sufficient to obtain converged results with small statistical uncertainty. We report the results of these simulations in Fig. 4.13, by the empty symbols. The differences due to the finite electron temperature grow steadily between 10,000K and 50,000K, and at the highest temperature, they range between 4% at 1,600GPa and 10% at 400GPa, providing an indication of the impact of finite-$T^{\mathrm{el}}$ in this range of pressure and density.

We then run two analogous sets of trajectories based on the finite-$T^{\mathrm{el}}$ MLIP, temperatures as for the FPMD, and pressures spanning the range between 400GPa and 1,600GPa in intervals of 100GPa. As for the case of FPMD, the first set of simulations does not include any finite temperature effects (dashed lines in Fig. 4.13), while the second incorporates them (solid lines). Our ML EOSs are in excellent agreement with the reference curves obtained with explicit finite-$T^{\mathrm{el}}$ FPMD, up to the statistical uncertainties. We also observe a small shift in the radial distribution at the lower pressure and higher temperature range, corresponding to the difference in particle densities. As an additional demonstration of the importance of incorporating finite-$T^{\mathrm{el}}$ effects, we compute constant-pressure heat capacities, $C_p = \left(\frac{\partial H}{\partial T}\right)_p$ that we obtain as finite differences of the enthalpy $H = \langle K \rangle + \langle A(T^{\mathrm{el}}) \rangle + T^{\mathrm{el}} \langle S^0(T^{\mathrm{el}}) \rangle + p \langle V \rangle$, Here $K$ is the kinetic energy of the ions, and the averages $\langle \ldots \rangle$ are computed over finite-$T^{\mathrm{el}}$ $NpT$ sampling. Fig. 4.14 compares the heat capacity computed from $T^{\mathrm{el}} = 0$K simulations (blue) with that computed including the electronic contributions (green) - which amounts to almost 50% at the highest temperature considered. DFT and ML simulations agree with each other within their statistical uncertainty. The a-posteriori incorporation of electronic excitation by adding $C_{el} \equiv \langle \frac{\partial \Delta E_{\mathrm{band}}}{\partial T} \rangle$ (orange) on top of the $T^{\mathrm{el}} = 0$K ionic contribution, as done in Ref. [136] and explained in Section 4.1, cannot reproduce accurately the finite-$T^{\mathrm{el}}$ results.

These results demonstrate the accuracy of an ML model based on the ground-state DOS approximation in sampling the finite-$T^{\mathrm{el}}$ thermophysical properties of hydrogen in a challenging portion of its phase diagram. By treating explicitly the ionic and electronic degrees of freedom, our ML models eliminate one of the most glaring limitations of traditional MLIPs, which are restricted to performing simulations at a single (usually zero) electron temperature. We remark that no restriction occurs in applying our machinery to a two-temperature model where the electrons and the nuclei are thermalised at different temperatures, i.e. $T^{\mathrm{el}} \neq T^i$, even though a more realistic scenario would incorporate some coupling term between the electronic and

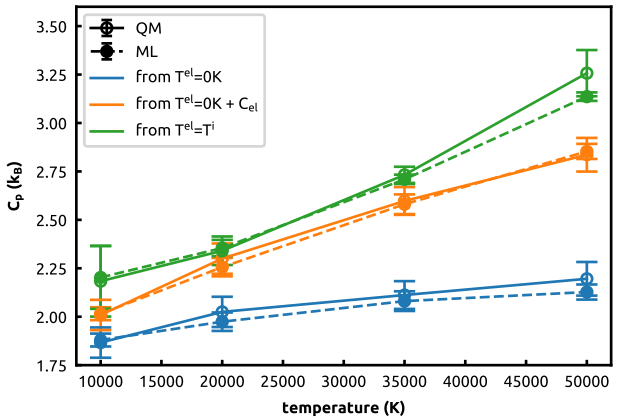


Figure 4.14: Specific heat capacity $C_p$ of hydrogen from $NpT$ simulations at 400GPa. The solid lines represent the DFT calculations and the dashed lines represent the ML calculations. Blue: $C_p$ from the fluctuations of the ions' enthalpy at $T^{\text{el}} = 0$K; orange: $C_p$ same as the blue curves in addition to a correction term computed from the average band energy of the electrons over the trajectories; green: $C_p$ from the finite-$T^{\text{el}}$ sampling. The error bars are computed from standard block analysis.

the nuclear subsystems. Our approach can be easily extended to any electronic structure method based on the Kohn-Sham mapping and can be naturally used also for multiple-species systems, opening the possibility of studying the complex phase diagram of metallic mixtures at high-pT conditions, which dictates the evolution of giant planets [174]. On a conceptual level, the idea of using a physical approximation in synergy with data-driven predictions emerges as a promising research direction to further extend the scope of applicability of predictive atomic-scale simulations.

# 5 Conclusions

The electronic density of states (DOS) is a central quantity in the electronic structure theory. It can be used to estimate several properties of materials, including thermal and optical properties. Traditionally, the DOS is determined by performing expensive first-principle calculations, which significantly limits the size of problems that can be investigated, even with modern computing platforms. This makes the development of cheap and accurate models a necessity for large-scale calculations. Machine-learning approaches are a good candidate for building these surrogate models, as one only needs to perform calculations on a small number of structures. Then the model performs the generalisation to other configurations.

In this thesis, we tried to tackle some of the issues linked to building machine-learning for the DOS, including the optimal methods to represent it as a target for a machine-learning algorithm, in order to reduce the complexity of the learning task. The success of the model led us to explore its use as a building block in integrated machine-learning approaches, where we used the DOS to account for the effects of the (thermally-excited) electrons.

However, constructing machine-learning models for the DOS can be challenging because the DOS of a material is a smooth vector-valued function of the energy levels. The dimension of the output space can become arbitrarily large when accurately representing the DOS variations. Nevertheless, the smooth construction of the DOS hints towards possible correlations between the values of the DOS. In particular, we investigated three approaches to represent the DOS while keeping in mind the smoothness condition for the machine-learned DOS. We found that it is possible to leverage the correlations between the values of the DOS to increase the accuracy of the model by first predicting its cumulative distribution function when applied to a challenging silicon dataset.

Alongside the proposed representations for the DOS, we introduced an atom-centred model

and applied it to different classes of materials ranging from semiconductors, metals and matter in extreme conditions. Its additivity property guarantees the scalability of the model trained on small atomic configurations to larger and more complex structures. We demonstrated its generalisation ability by predicting the electronic fingerprints in large silicon structures. The atom-centred approach also gave us access to locally-defined DOS that, despite not being a physical observable, provided us with insights into the electronic transitions in disordered silicon phases and their interplay with the structural transitions.

Then, we explored how to use the DOS models in an integrated machine-learning framework, arising from a physics-inspired modelling approach, to compute macroscopic properties of materials. As a first step, we combined the DOS model with traditional ML potentials to compute the heat capacity of nickel near its melting temperature, by assuming that the thermal excitations of electrons do not affect the dynamics of the ions. Hence, we were able to estimate the electronic contribution to the heat capacity of molten nickel, which explained the discrepancy between the experimental and computational values. This approach was successful because the difference between finite temperature forces and the ground state forces used to train the potential energy surface model was negligible at the temperature regime investigated. However, this assumption could not be maintained at higher temperatures, like in warm dense matter conditions. In these cases, the finite temperature electronic free energy must be used to perform finite temperature Born-Oppenheimer simulations. Machine-learning techniques could reduce computational costs, but prior to our work, traditional approaches to constructing the potential energy surface suffered from non-transferability between electronic temperatures.

This thesis proposed a framework to approximate the electronic free energy from ground-state data exclusively, which solved the temperature transferability problem. We combined over-the-shelf machine-learning interatomic potentials with DOS models in a Mermin-like expression for the electronic free energy and its gradients. In our framework, the electronic temperature is an external parameter independent of the training phase of the models. Therefore, it is guaranteed that the approximated free energy can be used over a wide range of temperatures. We implemented these models in existing atomistic modelling software and showcased them on metallic hydrogen in warm dense matter conditions. We successfully recovered the equation of state and the heat capacity, both computed from first-principle molecular dynamics, with our machine-learning framework at a fraction of the cost. This is a powerful example of the potential of physics-driven modelling of materials combining structural and electronic descriptions of atomic configurations.

Moving forward, one could think of different ways to enhance the atom-centred models of the

DOS. One possible approach is to exploit the freedom in defining the local density of states and, subsequently, the local atomic charges. They can be used within a self-consistent scheme to align the local DOS so that the atomic charges can reconstruct a particular global field like the Hartree potential within the simulation box. This charge equilibration scheme might be useful to construct models for charge flow when two materials with different Fermi energies are in contact, like semiconductors heterojunctions. Another aspect that could benefit from accelerating the electronic DOS is two-temperature experiments mimicking light shining on metals that can excite electrons to higher temperatures while keeping the ions "cold". This model can be combined with nonadiabatic simulation techniques to study several phenomena like thermal transport across metal/non-metal interfaces [175] for example. Finally, one could explore further the underlying assumption of decomposing the DOS of a structure into contributions from atomic centres since the DOS is a global property of the entire structure. The locality investigation of the DOS is part of an effort to study the locality effects of the machine-learning models used in atomistic modelling. While these local contributions are the result of a modelling exercise, they may be linked to experimental observables, hence, validate our (machine-learning) models.

More broadly, this thesis demonstrates the advantages of combining ML predictions with physics-based approximations, providing an example of hybrid modelling paradigm that combines the flexibility of data-driven techniques with the transferability and interpretability of physical models.

# A Appendix

In these three sections, we present and discuss some of the technical challenges in building machine-learning (ML) models for the electronic density of states (DOS). In particular, we look into the effect of the energy reference on the indirect learning of some derived quantities from the DOS, and on the principal component (PC) representation of the DOS. We will use the two data sets of silicon (Section 3.2) and hydrogen (Section 4.3) as examples to illustrate these challenges and provide general guidelines for learning strategies for the DOS. Also, we provide extra results about the interpretation of the local atomic charges defined from the local density of states (LDOS) obtained from the DOS representations introduced in Section 3.1.

## A.1 Alignment of the DOS

In this paragraph, we discuss some methods one could use to align the DOS as a target for an ML approach and their effect on the calculation of the value of the DOS at the Fermi energy (DOS($\varepsilon_F$)), the Fermi energy ($\varepsilon_F$) and the band energy.

The diversity of structures in the silicon dataset used in Section 3.2, in terms of densities and type of structures (bulk vs clusters), presents one main challenge for modelling the DOS, that is the choice of the energy reference to construct the DOS of the training set. We refer to this problem as the DOS alignment.One could state that this problem is purely a mathematical issue, as we see in Section 4.2, where the finite-temperature correction is independent of the alignment of the DOS, i.e. the only constraint here is to have a consistent method to perform the alignment for the entire data set. Therefore, the optimal choice should be left to the data and the machine learning models.

In particular, we are interested in four strategies: aligning to the $G = 0$ component of the Hartree potential (MH), the Fermi energy, the deepest core level (DS), and the valence band

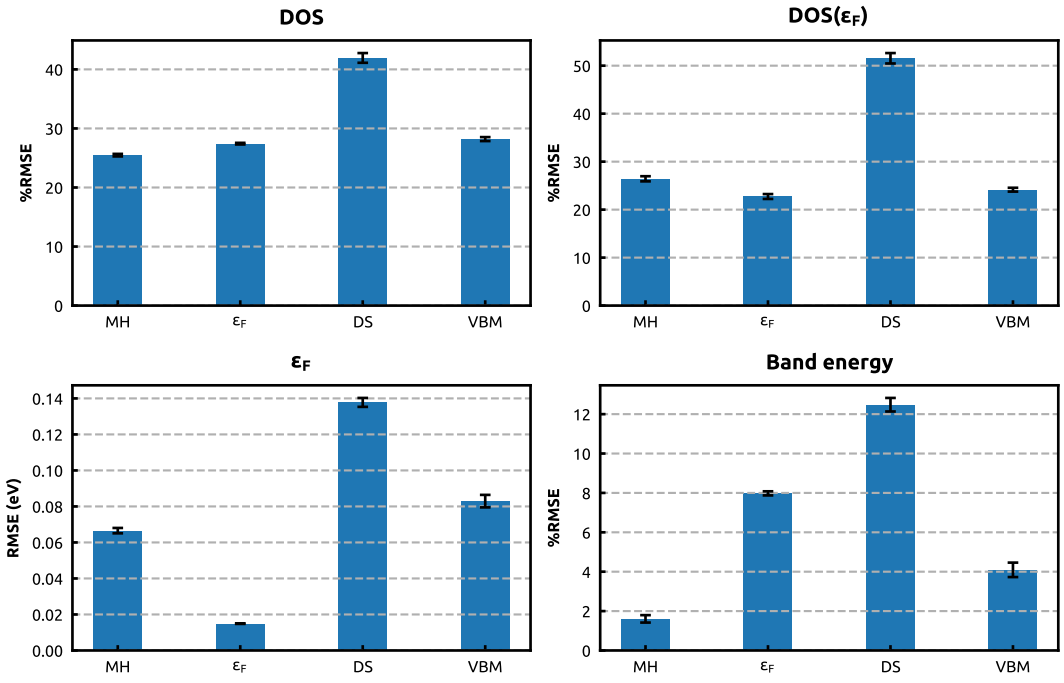


Figure A.1: Average errors in the density of states over 8 splits in the silicon dataset for different band alignment strategies. The reference DOS is constructed using a Gaussian broadening $g_b = 0.1$eV. MH stands for average Hartree potential, $\epsilon_F$ stands for Fermi energy, DS stands for the deepest core state, and VBM stands for valence band minimum.

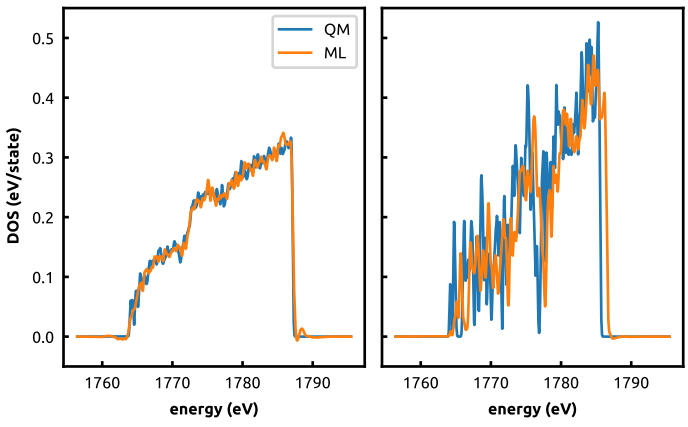


Figure A.2: Examples of the DFT and ML-predicted DOS of two silicon structures: (left) a liquid structure; (right) a cluster structure. The energy reference is the deepest core level.

minimum (VBM). We test these four strategies on the silicon data set of Section 3.2 using the pointwise representation of the DOS targeting a Gaussian broadening of $g_b = 0.1$eV. We compute the errors by performing 8 train/test splits of the data set and report the test errors only. We estimate the uncertainty of each strategy by computing the standard error of the mean from the 8 splits. The top left panel of Fig. A.4 shows that the best method to align the DOS is the $G = 0$ component of the Hartree potential used by default in FHI-aims, followed by the Fermi energy and the VBM alignments. The DS alignment performs the worst in this data set, even though the intuitive understanding states that these core levels should be isolated from external effects; hence, they should provide a robust energy reference. Fig. A.2 might provide an explanation for the high errors. We see that this model is prone to predicting non-zero occupation for states in the conduction band while their DFT reference is zero. In Section A.2, we discuss this problem further.
In the other panels of Fig. A.4, we report the prediction errors for up-mentioned derived quantities from the different aligning strategies. We find that the DS alignment yields the worst prediction errors for all the properties calculated, especially for the Fermi energy and the band energy. However, it seems that MH alignment is the best approach to use with the lowest errors in all quantities, followed by the VBM alignment. The latter has a straightforward implementation from the eigenenergies themselves and does not require "hacking" DFT packages to extract the mean Hartree potential value. As expected, we find that the alignment to the Fermi energy performs the best in predicting the Fermi energy. This is explained by the fact the Fermi energy is now a constant of the learning problem and is expected to be close to zero. In fact, the RMSE of the Fermi energy, in this case, is lower than the energy grid spacing of $\delta\epsilon = 0.05$eV.

Another direction to tackle this problem is to adjust the DOS in such a way that it has the minimum possible variance at each energy channel without changing its shape, within a rigid DOS approximation. This approach could be useful in situations where it is necessary to align the DOS with certain physical constraints or requirements.

## A.2 Unphysical unoccupied states

In Section A.1, we saw that the alignment of the DOS could play an important role in the accuracy of the ML DOS models, and it could also affect the quantities that we compute with DOS. One problem that we noticed was the prediction of non-zero DOS($\epsilon$) when its reference DFT value is zero, as seen in Fig. A.2. The ML DOS model predicts the existence of the non-zero DOS($\epsilon$) towards the edge of the computed conduction band. One could argue that this arises from the construction of the reference DFT DOS. When performing DFT calculations on the

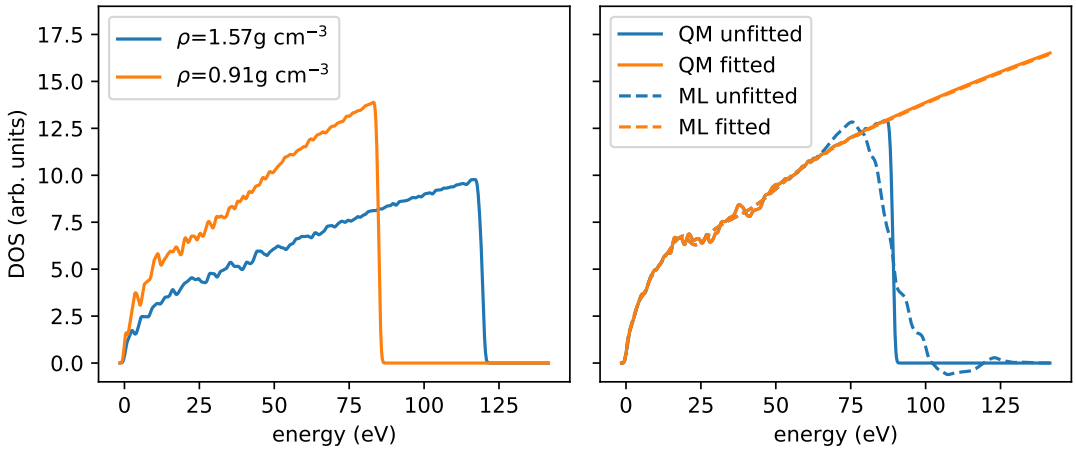


Figure A.3: Left: Example of the unfitted DFT DOS of two hydrogen structures at different densities computed using 3 bands per atom. Blue: density is 1.57g cm$^{-3}$; orange: density is 0.91g cm$^{-3}$. Right: Example of the effect of unphysical discontinuity of the DFT DOS on the ML DOS (in blue) and the stabilised model after fitting the missing bands (in orange).

training set, one only provides the number of bands to be computed by the software and not their energy range. This means that if we performed DFT calculations with more electronic states per atom, the zero-occupation region in the right panel of Fig. A.2 would have states that can be occupied. Therefore, we could say that these zero-values of the reference DOS are not physical.

As mentioned earlier, we would like to assess the effects of these unphysical discontinuities of the DFT DOS on the PC representation of the DOS and on the indirect learning of the DOS-dependent quantities: the Fermi energy, the DOS($\varepsilon_F$), and the band energy for the silicon data set. We use the VBM alignment and a Gaussian broadening of $g_b = 0.3$eV. We truncate the DOS to an energy level in the conduction band on which the DOS($\epsilon$) is non-zero because we are not able to fit a function for the missing bands. We also compare to the models trained on DOS presenting the unphysical discontinuities, that we truncate after obtaining the ML predictions. We report in Fig. A.4 the evolution of the prediction errors as a function of the number of principal components used to construct the DOS. These results are obtained by performing the same 8 splits of train/test of the data set from Section A.1. We notice that all prediction errors are saturated after using $\approx 15$ PCs. Both strategies to train the DOS yield compatible results on the derived quantities, which hints that the effect of the "unlearnable" PCs, discussed in Section 3.2, is more important in this indirect learning approach.

We also see the same problem of the non-physical zero-DOS when dealing with the hydrogen data set of Section 4.3. The data set contains configurations at different volumes and hence

their DOS can span a different energy range, even after VBM alignment, for the same number of the calculated electronic states (c.f. left panel of Figure A.3). The calculated DOS drops to zero above the highest computed energy level. Even though the chosen number of energy levels is always such that the Fermi-Dirac occupation of the highest level is negligible even at the largest $T^{\mathrm{el}}$ considered, a sudden, unphysical drop in the DOS may negatively affect the learning, as discussed in Section A.1. To solve this issue, we fit the values of the missing (empty) states to a square root behaviour, where we use a $\propto \sqrt{\epsilon - \epsilon_0}$ filling of the unphysically-zero DOS($\epsilon$), to ensure that the targets for our ML DOS model do not involve unphysical discontinuities. The right panel of Figure A.3 shows an example of the effect of such discontinuities on the learning of the DOS in a liquid hydrogen structure and how fitting the missing bands provides much-needed stability to the ML DOS model. We want to stress that these fitted occupations do not contribute to the finite-$T^{\mathrm{el}}$ correction, even at $T^{\mathrm{el}} = 50,000$K, despite the long tail of the Fermi-Dirac distribution at this temperature. The RMSE of the finite-$T^{\mathrm{el}}$ correction to the atomic forces computed with the DFT fitted and unfitted DOS is 0.00037eV/Å and can be neglected.

We also perform a similar study using the hydrogen data set. We include the finite-temperature correction to the electronic free energy $A(T^{\mathrm{el}})$ at $T^{\mathrm{el}} = 50,000$K for the hydrogen data set. We use the VBM alignment and a Gaussian broadening of $g_{\mathrm{b}} = 0.5$eV. We compare the performance of the pointwise and the PC representations of the DOS. We also include results from one extra strategy for building the DOS, which is dividing each the DOS of every structure $A$, DOS($A, \epsilon$), by the volume of the structure. In a preliminary study, we found that this approach might be justified by the fact the slope of the DOS in the conduction band was strongly dependent on the volume. Due to the large size of this set ($\approx 28,500$ structures), we only perform these predictions on a single train/test split. The train set size is 26,500 structures. We report the results in Fig. A.5. We find that the derived quantities are almost insensitive to the model's building strategy, even though one would prefer to scale by the structures' volume as a first step in the fitting process and also fit the missing states from the DOS. In this example, the PC representation of the DOS seems to outperform the pointwise representation.

## A.3 Effect of the representation of the atomic charges

In Section 3.3, we show that the different representations of the DOS yield different ML LDOS compared to the LOBSTER-defined LDOS. Consequently, the differences in the predicted ML LDOS lead to different values of the atomic charges $Q(A_i)$. In Fig. A.6, we report three scatter plots of the atomic charges for a 512-atom amorphous silicon structure obtained from the pointwise, principal components, and CDF representations of the DOS. On one

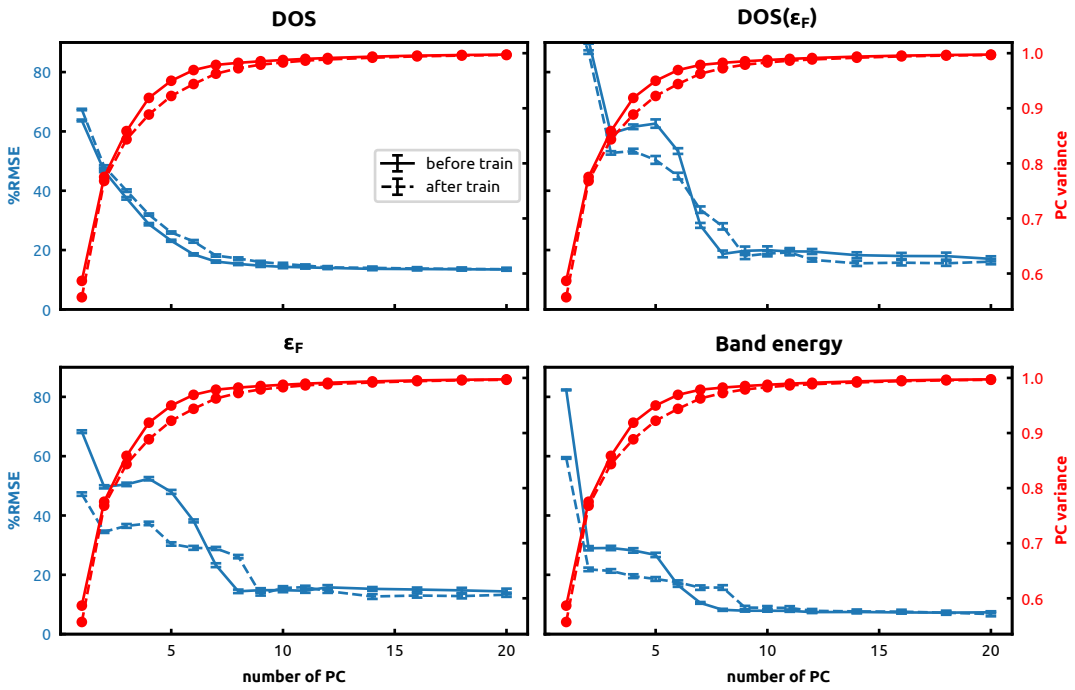


Figure A.4: (Blue curves) Evolution of the prediction errors of the PC representation of the VBM-aligned DOS and its derived quantities: the DOS($\varepsilon_F$), the Fermi energy and the band energy, as a function of the number of principal components. Solid lines describe truncating the DOS before training the ML model. The dashed lines describe truncating the DOS after constructing the ML models. (Red curves) The evolution of the variance explained by the number of principal components.

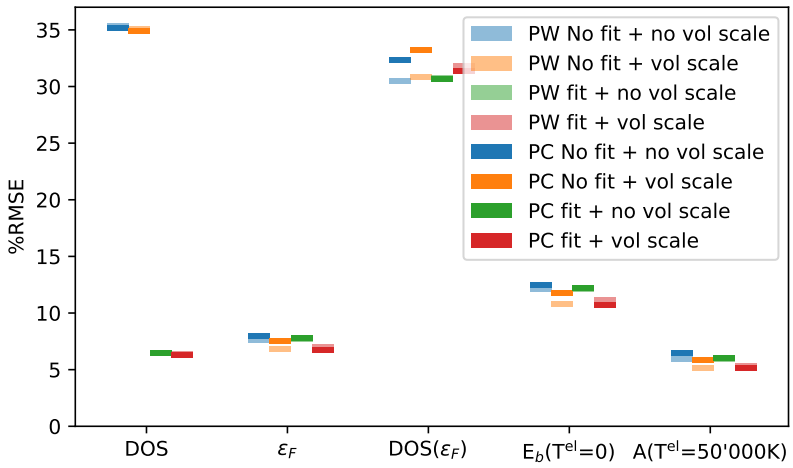


Figure A.5: Prediction errors of several strategies to represent the VBM-aligned DOS. The quantities shown are the DOS, the Fermi energy $\varepsilon_F$, the DOS($\varepsilon_F$), the band energy, and the finite-temperature correction to the electronic free energy A($T^{\text{el}}$= 50,000K). Transparent colours: the pointwise representation. Opaque colours: the PC representation. Blue: no fitting of the missing band and no normalising by the volume of the structures. Orange: no fitting of the missing band and normalising by the volume of the structures. Green: fitting of the missing band and no normalising by the volume of the structures. Red: fitting of the missing band and normalising by the volume of the structures

side, we notice that the charges obtained from the pointwise and CDF representations are correlated. Conversely, the charges from the PC representation show little correlation with their counterpart from the pointwise or the CDF representations. This could be explained by the difficulty in learning several PCs, which results in a predicted LDOS that is different from the one obtained from the pointwise or the CDF representations.

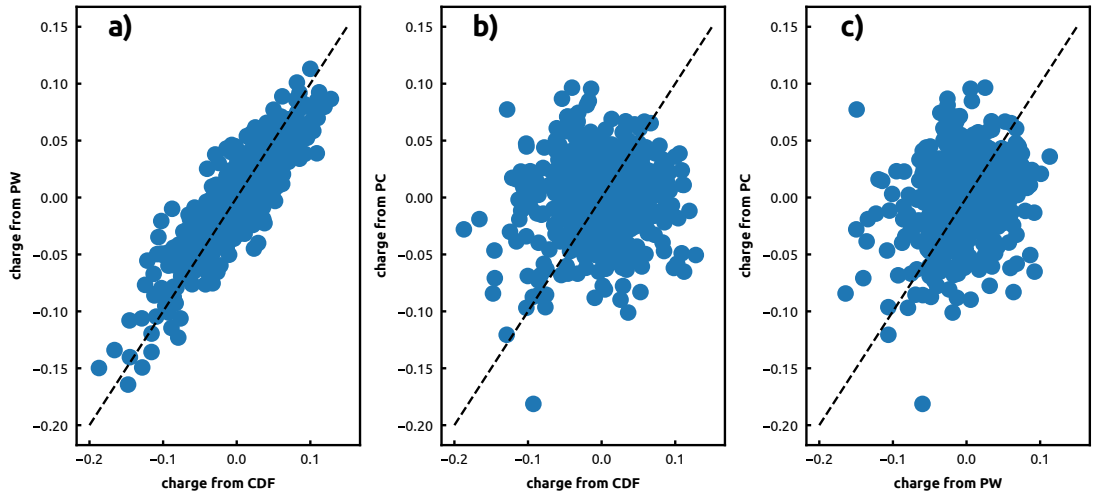


Figure A.6: Scatter plots of atomic charges defined from the LDOS of a 512-atom amorphous silicon structure: comparison of atomic charges computed from (a) CDF and PW representations, (b) CDF and PC representations, and (c) PW and PC representations. The reference DOS is constructed using $g_b$=0.1eV. We notice that PW and CDF yielded correlated charges. Both are not correlated with charges derived from the PC representation.

# Bibliography


[1] Wendy S. Parker. “Evidence and Knowledge from Computer Simulation”. In: *Erkenntnis* 87.4 (Aug. 2022), pp. 1521–1538. DOI: 10.1007/s10670-020-00260-1.

[2] Kota Murakami, Yuta Mizutani, Hiroshi Sampei, Atsushi Ishikawa, Yuta Tanaka, Sasuga Hayashi, Sae Doi, Takuma Higo, Hideaki Tsuneki, Hiromi Nakai, and Yasushi Sekine. “Theoretical Prediction by DFT and Experimental Observation of Heterocation-Doping Effects on Hydrogen Adsorption and Migration over the $CeO_2$ (111) Surface”. In: *Physical Chemistry Chemical Physics* 23.8 (2021), pp. 4509–4516. DOI: 10.1039/D0CP05752E.

[3] Andriy Zakutayev, Xiuwen Zhang, Arpun Nagaraja, Liping Yu, Stephan Lany, Thomas O. Mason, David S. Ginley, and Alex Zunger. “Theoretical Prediction and Experimental Realization of New Stable Inorganic Materials Using the Inverse Design Approach”. In: *Journal of the American Chemical Society* 135.27 (July 2013), pp. 10048–10054. DOI: 10.1021/ja311599g.

[4] Zhen-Yu Wu, Feng-Yang Chen, Boyang Li, Shen-Wei Yu, Y. Zou Finfrock, Debora Motta Meira, Qiang-Qiang Yan, Peng Zhu, Ming-Xi Chen, Tian-Wei Song, Zhouyang Yin, Hai-Wei Liang, Sen Zhang, Guofeng Wang, and Haotian Wang. “Non-Iridium-Based Electrocatalyst for Durable Acidic Oxygen Evolution Reaction in Proton Exchange Membrane Water Electrolysis”. In: *Nature Materials* (Oct. 2022). DOI: 10.1038/s41563-022-01380-5.

[5] Alaeddin Burak Irez, Emin Bayraktar, and Ibrahim Miskioglu. “Fracture Toughness Analysis of Epoxy-Recycled Rubber-Based Composite Reinforced with Graphene Nanoplatelets for Structural Applications in Automotive and Aeronautics”. In: *Polymers* 12.2 (Feb. 2020), p. 448. DOI: 10.3390/polym12020448.

[6] J. C. Slater and G. F. Koster. “Simplified LCAO Method for the Periodic Potential Problem”. In: *Physical Review* 94.6 (June 1954), pp. 1498–1524. DOI: 10.1103/PhysRev.94.1498.

[7] F. Coester. “Bound States of a Many-Particle System”. In: *Nuclear Physics* 7 (June 1958), pp. 421–424. DOI: 10.1016/0029-5582(58)90280-3.

[8] Atsushi Togo and Isao Tanaka. “First Principles Phonon Calculations in Materials Science”. In: *Scripta Materialia* 108 (Nov. 2015), pp. 1–5. DOI: 10.1016/j.scriptamat.2015.07.021.

[9] Eliano Diana and Edoardo Marchese. “Vibrational and DFT Analysis of Perfluoro-o-Phenylenemercury Compounds”. In: *Journal of Organometallic Chemistry* 695.12-13 (June 2010), pp. 1651–1656. DOI: 10.1016/j.jorganchem.2010.03.031.

[10] Francisco Colmenero, Laura. J. Bonales, Joaquín Cobos, and Vicente Timón. “Density Functional Theory Study of the Thermodynamic and Raman Vibrational Properties of $\gamma$-UO $_3$ Polymorph”. In: *The Journal of Physical Chemistry C* 121.27 (July 2017), pp. 14507–14516. DOI: 10.1021/acs.jpcc.7b04389.

[11] Jinxiang You, Jing Wang, Shuhui Zhang, Jun Luo, Zhiwei Peng, Mingjun Rao, and Guanghui Li. “Thermodynamic Properties of Na2MgSiO4: DFT Calculation and Experimental Validation”. In: *Calphad* 79 (Dec. 2022), p. 102480. DOI: 10.1016/j.calphad.2022.102480.

[12] Kieron Burke, Roberto Car, and Ralph Gebauer. “Density Functional Theory of the Electrical Conductivity of Molecular Devices”. In: *Physical Review Letters* 94.14 (Apr. 2005), p. 146803. DOI: 10.1103/PhysRevLett.94.146803.

[13] Evan Kiely, Reabetswe Zwane, Robert Fox, Anthony M. Reilly, and Sarah Guerin. “Density Functional Theory Predictions of the Mechanical Properties of Crystalline Materials”. In: *CrystEngComm* 23.34 (2021), pp. 5697–5710. DOI: 10.1039/D1CE00453K.

[14] Mahdi Faghihnasiri, Morteza Izadifard, and Mohammad Ebrahim Ghazi. “DFT Study of Mechanical Properties and Stability of Cubic Methylammonium Lead Halide Perovskites (CH $_3$ NH $_3$ PbX $_3$ , X = I, Br, Cl)”. In: *The Journal of Physical Chemistry C* 121.48 (Dec. 2017), pp. 27059–27070. DOI: 10.1021/acs.jpcc.7b07129.

[15] P. Hohenberg and W. Kohn. “Inhomogeneous Electron Gas”. In: *Physical Review* 136.3B (Nov. 1964), B864–B871. DOI: 10.1103/PhysRev.136.B864.

[16] W. Kohn and L. J. Sham. “Self-Consistent Equations Including Exchange and Correlation Effects”. In: *Physical Review* 140.4A (Nov. 1965), A1133–A1138. DOI: 10.1103/PhysRev.140.A1133.

[17] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. “Generalized Gradient Approximation Made Simple”. In: *Physical Review Letters* 77.18 (Oct. 1996), pp. 3865–3868. DOI: 10/bppfwt.

[18] Matthias Ernzerhof and Gustavo E. Scuseria. “Assessment of the Perdew–Burke–Ernzerhof Exchange-Correlation Functional”. In: *The Journal of Chemical Physics* 110.11 (Mar. 1999), pp. 5029–5036. DOI: 10.1063/1.478401.

[19] Jochen Heyd, Gustavo E. Scuseria, and Matthias Ernzerhof. "Hybrid Functionals Based on a Screened Coulomb Potential". In: *The Journal of Chemical Physics* 118.18 (May 2003), pp. 8207–8215. DOI: 10.1063/1.1564060.

[20] Aliaksandr V. Krukau, Oleg A. Vydrov, Artur F. Izmaylov, and Gustavo E. Scuseria. "Influence of the Exchange Screening Parameter on the Performance of Screened Hybrid Functionals". In: *The Journal of Chemical Physics* 125.22 (Dec. 2006), p. 224106. DOI: 10/fn9p79.

[21] Pedro Borlido, Thorsten Aull, Ahmad W. Huran, Fabien Tran, Miguel A. L. Marques, and Silvana Botti. "Large-Scale Benchmark of Exchange–Correlation Functionals for the Determination of Electronic Band Gaps of Solids". In: *Journal of Chemical Theory and Computation* 15.9 (Sept. 2019), pp. 5069–5079. DOI: 10/gj8q2q.

[22] Jess Wellendorff, Keld T. Lundgaard, Andreas Møgelhøj, Vivien Petzold, David D. Landis, Jens K. Nørskov, Thomas Bligaard, and Karsten W. Jacobsen. "Density Functionals for Surface Science: Exchange-correlation Model Development with Bayesian Error Estimation". In: *Physical Review B* 85.23 (June 2012), p. 235149. DOI: 10.1103/PhysRevB.85.235149.

[23] David B. Williams and C. Barry Carter. *Transmission Electron Microscopy*. Boston, MA: Springer US, 2009. DOI: 10.1007/978-0-387-76501-3.

[24] Warren J. Hehre. *A Guide to Molecular Mechanics and Quantum Chemical Calculations*. Irvine, CA: Wavefunction, Inc, 2003.

[25] Stefan Chmiela, Huziel E. Sauceda, Klaus-Robert Müller, and Alexandre Tkatchenko. "Towards Exact Molecular Dynamics Simulations with Machine-Learned Force Fields". In: *Nature Communications* 9.1 (Dec. 2018), p. 3887. DOI: 10.1038/s41467-018-06169-2.

[26] Albert P. Bartók, James Kermode, Noam Bernstein, and Gábor Csányi. "Machine Learning a General-Purpose Interatomic Potential for Silicon". In: *Physical Review X* 8.4 (Dec. 2018), p. 041048. DOI: 10/gfrfqd.

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

[28] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. "Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering". In: *Science* 361.6400 (July 2018), pp. 360–365. DOI: 10.1126/science.aat2663.

[29] Addis S. Fuhr and Bobby G. Sumpter. “Deep Generative Models for Materials Discovery and Machine Learning-Accelerated Innovation”. In: *Frontiers in Materials* 9 (Mar. 2022), p. 865270. DOI: 10.3389/fmats.2022.865270.

[30] E. Prodan and W. Kohn. “Nearsightedness of Electronic Matter”. In: *Proceedings of the National Academy of Sciences* 102.33 (Aug. 2005), pp. 11635–11638. DOI: 10/d3sjht.

[31] E-Wen Huang, Wen-Jay Lee, Sudhanshu Shekhar Singh, Poresh Kumar, Chih-Yu Lee, Tu-Ngoc Lam, Hsu-Hsuan Chin, Bi-Hsuan Lin, and Peter K. Liaw. “Machine-Learning and High-Throughput Studies for High-Entropy Materials”. In: *Materials Science and Engineering: R: Reports* 147 (Jan. 2022), p. 100645. DOI: 10.1016/j.mser.2021.100645.

[32] Xiaobo Li, Phillip M. Maffettone, Yu Che, Tao Liu, Linjiang Chen, and Andrew I. Cooper. “Combining Machine Learning and High-Throughput Experimentation to Discover Photocatalytically Active Organic Molecules”. In: *Chemical Science* 12.32 (2021), pp. 10742–10754. DOI: 10.1039/D1SC02150H.

[33] Julia Westermayr, Joe Gilkes, Rhyan Barrett, and Reinhard J. Maurer. *High-Throughput Property-Driven Generative Design of Functional Organic Molecules*. July 2022. arXiv: 2207.01476 [physics].

[34] Robert Schade, Tobias Kenter, Hossam Elgabarty, Michael Lass, Ole Schütt, Alfio Lazzaro, Hans Pabst, Stephan Mohr, Jürg Hutter, Thomas D. Kühne, and Christian Plessl. “Towards Electronic Structure-Based Ab-Initio Molecular Dynamics Simulations with Hundreds of Millions of Atoms”. In: *Parallel Computing* 111 (July 2022), p. 102920. DOI: 10.1016/j.parco.2022.102920.

[35] Joe D. Morrow and Volker L. Deringer. “Indirect Learning and Physically Guided Validation of Interatomic Potential Models”. In: *The Journal of Chemical Physics* 157.10 (Sept. 2022), p. 104105. DOI: 10.1063/5.0099929.

[36] Weile Jia, Han Wang, Mohan Chen, Denghui Lu, Lin Lin, Roberto Car, E Weinan, and Linfeng Zhang. “Pushing the Limit of Molecular Dynamics with Ab Initio Accuracy to 100 Million Atoms with Machine Learning”. In: *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. Atlanta, GA, USA: IEEE, Nov. 2020, pp. 1–14. DOI: 10.1109/SC41405.2020.00009.

[37] Pablo M. Piaggi, Jack Weis, Athanassios Z. Panagiotopoulos, Pablo G. Debenedetti, and Roberto Car. “Homogeneous Ice Nucleation in an Ab Initio Machine-Learning Model of Water”. In: *Proceedings of the National Academy of Sciences* 119.33 (Aug. 2022), e2207294119. DOI: 10.1073/pnas.2207294119.

[38] Volker L. Deringer, Noam Bernstein, Gábor Csányi, Chiheb Ben Mahmoud, Michele Ceriotti, Mark Wilson, David A. Drabold, and Stephen R. Elliott. "Origins of Structural and Electronic Transitions in Disordered Silicon". In: *Nature* 589.7840 (Jan. 2021), pp. 59–64. DOI: 10/ghsb84.

[39] Benjamin A. Helfrecht, Rose K. Cersonsky, Guillaume Fraux, and Michele Ceriotti. "Structure-Property Maps with Kernel Principal Covariates Regression". In: *Machine Learning: Science and Technology* 1.4 (Nov. 2020), p. 045021. DOI: 10/gh7sdr.

[40] Johannes Hoja, Leonardo Medrano Sandonas, Brian G. Ernst, Alvaro Vazquez-Mayagoitia, Robert A. DiStasio, and Alexandre Tkatchenko. "QM7-X, a Comprehensive Dataset of Quantum-Mechanical Properties Spanning the Chemical Space of Small Organic Molecules". In: *Scientific Data* 8.1 (Dec. 2021), p. 43. DOI: 10.1038/s41597-021-00812-2.

[41] Colin R. Groom, Ian J. Bruno, Matthew P. Lightfoot, and Suzanna C. Ward. "The Cambridge Structural Database". In: *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials* 72.2 (Apr. 2016), pp. 171–179. DOI: 10.1107/S2052520616003954.

[42] Albert P. Bartók, Risi Kondor, and Gábor Csányi. "On Representing Chemical Environments". In: *Physical Review B* 87.18 (May 2013), p. 184115. DOI: 10/gft56g.

[43] Jörg Behler and Michele Parrinello. "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces". In: *Physical Review Letters* 98.14 (Apr. 2007), p. 146401. DOI: 10/c7kbsq.

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

[45] 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 (Apr. 2011), p. 153101. DOI: 10.1103/PhysRevB.83.153101.

[46] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller. "SchNetPack: A Deep Learning Toolbox For Atomistic Systems". In: *Journal of Chemical Theory and Computation* 15.1 (Jan. 2019), pp. 448–455. DOI: 10.1021/acs.jctc.8b00908.

[47] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, Mass: MIT Press, 2006.

[48] Linfeng Zhang, Mohan Chen, Xifan Wu, Han Wang, Weinan E, and Roberto Car. "Deep Neural Network for the Dielectric Response of Insulators". In: *Physical Review B* 102.4 (July 2020), p. 041121. DOI: 10.1103/PhysRevB.102.041121.

[49] David J. Tozer, Victoria E. Ingamells, and Nicholas C. Handy. “Exchange-correlation Potentials”. In: *The Journal of Chemical Physics* 105.20 (Nov. 1996), pp. 9200–9213. DOI: 10.1063/1.472753.

[50] John C. Snyder, Matthias Rupp, Katja Hansen, Klaus-Robert Müller, and Kieron Burke. “Finding Density Functionals with Machine Learning”. In: *Physical Review Letters* 108.25 (June 2012), p. 253002. DOI: 10.1103/PhysRevLett.108.253002.

[51] John C. Snyder, Matthias Rupp, Katja Hansen, Leo Blooston, Klaus-Robert Müller, and Kieron Burke. “Orbital-Free Bond Breaking via Machine Learning”. In: *The Journal of Chemical Physics* 139.22 (Dec. 2013), p. 224104. DOI: 10.1063/1.4834075.

[52] Anand Chandrasekaran, Deepak Kamal, Rohit Batra, Chiho Kim, Lihua Chen, and Rampi Ramprasad. “Solving the Electronic Structure Problem with Machine Learning”. In: *npj Computational Materials* 5.1 (Dec. 2019), p. 22. DOI: 10.1038/s41524-019-0162-7.

[53] J. Austin Ellis, Attila Cangi, Normand A. Modine, J. Adam Stephens, Aidan P. Thompson, and Sivasankaran Rajamanickam. “Accelerating Finite-temperature Kohn-Sham Density Functional Theory with Deep Neural Networks”. In: *arXiv:2010.04905 [cond-mat, physics:physics]* (Oct. 2020). arXiv: 2010.04905 [cond-mat, physics:physics].

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

[55] John M. Alred, Ksenia V. Bets, Yu Xie, and Boris I. Yakobson. “Machine Learning Electron Density in Sulfur Crosslinked Carbon Nanotubes”. In: *Composites Science and Technology* 166 (Sept. 2018), pp. 3–9. DOI: 10.1016/j.compscitech.2018.03.035.

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

[57] Andrew T Fowler, Chris J Pickard, and James A Elliott. “Managing Uncertainty in Data-Derived Densities to Accelerate Density Functional Theory”. In: *Journal of Physics: Materials* 2.3 (Apr. 2019), p. 034001. DOI: 10.1088/2515-7639/ab0b4a.

[58] Victor Fung, P. Ganesh, and Bobby G. Sumpter. “Physically Informed Machine Learning Prediction of Electronic Density of States”. In: *Chemistry of Materials* 34.11 (June 2022), pp. 4848–4855. DOI: 10.1021/acs.chemmater.1c04252.

[59] Shufeng Kong, Francesco Ricci, Dan Guevarra, Jeffrey B. Neaton, Carla P. Gomes, and John M. Gregoire. "Density of States Prediction for Materials Discovery via Contrastive Learning from Probabilistic Embeddings". In: *Nature Communications* 13.1 (Feb. 2022), p. 949. DOI: 10.1038/s41467-022-28543-x.

[60] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross. "How to Represent Crystal Structures for Machine Learning: Towards Fast Prediction of Electronic Properties". In: *Physical Review B* 89.20 (May 2014), p. 205118. DOI: 10.1103/PhysRevB.89.205118.

[61] S. R. Broderick and K. Rajan. "Eigenvalue Decomposition of Spectral Features in Density of States Curves". In: *EPL (Europhysics Letters)* 95.5 (Sept. 2011), p. 57005. DOI: 10.1209/0295-5075/95/57005.

[62] Kihoon Bang, Byung Chul Yeo, Donghun Kim, Sang Soo Han, and Hyuck Mo Lee. "Accelerated Mapping of Electronic Density of States Patterns of Metallic Nanoparticles via Machine-Learning". In: *Scientific Reports* 11.1 (Dec. 2021), p. 11604. DOI: 10.1038/s41598-021-91068-8.

[63] Byung Chul Yeo, Donghun Kim, Chansoo Kim, and Sang Soo Han. "Pattern Learning Electronic Density of States". In: *Scientific Reports* 9.1 (Dec. 2019), p. 5879. DOI: 10.1038/s41598-019-42277-9.

[64] Andrea Grisafi, Alan M. Lewis, Mariana Rossi, and Michele Ceriotti. *Electronic-Structure Properties from Atom-Centered Predictions of the Electron Density*. June 2022. arXiv: 2206.14087 [cond-mat, physics:physics, stat].

[65] Olexandr Isayev, Denis Fourches, Eugene N. Muratov, Corey Oses, Kevin Rasch, Alexander Tropsha, and Stefano Curtarolo. "Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints". In: *Chemistry of Materials* 27.3 (Feb. 2015), pp. 735–743. DOI: 10.1021/cm503507h.

[66] Martin Kuban, Santiago Rigamonti, Markus Scheidgen, and Claudia Draxl. "Density-of-States Similarity Descriptor for Unsupervised Learning from Materials Data". In: *Scientific Data* 9.1 (Oct. 2022), p. 646. DOI: 10.1038/s41597-022-01754-z.

[67] Jigyasa Nigam, Michael J. Willatt, and Michele Ceriotti. "Equivariant Representations for Molecular Hamiltonians and $N$ -Center Atomic-Scale Properties". In: *The Journal of Chemical Physics* 156.1 (Jan. 2022), p. 014115. DOI: 10.1063/5.0072784.

[68] Liwei Zhang, Berk Onat, Geneviève Dusson, Adam McSloy, G. Anand, Reinhard J. Maurer, Christoph Ortner, and James R. Kermode. "Equivariant Analytical Mapping of First Principles Hamiltonians to Accurate and Transferable Materials Models". In: *npj Computational Materials* 8.1 (July 2022), p. 158. DOI: 10.1038/s41524-022-00843-2.

[69] Jigyasa Nigam, Sergey Pozdnyakov, and Michele Ceriotti. "Recursive Evaluation and Iterative Contraction of N-body Equivariant Features". In: *The Journal of Chemical Physics* 153.12 (Sept. 2020), p. 121101. DOI: 10/ghj4bv.

[70] Michael J. Willatt, Félix Musil, and Michele Ceriotti. "Atom-Density Representations for Machine Learning". In: *The Journal of Chemical Physics* 150.15 (Apr. 2019), p. 154110. DOI: 10/ggbs9k.

[71] Raju P. Gupta. "Lattice Relaxation at a Metal Surface". In: *Physical Review B* 23.12 (June 1981), pp. 6265–6270. DOI: 10.1103/PhysRevB.23.6265.

[72] Oliver Fleetwood, Marina A. Kasimova, Annie M. Westerlund, and Lucie Delemotte. "Molecular Insights from Conformational Ensembles via Machine Learning". In: *Biophysical Journal* 118.3 (Feb. 2020), pp. 765–780. DOI: 10.1016/j.bpj.2019.12.016.

[73] Stefan Goedecker. "Linear Scaling Electronic Structure Methods". In: *Reviews of Modern Physics* 71.4 (July 1999), pp. 1085–1123. DOI: 10.1103/RevModPhys.71.1085.

[74] Sergey N. Pozdnyakov, Michael J. Willatt, Albert P. Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. "Incompleteness of Atomic Structure Representations". In: *Physical Review Letters* 125.16 (Oct. 2020), p. 166001. DOI: 10.1103/PhysRevLett.125.166001.

[75] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning". In: *Physical Review Letters* 108.5 (Jan. 2012), p. 058301. DOI: 10.1103/PhysRevLett.108.058301.

[76] Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O. Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. "Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space". In: *The Journal of Physical Chemistry Letters* 6.12 (June 2015), pp. 2326–2331. DOI: 10.1021/acs.jpclett.5b00831.

[77] Haoyan Huo and Matthias Rupp. "Unified Representation of Molecules and Crystals for Machine Learning". In: *Machine Learning: Science and Technology* 3.4 (Dec. 2022), p. 045017. DOI: 10.1088/2632-2153/aca005.

[78] Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. "Comparing Molecules and Solids across Structural and Alchemical Space". In: *Physical Chemistry Chemical Physics* 18.20 (May 2016), pp. 13754–13769. DOI: 10/gfx8f9.

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

[80] A.P. Thompson, L.P. Swiler, C.R. Trott, S.M. Foiles, and G.J. Tucker. "Spectral Neighbor Analysis Method for Automated Generation of Quantum-Accurate Interatomic Potentials". In: *Journal of Computational Physics* 285 (Mar. 2015), pp. 316–330. DOI: 10.1016/j.jcp.2014.12.018.

[81] Felix Musil, Andrea Grisafi, Albert P. Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. "Physics-Inspired Structural Representations for Molecules and Materials". In: *Chemical Reviews* 121.16 (Aug. 2021), pp. 9759–9815. DOI: 10.1021/acs.chemrev.1c00021.

[82] Michael J. Willatt, Félix Musil, and Michele Ceriotti. "Feature Optimization for Atomistic Machine Learning Yields a Data-Driven Construction of the Periodic Table of the Elements". In: *Physical Chemistry Chemical Physics* 20.47 (Dec. 2018), pp. 29661–29668. DOI: 10/gfz26d.

[83] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. "Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics". In: *Physical Review Letters* 120.14 (Apr. 2018), p. 143001. DOI: 10.1103/PhysRevLett.120.143001.

[84] Sergey N Pozdnyakov and Michele Ceriotti. "Incompleteness of Graph Neural Networks for Points Clouds in Three Dimensions". In: *Machine Learning: Science and Technology* 3.4 (Dec. 2022), p. 045020. DOI: 10.1088/2632-2153/aca1f8.

[85] Albert P. Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R. Kermode, Gábor Csányi, and Michele Ceriotti. "Machine Learning Unifies the Modeling of Materials and Molecules". In: *Science Advances* 3.12 (Dec. 2017), e1701816. DOI: 10/gcqj8b.

[86] 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 (Mar. 2013), pp. 1521–1532. DOI: 10.1021/ct3010563.

[87] Michael W. Mahoney and Petros Drineas. "CUR Matrix Decompositions for Improved Data Analysis". In: *Proceedings of the National Academy of Sciences* 106.3 (Jan. 2009), pp. 697–702. DOI: 10.1073/pnas.0803205106.

[88] Giulio Imbalzano, Andrea Anelli, Daniele Giofré, Sinja Klees, Jörg Behler, and Michele Ceriotti. "Automatic Selection of Atomic Fingerprints and Reference Configurations for Machine-Learning Potentials". In: *The Journal of Chemical Physics* 148.24 (June 2018), p. 241730. DOI: 10/gds5hz.

[89] N. Aronszajn. "Theory of Reproducing Kernels". In: *Transactions of the American Mathematical Society* 68.3 (1950), pp. 337–404. DOI: 10.1090/S0002-9947-1950-0051437-7.

[90] Federico M. Paruzzo, Albert Hofstetter, Félix Musil, Sandip De, Michele Ceriotti, and Lyndon Emsley. "Chemical Shifts in Molecular Solids by Machine Learning". In: *Nature Communications* 9.1 (Dec. 2018), p. 4501. DOI: 10.1038/s41467-018-06972-x.

[91] Annika Stuke, Milica Todorović, Matthias Rupp, Christian Kunkel, Kunal Ghosh, Lauri Himanen, and Patrick Rinke. "Chemical Diversity in Molecular Orbital Energy Predictions with Kernel Ridge Regression". In: *The Journal of Chemical Physics* 150.20 (May 2019), p. 204121. DOI: 10.1063/1.5086105.

[92] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. "Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons". In: *Physical Review Letters* 104.13 (Apr. 2010), p. 136403. DOI: 10/dkchb7.

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

[94] Davis Unruh, Reza Vatan Meidanshahi, Stephen M. Goodnick, Gábor Csányi, and Gergely T. Zimányi. "Gaussian Approximation Potential for Amorphous Si : H". In: *Physical Review Materials* 6.6 (June 2022), p. 065603. DOI: 10.1103/PhysRevMaterials.6.065603.

[95] Volker L. Deringer, Miguel A. Caro, and Gábor Csányi. "A General-Purpose Machine-Learning Force Field for Bulk and Nanostructured Phosphorus". In: *Nature Communications* 11.1 (Dec. 2020), p. 5461. DOI: 10.1038/s41467-020-19168-z.

[96] Ganesh Sivaraman, Jicheng Guo, Logan Ward, Nathaniel Hoyt, Mark Williamson, Ian Foster, Chris Benmore, and Nicholas Jackson. "Automated Development of Molten Salt Machine Learning Potentials: Application to LiCl". In: *The Journal of Physical Chemistry Letters* 12.17 (May 2021), pp. 4278–4285. DOI: 10.1021/acs.jpclett.1c00901.

[97] James Barker, Johannes Bulin, Jan Hamaekers, and Sonja Mathias. "LC-GAP: Localized Coulomb Descriptors for the Gaussian Approximation Potential". In: *Scientific Computing and Algorithms in Industrial Simulations*. Ed. by Michael Griebel, Anton Schüller, and Marc Alexander Schweitzer. Cham: Springer International Publishing, 2017, pp. 25–42. DOI: 10.1007/978-3-319-62458-7_2.

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

[99] Hans Wackernagel. *Multivariate Geostatistics: An Introduction with Applications*. Berlin ; New York: Springer, 1995.

[100] Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. "Safe Model-Based Reinforcement Learning with Stability Guarantees". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017.

[101] Haitao Liu, Jianfei Cai, and Yew-Soon Ong. "Remarks on Multi-Output Gaussian Process Regression". In: *Knowledge-Based Systems* 144 (Mar. 2018), pp. 102–121. DOI: 10.1016/j.knosys.2017.12.034.

[102] George Kimeldorf and Grace Wahba. "Some Results on Tchebycheffian Spline Functions". In: *Journal of Mathematical Analysis and Applications* 33.1 (Jan. 1971), pp. 82–95. DOI: 10.1016/0022-247X(71)90184-3.

[103] Pierre Goovaerts. *Geostatistics for Natural Resources Evaluation*. Applied Geostatistics Series. New York: Oxford University Press, 1997.

[104] Rupeng Li, Igor Shikhov, and Christoph H. Arns. "Bayesian Optimization With Transfer Learning: A Study on Spatial Variability of Rock Properties Using NMR Relaxometry". In: *Water Resources Research* 58.9 (Sept. 2022). DOI: 10.1029/2021WR031590.

[105] Marc Soutter and Yvan Pannatier. "Groundwater Vulnerability to Pesticide Contamination on a Regional Scale". In: *Journal of Environmental Quality* 25.3 (May 1996), pp. 439–444. DOI: 10.2134/jeq1996.00472425002500030009x.

[106] Jin Li and Andrew D. Heap. "A Review of Comparative Studies of Spatial Interpolation Methods in Environmental Sciences: Performance and Impact Factors". In: *Ecological Informatics* 6.3-4 (July 2011), pp. 228–241. DOI: 10.1016/j.ecoinf.2010.12.003.

[107] Philipp Schneider, Nuria Castell, Matthias Vogt, Franck R. Dauge, William A. Lahoz, and Alena Bartonova. "Mapping Urban Air Quality in near Real-Time Using Observations from Low-Cost Sensors and Model Information". In: *Environment International* 106 (Sept. 2017), pp. 234–247. DOI: 10.1016/j.envint.2017.05.005.

[108] A. G. Journel and Ch J. Huijbregts. *Mining Geostatistics*. Caldwell, N.J: Blackburn Press, 2003.

[109] Chiheb Ben Mahmoud, Andrea Anelli, Gábor Csányi, and Michele Ceriotti. "Learning the Electronic Density of States in Condensed Matter". In: *Physical Review B* 102.23 (Dec. 2020), p. 235130. DOI: 10.1103/PhysRevB.102.235130.

[110] Juan S Gómez-Jeria. "ON THE USE OF THE WHOLE EIGENVALUE SPECTRUM TO OBTAIN SINGLE MOLECULE BAND STRUCTURES AND SOLID BAND GAPS FOR MOLECULAR ELECTRONICS STUDIES". In: *Journal of the Chilean Chemical Society* 51.2 (June 2006). DOI: 10.4067/S0717-97072006000200014.

[111] Nicola Marzari, David Vanderbilt, and M. C. Payne. "Ensemble Density-Functional Theory for Ab Initio Molecular Dynamics of Metals and Finite-Temperature Insulators". In: *Physical Review Letters* 79.7 (Aug. 1997), pp. 1337–1340. DOI: 10/bhzwsv.

[112] Félix Musil, Michael J. Willatt, Mikhail A. Langovoy, and Michele Ceriotti. "Fast and Accurate Uncertainty Estimation in Chemical Machine Learning". In: *Journal of Chemical Theory and Computation* 15.2 (Feb. 2019), pp. 906–915. DOI: 10/gft4fv.

[113] Giulio Imbalzano, Yongbin Zhuang, Venkat Kapil, Kevin Rossi, Edgar A. Engel, Federico Grasselli, and Michele Ceriotti. "Uncertainty Estimation for Molecular Dynamics and Sampling". In: *The Journal of Chemical Physics* 154.7 (Feb. 2021), p. 074102. DOI: 10/gh4k2w.

[114] Y. Rubner, C. Tomasi, and L.J. Guibas. "A Metric for Distributions with Applications to Image Databases". In: *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*. Bombay, India: Narosa Publishing House, 1998, pp. 59–66. DOI: 10.1109/ICCV.1998.710701.

[115] Volker Blum, Ralf Gehrke, Felix Hanke, Paula Havu, Ville Havu, Xinguo Ren, Karsten Reuter, and Matthias Scheffler. "Ab Initio Molecular Simulations with Numeric Atom-Centered Orbitals". In: *Computer Physics Communications* 180.11 (Nov. 2009), pp. 2175–2196. DOI: 10/bzxqhn.

[116] Félix Musil, Max Veit, Alexander Goscinski, Guillaume Fraux, Michael J. Willatt, Markus Stricker, Till Junge, and Michele Ceriotti. "Efficient Implementation of Atom-Density Representations". In: *The Journal of Chemical Physics* 154.11 (Mar. 2021), p. 114109. DOI: 10/gjg7k4.

[117] Michele Ceriotti, Silvia Cereda, Francesco Montalenti, Leo Miglio, and Marco Bernasconi. "*Ab Initio* Study of the Diffusion and Decomposition Pathways of SiH x Species on Si(100)". In: *Physical Review B* 79.16 (Apr. 2009), p. 165437. DOI: 10.1103/PhysRevB.79.165437.

[118] Volker L. Deringer, Noam Bernstein, Albert P. Bartók, Matthew J. Cliffe, Rachel N. Kerber, Lauren E. Marbella, Clare P. Grey, Stephen R. Elliott, and Gábor Csányi. "Realistic Atomistic Structure of Amorphous Silicon from Machine-Learning-Driven Molecular Dynamics". In: *The Journal of Physical Chemistry Letters* 9.11 (June 2018), pp. 2879–2885. DOI: 10.1021/acs.jpclett.8b00902.

[119] David M. Wilkins, Andrea Grisafi, Yang Yang, Ka Un Lao, Robert A. DiStasio, and Michele Ceriotti. "Accurate Molecular Polarizabilities with Coupled Cluster Theory and Machine Learning". In: *Proceedings of the National Academy of Sciences* 116.9 (Feb. 2019), pp. 3401–3406. DOI: 10.1073/pnas.1816132116.

[120] Max Veit, David M. Wilkins, Yang Yang, Robert A. DiStasio, and Michele Ceriotti. "Predicting Molecular Dipole Moments by Combining Atomic Partial Charges and Atomic Dipoles". In: *The Journal of Chemical Physics* 153.2 (July 2020), p. 024113. DOI: 10.1063/5.0009106.

[121] Guillaume Fraux, Rose Cersonsky, and Michele Ceriotti. "Chemiscope: Interactive Structure-Property Explorer for Materials and Molecules". In: *Journal of Open Source Software* 5.51 (July 2020), p. 2117. DOI: 10.21105/joss.02117.

[122] P.-L. Chau and A. J. Hardwick. "A New Order Parameter for Tetrahedral Configurations". In: *Molecular Physics* 93.3 (Feb. 1998), pp. 511–518. DOI: 10.1080/002689798169195.

[123] Jeffrey R. Errington and Pablo G. Debenedetti. "Relationship between Structural Order and the Anomalies of Liquid Water". In: *Nature* 409.6818 (Jan. 2001), pp. 318–321. DOI: 10.1038/35053024.

[124] Noam Bernstein, Bishal Bhattarai, Gábor Csányi, David A. Drabold, Stephen R. Elliott, and Volker L. Deringer. "Quantifying Chemical Structure and Machine-Learned Atomic Energies in Amorphous and Liquid Silicon". In: *Angewandte Chemie International Edition* 58.21 (2019), pp. 7057–7061. DOI: 10/gfwgnz.

[125] Stefan Maintz, Volker L. Deringer, Andrei L. Tchougréeff, and Richard Dronskowski. "LOBSTER: A Tool to Extract Chemical Bonding from Plane-Wave Based DFT: Tool to Extract Chemical Bonding". In: *Journal of Computational Chemistry* 37.11 (Apr. 2016), pp. 1030–1035. DOI: 10.1002/jcc.24300.

[126] Sudip K. Deb, Martin Wilding, Maddury Somayazulu, and Paul F. McMillan. "Pressure-Induced Amorphization and an Amorphous–Amorphous Transition in Densified Porous Silicon". In: *Nature* 414.6863 (Nov. 2001), pp. 528–530. DOI: 10.1038/35107036.

[127] Paul F. McMillan, Mark Wilson, Dominik Daisenberger, and Denis Machon. "A Density-Driven Phase Transition between Semiconducting and Metallic Polyamorphs of Silicon". In: *Nature Materials* 4.9 (Sept. 2005), pp. 680–684. DOI: 10.1038/nmat1458.

[128] Dominik Daisenberger, Thierry Deschamps, Bernard Champagnon, Mohamed Mezouar, Raúl Quesada Cabrera, Mark Wilson, and Paul F. McMillan. "Polyamorphic Amorphous Silicon at High Pressure: Raman and Spatially Resolved X-ray Scattering and Molecular Dynamics Studies". In: *The Journal of Physical Chemistry B* 115.48 (Dec. 2011), pp. 14246–14255. DOI: 10.1021/jp205090s.

[129] K. K. Pandey, Nandini Garg, K. V. Shanavas, Surinder M. Sharma, and S. K. Sikka. "Pressure Induced Crystallization in Amorphous Silicon". In: *Journal of Applied Physics* 109.11 (June 2011), p. 113511. DOI: 10.1063/1.3592963.

[130] Nandini Garg, K. K. Pandey, K. V. Shanavas, C. A. Betty, and Surinder M. Sharma. "Memory Effect in Low-Density Amorphous Silicon under Pressure". In: *Physical Review B* 83.11 (Mar. 2011), p. 115202. DOI: 10.1103/PhysRevB.83.115202.

[131] N. F. Mott and E. A. Davis. *Electronic Processes in Non-Crystalline Materials*. 2nd ed. International Series of Monographs on Physics. Oxford: Clarendon Press, 2012.

[132] Martin Beye, Florian Sorgenfrei, William F. Schlotter, Wilfried Wurth, and Alexander Föhlisch. "The Liquid-Liquid Phase Transition in Silicon Revealed by Snapshots of Valence Electrons". In: *Proceedings of the National Academy of Sciences* 107.39 (Sept. 2010), pp. 16772–16776. DOI: 10.1073/pnas.1006499107.

[133] O. I. Barkalov, V. G. Tissen, P. F. McMillan, M. Wilson, A. Sella, and M. V. Nefedova. "Pressure-Induced Transformations and Superconductivity of Amorphous Germanium". In: *Physical Review B* 82.2 (July 2010), p. 020507. DOI: 10.1103/PhysRevB.82.020507.

[134] J.M. Mignot, G. Chouteau, and G. Martinez. "High Pressure Superconductivity of Silicon". In: *Physica B+C* 135.1-3 (Dec. 1985), pp. 235–238. DOI: 10.1016/0378-4363(85)90473-5.

[135] Ali Alavi, Jorge Kohanoff, Michele Parrinello, and Daan Frenkel. "Ab Initio Molecular Dynamics with Excited Electrons". In: *Physical Review Letters* 73.19 (Nov. 1994), pp. 2599–2602. DOI: 10.1103/PhysRevLett.73.2599.

[136] Nataliya Lopanitsyna, Chiheb Ben Mahmoud, and Michele Ceriotti. "Finite-Temperature Materials Modeling from the Quantum Nuclei to the Hot Electron Regime". In: *Physical Review Materials* 5.4 (Apr. 2021), p. 043802. DOI: 10/gjr8gv.

[137] Chiheb Ben Mahmoud, Federico Grasselli, and Michele Ceriotti. "Predicting Hot-Electron Free Energies from Ground-State Data". In: *Physical Review B* 106.12 (Sept. 2022), p. L121116. DOI: 10.1103/PhysRevB.106.L121116.

[138] G.P. Purja Pun and Y. Mishin. "Development of an Interatomic Potential for the Ni-Al System". In: *Philosophical Magazine* 89.34-36 (Dec. 2009), pp. 3245–3267. DOI: 10.1080/14786430903258184.

[139] F. Körmann, A. Dick, T. Hickel, and J. Neugebauer. "Role of Spin Quantization in Determining the Thermodynamic Properties of Magnetic Transition Metals". In: *Physical Review B* 83.16 (Apr. 2011), p. 165114. DOI: 10.1103/PhysRevB.83.165114.

[140] Ulf R. Pedersen, Felix Hummel, Georg Kresse, Gerhard Kahl, and Christoph Dellago. "Computing Gibbs Free Energy Differences by Interface Pinning". In: *Physical Review B* 88.9 (Sept. 2013), p. 094101. DOI: 10.1103/PhysRevB.88.094101.

[141] Li-Fang Zhu, Fritz Körmann, Andrei V. Ruban, Jörg Neugebauer, and Blazej Grabowski. "Performance of the Standard Exchange-Correlation Functionals in Predicting Melting Properties Fully from First Principles: Application to Al and Magnetic Ni". In: *Physical Review B* 101.14 (Apr. 2020), p. 144108. DOI: 10/gh5dpr.

[142] Roberto Scipioni, Lars Stixrude, and Michael P. Desjarlais. "Electrical Conductivity of SiO2 at Extreme Conditions and Planetary Dynamos". In: *Proceedings of the National Academy of Sciences* 114.34 (Aug. 2017), pp. 9009–9013. DOI: 10/gbvq7t.

[143] D. I. Mihaylov, V. V. Karasiev, S. X. Hu, J. R. Rygg, V. N. Goncharov, and G. W. Collins. "Improved First-Principles Equation-of-State Table of Deuterium for High-Energy-Density Applications". In: *Physical Review B* 104.14 (Oct. 2021), p. 144104. DOI: 10.1103/PhysRevB.104.144104.

[144] Valentin V. Karasiev, James W. Dufty, and S. B. Trickey. "Nonempirical Semilocal Free-Energy Density Functional for Matter under Extreme Conditions". In: *Physical Review Letters* 120.7 (Feb. 2018), p. 076401. DOI: 10.1103/PhysRevLett.120.076401.

[145] Jeffrey M. McMahon, Miguel A. Morales, Carlo Pierleoni, and David M. Ceperley. "The Properties of Hydrogen and Helium under Extreme Conditions". In: *Reviews of Modern Physics* 84.4 (Nov. 2012), pp. 1607–1653. DOI: 10.1103/RevModPhys.84.1607.

[146] M. Bonitz, T. Dornheim, Zh. A. Moldabekov, S. Zhang, P. Hamann, H. Kählert, A. Filinov, K. Ramakrishna, and J. Vorberger. "Ab Initio Simulation of Warm Dense Matter". In: *Physics of Plasmas* 27.4 (Apr. 2020), p. 042710. DOI: 10/ghnbz6.

[147] B. Grabowski, L. Ismer, T. Hickel, and J. Neugebauer. "*Ab Initio* up to the Melting Point: Anharmonicity and Vacancies in Aluminum". In: *Physical Review B* 79.13 (Apr. 2009), p. 134106. DOI: 10/b827jj.

[148] Duancheng Ma, Blazej Grabowski, Fritz Körmann, Jörg Neugebauer, and Dierk Raabe. "Ab Initio Thermodynamics of the CoCrFeMnNi High Entropy Alloy: Importance of Entropy Contributions beyond the Configurational One". In: *Acta Materialia* 100 (Nov. 2015), pp. 90–97. DOI: 10.1016/j.actamat.2015.08.050.

[149] Yuzhi Zhang, Chang Gao, Qianrui Liu, Linfeng Zhang, Han Wang, and Mohan Chen. "Warm Dense Matter Simulation via Electron Temperature Dependent Deep Potential Molecular Dynamics". In: *Physics of Plasmas* 27.12 (Dec. 2020), p. 122704. DOI: 10.1063/5.0023265.

[150] N. David Mermin. "Thermal Properties of the Inhomogeneous Electron Gas". In: *Physical Review* 137.5A (Mar. 1965), A1441–A1443. DOI: 10/dwfk8s.

[151] Renata M. Wentzcovitch, José Luís Martins, and Philip B. Allen. "Energy versus Free-Energy Conservation in First-Principles Molecular Dynamics". In: *Physical Review B* 45.19 (May 1992), pp. 11372–11374. DOI: 10.1103/PhysRevB.45.11372.

[152] Valentin V. Karasiev, Joshua Hinz, S. X. Hu, and S. B. Trickey. "Elucidation of the Subcritical Character of the Liquid–Liquid Transition in Dense Hydrogen". In: *arXiv:2012.13835 [cond-mat]* (Dec. 2020). arXiv: 2012.13835 [cond-mat].

[153] Aurora Pribram-Jones, Stefano Pittalis, E. K. U. Gross, and Kieron Burke. "Thermal Density Functional Theory in Context". In: *arXiv:1309.3043 [cond-mat, physics:physics, physics:quant-ph]* 96 (2014), pp. 25–60. DOI: 10/gjg4rp. arXiv: 1309.3043 [cond-mat, physics:physics, physics:quant-ph].

[154] M. Weinert and J. W. Davenport. "Fractional Occupations and Density-Functional Energies and Forces". In: *Physical Review B* 45.23 (June 1992), pp. 13709–13712. DOI: 10.1103/PhysRevB.45.13709.

[155] S. Goedecker and K. Maschke. "Operator Approach in the Linearized Augmented-Plane-Wave Method: Efficient Electronic-Structure Calculations Including Forces". In: *Physical Review B* 45.4 (Jan. 1992), pp. 1597–1604. DOI: 10.1103/PhysRevB.45.1597.

[156] Dávid Péter Kovács, Cas van der Oord, Jiri Kucera, Alice E. A. Allen, Daniel J. Cole, Christoph Ortner, and Gábor Csányi. "Linear Atomic Cluster Expansion Force Fields for Organic Molecules: Beyond RMSE". In: *Journal of Chemical Theory and Computation* 17.12 (Dec. 2021), pp. 7696–7711. DOI: 10.1021/acs.jctc.1c00647.

[157] Xi Zhang, Blazej Grabowski, Fritz Körmann, Christoph Freysoldt, and Jörg Neugebauer. "Accurate Electronic Free Energies of the $3\mathit{d},4\mathit{d}$, and $5\mathit{d}$ Transition Metals at High Temperatures". In: *Physical Review B* 95.16 (Apr. 2017), p. 165126. DOI: 10/ghtrv6.

[158] N. Nettelmann, A. Becker, B. Holst, and R. Redmer. "JUPITER MODELS WITH IMPROVED AB INITIO HYDROGEN EQUATION OF STATE (H-REOS.2)". In: *The Astrophysical Journal* 750.1 (Apr. 2012), p. 52. DOI: 10/ggvszg.

[159] Bingqing Cheng, Guglielmo Mazzola, Chris J. Pickard, and Michele Ceriotti. "Evidence for Supercritical Behaviour of High-Pressure Liquid Hydrogen". In: *Nature* 585.7824 (Sept. 2020), pp. 217–220. DOI: 10/gg99h2.

[160] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L. Chiarotti, Matteo Cococcioni, Ismaila Dabo, Andrea Dal Corso, Stefano de Gironcoli, Stefano Fabris, Guido Fratesi, Ralph Gebauer, Uwe Gerstmann, Christos Gougoussis, Anton Kokalj, Michele Lazzeri, Layla Martin-Samos, Nicola Marzari, Francesco Mauri, Riccardo Mazzarello, Stefano Paolini, Alfredo

Pasquarello, Lorenzo Paulatto, Carlo Sbraccia, Sandro Scandolo, Gabriele Sclauzero, Ari P. Seitsonen, Alexander Smogunov, Paolo Umari, and Renata M. Wentzcovitch. "QUANTUM ESPRESSO: A Modular and Open-Source Software Project for Quantum Simulations of Materials". In: *Journal of Physics: Condensed Matter* 21.39 (Sept. 2009), p. 395502. DOI: 10/d7spb8.

[161] P Giannozzi, O Andreussi, T Brumme, O Bunau, M Buongiorno Nardelli, M Calandra, R Car, C Cavazzoni, D Ceresoli, M Cococcioni, N Colonna, I Carnimeo, A Dal Corso, S de Gironcoli, P Delugas, R A DiStasio, A Ferretti, A Floris, G Fratesi, G Fugallo, R Gebauer, U Gerstmann, F Giustino, T Gorni, J Jia, M Kawamura, H-Y Ko, A Kokalj, E Küçükbenli, M Lazzeri, M Marsili, N Marzari, F Mauri, N L Nguyen, H-V Nguyen, A Otero-de-la-Roza, L Paulatto, S Poncé, D Rocca, R Sabatini, B Santra, M Schlipf, A P Seitsonen, A Smogunov, I Timrov, T Thonhauser, P Umari, N Vast, X Wu, and S Baroni. "Advanced Capabilities for Materials Modelling with Quantum ESPRESSO". In: *Journal of Physics: Condensed Matter* 29.46 (Nov. 2017), p. 465901. DOI: 10.1088/1361-648X/aa8f79.

[162] Paolo Giannozzi, Oscar Baseggio, Pietro Bonfà, Davide Brunato, Roberto Car, Ivan Carnimeo, Carlo Cavazzoni, Stefano de Gironcoli, Pietro Delugas, Fabrizio Ferrari Ruffino, Andrea Ferretti, Nicola Marzari, Iurii Timrov, Andrea Urru, and Stefano Baroni. "Q UANTUM ESPRESSO toward the Exascale". In: *The Journal of Chemical Physics* 152.15 (Apr. 2020), p. 154105. DOI: 10.1063/5.0005082.

[163] Martin Schlipf and François Gygi. "Optimization Algorithm for the Generation of ONCV Pseudopotentials". In: *Computer Physics Communications* 196 (Nov. 2015), pp. 36–44. DOI: 10/f7tvm7.

[164] Jiming Sun, Bryan K. Clark, Salvatore Torquato, and Roberto Car. "The Phase Diagram of High-Pressure Superionic Ice". In: *Nature Communications* 6.1 (Nov. 2015), p. 8156. DOI: 10.1038/ncomms9156.

[165] Kristian Berland, Valentino R. Cooper, Kyuho Lee, Elsebeth Schröder, T. Thonhauser, Per Hyldgaard, and Bengt I. Lundqvist. "Van Der Waals Forces in Density Functional Theory: A Review of the vdW-DF Method". In: *Reports on Progress in Physics* 78.6 (May 2015), p. 066501. DOI: 10/f3n3tz.

[166] T. Thonhauser, S. Zuluaga, C. A. Arter, K. Berland, E. Schröder, and P. Hyldgaard. "Spin Signature of Nonlocal Correlation Binding in Metal-Organic Frameworks". In: *Physical Review Letters* 115.13 (Sept. 2015), p. 136402. DOI: 10.1103/PhysRevLett.115.136402.

[167] D C Langreth, B I Lundqvist, S D Chakarova-Käck, V R Cooper, M Dion, P Hyldgaard, A Kelkkanen, J Kleis, Lingzhu Kong, Shen Li, P G Moses, E Murray, A Puzder, H Rydberg, E Schröder, and T Thonhauser. "A Density Functional for Sparse Matter". In: *Journal*

*of Physics: Condensed Matter* 21.8 (Feb. 2009), p. 084203. DOI: 10.1088/0953-8984/21/8/084203.

[168] T. Thonhauser, Valentino R. Cooper, Shen Li, Aaron Puzder, Per Hyldgaard, and David C. Langreth. "Van Der Waals Density Functional: Self-consistent Potential and the Nature of the van Der Waals Bond". In: *Physical Review B* 76.12 (Sept. 2007), p. 125112. DOI: 10.1103/PhysRevB.76.125112.

[169] Nicola Marzari, David Vanderbilt, Alessandro De Vita, and M. C. Payne. "Thermal Contraction and Disordering of the Al(110) Surface". In: *Physical Review Letters* 82.16 (Apr. 1999), pp. 3296–3299. DOI: 10/fg796x.

[170] Aidan P. Thompson, H. Metin Aktulga, Richard Berger, Dan S. Bolintineanu, W. Michael Brown, Paul S. Crozier, Pieter J. in 't Veld, Axel Kohlmeyer, Stan G. Moore, Trung Dac Nguyen, Ray Shan, Mark J. Stevens, Julien Tranchida, Christian Trott, and Steven J. Plimpton. "LAMMPS - a Flexible Simulation Tool for Particle-Based Materials Modeling at the Atomic, Meso, and Continuum Scales". In: *Computer Physics Communications* 271 (Feb. 2022), p. 108171. DOI: 10.1016/j.cpc.2021.108171.

[171] Giovanni Bussi, Davide Donadio, and Michele Parrinello. "Canonical Sampling through Velocity Rescaling". In: *The Journal of Chemical Physics* 126.1 (Jan. 2007), p. 014101. DOI: 10.1063/1.2408420.

[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 (Feb. 2009), p. 074101. DOI: 10.1063/1.3073889.

[173] Michele Ceriotti, Giovanni Bussi, and Michele Parrinello. "Colored-Noise Thermostats à La Carte". In: *Journal of Chemical Theory and Computation* 6.4 (Apr. 2010), pp. 1170–1180. DOI: 10.1021/ct900563s.

[174] Ravit Helled, Guglielmo Mazzola, and Ronald Redmer. "Understanding Dense Hydrogen at Planetary Conditions". In: *Nature Reviews Physics* 2.10 (Oct. 2020), pp. 562–574. DOI: 10/gk5m37.

[175] Yan Wang, Xiulin Ruan, and Ajit K. Roy. "Two-Temperature Nonequilibrium Molecular Dynamics Simulation of Thermal Transport across Metal-Nonmetal Interfaces". In: *Physical Review B* 85.20 (May 2012), p. 205311. DOI: 10.1103/PhysRevB.85.205311.

# Curriculum Vitae

## Personal Information

Chiheb BEN MAHMOUD
Doctoral Assistant
Ecole Polytechnique Fédérale de Lausanne
COSMO - STI 1015 Lausanne
Switzerland

Email: bmahmoud.chiheb@gmail.com
G. Scholar: https://scholar.google.com/citations?user=hHRmxPgAAAAJ&hl
ORCiD: https://orcid.org/0000-0002-6695-1402

## Education

09/2018–now **Ph.D., Materials Science and Engineering**
Ecole Polytechnique Fédérale de Lausanne
**Advisor:** Prof. Michele Ceriotti
I am working under the supervision of Prof. Michele Ceriotti on applying machine-learning techniques to learn and predict electronic structure properties. In particular, my research project focuses on modelling the electronic density of states (DOS) as a multi-output target using geometrical descriptors. The main project lines are:

- building an atom-centred model for the DOS [1] using the SOAP descriptor and applying it to identify possible correlations between structural and electronic features in amorphous Silicon [1,2]
- utilizing the atom-centred model for the DOS to account for the electronic thermal excitations as a correction term for ground-state properties [3], and to build temperature-dependent machine-learning interatomic potentials by exclusively training on ground-state data and then extrapolating to any given temperature to perform finite-temperature simulations of materials at the warm dense matter conditions [4]

09/2014–12/2017 **MSc in Engineering, Physics and Applications**
*Ecole CentraleSupelec (ex Ecole Centrale Paris), France*
One of the top French engineering schools. I joined the research-optimized program in January 2015. Some relevant courses are quantum physics, non-equilibrium statistical physics, solid state physics, numerical atomistic simulations, nanomagnetism and spintronics. Average: 15/20.

09/2016–09/2017 **MSc, Nanosciences**
*University of Paris-Sud (Orsay), France*
One of the top universities in France. I am enrolled in a double degree program with Ecole CentraleSupélec. Some of the relevant courses are quantum optics, mesoscopic physics and nanophotonics. End of year project is validated with high honours. Average: 14.41/20.

09/2011–07/2014 **Classe preparatoire aux grandes écoles**
*Institut Préparatoire aux Etudes Scientifiques et Techniques, La Marsa, Tunisia*

High-level academic training in mathematics and physics (MP). The main subjects were Analysis, Algebra, Mechanics, Electromagnetism and Chemistry with additional courses in philosophy and English. At the end of my studies, I sat for a highly selective pathway to Ecole Centrale Paris.

## Internships and projects

05/2017–02/2018 **Ecole Polytechnique Fédérale de Lausanne, THEOS**
Supervisor: Prof. Nicola Marzari
Project: My project was about comparing the branching method vs the use of several thermostats (GLE and SVR) in extracting diffusion coefficients in the TIP4P water model and LLZO using MD. The effect of these thermostats on the tracer and charge diffusion coefficients and the Haven ratio was investigated.

03/2016–06/2016 **The Hong Kong University of Science and Technology**
Supervisor: Prof. Francesco Ciucci
Project: I studied the effect of oxygen vacancies on proton diffusion in Yttrium-doped Barium Zirconate using ReaxFF potential by LAMMPS MD simulations. I tried to look for possible correlations. The structure has become too soft.

## Tutoring experience

During my thesis preparation, I have been a TA in the following courses:

- Algèbre linéaire (MATH-111(g)): fall semesters of 2018, 2019 and 2020
- Statistical mechanics (MSE-421): spring semesters of 2019, 2020 and 2021
- MARVEL summer camp (for high school students): June 2021 and June 2022

I was also supervising the following Master student(s) for their semester projects:

- Rachel Wang, *Data-driven materials modelling: a primer*: Autumn semester 2019

## Journal Publications

1. **Ben Mahmoud, C.**, Anelli, A., Csányi, G., Ceriotti, M., 2020. Learning the electronic density of states in condensed matter. Phys. Rev. B 102, 235130.
2. Deringer, V.L., Bernstein, N., Csányi, G., **Ben Mahmoud, C.**, Ceriotti, M., Wilson, M., Drabold, D.A., Elliott, S.R., 2021. Origins of structural and electronic transitions in disordered silicon. Nature 589, 59–64.
3. Lopanitsyna, N., **Ben Mahmoud, C.**, Ceriotti, M., 2021. Finite-temperature materials modelling from the quantum nuclei to the hot electrons regime. Phys. Rev. Materials 5, 043802
4. **Ben Mahmoud, C.**, Grasselli, F., Ceriotti, M., 2022. Predicting hot-electron free energies from ground-state data. Phys. Rev. B 106, L121116.

## Conferences and oral presentations

- APS March Meeting 2021 (online)
- Psi-k 2022 conference
- DPG conference Regensburg 2022

## Honors, Awards & Scholarships

- I was awarded a 4-year-merit-based scholarship by the French government to pursue my studies at Ecole Centrale Paris (2014-2017)

## Extra tarining

- CAMD summer school 2022 (Denmark): a week-long summer school with a focus on the use of electronic structure theory in materials design.

## Languages

- Arabic (native)
- French (full proficiency)
- English (full proficiency)
- German (Notions)

## Digital competence

- Languages: Python, C++, HTML/CSS/Javascript
- Libraries: NumPy, SciPy, PyTorch, librascal, Chemiscope, ASE
- Programs: Quantum Espresso, FHI-aims, i-PI, LAMMPS