The structure and dynamics of materials using machine learning

Dissertation

zur Erlangung des Doktorgrades der Naturwissenschaften
Dr. rer. nat.

der

Naturwissenschaftlichen Fakultä II
Chemie, Physik und Mathematik

der Martin-Luther-Universität
Halle-Wittenberg

vorgelegt von

Herrn Gonçalves Marques, Mário Rui

geb. am 24.10.1992 in Coimbra, Portugal

Betreuer: Prof. Dr. Miguel A. L. Marques

1. Gutachter: Prof. Dr. Miguel A. L. Marques
2. Gutachter: Prof. Dr. Wolfgang Paul
3. Gutachter: Prof. Dr. Patrick Rinke

Datum der Abgabe: 17. Dezember 2019
Datum der öffentlichen Verteidigung: 5. Mai 2020
Vorsitzender der Promotionskommision: Prof. Dr. Jochen Balbach

Contents

Introduction **5**

1 Many body problem and density functional theory **9**
- 1.1 Many body problem . . . 10
- 1.2 Density functional theory . . . 13
  - 1.2.1 Hohenberg-Kohn theorems . . . 14
  - 1.2.2 Kohn-Sham scheme . . . 16
  - 1.2.3 Jacob's ladder . . . 17

2 Structural prediction and molecular dynamics **21**
- 2.1 Structure prediction . . . 22
  - 2.1.1 Minima hopping method . . . 24
  - 2.1.2 Genetic algorithms . . . 25
- 2.2 Molecular dynamics . . . 26
  - 2.2.1 Overview of molecular dynamics simulations . . . 26
  - 2.2.2 Berendsen thermostat and barostat . . . 27

3 Machine learning in material science **29**
- 3.1 Overview of machine learning . . . 29
- 3.2 Data . . . 32
- 3.3 Features . . . 33
- 3.4 Algorithms . . . 36
  - 3.4.1 Neural networks . . . 36
- 3.5 Recent applications . . . 40
  - 3.5.1 Applications to solids . . . 40
  - 3.5.2 Force-fields . . . 42

4 Neural Networks force fields **47**
- 4.1 Behler and Parrinelo neural networks in the ÆNET package . . . 47
- 4.2 Stress tensor . . . 48
- 4.3 Training neural networks for forces and stresses . . . 51
  - 4.3.1 Genetic algorithms . . . 52
  - 4.3.2 Back-propagation algorithm . . . 53
- 4.4 Example force-fields . . . 55
  - 4.4.1 Data . . . 56

4.4.2 Details on the features, algorithm and training 59
4.4.3 Validation 60
4.5 Applications 66
4.5.1 Phonon dispersion 66
4.5.2 Molecular dynamics and melting temperature 67
4.5.3 Structural prediction and defects 71
4.6 Interpretability of the neural networks 75

5 Copper based materials and cluster expansions 77
5.1 Cluster expansions 77
5.2 CZTS 79
5.2.1 Genetic algorithms and stability 81
5.2.2 Kesterite or stannite 84
5.3 Cuprous iodide 84
5.3.1 Copper vacancy complexes and stability 86
5.3.2 Phonon dispersion 89
5.3.3 Density of states 89

Conclusions and outlook 93

A DFT calculations with VASP 95

Introduction

> All animals are equal, but some animals are more equal than others.
>
> George Orwell
> Animal Farm

Equal... The search for new materials has been a quest present in every age since the start of our recorded history. Nowadays, we seek materials for various reasons. Here are a few examples: In terms of energy production, the search for renewable, efficient, and sustainable sources of energy led to the development of photovoltaics. Yet, in 2018, fewer than 4% of the total world energy consumption come from this source. Several factors contribute to this low percentage, however from a technical point of view, the efficiency of photovoltaic cells can still be increased with the improvement of the design of the devices and with the application of materials that possess optimal properties for the construction of these cells. An example of the latter is a semiconductor material with a direct band gap that absorbs in the visible range of the spectra and possesses high conductivity.

In terms of everyday gadgets, integrated circuits, and other electronic devices, the industry, and even consumers, expect developments at the pace of Moore's Law. This means that every year we expect a decrease in the size of transistors. Additionally, for many years, it was possible to increase the off and on switch rate of the transistors, which can be translated into higher computer (or processor) performance at every new generation of devices. However, this is no longer a possibility since some integrated circuits already reached fundamental thermal limits due to the ever increasing power consumption of these circuits. This promoted the research of new transistor designs with different materials and new architectures over the past 10 years.

Finally, we live in a digital universe characterized by an insatiable demand for information and by the creation of 2.5 quintillion bytes of data each day. Moreover, the pace at which this information is created has been increasing every year due the improved availability of electronic devices, i.e. more people are getting access to the internet and there is also an increase of the number of devices per person. Furthermore, the pace is also increasing with the development of new technologies, such as the internet of things, virtual reality, 5G, video surveillance, among others... This led to the creation of huge data centers for the purpose of storing and processing this information. They require complex cooling systems and when the volume of information increases, so does the power consumption. This motivated researchers to look for more efficient devices, which in many cases, involve the application of different materials.

One can see that the demand is high for materials exhibiting properties desired for the construction of several electronic devices. Although several materials already display some of these properties, in many cases their performance, relative abundance, price, and toxicity limit their large scale application. Remember that some materials are more equal than others. In other cases, the search for new materials just comes from the desire to generate more efficient devices.

Here we are going to provide a humble contribution towards the solution of this problem. In the past, the discovery of new materials came only from their experimental synthesis and characterization. This type of research was slow since experiments required expensive resources and were time consuming. Recently, the cost and time of materials design was greatly diminished due to the combination of experiments with computer simulations, in particular computational structural prediction methods [1–4]. This revolution was possible due to the increase of computer resources, and the development of electronic structure methods and their efficient implementation in computer packages. All of these allowed for innumerous high-throughput studies and for the creation of several databases containing information on known crystal structures and chemical substances. To give some examples, the Inorganic Crystal Structure Database contains information on 210,229 crystal structures [5], the Cambridge Structural Database on 1 million small-molecule organic and metal-organic crystal structures [6], and the Chemical Abstract Service registry on 157 million unique organic and inorganic chemical substances [7]. If we remove duplicates and alloys from the Inorganic Crystal Structure Database, we realize that, nowadays, we have information on 50 000 different inorganic materials. This number probably includes the most possible elemental substances and binary compounds, however it lacks many complex compounds, such as quaternary compounds.

The creation of all of these data is paving the way to yet another revolution in the field of material science: that of machine learning (ML). ML techniques take advantage of large amounts of data to find hidden patterns and a relation between input data and a certain target property. The application of these techniques to material science problems is recent and lacks the complexity exhibited in other fields. However, it has been shown already that ML methods further decrease the time necessary to find new stable phases and allow for a more efficient exploration of all estimated possible materials, which number around 10^{100} [8].

Often, structural prediction simulations require the evaluation of the total energies (forces and stresses) of millions of phases in order to search through the potential energy surface of a system. Among these, only a few points are indeed interesting: those corresponding to the minima of the energy surface, which might correspond to the ground-state structures. Many methods were developed to study the intricacies of the potential energy surface of a certain system, such as the minima hopping method [9, 10], random sampling [11, 12], and evolutionary algorithms [13, 14], to name just a few. Usually these methods use density functional theory [15–17] for the energy evaluations. However, studies with such methods are limited in the number of atoms included in the unit cell (usually no more than 10) since the number of minima, and therefore, the number of calculations required, grows exponentially with the number of atoms.

Similarly, molecular dynamics [18] simulations, used to calculate different properties of these materials, also require millions of evaluations of total energies of a system. Not only because they entail systems with more than millions of atoms, but also because they require long simulation times in order for the systems to reach equilibrium and to behave according to the ensembles of statistical mechanics. For these reasons, normally they employ classical force-fields for the energy evaluations (which come with a loss in accuracy), although we can find some smaller examples using density functional theory.

The aim of this thesis is to develop strategies to counter these obstacles using machine learning techniques. In particular, we resort to neural networks, genetic algorithms, and cluster expansions to speed-up first principles studies and to construct efficient, yet accurate, methods for energy evaluation. Although technically not part of the machine learning repertoire of tools, genetic algorithms and cluster expansions share some of their characteristics. Actually, a good introductory book on machine learning, Ref. [19], includes genetic algorithms as a type of machine learning since biological evolution can be seen as a learning process. Additionally, cluster expansions consist of a least-square fit of the total energies of a system based on a correlation matrix, which looks similar to several algorithms of supervised learning, which is the most common type of machine learning. Neural networks are, without a doubt, the most famous and the most successful of the machine learning algorithms.

This thesis is organized as follows. In chapter 1 we start our discussion from its foundation: the many body problem [20] and one of its most successful solutions: density functional theory. An efficient, accurate theory, that relies on an hypothesis and on an approximation. Afterwards, in chapter 2, we discuss the problems we intend to solve with density functional theory, namely structural prediction and molecular dynamics. These are the basis for the calculation of many interesting and important properties of materials. The obstacles that arise from these concern both simulation time and size, as well as the time required for a single calculation. An attempt to surpass them, by finding methods that are both accurate and efficient, revolves around machine learning [19, 21, 22], that we promptly discuss in chapter 3. In particular, we discuss applications of machine learning in material science [23]. This leads to the core of this work: neural networks force-fields [24]. We discuss them from their inception to the most recent research and then, in chapter 4, we present our methodology to construct neural network force-fields capable of describing the potential energy surface (PES) of solids using relatively small, unbiased training sets. To apply these force-fields in molecular dynamics and structural prediction simulations they have to provide accurate forces and stresses. Unfortunately, the high accuracy of the energy is not a sufficient condition to assure an appropriate accuracy for these derivatives of the energy. This led us to develop and implement methodologies to optimize the neural networks with respect to energies, forces, and stresses. Additionally, we use our results to show the challenges, limitations, and potential of such force-fields, and we discuss their interpretability. Moreover, our methodology permitted the study of large complex systems, such as the formation of defects in Si and the melting of metals with an accuracy comparable to density functional theory. We take on a different route in chapter 5, where we discuss and use cluster expansions to tackle the structural prediction of copper based materials. In particular, we use genetic algorithms to identify secondary phases of $Cu_2ZnSn(S,Se)_4$ (CZTS), which usually hinder the efficiency of solar cells made out of this photovoltaic material. Moreover we study the transition between the kesterite and the stannite phases in $Cu_2Zn_{1-x}SnFe_xSe_4$ compounds. Our last study involved the formation of complexes of defects in CuI [25], a transparent conducting semiconductor (TCS). Although the role played by Cu vacancies in the p-type transparent conductivity of CuI has been properly acknowledged, the way they arrange themselves, as well as their optimal and maximum concentrations remained unclear. Our objective was to provide an answer to these unresolved questions. We note that usually these three studies are too taxing to treat only with density functional theory.

Finally, at the end of this thesis we present our relevant conclusions.

Chapter 1

Many body problem and density functional theory

Who in the world am I? Ah, that's the great puzzle.

Lewis Carroll
Alice in Wonderland

Puzzle... The many-body problem [20, 26] represents perfectly how both complicated and unrewarding physics can be. For simplicity, think of an atom, or even a molecule, placed in any region of space and time, subject to whichever field, or fields, and basically try to name all the interactions that maintain its structure and keep the electrons bound to the nuclei, or not, if the field is strong enough... Now, neglect most of them. Keep only the interactions between electrons and nuclei, and their self-interactions. Oh, and consider that electrons and nuclei can move, but not very fast. One should try to avoid all those relativistic shenanigans as much as possible. In fact, for most of it, think of the movement of the electrons as slow, but instantaneous when compared to the movement of the nuclei. Moreover, remove most of the intricacies related to spin, specially its complicated interactions with orbitals or colinearity. At most consider that electrons can have spin up or down. Ah, what remains is the puzzle usually refereed to as the many-body problem. Now, try to solve it! The peculiarity of this problem is that, even after so many simplifications, it remains rather unsolvable... Strangely this is what motivates physicists the most. The joy of the challenge that dwell within every phenomena exhibited by matter. Not only that, but the elegance of nature, hidden behind the mysteries of the universe, and unveiled by that undoubtedly extraordinary approach. It is quite remarkable. [1]

In this chapter we introduce the many body problem and the Born-Oppenheimer approximation. Furthermore we explain how to obtain the most important properties in electronic structure theory. Afterwards, we present the most reasonable theory to solve the many body problem: density functional theory (DFT). We start with its foundation, the Hohenberg-Kohn theorem, and the hypothesis that lead to its most commonly used form: Kohn-Sham DFT. We finish the chapter with the Jacob's ladder and the description of the most used approximation to the exchange and correlation energy functional in material science.

---

[1] Adaptaded from Ref. [27]

1.1 Many body problem

The description of matter and its properties using theoretical methods starts with the interaction between N electrons and M nuclei, which can be cast as the Hamiltonian

$$\begin{aligned}\hat{H} = &-\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 + \sum_{i,I}\frac{Z_I e^2}{|\mathbf{r}_i - \mathbf{R}_I|} + \frac{1}{2}\sum_{i\neq j}\frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} \\ &-\sum_I \frac{\hbar^2}{2M_I}\nabla_I^2 + \frac{1}{2}\sum_{I\neq J}\frac{Z_I Z_J e^2}{|\mathbf{R}_I - \mathbf{R}_J|},\end{aligned} \tag{1.1}$$

where the lowercase subscripts denote electrons with mass m_e and charge e at position $\mathbf{r}_i$, and the uppercase subscripts denote nuclei with mass M_I and charge Z_I at position $\mathbf{R}_I$. Throughout this thesis we will adopt Hartree atomic units ($e = m_e = \hbar = 4\pi\epsilon_0 = 1$) to simplify equations. Under these units we can describe the terms of the aforementioned Hamiltonian as follows. The first term represents the kinetic energy operator for electrons

$$\hat{T} \; = \; -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2, \tag{1.2}$$

while the second term, the potential operator, represents the interaction between electrons and nuclei,

$$\hat{V} \; = \; \sum_{i,I=1}^{\text{N,M}} v(\mathbf{r}_i, \mathbf{R}_I), = -\sum_{i,I=1}^{\text{N,M}}\frac{Z_I}{|\mathbf{r}_i - \mathbf{R}_I|}, \tag{1.3}$$

and the third, is the electron-electron interaction operator,

$$\hat{W} \; = \; \frac{1}{2}\sum_{i\neq j}^{N} w(\mathbf{r}_i, \mathbf{r}_j) \; = \; \frac{1}{2}\sum_{i\neq j}^{N}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}. \tag{1.4}$$

The last two terms are the nuclei kinetic operator and the nuclei-nuclei interaction operator

$$\hat{T}_{\text{N}} \; = \; -\frac{1}{2m_I}\sum_{I=1}^{M}\nabla_I^2\;, \qquad \hat{W}_{\text{N}} \; = \; \frac{1}{2}\sum_{I\neq J}^{M}\frac{1}{|\mathbf{R}_I - \mathbf{R}_J|}. \tag{1.5}$$

As the Hamiltonian in eq. (1.1) is time-independent, the eigenstates of the fundamental equation governing a non-relativistic quantum system, i.e., the time dependent Schrödinger equation, consist on a phase modulation factor (e^{-iEt}) times the solution of the time-independent Schrödinger equation

$$\hat{H}(\{\mathbf{r}\},\{\mathbf{R}\})\Psi(\{\mathbf{r}\},\{\mathbf{R}\}) = E\Psi(\{\mathbf{r}\},\{\mathbf{R}\}), \tag{1.6}$$

where the abbreviation $\{\mathbf{r}\} = (\mathbf{r}_1, ..., \mathbf{r}_N)$ was used. The many-body wavefunction is a function of $3(N+M)$ spacial coordinates and $N+M$ spin coordinates. Unfortunately, the solution of this equation is usually an impossible task. However, close inspection of the many-body Hamiltonian provides a clue towards the simplification of the many-body problem. The nuclei kinetic term in the eq. (1.1) provides a small contribution to the energy, as the inverse mass of the nuclei $\frac{1}{m_I}$ stands as rather "small". As the electrons' mass is much smaller than the mass of the nuclei, when the nuclei move, the electrons appear to adjust their positions instantaneously. Therefore, the electrons move adiabatically with the nuclei. This is the reasoning behind the adiabatic or Born-Oppenheimer approximation [20].

The full solutions for the coupled system of electrons and nuclei $\Psi(\{\mathbf{r}\}, \{\mathbf{R}\})$ can be written in terms of functions of the nuclear coordinates $\xi_i(\{\mathbf{R}\})$ and electron wavefunctions $\Psi_i(\{\mathbf{r}\} : \{\mathbf{R}\}$, which depend upon the nuclear positions as parameters:

$$\Psi(\{\mathbf{r}\}, \{\mathbf{R}\}) = \sum_i \xi_i(\{\mathbf{R}\})\Psi_i(\{\mathbf{r}\} : \{\mathbf{R}\}). \tag{1.7}$$

In this manner, the time-independent Schrödinger equation 1.6 can be decoupled into an equation for the electrons

$$[\hat{T}(\{\mathbf{r}\}) + \hat{W}(\{\mathbf{r}\}) + \hat{V}(\{\mathbf{r}\}, \{\mathbf{R}\})]\Psi_i(\{\mathbf{r}\} : \{\mathbf{R}\}) \ = \ E_i \ \Psi_i(\{\mathbf{r}\} : \{\mathbf{R}\}) \tag{1.8}$$

and into a purely nuclear equation for each electronic state i:

$$[\hat{T}_{\mathrm{N}}(\{\mathbf{R}\}) + \hat{U}_i(\{\mathbf{R}\})] \ \xi_i(\{\mathbf{R}\}) \ = \ E \ \xi_i(\{\mathbf{R}\}) \tag{1.9}$$

In the previous equations $E_i(\{\mathbf{R}\})$ stands for the eigenvalues of the electron equation and $\hat{U}_i(\{\mathbf{R}\})$ represents a modified potential function for the nuclear motion that includes the interactions between nuclei and $E_i(\{\mathbf{R}\})$, among other terms (see Ref [20]).

In spite of the simplifications that arose from the adiabatic approximation, these equations are far too difficult to solve for a reasonable number of electrons and nuclei. Normally, in electronic structure, the equation for the nuclei is neglected in favour of a classical one while the equation for the electrons is further simplified. Before even attempting to solve eq. (1.8), it is necessary to understand how to extract the most important properties in electronic structure theory. These are the ground state total energy, the electron density, and excitations. Additionally, one can include two energy derivatives: the forces and the stress tensor.

The total energy of a system is given by the expectation value of the Hamiltonian

$$E[\Psi] = \frac{\langle\Psi| \hat{H} |\Psi\rangle}{\langle\Psi|\Psi\rangle}. \tag{1.10}$$

By definition, the ground state Ψ_0 is associated with the lowest energy. Consequently, the variation of the energy functional ($E[\Psi]$) with respect to the wavefunction has to be stationary for the ground state. In fact, this variation leads to the Schrödinger equation:

$$\frac{\delta E[\Psi_0]}{\delta \Psi_0^*} \ = \ \frac{\hat{H}\Psi_0}{\langle\Psi_0|\Psi_0\rangle} - \frac{\langle\Psi_0| \hat{H} |\Psi_0\rangle \ \Psi_0}{\langle\Psi_0|\Psi_0\rangle^2} = 0 \quad \text{hence} \quad \hat{H}\Psi_0 \ = \ E_0\Psi_0. \tag{1.11}$$

This can also be found by varying the energy subject to the constrain of orthonormality using the method of Lagrange multipliers

$$\delta[\langle\Psi| \hat{H} |\Psi\rangle - E(\langle\Psi|\Psi\rangle - 1)] = 0, \tag{1.12}$$

which is equivalent to the variational (Rayleigh-Ritz) principle. Additionally, a small deviation from the ground state wave function wields

$$E\left[\Psi_0+\delta\Psi\right]=\frac{\langle\Psi_0+\delta\Psi|\,\hat{H}\,|\Psi_0+\delta\Psi\rangle}{\langle\Psi_0+\delta\Psi|\Psi_0+\delta\Psi\rangle}=\frac{E_0\,\langle\Psi_0|\Psi_0\rangle+\langle\delta\Psi|\hat{H}|\delta\Psi\rangle}{\langle\Psi_0|\Psi_0\rangle+\langle\delta\Psi|\delta\Psi\rangle}=E_0+O\left(\delta\Psi^2\right). \tag{1.13}$$

Thus, the variational principle provides a strategy to discover approximations to the ground state wavefunction through energy minimization. Furthermore, the error for such approximation for the ground state converges with second order of the deviation.

In a similar fashion, the expectation value of the density operator

$$\hat{n}(\mathbf{r})=\sum_{i=1,N}\delta\left(\mathbf{r}-\mathbf{r}_i\right) \tag{1.14}$$

provides the electron density

$$n(\mathbf{r})=\frac{\langle\Psi|\hat{n}(\mathbf{r})|\Psi\rangle}{\langle\Psi|\Psi\rangle}=N\frac{\int\mathrm{d}^3r_2\cdots\mathrm{d}^3r_N\sum_{\sigma_1}\left|\Psi\left(\mathbf{r},\mathbf{r}_2,\mathbf{r}_3,\ldots,\mathbf{r}_N\right)\right|^2}{\int\mathrm{d}^3r_1\mathrm{d}^3r_2\cdots\mathrm{d}^3r_N\left|\Psi\left(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3,\ldots\mathbf{r}_N\right)\right|^2}. \tag{1.15}$$

As we will see later in this chapter, the electron density can be used to express all quantum mechanical observables as a functional of this real, scalar function of three variables. For example the expectation value of the interaction between electrons and nuclei (from eq. (1.3)) can be written as:

$$\langle\Psi|\hat{V}(\mathbf{r})|\Psi\rangle\;=\;\int\mathrm{d}^3r\;V_{\mathrm{ext}}(\mathbf{r})\;n(\mathbf{r}). \tag{1.16}$$

In condensed matter, excitations are nothing more than small perturbations of a system. Examples of excitations are variations of the ground state (e.g. small displacements of the ions in phonon modes) or true electronic excitations (e.g. optical electronic excitations). Therefore, after finding the ground state, perturbation theory techniques provide the appropriate tools to calculate excitations, such as excitation spectra and the real and imaginary parts of response functions.

Forces, on the other hand, can be calculated as in classical mechanics, as was noted by Ehrenfest [28] already in 1927, and by many that followed [20, 29–32]. The force (Hellman-Feynmann) theorem [33] was derived by Feynman in 1939 however, who explicitly demonstrated that the force on a nucleus depends strictly on the charge density and not on the electron kinetic energy, exchange, and correlation. The force conjugate to any parameter describing a system (e.g. the position of a nucleus $\mathbf{R}_I$) can always be expressed as

$$\mathbf{F}_I\;=\;-\frac{\partial E}{\partial\mathbf{R}_I} \tag{1.17}$$

Performing the differentiation, assuming $\langle\Psi|\Psi\rangle=1$ in eq. (1.10) for convenience, leads to

$$\frac{\partial E}{\partial\mathbf{R}_I}\;=\;\langle\Psi|\,\frac{\partial\hat{H}}{\partial\mathbf{R}_I}\,|\Psi\rangle\;+\;\langle\frac{\partial\Psi}{\partial\mathbf{R}_I}|\hat{H}|\Psi\rangle\;+\;\langle\Psi|\hat{H}|\frac{\partial\Psi}{\partial\mathbf{R}_I}\rangle\;=\;\langle\Psi|\,\frac{\partial\hat{H}}{\partial\mathbf{R}_I}\,|\Psi\rangle \tag{1.18}$$

where we used the fact that the Hamiltonian is Hermitian. Furthermore, we can explicitly write the Hamiltonian terms to reach

$$\mathbf{F}_I = \frac{\partial E}{\partial \mathbf{R}_I} = -\langle\Psi|\frac{\partial \hat{V}}{\partial \mathbf{R}_I}|\Psi\rangle = -\int \mathbf{d}^3 r n(\mathbf{r}) \frac{\partial V(\mathbf{r})}{\partial \mathbf{R}_I}. \tag{1.19}$$

Similarly, the stress (generalized virial) theorem [34–36] provides us a different type of variation. For a system in equilibrium, the stress tensor $\sigma_{\alpha\beta}$ derivation requires the application of an infinitesimal homogeneous scaling to the ground state, and the derivative of the energy with respect to the symmetric strain $\epsilon_{\alpha\beta}$

$$\sigma_{\alpha\beta} = -\frac{1}{\Omega}\frac{\partial E}{\partial \epsilon_{\alpha\beta}}. \tag{1.20}$$

Here Ω is the volume, and α and β Cartesian indices that come from the scaling of the space

$$\mathbf{r}_\alpha \to (\delta_{\alpha\beta} + \epsilon_{\alpha\beta})\, \mathbf{r}_\beta, \tag{1.21}$$

where $\mathbf{r}$ represents particle positions and translation vectors. Under this scaling the wavefunction changes [37] to

$$\Psi_\epsilon(\{\mathbf{r}_i\}) = \det(\delta_{\alpha\beta} + \epsilon_{\alpha\beta})^{-1/2}\, \Psi\left(\left\{(\delta_{\alpha\beta} + \epsilon_{\alpha\beta})^{-1} \mathbf{r}_{i\beta}\right\}\right). \tag{1.22}$$

Actually, the wavefunctions and the nuclear positions subjected to an expansion or a compression can change in different ways, however these other ways do not contribute to the energy (to first order) since the wavefunction and the nuclear positions are at their variational minima. The combination of the previous equations with eq. (1.10) and performing the integrations by changing variables results in

$$\sigma_{\alpha\beta} = -\langle\Psi| \sum_k \frac{\hbar^2}{2m_k}\nabla_{k\alpha}\nabla_{k\beta} - \frac{1}{2}\sum_{\substack{k,k' \\ k\neq k'}} \frac{(\mathbf{r}_{kk'})_\alpha (\mathbf{r}_{kk'})_\beta}{r_{kk'}} \left(\frac{\mathrm{d}}{\mathrm{d}r_{kk'}}\hat{V}\right)|\Psi\rangle, \tag{1.23}$$

where k and k' represent particles, and $r_{kk'}$ is the distance between them. The trace of the previous equation $P = -\sum_\alpha \sigma_{\alpha\alpha}$ amounts to the well known virial theorem for the pressure [38, 39]. This derivation in terms of the stretching of the ground-state was first derived by Fock in 1930 [40]. If only Coulomb interactions are present and if we include all the contributions from the nuclei and electrons in the potential energy, we can obtain the relation

$$3P\Omega = 2E_{\text{kinetic}} + E_{\text{potential}}. \tag{1.24}$$

1.2 Density functional theory

We must now explain why solving eq. (1.8) is unfeasible. Firstly, the computational methods required to solve this equation scale exponentially with the number of electrons. And that is just for finding the ground-state. Secondly the many-body wavefunction contains much more information than necessary. This can be further understood if we consider the example of

Ref. [41]. Consider the solution of an oxygen atom. Even neglecting spin, the wavefunction depends on 24 coordinates: 3 spatial coordinates for each of the 8 electrons. The solution of this problem usually requires a basis set or the discretization of space. So let us consider a small grid of 10 points per coordinate. This means that to store the wavefunction we need 10^{24} numbers. In scientific calculations, these numbers are usually stored as double precision floating points, with each requiring 64 bits. This means that to store the wavefunction of the oxygen atom we need roughly 8×10^{12} TB... From here we can speculate how tremendous the task of calculating additional properties would be, such as phonons or critical temperatures, or even solving the time dependent equation.

Luckily, there is a way to bypass these problems by considering the density as a basic variable and re-writing the energy of a quantum system as a functional of the density. This idea was first proposed by Thomas [42] and Fermi [43] in 1927. In their method, the system of electrons is described as a classical liquid and the kinetic energy is approximated as an explicit functional of the density. The Thomas-Fermi method was further developed in the following years. For example in 1930 Dirac [44] formulated the local approximation for the exchange, which was neglected in the original method. However, the formulation of density functional theory as we know today, i.e., as an exact theory of many-body systems only appeared in 1964 with the work of Hohenberg and Kohn [45].

Finally, we would like to note that DFT is not the only (approximate) solution to the many-body problem, although it is the most efficient for solids. Other theories worth mention are many-body perturbation theory [46–49] (in particular the *GW* approximation [50] and the Bethe-Salpeter equation [51]), coupled cluster [26], Hartree-Fock [20, 26], and full configuration interaction [26].

1.2.1 Hohenberg-Kohn theorems

At the heart of DFT lies the Hohenberg-Kohn theorems [15, 20, 52, 53], first proved by Hohenberg and Kohn by *reductio ad absurdum* and using the variational principle. The first of these theorems states that the external potential $\hat{V}(\mathbf{r})$ of a system of interacting particles is a unique functional of the ground state density, apart from a trivial additive constant. This means that the mappings A and B between ground state density, wavefunction, and the external potential are bijective (fully invertible)

$$\{\hat{V}\} \xleftrightarrow{A} \{\Psi_0\} \xleftrightarrow{B} \{n_0\}. \tag{1.25}$$

In other words, that there is a one to one correspondence between the ground state electron density and the external potential. Proofs of this theorem can be found in the aforementioned references or in any other DFT textbook [15–17]. Since the Hamiltonian and the wavefunctions are fully determined given the knowledge of the ground state density (up to a constant shift in the energy), all properties of the system are completely determined. This means that the nondegenerate ground-state wave function and the expectation value of any observable $\hat{O}$ are functionals of the ground-state density

$$O_0 = O[n_0] = \langle \Psi_0[n_0] | \, \hat{O} \, | \Psi_0[n_0] \rangle \, . \tag{1.26}$$

This is true in particular for the energy functional. The second of the Hohenberg-Kohn theorems reveals that an universal functional of the density $F[n(\mathbf{r})]$ can be defined for any number of particles or external potential. Then, for a certain external potential $V(\mathbf{r})$, an energy functional of the density $E[n(\mathbf{r})]$ can also be defined. Moreover, the ground-state density is the density that minimizes this functional, for which the functional provides the ground-state energy. Once again, proof of this theorem can be found in any other DFT textbook [15–17]. Here we will just show the definitions of these functionals:

$$E[n(\mathbf{r})] = \int d^3r\ v(\mathbf{r})n(\mathbf{r}) + F[n(\mathbf{r})] = \int d^3r\ v(\mathbf{r})n(\mathbf{r}) + \langle\Psi|\,\hat{T} + \hat{W}\,|\Psi\rangle \tag{1.27}$$

Clearly, for the ground-state density, the functional equals the ground-state energy. We note that the original proof of Hohenberg and Kohn is restricted to V-representable densities, i.e., densities that are the ground-state densities of the electron Hamiltonian with a particular potential. However, many reasonable densities have been shown to be non-V-representable. So, it is noteworthy to mention the constrained-search independent formulations of Levy [54–56] and Lieb [56–58] for the Hohenberg-Kohn functional. Their approach consists in defining a two-step minimization procedure, where the first step consists in minimizing the energy over the class of wavefunctions with the same density $n(\mathbf{r})$:

$$E_{\mathrm{LL}}[n(\mathbf{r})] = \int d^3r\ v(\mathbf{r})n(\mathbf{r}) + \min_{\Psi\to n(\mathbf{r})} \langle\Psi|\,\hat{T} + \hat{W}\,|\Psi\rangle\,. \tag{1.28}$$

This leads to an unique lowest energy for that density, and the ground-state density is found by minimizing this functional of the density

$$E_0 = \min_{\{n\}}\ E_{\mathrm{LL}}[n]. \tag{1.29}$$

Formally, this can be achieved by varying the energy subject to the particle number constrain using the Lagrange multiplier (μ):

$$\delta E = \delta\Big\{F[n] + \int d^3r\ v(\mathbf{r})n(\mathbf{r}) - \mu\ \Big(\int d^3r\ n(\mathbf{r})\ -\ N\Big)\Big\}\ =\ 0. \tag{1.30}$$

On top of that, carrying out the functional derivatives results in a Euler equation

$$\frac{\delta F[n]}{\delta n(\mathbf{r})} + v(\mathbf{r}) - \mu =\ 0. \tag{1.31}$$

In this formulation the energy functional is defined for any density obtained from a wavefunction Ψ_N for N-electrons, or in other words, for any N-representable density, which is fantastic since any reasonable density satisfy the N-representability condition [20, 59].

The Hohenberg-Kohn theorems reveal how to reformulate the many-body problem for N electrons resorting to the density instead of the wavefunction, and ensure that that density is sufficient to determine all properties of a system. However this relation is rather subtle and it still remains unclear how one can extract a set of properties directly from the density. This could have been an huge setback for density functional theory, if not for the brilliant Kohn-Sham approach [60].

1.2.2 Kohn-Sham scheme

The success of DFT and the reason why it is one of the most widely used methods for electronic structure calculations originates from the work of Kohn and Sham. Their approach consists in replacing the original many-body problem by an auxiliary independent particle problem [20]. To do so, however, we are forced to consider a hypothesis: the existence of an auxiliary non-interacting system whose ground-state density represents the ground-state density of an interacting system.

Now the ground-state wavefunction of such a system of N non-interacting electrons can be written as a Slater determinant of single-particle orbitals like

$$\psi(\mathbf{x_1},\mathbf{x_2}...,\mathbf{x_N}) = \frac{1}{\sqrt{N}}\begin{bmatrix}\varphi_1(\mathbf{x_1}) & \varphi_2(\mathbf{x_1}) & \dots & \varphi_N(\mathbf{x_1})\\ \varphi_1(\mathbf{x_2}) & \varphi_2(\mathbf{x_2}) & \dots & \varphi_N(\mathbf{x_2})\\ \vdots & \vdots & \ddots & \vdots\\ \varphi_1(\mathbf{x_N}) & \varphi_2(\mathbf{x_N}) & \dots & \varphi_N(\mathbf{x_N})\end{bmatrix},$$

where the orbitals that form the antisymmetric wavefunction satisfy the equation

$$\Big[-\frac{\nabla^2}{2} + v_{\rm s}[n](\mathbf{r})\Big]\varphi_i(\mathbf{r}) = \epsilon_i\varphi_i(\mathbf{r}). \tag{1.32}$$

Then, it is trivial to obtain the energy functional of such auxiliary system

$$E_{\rm s}[n] = T_{\rm s}[n] + \int d^3r\ v_{\rm s}(\mathbf{r})n(\mathbf{r}). \tag{1.33}$$

The constrained variation of this functional with respect to the density results in the Euler equation

$$\frac{\delta T_{\rm s}[n]}{\delta n(\mathbf{r})} + v_{\rm s}(\mathbf{r}) - \mu_{\rm s} = \ 0, \tag{1.34}$$

where $T_{\rm s}$ the non-interacting kinetic energy operator and m_s a Lagrange multiplier. The gist of the Kohn-Sham approach is to rearrange the terms of the energy functional in eq. (1.27) in a away that resembles the one above:

$$\begin{aligned} E[n] &= T[n] + W[n] + \int d^3r\ v(\mathbf{r})n(\mathbf{r})\\ &= T_{\rm s}[n] + (T[n] - T_{\rm s}[n]) + E_{\rm H}[n] + (W[n] - E_{\rm H}[n]) + \int d^3r\ v(\mathbf{r})n(\mathbf{r}) \qquad (1.35)\\ &= T_{\rm s}[n] + E_{\rm H}[n] + E_{\rm xc}[n] + \int d^3r\ v(\mathbf{r})n(\mathbf{r}). \end{aligned}$$

In this equation we used the definition of the self-interaction energy, that takes into account the classical Coulomb interaction between the electrons,

$$E_{\rm H}[n] = \frac{1}{2}\int d^3r \int d^3r'\ \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}. \tag{1.36}$$

also known as the Hartree energy, and that of the exchange and correlation energy functional

$$E_{\rm xc}[n] = T[n] - T_{\rm s}[n] + W[n] - E_{\rm H}[n]. \tag{1.37}$$

Again, we note that the rearrangement in equation (1.35) is only possible under the assumption that the ground state density of the interacting system can be represented as the ground state density of the non-interacting system. With this rearrangement, the Euler equation for the interacting system (eq. (1.27)) becomes

$$\frac{\delta T_{\mathrm{s}}[n]}{\delta n(\mathbf{r})} + v(\mathbf{r}) + v_{\mathrm{H}} + v_{\mathrm{xc}} - \mu = 0. \tag{1.38}$$

Turns out that the Euler equations for both systems (eqs. (1.34) and (1.38)) are equivalent if

$$v_{\mathrm{s}} \;=\; v(\mathbf{r}) + v_{\mathrm{H}} + v_{\mathrm{xc}} - (\mu - \mu_{\mathrm{s}}) \;=\; v(\mathbf{r}) + v_{\mathrm{H}} + v_{\mathrm{xc}}. \tag{1.39}$$

where the difference between Lagrange multipliers was absorbed by the exchange and correlation term. In this manner, one can calculate the variation of the energy functional in eq. (1.35) with respect to the wavefunctions of the auxiliary system and obtain the well known equations of Kohn and Sham

$$\left[-\tfrac{\nabla^2}{2} + v(\mathbf{r}) + v_{\mathrm{Hartree}}[n](\mathbf{r}) + v_{\mathrm{xc}}\left[n\right]\left(\mathbf{r}\right)\right]\psi_i(\mathbf{r}) = \varepsilon_i \psi_i(\mathbf{r}) \ . \tag{1.40}$$

The exchange and correlation potential and the Hartree potential are defined as the functional derivative of their energy counterparts, for example

$$v_{\mathrm{xc}}\left[n\right]\left(\mathbf{r}\right) = \frac{\delta E_{\mathrm{xc}}\left[n\right]}{\delta n(\mathbf{r})}. \tag{1.41}$$

Finally, the electronic density can be calculated as

$$n(\mathbf{r}) = \sum_{i}^{\mathrm{occ.}} \left|\psi_i(\mathbf{r})\right|^2 , \tag{1.42}$$

where the sum runs over all the occupied states.

The beauty of the Kohn-Sham equations is that the solution of the many-body problem no longer involves complicated wave-functions of many-body interacting electrons. The ground-state density is now obtained from several single-particle wave-functions that obey a Schrödinger like equation with an extremely complex potential. And if this potential were known, the self-consistent-cycle solution of the Kohn-Sham equations would yield the exact ground state density and energy for the interacting system.

1.2.3 Jacob's ladder

The elusive exchange and correlation energy functional of Kohn-Sham DFT is often split in two terms: the exchange functional (E_x) and the correlation functional (E_c):

$$E_{\mathrm{xc}}[n] = E_{\mathrm{x}}[n] + E_{\mathrm{c}}[n] = \int d^3r \; n(\mathbf{r})[\epsilon_x(n(\mathbf{r})) + \epsilon_c(n(\mathbf{r}))], \tag{1.43}$$

where ϵ_i represents energy per electron of the system. We use this notation throughout this section. Moreover, these functionals can be defined as [61]

$$\begin{aligned} E_x[n] &= \langle\Phi_s[n]|\,\hat{W}|\,|\Phi_s[n]\rangle - \frac{1}{2}\int d^3r d^3r' n(\mathbf{r})n\left(\mathbf{r}'\right) w\left(|\mathbf{r}-\mathbf{r}'|\right) \\ E_c[n] &= \langle\Psi[n]|\,\hat{T}+\hat{V}+\hat{W}\,|\Psi[n]\rangle - \langle\Phi_s[n]|\,\hat{T}+\hat{V}+\hat{W}\,|\Phi_s[n]\rangle \end{aligned} \tag{1.44}$$

where $|\Phi_s[n]\rangle$ is the Kohn-Sham wavefunction and $|\Psi[n]\rangle$ the true ground state wavefunction of the interacting system with density n. Nonetheless, the exact mathematical expression for these functionals of the density, that should describe many-body interactions, is not known. One could, in principle, construct an explicit form for the functional after solving all possible electronic systems. Yet this approach remains rather unfeasible. So, as it usually happens, approximations are required and, over the last decades, hundreds of approximations for the exchange and correlation functionals were proposed [62–65]. The major difficulty, however, concerns the impossibility of systematically improving these functionals of the density. Nothing guarantees that the addition of further ingredients, that satisfy more exact constrains or that make the functional more flexible, leads to an improvement in the description of all types of interactions across distinct chemical environments. Still, an hierarchy has been in development since the work of Kohn and Sham [60], an hierarchy coined by Perdew as "Jacob's ladder" [66]. The metaphor starts with the Hartree world at the bottom, where exchange and correlation energies are zero and classical electrostatics perfectly describes the interaction between electrons. Then every rung represents the inclusion of a different ingredient in the functional and the belief is that the ladder culminates in the Heaven of chemical accuracy. This concept of chemical accuracy was further discussed by Pople in his Nobel prize lecture [67], where he argues that a global accuracy of 1 kcal/mol (roughly 43 meV/atom) for energies with respect to experimental values would be appropriate.

The first rung consists on the local spin density approximation (LDA), where the functional depends on the density as in

$$E_{\mathrm{xc}}^{\mathrm{LDA}} = \int d^3r\, n(\mathbf{r})\, \epsilon_{\mathrm{xc}}^{\mathrm{LDA}}\left(|n(\mathbf{r})|\right). \tag{1.45}$$

The second rung sees the inclusion of the gradient of the density in the approximation

$$E_{\mathrm{xc}}^{\mathrm{GGA}} = \int d^3r\, n(\mathbf{r})\, \epsilon_{\mathrm{xc}}^{\mathrm{GGA}}\left(|n(\mathbf{r})|\,, |\nabla n(\mathbf{r})|\right) \tag{1.46}$$

This is the form of the generalized-gradient approximation (GGA), and functionals of this family are often regarded as semi-local due to the infinitesimal region around $\mathbf{r}$ spanned by the gradient. Following the trend, the ingredient required for the third rung is the Laplacian of the density $\nabla^2 n(\mathbf{r})$ and/or similar quantities, such as kinetic energy density

$$\tau(\mathbf{r}) = \frac{1}{2}\sum_i^{\mathrm{occ}} |\nabla\psi_i(\mathbf{r})|^2\,. \tag{1.47}$$

With these quantities the functional form changes to that of a meta-GGA (mGGA):

$$E_{\mathrm{xc}}^{\mathrm{mGGA}} = \int d^3r\, n(\mathbf{r})\, \epsilon_{\mathrm{xc}}^{\mathrm{mGGA}}\left(|n(\mathbf{r})|\,, |\nabla n(\mathbf{r})|\,, |\tau(\mathbf{r})|\,, \left|\nabla^2 n(\mathbf{r})\right|\right) \tag{1.48}$$

The fourth rung takes on a different approach: the inclusion of terms involving the dependence on the occupied Kohn-Sham orbitals, which can be achieved in different manners. For example, by including exact exchange and a compatible correlation, i.e. mixing a fraction α_{x} of exact (Hartree-Fock) exchange with another functional (either GGA or mGGA) as in

$$E_{\mathrm{xc}}^{\mathrm{hyb}} = -\frac{\alpha_{\mathrm{x}}}{2} \int d^3r \int d^3r' \frac{\psi_i(\mathbf{r})\psi_i^*(\mathbf{r}')\,\psi_j(\mathbf{r}')\,\psi_j^*(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} + E_{\mathrm{xc}}^{\mathrm{DFT}}[n]. \tag{1.49}$$

This is the form of a hybrid functional and it can truly be regarded as truly non-local due to the exact exchange term. The last known rung, for which we do not show the functional form, incorporates the dependence on all orbitals (both occupied and unoccupied). Functionals of this kind contain exact exchange and exact partial correlation. Examples include the inclusion of post-Hartree-Fock correlation in the approximation denoted as double-hybrids [68] and the random phase approximation plus corrections [66].

Among all these families of functionals, the GGA proposed by Perdew, Burk and Ernzerhof (PBE [69]) stands out, as it is the most used functional in material science. Many reasons contributed to this, for example while the most accurate approximation for the exchange and correlation energy, the hybrid functionals, require the computation of exact exchange, which is a rather demanding operation in plane wave codes, the PBE does not. It is in fact very computational efficient. Furthermore it provides a fairly accurate reproduction of crystal structures energies and lattice constants (among other properties). Furthermore, the PBE was presented at the correct time and witnessed the boom of electronic structure packages. It was used for many calculations and re-used for comparison reasons. Not only that, but it was the approximation of choice for many databases, such as the Materials Project [70].

The exchange part of this approximation is given by [16, 71]

$$\begin{aligned} \epsilon_x^{\mathrm{PBE}} &= \epsilon_x^{\mathrm{LDA}} F_x^{\mathrm{PBE}} \\ F_x^{\mathrm{PBE}} &= 1 + k - \frac{k}{1 + (\mu s^2/k)}, \end{aligned} \tag{1.50}$$

where $k = 0.804$ and $\mu = 0.21951$. The dimensionless gradient s is defined it terms of the gradient of the density, the density, and the local Fermi wave vector as in

$$s = \frac{|\nabla n|}{2k_F n}. \tag{1.51}$$

While the correlation part is chosen as [16, 71]

$$\epsilon_{xc}^{\mathrm{GGA}} = \epsilon_c^{\mathrm{LDA}} + H(r_s, \zeta, t), \tag{1.52}$$

where $\zeta = (n^{\uparrow} - n^{\downarrow})/n$ is the spin polarization, $t = |\nabla n|/(2\phi k_s n)$ is a dimensionless gradient, and r_s is the local Seitz radius. In the definition of t, k_s is the Thomas-Fermi screening wave vector and $\phi = [(1+\zeta)^{2/3} + (1-\zeta)^{2/3}]/2$ is just a spin scaling factor. Using these definitions, we can write H as

$$H = \frac{e^2\gamma\phi^3}{a_0} \log\big(1 + \frac{\beta t^2}{\gamma} \frac{1 + At^2}{1 + At^2 + A^2t^4}\big). \tag{1.53}$$

Here, γ and β are constants and the function A represents

$$A = \frac{\beta}{\gamma}\left[e^{\frac{-a_0\epsilon_c^{\mathrm{LDA}}}{e^2\gamma\phi^3}} - 1\right]^{-1}. \tag{1.54}$$

Chapter 2

Structural prediction and molecular dynamics

> You start a question, and it's like starting a stone. You sit quietly on the top of a hill; and away the stone goes, starting others...
>
> Robert L. Stevenson
> The Strange Case of Dr. Jekyll and Mr. Hyde

Starting... In the last chapter we explained Kohn-Sham DFT. In the words of Richard Martin [20]: "So long as the true many-body solution is sufficiently close to the independent-particle formulation, e.g. the states must have the same symmetry, then the Kohn-Sham approach provides insightful guidance and powerful methods for electronic structure theory." Kohn-Sham DFT is indeed a remarkable theory that allows for the determination of the ground-state of a system. Actually, how can we be sure that we reached the ground state of a system? Surely performing a self-consistent calculation for a random configuration of 8 copper atoms followed by a geometry relaxation should result in a supercell of its ground state fcc lattice. Unfortunately, this is not always the case: the minimization procedure might encounter a local minima of the potential energy surface (PES), or even a saddle point. This simple example reveals a particular difficult problem when considering solids, which pertains the determination of a ground-state crystal structure of a system from only its chemical composition.

From a different point of view, often the study of a system involves the determination of dynamical properties and the simulation of the system under a specific statistical ensemble. For example, the latter requires an accurate evaluation of forces of large supercells in order to properly simulate the evolution of a system, while keeping certain properties of the system constant, such as the total pressure or temperature.

The solution to these problems took many years to be found and as often occurs in science, such development came from many brilliant contributions that still continue to start others. In this chapter we describe two types of simulations particularly relevant in the field of materials science: structural prediction and molecular dynamics (MD). We focus more prominently it the flavours that we used for our applications, mainly the minima hopping method (MHM), genetic algorithms, and the Berendsen thermostat and barostat.

2.1 Structure prediction

In 1988 John Maddox draw attention to a considerable problem in material science in a Nature editorial. He wrote: "One of the continuing scandals of physical science is that it remains in general impossible to predict the structure of even the simplest crystalline solids from a knowledge of their chemical composition."[72] In fact, at the time, crystal structures were considered as unpredictable as the behaviour of the stock exchange. However, the situation change dramatically in 2003-2006 with the proposal of several approaches that aimed at the solution of this problem, only possible due to the explosive development of electronic structures methods. The idea behind structure prediction can be extract from Maddox text: find the stable crystal structure of a certain material knowing only its chemical compositions (for specific thermodynamic conditions).

Here stability pertains to both thermodynamic and dynamical stability. Thermodynamic stability means the minimum of the Gibbs free energy. So a structure will not decompose into another, if the free energy of the decomposition channel is positive. Frequently, many calculations resort to conditions of zero pressure and temperature. In these conditions, the relevant thermodynamic quantity is the total energy, and a structure is stable if its energy is lower than that of its individual constituents. For example, consider the formation energy of a binary compound $\mathrm{A}_i\mathrm{B}_j$:

$$E_F = E_{A_iB_j}/N - (x_iE_A + x_jE_B), \tag{2.1}$$

where E denotes energy, x concentrations, and N the total number of atoms in the structure. The formation energy is then an energy per atom. If the formation energy is negative, then $\mathrm{A}_i\mathrm{B}_j$ is more stable than its constituents. Then, to evaluate thermodynamic stability we just have to compare energies with respect to all possible decomposition channels. On the other hand, dynamical stability concerns the second derivatives of the energy and the identification of minima among stationary points of the PES (points whose first derivatives are zero). For example, local optimizers search for points that minimize the forces that a structure is subjected to. However these points might correspond to saddle points of the energy surface instead of minima. Their proper identification requires the second derivative test or the calculation of related properties, such as phonon frequencies. For instance, a structure incorrectly deemed as thermodynamic stable, due to the lack of information on all possible decomposition channels, might be proper identified as unstable if it displays imaginary phonon frequencies, which indicate dynamical instability.

But why is this problem of structure prediction so complex? Well, each chemical composition is associated with an infinite number of atomic arrangements, and from these no one knows how many correspond to a local minima of the free energy surface. Furthermore, among these, it is also unknown which ones are the most stable and most probable to be synthesised in a laboratory. Fortunately, several approaches were developed to tackle this problem, namely topological approaches [2], approaches based on empirical correlations using either structural diagrams [73–75] or data mining approaches [76, 77], and approaches based on computational optimization [2]. Topological approaches rely on known information of the chemistry and symmetry of the system. For example, knowledge of sp^3-hybridization of carbon atoms leads to the diamond structure. Approaches based on empirical correlations require large databases of know stable crystal structures, and employ machine learning techniques or structural diagrams of basic properties (such as ionic radius and Mendeleev number), to uncover patterns among the data of similar structures. On the other hand, computational optimization consists in explicitly performing calculations to explore the free energy surface with the objective of finding its minima. Contrary to the other approaches that are biased, this approach can lead to completely unforeseen results and novel structures.

Still this is not an easy task. First of all, exhaustive search optimization techniques have to be discarded since crystal structure prediction is an high-dimensional problem, in an extremely complex landscape, and admits an enormous number of feasible solutions. To make matters worse, this problem scales exponentially with the system size, as the number of points in the energy landscape can be obtained from

$$C = \binom{V/\delta^3}{N} \prod_i \binom{N}{n_i}, \tag{2.2}$$

where V is the volume of the unit cell that contains N atoms, δ is a discretization parameter, and n_i represents the number of atoms in the unit cell with type i. Moreover, the number of local minima depends exponentially on the dimentionality d of the energy landscape, which can be calculated as

$$d^* = 3N + 3 - \kappa, \tag{2.3}$$

where $3N - 3$ degrees of freedom come from the atomic positions, 6 from the lattice parameters, and the non-integer κ represents the number of correlated dimensions. Luckily, performing structure relaxations provides a simplification to the problem. In fact, for some cases structure relaxation greatly reduces the dimensionality of the problem. For example, during relaxations the interatomic distances adjust to sensible, physical values.

This implies that the best methods for structure prediction contain both a local and a global optimization procedure. Over the years, several methods were developed to tackle this exceptionally difficult problem, such as random sampling [11, 12, 78, 79], simulated annealing [80–82], metadynamics [83], minima hopping [9], and evolutionary algorithms [3, 13, 14, 84–88]. In the next sections we proceed with the discussion of both the minima hopping method (MHM) and genetic algorithms due to their importance in the scope of this thesis. However, before proceeding we are going to discuss extensions to the problem just described.

Often, the interest is not in the knowledge of a stable crystal structure, but on the stable crystal structure that exhibits a certain property. This can be achieved with an hybrid optimization procedure that combines a local optimization for the energy with a global optimization for the property of interest [3, 86].

Another extension to the problem consists in the addition of the other chemical compositions, and subsequent prediction of the entire set of stable chemical compositions. In this manner, the complexity of the problem increases, as the energy landscape depends now on both compositional and structural coordinates. Moreover, the end result will now include the minima for all compositions and the set of ground-states located on the convex hull of thermodynamic stability. This convex hull is the hyper-surface in composition space that contains all the thermodynamic stable materials, i.e., materials that lack any decomposition channel and that, therefore, will not decompose into other (more stable) phases. The distance to the convex hull represents the energy (or free energy) released in the decomposition.

2.1.1 Minima hopping method

The MHM [9, 10] is a global prediction method or global optimization procedure. This algorithm can be placed with genetic algorithms and particle swarm, as methods that do not rely on thermodynamic principles and Markov-based Monte Carlo methods, such as as simulated annealing, basin hopping, and multicanonical methods.

The efficiency of a global optimization procedure can be understood as how fast it can climb out of wrong basins (local minima) and consequently find the global minima of a complex function, such as a potential energy surface.

Figure 2.1 shows three distinct steps in a usual MHM run. The MHM consists in 2 parts: an inner part that performs jumps into the local minimum of different basins and an outer part that will accept or reject this local minimum.

Minima are accepted or rejected based on thresholding: if the difference in energies between the new and the current minimum is smaller than a certain variable tolerance, the minimum is accepted. In this manner, there is a preference for steps that lower the energy, yet steps that increase it can also be accepted.

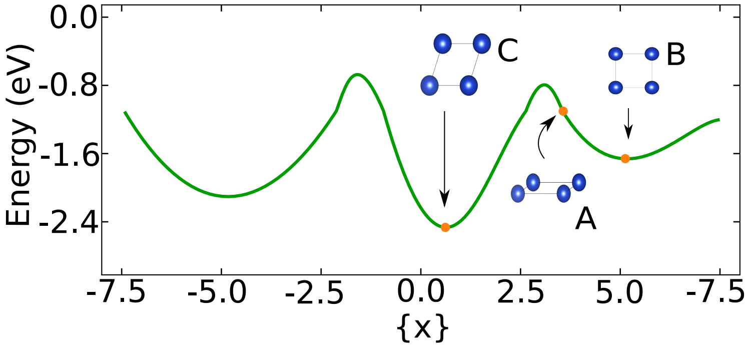


Figure 2.1: Scheme of a MHM run. A) Initial random structure that can be anywhere in the PES. B) Geometry optimization leads to a minimum of the PES. C) After several geometry optimizations and short molecular dynamics the method finds the global minimum of the PES.

On the other hand, the inner part relies on short molecular dynamics to escape from the current local minimum, followed by geometry relaxation to the closest local minimum. The geometry optimization can be done by standard steepest descent and conjugate gradient methods, or any other local optimization method. Usually we use the optimization procedure of the code we are using, in this case the geometry optimization of vasp. Now, the molecular dynamics simulations are not physical. The initial velocities of the atoms are chosen according to a Boltzmann distribution, in such a manner that the total kinetic energy is equal to a certain parameter γ_{kin}. In this manner, the system has sufficient energy to cross over any barrier of height up to the value of this parameter. Also in play here is the Bell-Evans-Polanyi principle. The principle states that highly exothermic chemical reactions have a low activation energy. In the context of global optimization this means that is more probable to find low energy minima if the jump between basins overcomes a low barrier rather than a high barrier. In practise, large γ_{kin} leads to a larger search space, while smaller values require extra MD simulations to find an escape path. It turns out that dynamically varying γ_{kin} during the simulation, such that half of the MD simulations find a new basin is almost optimal.

2.1.2 Genetic algorithms

Genetic algorithms (GA) are random-based classical evolutionary algorithms [89, 90]. As the name suggests, these algorithms draw inspiration from natural evolution processes, where a population competes for limited resources within an environment, which leads to natural selection (or survival of the fittest). Although initialy proposed by Holland [91] to study adaptive behaviour, genetic algorithms are widely known as optimization methods ever since their earlier successes reported in the works of Goldberg [92], De Jong [93], among others. Particularly in material science, genetic algorithms are renown global structure optimizers.

Traditional, genetic algorithms follow a rather fixed workflow. They start with the generation of a population of μ individuals. Though duplicate individuals might be present, the diversity of the initial population improves the efficiency of the algorithm. Each individual is characterized by a genotype and a fitness value. Actually, a more rigorous description would also include phenotype, the observable characteristics of the collection of genes (genotype).

For structural prediction each individual might represent a structure for a certain chemical composition, the phenotype can then be the list of different atoms in the structure (Si or Ge for example), while its encoding is the genotype (this could be an array with zeros for Si and ones for Ge).

After the generation comes the evaluation: the individual fitness value comes from the evaluation of the population using a fitness function, which represents the requirements the population should adapt to meet. Following our example, this can be the calculation of the formation energy of each structure from the collection of genes, and it can even include a local optimization, such as a geometry relaxation.

The next step involves the parent selection: pairs of individuals are selected to become the parents of the next generation. Normally, this selection depends on the fitness function, so that the best individuals (those with higher fitness value) have a higher chance of being selected. Several algorithms exist to perform this selection: fitness proportional selection, ranking selection, tournament selection, uniform parent selection, and over-selection for large populations. Their description can be found in any genetic algorithms book, such as Ref [89].

Afterwords, it is time to apply (with a certain probability) the variation operators, namely the mutation and the recombination (or crossover) operators. These operators can be implemented in a plethora of ways and the efficiency of the genetic algorithms is once again tied with their quality. As the name implies, the crossover operator merges the information from the genotypes of the parents in one or several offspring genotypes. For example, an one-point crossover divides the genotype of both parents at the same point and creates one or two offsprings by merging the portions from different parents (and keeping the same length). So, a parent with only Si atoms can recombine with another with just Ge atoms to form a structure with 3 Si and 1 Ge. On the other hand, the mutation operator involves a stochastic slight variation of a genotype and usually occurs with a low probability. When it occurs though, it increases the diversity of a population and it might stear the optimization procedure from persistent local minima. Figure 2.2 displays examples of well known mutation operators.

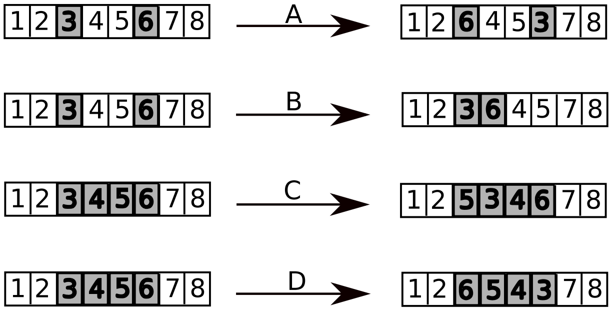


Figure 2.2: Swap (A), insert (B), scramble (C), and inversion (D) mutation examples.

Finally, a new population arises from survivor selection and the algorithm repeats itself until a certain number of populations have been created or until a certain fitness value is achieved. The new population of μ individuals is selected from the previous μ individuals and from the λ offprings. A handful of methods exist to perform this replacement based on age or on fitness: replace worst, elitism, round-robin tournament, $(\mu + \lambda)$ selection, and (μ, λ) selection. Again their description can be found in Ref. [89].

2.2 Molecular dynamics

Another kind of simulations that remain particularly relevant in the field of material science and quantum chemistry are molecular dynamics (MD) simulations, as they allow to obtain dynamic properties of many particle systems.

2.2.1 Overview of molecular dynamics simulations

MD simulations have been thoroughly used to model statistical ensembles and to study innumerous properties such as thermal conductivity [94, 95], critical temperatures [96–98], the folding of proteins [99–101], dynamical stability [102], among others. The term MD designates the solution of the classical equations of motion (usually Newton's equations) for a set of molecules [18] and was first associated with the simulation of a system of hard spheres by Alder and Wainwright [103, 104]. Usually the equations of motion are integrated using the Störmer-Verlet algorithm [105, 106] (in particular velocity Verlet), as this method to solve differential equations possess 3 important properties: reversibility, symmetry, and symplecticity. When solving equations of motion, reversibility broadly means that inverting the initial velocities only changes the direction of the movement, symmetry that if we reverse time at any step, we can return to the initial position, and symplecticity that the energy is nearly conserved provided that a sufficiently small time-step is used. It also implies that the volume in phase-space is conserved in the flow (one-step map).

The Störmer-Verlet integration of the equations of motion leads to simulations that preserve the volume, the total energy, and the number of atoms of a system, i.e. simulations that reproduce a microcanonical ensemble. In a similar fashion, other statistical ensembles can by reproduced by adding additional terms to the equations of motion, to allow for the preservation of other quantities, such as temperature and pressure. Physically this entails the coupling of a system to an external heat bath or its placement in a pressure bath. In practise these constrains are achieved with thermostats and barostats. To simulate a canonical ensemble, where the temperature, the volume and the number of particles remain constant, we can resort to the Langevin [18, 107], the Nosé-Hoover [108], the Andersen [109] or the Berendsen thermostats. On the other hand, the pressure can be preserved by resorting to barostats such as the Nosé-Hoover [18], the Berendsen [110], and the Parrinello-Rahman [111, 112] barostats. Furthermore, by coupling both a thermostat and a barostat, simulations can preserved the number of atoms, the pressure, and the temperature of a system, i.e., they can reproduce an isothermal-isobaric (or NPT) ensemble. The next section describes the thermostat and barostat used in the MD simulations described in section 4.5.2.

2.2.2 Berendsen thermostat and barostat

Coupling a system to a heat bath can be accomplished by inserting stochastic and friction terms in the equations of motion. Starting from the Newton's equation of motion, this results in the Langevin equation

$$m_i \dot{v}_i = F_i - m_i \gamma_i v_i + R_i(t), \tag{2.4}$$

where γ_i represents damping constants and R_i a Gaussian stochastic variable with zero mean and with intensity

$$\langle \boldsymbol{R}_i(t) \boldsymbol{R}_j(t+\tau) \rangle = 2 m_i \gamma_i \boldsymbol{k} T_0 \delta(\tau) \boldsymbol{\delta}_{ij}. \tag{2.5}$$

The idea behind the Berendsen thermostat is to consider how such coupling affects the temperature T of the system. According to the equipartition theorem, this can be obtained by calculating the time derivative of the kinetic energy. After some integrations this results in

$$\frac{dE_k}{dt} = \sum_{i=1}^{3N} v_i F_i + 2\gamma \left(\frac{3N}{2} k T_0 - E_k \right), \tag{2.6}$$

which contains the derivative of the potential energy and an additional term describing the coupling to the heat bath. In terms of temperature, this extra term can be written as

$$\left(\frac{dT}{dt} \right)_{bath} = 2\gamma \left(T_0 - T \right). \tag{2.7}$$

However, a similar variation can be obtain by just changing the equations of motion to

$$m_i \dot{v}_i = F_i + m_i \gamma \left(\frac{T_0}{T} - 1 \right) v_i. \tag{2.8}$$

where 2γ can be see as the inverse of a coupling constant ($\tau_T = 1/2\gamma$). Finally, this corresponds to a proportional scaling of the velocities at every time step ($v \to \lambda v$) with

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T} \left(\frac{T_0}{T} - 1 \right) \right]^{1/2}. \tag{2.9}$$

Similarly, the coupling to an constant pressure bath can be achieved by adding to the equations of motion a term that alters the pressure according to

$$\left(\frac{dP}{dt} \right)_{\text{bath}} = \frac{P_0 - P}{\tau_P}. \tag{2.10}$$

This term is nothing more than a simple proportional coordinate scaling that changes the equations of motion to

$$\dot{x} = v + \alpha x, \tag{2.11}$$

where α is obtained from the time derivative of the pressure

$$\frac{dP}{dt} = -\frac{1}{\beta V}\frac{dV}{dt} = -\frac{3\alpha}{\beta}, \tag{2.12}$$

where β is the is the isothermal compressibility. We note that the pressure is calculated using the virial theorem [38, 39]

$$P = \frac{1}{V}\left(Nk_{\mathrm{B}}T + \frac{1}{3}\sum_i \langle \boldsymbol{r}_i \cdot \boldsymbol{f}_i \rangle\right). \tag{2.13}$$

Thus $\alpha = -\beta\left(P_0 - P\right)/3\tau_P$. Finally, this change in the equation of motion is equivalent to a proportional scaling of the coordinates and box length at every time step ($x \to \mu x$) with (up to first order)

$$\mu = 1 - \frac{\beta \Delta t}{3\tau_P}\left(P_0 - P\right). \tag{2.14}$$

Here we showed the equations for the barostat and the thermostat of Berendsen, but not the algorithm to incorporate them successfully. This can be found in Ref. [110].

The coupling constants of these thermostat and barostat represent double-edged-swords and the choice of their values is quite critical for the success of the simulation. Low values can be used for thermalization purposes. However they can lead to instabilities or wrong fluctuations. On the other hand, large values can lead to wrong oscillations. In particular, the Berendsen thermostat suppresses fluctuations of the kinetic energy and the corresponding error scales with the inverse of the number of atoms. So, most of the ensemble averages will remain unaffected for very large systems. However, this is not the case for fluctuation properties, such as the heat capacity [113]. Meanwhile, the Berendsen barostat provides correct average pressures (even with only a rough estimate for the isothermal compressibility) but may not reproduce the correct NPT ensemble. Therefore, the values for the coupling constants have to be chosen with care in order to obtain realistic fluctuations [18, 110, 113].

Chapter 3

Machine learning in material science

> The fact that the price must be paid is proof it is worth paying.
> Robert Jordan
> The Eye of the World

Price... In the previous chapter we discussed two problems, or types of simulations, that require the accurate and efficient calculation of energies and its derivatives. Furthermore, we mentioned that the amount of calculations required to solve these problems might become too cumbersome even for a method as efficient as DFT. Materials science researchers have faced these problems before and found several satisfying solutions, which usually involve a price. Often an increase of efficiency can be gained from the application of methods constructed with more lenient conditions and that are less accurate than DFT. For example classical force-fields are the standard tool to calculate energies and forces in MD simulations. However, the quest for more accurate and efficient methods led researchers to the field of machine learning. Furthermore, machine learning provides tools to extract information from huge chunks of data in ways that far surpasses the capabilities of the human mind.

In this chapter we discuss machine learning. We start with an overview where we present the basics, the most successful applications, and the different categories of machine learning. We follow it by the applications in the field of material science. We describe its basic tenants: data, features, and algorithms (in particular neural networks). Finally we present the most recent or successful applications in the field of material science. The research presented here was published in Ref. [23].

3.1 Overview of machine learning

Machine learning [19, 21] consists of a collection of statistical models that search for patterns in order to extract the necessary information to perform a specific task, without resorting to explicit instructions. Currently, machine learning algorithms are highly sought and well regarded due to their accomplishments while performing regression, classification, clustering, dimensionality reduction, optimal experimental design, or decision making. In fact, machine learning algorithms proved extremely successful at extracting information and relations from big chunks of high dimensional data. Furthermore, these algorithms manage to surpass human capabilities and obtain outstanding results in several fields, such as face recognition [114–117], image classification [118], driving cars [119], and playing games: Atari [120], Go [121], chess [122], and others [123, 124]. These algorithms even spread to many of our daily life activities, for example to image and speech recognition [125, 126], credit scores [127], fraud detection [128], web-searches [129], email/spam filtering [130], and many, many others.

Meanwhile, the application of these algorithms also extended to the fields of biology and chemistry, where they obtained fantastic results [131, 132]. However in solid-state material science, this integration has been slower. Machine learning algorithms also boast several accomplishments in this field [133–138], in particular in chemical sciences [139], in materials design of thermoelectrics and photovoltaics [140], in the development of lithium-ion batteries [141], and in atomistic simulations [142]. Nonetheless, many of the published applications remain very basic and lack complexity. Often applications involve small training sets or simple fitting procedures that do not require this kind of techniques. This means that these applications do not take full advantage of the power of machine learning, and consequently, are unable to replicate their accomplishments in other fields. Fortunately, this is changing quite rapidly, on par with the growing interest in these kind of techniques. In section 3.5 we mention more applications of machine learning in the field of material science.

Different types of machine learning algorithms deviate from each other based on their approach to solve a certain problem or task, the task itself, and their input and output. This offers a way to organize the machine learning algorithms under three main categories: supervised learning, unsupervised learning, and reinforcement learning.

The most widespread type of learning is certainly supervised learning, which comes as no surprised, due to its similarity to a simple fitting procedure. In supervised learning, models take advantage of a collection of data, containing both the input and the target property or response, to find hidden patterns among the data and construct a function that can extrapolate an unknown target property based on a given input. On the other hand, unsupervised learning includes algorithms whose objective consists in the categorization of data based on the similarities found among the inputs. This type of learning is not conditioned, in the sense that the outputs (or categories) are not known *a priori*. Finally, reinforcement learning encompasses goal-oriented algorithms [143, 144]. In these algorithms an agent in a certain state interacts with its environment by performing an action. From this interaction results a reward, which the agent intends to maximize. With successive actions, the agent improves its policy, i.e., the strategy to determine the next action based on the current state and the maximum reward.

Worth mentioning is the possibility of overlap between techniques of different categories. For example semi-supervised learning exists between supervised and unsupervised learning. These are algorithms that take advantage of some data containing the target properties to improve its performance on the identification of unlabelled data (as in unsupervised learning). Usually this is particularly useful to learn representations [145].

Here, we will focus on the explanation of the workflow of supervised learning algorithms (see fig. 3.1). As mentioned above, this is the most common type of learning, specially in the field of material science. Furthermore, it represents the type of algorithms used for the work described in this thesis.

The first step in any machine learning algorithm consists in the generation of data. This can occur in a plethora of ways, such as performing calculations or just from observations.

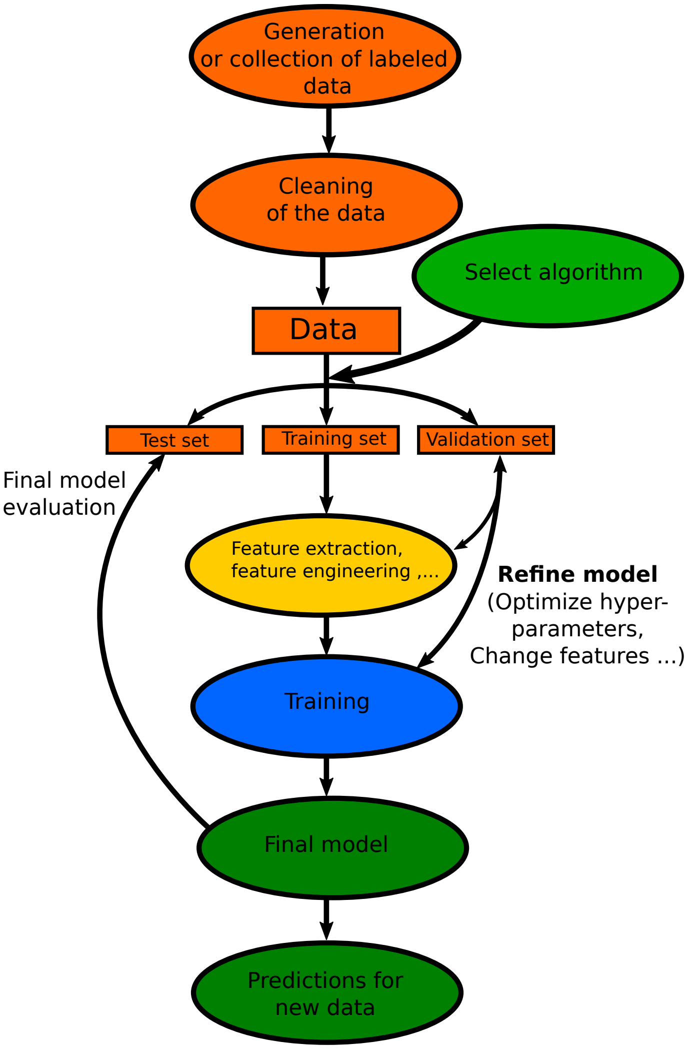

Generation or collection of labeled data
Cleaning of the data
Select algorithm
Data
Test set
Training set
Validation set
Final model evaluation
Feature extraction, feature engineering ,...
Refine model (Optimize hyper-parameters, Change features ...)
Training
Final model
Predictions for new data

Figure 3.1: Supervised learning workflow.

Usually this step involves a lot of work, as the data has to be collected and cleaned, before one can generate varied, consistent, and accurate data sets. In section 3.2 we further discuss this topic in association with material science. As mentioned, for supervised learning, this data is labelled, this means that it contains both the target property as well as the information required (by an algorithm) to compute said property. Often, as the data is generated, the machine learning algorithm that will fit it is also chosen. We mention some of these algorithms in section 3.4 and in particular we describe neural networks. After this selection, follows the extraction of the relevant information from the data and its processing, in order to provide the chosen algorithm with suitable inputs, called features or descriptors, for the task at hand. A more detailed discussion of these features follows in section 3.3. For now it is enough to understand that the raw information requires processing so that it can be understood by the machine (learning model).

Once all of these tasks are accomplished, the model is trained to learn, i.e. identify and reproduce, the intricacies of the data. Normally this entails the minimization of a cost function that evaluates the performance of the model in a subset of the data, the training set.

Another subset of the data, the validation set, is chosen to adjust certain hyper-parameters involved in the training process or parameters of the model itself, such as learning rates of an optimization method or the number of layers of a neural network. Finally, the predictive power of the trained model is ascertained using yet another unseen subset of the data, the test set.

Actually, several techniques from statistics exist to assess the extrapolation ability of a model to an independent data set, the so called cross-validation techniques, such as holdout, over k-fold cross-validation, leave-one-out cross-validation, Monte Carlo cross-validation [146], up to leave-one-cluster-out cross-validation [147], among others. The technique described above consists on the holdout, where the data is separated once into the different subsets. For example, in k-fold cross-validation, the data is divided into k sets. Then, each of the sets are hold out independently for the training of the model, i.e., the sets are selected, one by one, as the test set while the other $k-1$ sets are used for the training of independent models.

When discussing the error of predicative models, the trade-off between bias, variance and the irreducible error usually follows [148, 149]. Bias errors concerns incorrect assumptions in the training of an algorithm, while variance errors normally relate to the capture of small fluctuations or noise. Irreducible errors, as the name implies, represent errors from the problem itself that can not be minimized. The combination of high variance and low bias means that the model identifies noise along with the underlying patterns of the training set, which leads to over-fitting. Usually, reducing this error requires the increase of the training set or the simplification of the model. Meanwhile, low variance and a high bias indicate the incapability of the model to find the underlying patterns of the training set. This is the definition of under-fitting and it means that the model is too simple. The most common solution to this problem involves the increase of the complexity of the model.

3.2 Data

The success of machine learning methods highly depends on the quantity and quality of the data at its disposal.

As such, researches that wish to employ machine learning techniques to study a certain target property, start by gathering relevant data. This process may include several experiments in a high-quality lab, a plethora of computer simulations/calculations, or it can be as simple as accessing a database and downloading its contents. Obviously, the latter constitutes the least time consuming approach, provided a FAIR treatment of the data [150, 151], and that the database contains such data. FAIR treatment means: findability, accessibility, interoperability, and repurposability. This conveys the necessity to store data in effortless ways to find and access, in a format understood by different software, so that it can be applied to new purposes. Furthermore, the storage of data suggests that calculations do not have to be repeated. This means that the resources that would be wastefully used to generate some data, can be devoted to some other application. This reveals the importance of constructing, maintaining, and improving databases in material science and informatics, such as those in Ref. [5, 6, 70, 152–166].

Additionally, the machine learning approach intends to change even more the material science community. Usually, negative, unsuccessful, or even intermediate results are deemed unsuitable for publication. However, this results contribute as much as positive ones for the optimization of the machine learning algorithms [167, 168]. So, it is imperative to save these results in databases.

3.3 Features

After gathering the data set, machine learning methods require the extraction or engineering of features (or descriptors). This means that the relevant information among the data has to be found and represented in a way that is suitable and understood by the desired algorithm.

Naturally, the quality of the features depends on their capability to distinguish between two different elements of the data set, for example, two crystal environments. Furthermore, the accuracy and efficiency of the machine learning model highly depends on several properties of the features. In this regard, ideal features should be uncorrelated and have the lowest dimension possible. Moreover, the cost of feature extraction should not surpass that of the evaluation of the target property [169]. When too many features are correlated, feature selection can improve the efficiency and interpretability of the model, and avoid the curse of dimensionality [170]. For example, a machine learning model can resort to several elemental properties to determine the energy of a structure (or its distance to the convex hull of stability). Yet, only two, the group and the period in the periodic table, are necessary to obtain a reasonable accuracy [171]. Some algorithms even perform this feature selection automatically (see section 3.4).

In fact, feature extraction might just entail the selection of elemental properties, such as the atomic number, ionization potential, covalent radius, or others. This is the case for problems that are restricted to only one possible crystal structure and stoichiometry [171–175]. Similarly, for other constrained problems, feature extraction might consist in the identification of building blocks, for example the number of certain molecules in polymers [176] or molecular crystals. On the other hand, if the constrains are lifted, and a machine learning model is intended to describe a complete potential energy surfaced based solely on atomic positions and element types, feature selection may involve more complex transformations, such as an expansion of angular distribution functions in a certain basis [177] or the construction of crystal graphs [178]. Another approach for feature construction involves aggregations based on statistics. Usually, more features can be obtained from the calculation of averages or minimum values. Finally, a complement to these approaches are the crude estimations of properties [179]. These consist in using a target property calculated within a certain methodology (for example PBE band gap) to compute the same property under a different methodology (experimental band gap). In this manner, the model predicts a difference or error rather than a target property. These examples imply that the amount of processing required for the construction of the descriptors depends on the problem considered. Furthermore, the selected algorithm may play a role, as several algorithms already contain a feature extraction phase [180], e.g. deep learning neural networks.

Over the years, several features were specially designed and proposed for material science problems [169, 171–178, 181–204]. We note that most of them concern the reconstruction of potential energy surfaces and, therefore, the creation of machine learning force-fields (see section 3.5.2). A thorough study of most of these descriptors and the properties they must satisfy can be found in the review of Bartók *et al.*[181]. Essentially, the quality of these features improves if they form a complete representation with fixed dimension, if they remain invariant under symmetry operations (such as translation, rotation, reflection, and permutation of equivalent atoms), and they are differentiable with respect to atomic positions. While the latter guarantees the calculation of energy derivatives, such as forces, completeness ensures that the representation includes the necessary features, no more, no less. To clarify, it is possible to neglect these conditions, however this hinders the efficiency of the machine learning model. For example, nuclear charges and atomic positions should in principle constitute sufficient features, as they fully define the Hamiltonian of a system. Yet permutation of atomic positions may result in different outcomes of the model. Moreover, most machine learning algorithms demand a fixed number of inputs, so two structures with different number of atoms would have to be treated differently. Another possibility involves the padding of the input vector of the model with zeros. Other well known descriptions that neglect some of the above mention properties are transformations of pairwise distances [205–207], Weyl matrices [208], Z-matrices [209], and Coulomb matrices [182].

Before proceeding, we would like to describe some of the most used or insightful descriptors in material science, namely the Behler and Parrinelo [185] symmetry functions, the smooth overlap of atomic positions (SOAP) kernel of Bartók *et al.* [181], and the Chebyshev polynomials based descriptor of Artrith *et al.* [177]. Another successful descriptor that will only be mentioned are the crystal graphs of Xie *et al.* [178], which require convolutional neural networks and can be understood as a message passing neural network [210].

Behler and Parrinelo created representations of the chemical environment based on radial and angular symmetry functions. Usually, both of these types of functions are centered around atom i and provide information on its interaction with all the neighbouring atoms, within a certain radius R_c. However they can also be pair centered [211]. While the radial functions [212] map the distributions of distances $R_{ij} = |\mathbf{R}_j - \mathbf{R}_i|$,

$$G_i^\mathrm{r}(\{\boldsymbol{R}_i\}) = \sum_{j\neq i}^{\text{neighbors}} g^\mathrm{r}(R_{ij})\,, \tag{3.1}$$

the angular functions complement them with information on the distribution of bond angles $\theta_{ijk} = \angle\,(\mathbf{R}_j - \mathbf{R}_i, \mathbf{R}_k - \mathbf{R}_i)$:

$$G_i^\mathrm{a}(\{\boldsymbol{R}_i\}) = \sum_{j\neq i}^{\text{neighbors}} g^\mathrm{a}(\theta_{ijk}). \tag{3.2}$$

Several forms of these functions were proposed [213], for example, the radial symmetry function

$$G_i^\mathrm{r} = \sum_{j\neq i}^{\text{neighbors}} f_\mathrm{c}(R_{ij})\,\mathrm{e}^{-\eta(R_{ij}-R_\mathrm{s})^2} \tag{3.3}$$

where f_c represents the cutoff function that neglects the pair-wise contributions above R_c, η a parameter that controls the width of the Gaussians, and R_s another parameter that introduces a shift to the Gaussians. Similarly, an example of an angular function is

$$G_i^{\mathrm{a}} = 2^{1-\zeta} \sum_{\substack{jk \\ i\neq j\neq k}}^{\text{neighbors}} (1+\lambda\cos\theta_{ijk})^{\zeta}\, \mathrm{e}^{-\eta(R_{ij}^2+R_{ik}^2+R_{jk}^2)} \times\, f_{\mathrm{c}}(R_{ij}) f_{\mathrm{c}}(R_{ik}) f_{\mathrm{c}}(R_{jk})\,. \tag{3.4}$$

where the parameters λ and ζ, determine the positions of the extrema of the cosine and control the angular resolution, respectively. The features thus consist of 20 to 100 of these symmetry functions, obtained for different values of the parameters mentioned.

The SOAP descriptor of Bartók *et al.* calculates the similarity measurement between two sets of atomic configurations based on

$$K(\rho,\rho') = \left[\frac{k(\rho,\rho')}{\sqrt{k(\rho,\rho)k(\rho',\rho')}}\right]^{\zeta}. \tag{3.5}$$

Here, the parameter ζ enhances the sensitivity of the kernel to slight variations of the atomic positions, the denominator is just a normalization factor that ensures the comparison between the same structure is one, and ρ represents the Gaussian-smeared atomic neighbor densities

$$\rho(\mathbf{r}) = \sum_i \mathrm{e}^{-\alpha|\mathbf{r}-\mathbf{r}_i|^2}, \tag{3.6}$$

which is usually expanded in terms of spherical harmonics. Finally, $k(\rho,\rho')$ amounts to a rotationally invariant kernel, constructed from the overlap of an atomic environment and all the other rotated environments

$$k(\rho,\rho') = \int d\hat{R} \int d\mathbf{r}\; \rho(\mathbf{r})\rho'(\hat{R}\mathbf{r}). \tag{3.7}$$

This descriptor can be understood as a three dimensional generalization of the radial atom-centered symmetry functions [181].

In an attempt to devise a descriptor whose dimension remains constant with the increasing number of elemental species, Artrith *et al.* proposed a descriptor that consists on the union of two sets of invariant coordinates: one to map the structure and another for the compositions. Each then consist on the expansion of radial distribution functions (RDF)

$$\mathrm{RDF}_i(r) = \sum_\alpha c_\alpha^{\mathrm{RDF}} \phi_\alpha(r) \quad \text{for} \quad 0 \leq r \leq R_{\mathrm{c}} \tag{3.8}$$

and angular distribution functions (ADF)

$$\mathrm{ADF}_i(\theta) = \sum_\alpha c_\alpha^{\mathrm{ADF}} \phi_\alpha(\theta) \quad \text{for} \quad 0 \leq r \leq R_{\mathrm{c}}. \tag{3.9}$$

in a complete basis set ϕ_α, such as the Chebyshev polynomials. The expansion coefficients can then be obtained from

$$c_\alpha^{\mathrm{RDF}} = \sum_{R_j} \phi_\alpha(R_{ij}) f_{\mathrm{c}}(R_{ij}) w_{tj} \tag{3.10}$$

and

$$c_\alpha^{\mathrm{ADF}} = \sum_{R_j, R_k} \phi_\alpha(\theta_{ijk}) f_c(R_{ij}) f_c(R_{ij}) w_{tj} w_{tk}, \, . \tag{3.11}$$

where f_c is a cut-off function. Furthermore, the values of the weights w_{tj} and w_{tk} depend on the map. For the structure maps these are just 1, while for the composition maps these depend on the chemical species of the atom they are describing, following the pseudo-spin convention of the Ising model. Clearly, this descriptor was influenced by both the Behler symmetry functions and the SOAP method, and one of its great advantages consists on its systematic refinement, which only requires more terms in the polynomial expansion.

After mentioning so many descriptors and explaining these three, we should comment on the selection of the features. In the end, the selection of the best features can turn into a rather difficult task, and it surely depends on the desired target quantity and the space of the problem. A careful methodology to solve this problem of finding the best representation surely involves the creation of libraries with the implementation of all possible features [214–218], and the calculation of rigorous benchmarks. Unfortunately, only a few studies actually compare quantitatively different descriptors. However, the importance of these studies have been acknowledged and their number is increasing [169, 181, 219, 220].

3.4 Algorithms

From the numerous algorithms supplied by machine learning, only a few have been successfully applied to material science. Several of these applications concern linear regression and classification methods such as ridge regression [221], support vector machines [222], and Gaussian process regression [223–225]. Typically these algorithms resort to the kernel trick [226] to tackle more complex problems that require non-linear models. This trick consists in applying a kernel that maps the feature space into an higher dimensional one and then solving the linear problem in this higher dimensional space.

The most prominent applications regarding variable selection and extraction algorithms consist of the least absolute shrinkage and selection operator [227–229] (LASSO), the sure independence screening and sparsifying operator [230] (SISSO), the bootstrapped projected gradient descent, and the principal component analysis [231, 232] algorithm.

Finally, only two completely non-linear machine learning models have been used in material science problems, namely neural networks and decision tree based methods, like random forests and extremely randomized trees.

Information on these algorithms can be found in the provided references and in Refs. [19, 233–237]. We now proceed with the explanation of neural networks, in particular, fully connected feed-forward neural networks, due to their application in the research discussed in chapter 4.

3.4.1 Neural networks

Neural networks are machine learning algorithms inspired by biological neurons. As such they can be described as a collection of neurons, grouped into layers, and that interact with each other in some way. Ultimately, these connections map a set of inputs into a set of outputs. Now, the configuration of the neurons and layers and the type of operation performed between each layer varies and gives rise to several neural networks structures [238]. Examples are perceptrons [239], Boltzmann (restricted) machines [240, 241], recurrent neural networks [242], (variational) auto enconders [243–245], generative adversarial networks [246], deep convolutional neural networks [180, 247, 248], among others.

For example, the perceptron, which amounts to a collection of McCulloch and Pitts neurons together [19, 249], represents a neural network with just two layers, the input layer and the output layer. A weight represents the connection between each input and output node, and the determination of the output nodes requires the application of an activation function over the weighted sum of the input neurons. The term activation function comes from the use of the Heaviside step function that returns 1 when the sum reaches a certain threshold, signalling that the neuron fired or got activated with that input. We note that the value of the threshold should be adjustable, in order to provide more control of the firing of a node, without providing an additional parameter to the model. This is achieved with a trick that consists in the introduction of an additional constant neuron (with value ±1), usually denoted bias neuron. In this manner, the weight of the bias neuron will shift the sum and provide an additional degree of freedom in the determination of a target property.

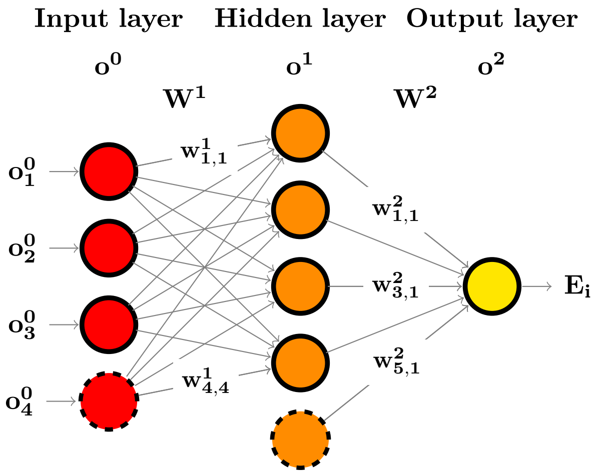


Figure 3.2: Example of a multilayer perceptron feedforward neural network as used in this work. The bias nodes are shown with a dashed contour. Their standard value is 1.

A simple extension of the perceptron consists in adding more layers, which gives rise to a multilayer perceptron feedforward neural network, schematically represented in fig. 3.2. These new layers are denoted as hidden layers, which perfectly describes their function to connect inputs to outputs with hidden, and fairly often not interpretable operations. In this neural network structure, the nodes are connected in only one way: forward. This means that the calculation of the value of the nodes o^{ν}_{ζ} of a certain layer require the values of the nodes of the previous layer and the weight matrix (w^{ν}) that connects them:

$$o^{\nu}_{\zeta} = \varphi(h^{\nu}_{\zeta}) = \varphi\left(\sum_j w^{\nu}_{j\zeta}\ o^{\nu-1}_j\right), \tag{3.12}$$

where φ represents an activation function, $w^{\nu}_{\zeta j}$ the weight between the node j in layer $\nu - 1$ and the node ζ in layer ν, and h^{ν}_{ζ} the argument of the activation function. For clarification, we define o^{ν} as the vector containing all the nodes of layer ν, therefore the input layer can be represented as o^0.

We note that all the magic behind the fitting capabilities on neural networks and their successive application to many problems stems from the hidden layers and the non-linearity of the activation functions. The combination of these two factors, ultimately change the neural network function from a simple linear combination to a highly complex non-linear function. Traditionally, sigmoid functions were used as activation functions for this type of neural networks, such as the logistic function

$$\varphi_{\text{logistic}}(x) \;=\; \frac{1}{1+\mathrm{e}^{-x}} \tag{3.13}$$

However, the vanishing gradient problem [250] and the search for more efficient implementations lead to the development of modern activation functions, such as the rectified linear units [251, 252] (ReLU), their smooth approximation: softplus, and the exponential linear units [253] (ELU)

$$\begin{aligned}
\varphi_{\text{ReLU}}(x) &= \max(0, x) && (3.14)\\
\varphi_{\text{leaky ReLU}}(x) &= \max(0.01x, x) && (3.15)\\
\varphi_{\text{softplus}}(x) &= \log(1+\mathrm{e}^{x}) && (3.16)\\
\varphi_{\text{ELU}}(x) &= \begin{cases} x & \text{if } x > 0 \\ \alpha(e^{x}-1) & \text{otherwise} \end{cases}. && (3.17)
\end{aligned}$$

Other activation functions fairly used in material science are the linear function and the hyperbolic tangent:

$$\begin{aligned}
\varphi_{\text{linear}}(x) &= x && (3.18)\\
\varphi_{\tanh}(x) &= \tanh(x) \;=\; \frac{1-\mathrm{e}^{-2x}}{1+\mathrm{e}^{-2x}}. && (3.19)
\end{aligned}$$

Going back to the structure of neural networks, with the increase of the number of hidden layers, it becomes tempting to call the networks ”deep” neural networks. Although lacking a precise definition, this terminology should refer to neural networks with 5 or more hidden layers [248], that can not only learn representations with different abstraction levels without human intervention, but also reuse them [180, 254].

First introduced in the field of image recognition [247, 248], and perhaps the most successful structure for neural networks nowadays, deep convolutional neural networks contain several hidden layers, some of which not fully connected. Further inspiration from biological processes, in particular from the organization of an animal visual cortex, lead to the creation of convolutional and pooling layers. Convolutional layers allow for the extraction of high-level features due to the application of filters that act on a certain receptive field, i.e. a small segment of the input nodes. The filters are applied across the entire input nodes, performing convolution operations to the nodes inside the receptive fields that can overlap. Usually the convolutional layer reduces the dimensionality of the features. However, the introduction of padding (or additional filters) results in features with the same or increased dimensionality. Meanwhile pooling layers downsample feature maps in certain receptive fields. This means that the neurons present in a certain region are combined by an operation, such as the maximum or the average, into a single neuron. This not only reduces the dimentionality of the features but makes them more robust. To summarize, convolutional and pooling layers allow for a reduction of the dimensionality of the features without losing critical information, which allows for a more efficient training while keeping, at least, the same level of accuracy in the predictions.

Concerning the accuracy of the predictions of neural networks, we would like to point out the universal approximation theorem [19, 255, 256]. This theorem states that a neural network with just one hidden layer and a finite number of nodes, can reproduce any function, provided a suitable activation function. However, the theorem does not mention how many nodes are required or details on how to train the neural network. And training a neural network is not an easy task! The weights constitute the only parameters that have to be optimized during the training process. Nevertheless, the performance of a neural network can be further improved with the optimization of its hyper-parameters, such as the topology and the architecture, i.e., the number of layers, and the ways the neurons are connected and distributed in those layers. Similarly to other machine learning methods, the training of a neural network involves the optimization of a high-dimensional cost function that measures its performance in the training set. Typically, this cost function contains a L_2 norm and a L_2 (or L_1) regularization term:

$$\epsilon = \frac{1}{2\alpha}\left[\sum_{\sigma}^{\alpha}\left(E(W,o^0) - E^{\text{ref}}\right)_{\sigma}^{2} + \lambda\sum_{i}^{k}|w_i|^2\right]. \tag{3.20}$$

Here, α represents the number of elements σ in the training set, k the number of weights w_i in the set of all weight matrices $W = \{w^\nu\}$, E^{ref} the value of the target property for element σ, obtained with a reference method, and $E(W,o^0)$ the neural network function.

In principle, the optimum values for the weights, those that minimize the cost function, can be found with any optimization method, for example genetic algorithms [257]. At the present time, the standard algorithm to perform this task is the back-propagation algorithm [19, 258]. This iterative, gradient-based optimization algorithm consists in deriving the cost function with respect to the weights, while exploiting the chain rule to obtain the derivatives of the nodes of each layer, starting from the output layer and going back until the input layer. By defining $\dot{\varphi}^n$ as a diagonal matrix that stores the derivatives of the nodes of the n layer, the derivative of the output of the neural network can be expressed as

$$\frac{\partial}{\partial w^n}E(W,o^0) = \left(\left[\prod_{m>n}^{N_h+1}\dot{\varphi}^m w^m\right]\dot{\varphi}^n \otimes o^{n-1}\right)_i \tag{3.21}$$

where N_h is the number of hidden layers, and the product inside square brackets is ordered in decreasing order of layers

$$\left[\prod_{m>1}^{3}\dot{\varphi}^m w^m\right] = \dot{\varphi}^3 w^3 \dot{\varphi}^2 w^2.$$

Having computed the Jacobian, the update for the weights can be calculated using any gradient based optimization method, such as gradient descent.

3.5 Recent applications

Machine learning algorithms tackle the problem of material discovery and structural prediction from different distinct directions. A conceptually simple approach consists in the replacement of first-principles methods by a machine learning model, in order to avoid costly energy evaluations and increase the speed of the computations. We will focus on these machine learning force-fields in section 3.5.2. Meanwhile, in section 3.5.1 we mention other approaches.

3.5.1 Applications to solids

The history of the development of DFT functionals with machine learning techniques started with the work of Tozer *et al.* [326] in 1996, with their mapping of the electronic density of some molecules to its exchange and correlation potential. Since then a few applications have been proposed, mainly regarding the Hohenberg-Kohn map between the potential and the density [327] in order to easily perform orbital-free DFT calculations, the approximation of the kinetic energy functional of the density for noninteracting spinless fermions [328] in 1D, or for diatomic molecules subjected to a soft Coulomb interaction [329], the determination of range-separation parameter in exchange and correlation functionals[330], and the projection from the charge density onto the Hartree-exchange-correlation potential [331]. More recently, a methodology was proposed to reproduce concurrently the exchange and correlation energy and potential functionals [332] of one-dimensional systems with two strongly correlated electrons.

Yet, one of the most prominent approaches involves the exploration of the composition space with the intent to find the most stable materials that crystallize into a certain structure. This is usually designated as component prediction [137] and essentially involves the determination of the thermodynamic stability of a huge number of crystal structures. Now, the gist is that machine learning techniques can infer the relations between the crystal structure and the different elements from a training set, that contains a small fraction of all possible combinations of elements, and predict the stability of all the other combinations. Therefore, machine learning techniques can be used to decrease the number of first principles calculations required, or to avoid altogether the attempt to experimentally synthesise many of these materials. This approach has been used with different algorithms to predict many stable materials, such as elpasolites [172], perovskites [171, 173–175, 333–335], ternary prototypes with stoichiometry AB_2C_2 [336], and Heusler compounds [337–339].

Moreover, if the machine learning algorithm is powerful enough and if the features are sufficiently processed, the constrain of the same structure can be removed and additional properties can be calculated. Examples of these more general approach comprise the random forests with Voronoi tessellations features of Ward *et al.* [340], the crystal graph convolutional neural networks by Xie *et al.* [178], the *MatErials Graph Networks* by Chen *et al.* [195], and the message passing neural networks by Jorgensen *et al.* [193].

Property	References
Curie temperature	[259–264]
Vibrational free energy and entropy	[265]
Band gap	[178, 191, 195, 228, 264, 266–277]
Dielectric breakdown strength	[278–280]
Lattice parameter	[277]
Debye temperature and heat capacity	[191, 281, 282]
Glass transition temperature	[283, 284]
Thermal expansion coefficient	[191]
Thermal boundary resistance	[285]
Thermal conductivity	[286–294]
Local magnetic moments	[192, 295]
Melting temperature	[291, 296, 297]
Magnetocaloric effects	[264]
Grain boundaries	[298]
Grain boundary energy	[299–302]
Grain boundary mobility	[302]
Interface energy	[277]
Seebeck coefficient	[289, 303, 304]
Thermoelectric figure of merit	[305]
Bulk and shear moduli	[178, 191, 195, 281, 306–308]
Electrical resistivity	[289]
Density of states	[184, 309, 310]
Fermi energy and Poisson ratio	[178]
Dopant solution energy	[311]
Metal-insulator classification	[230]
Topological invariants	[312–318]
Superconducting critical temperature	[147, 319–323]
Li-ion conductivity and battery state-of-charge	[141, 324, 325]

Table 3.1: Summary of material properties predicted with machine learning methods and corresponding references.

A contrasting approach is the differentiation between multiple crystal structures and their subsequent classification, usually denoted as just structure prediction. The name might be confusing, but it serves to categorized techniques similar to Pettifor structural maps [75, 341–343] that use properties of the elements to divide binary and ternary structures in a two-dimensional plot. This map can then be used to predict stable structures with two or three elements. Machine learning examples of this approach are the the prediction of binary structures using the cumulant expansion method [194, 344], the cluster resolution feature selection [345], SISSO structural maps [230], the classification of perovskites [261], and the classification of different crystal structures using random forests [346], variational autoencoders [168], or generative adversarial networks [347–349]. While the previous works rely on elemental properties for the classification, others rely on X-ray diffraction patterns [350] or a simulated two dimensional diffraction fingerprint [351], or even on machine learning based image processing [352–357].

Similarly, machine learning algorithms were successfully employed to predict a good deal of material properties, as shown in table 3.1 and its references.

Finally, the last approach we wish to mention intertwines surrogate-based optimization [358, 359] and active learning, and is commonly denoted as adaptive design process. This process consists on the development of a surrogate model at the same time that its predictions are used to search for the best possible points, in the space of the target properties. After that, these points are included in the training of the model. This cycle is repeated until the optimum value is found. The challenge of this process pertains then the balance between two tasks: the exploration of the space in order to improve the model and the determination of the material that exhibits the best value for a certain property, or set of properties [360].

The first question that arises from such process is related to the strategy to choose the best points at each iteration of the cycle. Contrary to popular belief, pure exploitation, i.e. always choosing the point that results in the minimum value for the target property in a minimization, is not the best approach [361–364]. Indeed, a much better strategy involves the calculation of the maximum expected improvement. However this requires machine learning models that provide both the predictions and the uncertainty in those predictions, such as Gaussian processes [223, 365–369], decision tree methods or SVM regressors combined with bootstrapping methods [360, 369, 370], random forests [323, 371], and Monte Carlo tree searches [372–375] combined with neural networks. Moreover similar techniques were proposed involving genetic algorithms [85, 376] or even other techniques [377–379].

3.5.2 Force-fields

Even with the development of several electronic structure methods [20], the improvement of their implementations in computer codes, and the availability of faster supercomputers every year, a plethora of problems and systems remains out-of-reach, due to the high computational price required to simulate them. Here, we are referring to molecular dynamics, Monte Carlo, global structural prediction, or any other simulations that require either a numerous number of atoms, long simulation times, or frequent evaluations of energies and its gradients: forces and stresses.

Even DFT, perhaps the most successful of the electronic structure methods, due to its unrivaled combination of accuracy and computational efficiency, a theory used to describe millions of compounds and the current backbone of high-throughput and accelerated material design efforts [156, 380–386] lacks efficiency and suffers from a number of limitations. Currently, DFT records include around 2 000 000 atoms in a single total energy evaluation [387–389], a few picoseconds long molecular dynamic simulations with fewer than ten thousand atoms [389], and a few tens of atoms in global structure prediction searches [390, 391].

These problems make researchers look elsewhere for answers. In fact, MD simulations usually resort to classical force-fields [392–398] to solve complex problems, such as protein folding [99–101]. Similarly, several structure prediction studies that encountered these limitations frequently end with a density-functional-based tight binding [399–401] description. Both these approaches offer faster energy evaluations, and also larger and longer simulations than DFT, and for this, researchers normally overlook the loss in accuracy. A comparison of these three methods places tight-binding in the middle, with classical force-fields on the side of least accuracy and best efficiency while DFT takes the opposite side.

Meanwhile, the past fantastic accomplishments of machine learning, and the future promise of more, led researches to believe that its algorithms could combine the best qualities of the approaches mentioned above. By this we mean a linear scaling with the number of atoms (or electrons), such as the scaling classical force-fields, and the same accuracy as the reference method employed in the training of the machine learning force-fields, such as the accuracy of DFT.

The first combination of machine learning with the construction of potential energy surfaces occurred in 1992, with the neural networks of Sumper *et al.* [402], which mapped the energy with the vibration spectra of polyethylene molecules. However, technical problems judged the approach as too difficult and inefficient to apply to other systems. The proof that neural networks could be used to accurately and efficiently predict potential energy surfaces only come a few years later, in 1995, with the influential study of the surface diffusion of CO/Ni(111) by Blank *et al.* [403].

Since then, a myriad of machine learning potentials have been proposed and discussed in the literature [211, 213, 404, 405]. For this reason, we will center this discussion around the most influential methodologies applied in the field of materials science, mainly the Behler and Parrinelo approach [406], the Gaussian approximation potentials by Bartók *et al.* [187], and the spectral neighbor analysis potential from Thompson *et al.* [407].

The Behler and Parrinelo approach uses multilayer perceptron feedforward neural networks to describe potential energy surfaces. In this approach, a set of radial and angular symmetry functions represent each atom, in a certain chemical environment. Then, each set serves as input to a neural network that returns the atom's contribution to the energy E_i. Different elements require different atomic neural networks. Subsequently, the sum of all atomic contributions corresponds to the total energy of the system. This is now the standard for all machine learning force-fields since it allows for their application to very large systems. Furthermore, the calculation of forces and (static) stresses requires only the analytical differentiation of the neural network function with respect to the atomic positions and the infinitesimal strains, respectively.

Since its first application to bulk silicon, this approach was applied to study many materials, for example carbon [408], sodium [409], zinc oxide [410], titanium dioxide [212], germanium telluride [411], copper [412], gold [413], lithium–silicon [414], and Al-Mg-Si alloys [415].

Moreover, several contributions improved upon this approach. Initially, the cost function only included energy terms. However starting in 2011, force terms were also included. This followed from the works of Witkoskie *et al.* [416] and Pukrittayakamee *et al.* [417, 418], that reported an increase in the accuracy of the force-fields with the inclusion of the gradients of the neural network function in the training. Basically, this is equivalent to the increase of the size of the training set, and to training under more restrictions. Artrith *et al.* [414] replaced the symmetry functions by descriptors based on Chebyshev polynomials [177], that can be systematically improved and that allow for the creation of potentials with constant complexity in the number of chemical species. Ghasemi *et al.* proposed a charge equilibration technique via neural networks [419, 420], where neural networks return the electronegativity and a charge equilibration method provides the total energy. Finally, Hajinazar *et al.* [421] proposed a strategy to train hierarchically multicomponent systems.

When first introduced in 2010, the Gaussian approximation potentials mapped the atomic energy with the bispectrum descriptor using Gaussian process regression, and described quite accurately the potential energy surface of iron and some semiconductors. However, studies conducted with this descriptor found it lacking for some systems, such as Si clusters with more than 13 atoms. This was solved with its substitution by the SOAP descriptor [181]. Further advances of the methodology include the improvement of the training process [422], the addition of two- and three-body descriptors to improve the description of solids [423], and the comparison of structures with multiple chemical species [424]. These improvements allow for the Gaussian approximation potentials to describe the potential energy surface and to study properties like bulk point defects, phonons, and Γ surfaces [425] of many materials or molecules. Noteworthy to mention are the studies of tungsten, carbon and silicon, iron [426], graphene [427], and formaldehyde [428]. Moreover, the Gaussian approximation potentials were used to accelerate the geometry optimization for some molecules [429], and to simultaneously explore and fit a complex potential energy surface [430, 431].

The spectral neighbor analysis potential consists on the description of a potential energy surface using the 4D bispectrum components and a simple linear fit. Its first application showed that a linear fit was sufficient to correctly reproduce the relative energy of different phases of tantalum. Nonetheless, improvements of this methodology include the extension of the model with the inclusion of quadratic terms in the bispectrum components [432]. Additionally, a two-step model fitting work-flow for multi-component systems [433] was introduced to study the binary alloy Ni–Mo, and PCA was used to examine the distribution of the features, which increases the efficiency of the fitting.

Before finishing this section, we would like to at least mention other methodologies to construct potential energy surfaces and their applications. For example the structure optimization technique based on evolutionary algorithms and kernel ridge regression potentials [434], the molecular dynamics scheme using either quantum mechanical calculations or gaussian process regression potentials [435, 436], the Gaussian process [437, 438] force-fields, the potentials based on kernel ridge regression and LASSO [186] and elastic net regression [439, 440], the (unconventional) deep neural network potentials [441, 442], and the moment tensor potentials [443].

Finally, we would like to draw attention to two of the most revolutionary approaches, developed in recent years, concerning the construction of potential energy surfaces: the accurate neural networK engine for molecular energies of *et al.* [444–447] and the deep learning architecture SchNet of Schütt *et al.* [204, 218, 448].

While influenced by the Behler and Parrinelo approach, the ANI approach introduces heavy modifications to it and takes advantage of deep neural network architecture to produce a truly transferable neural network potential. Noteworthy to mention is the GPU implementation that facilitates the training of a neural network with a pyramidal architecture containing 124033 weights, the construction of an atomic environment vector for each atom (using modified symmetry functions), which is then fed to a single neural network, and the application of active learning techniques in the pursuit of an automatic generation of datasets [447]. Unfortunately, it has only been applied to molecules.

Meanwhile, in the SchNet architecture, continuous-filter convolution layers and filter-generating networks model the interaction between atoms described as a tuple containing atomic numbers and atom positions. Convolution layers are even used to include periodic boundary conditions in order to describe solids.

Chapter 4

Neural Networks force fields

> It is no easy thing to slay a dragon, but it can be done.
>
> George R.R. Martin
> A Storm of Swords

Dragon... Our main objective for this work can be metaphorically described as slaying dragons. These dragons represent simulations with numerous atoms, simulations that require several energy calculations, and simulations that last a long period of time.

In this chapter we present our work in the construction of machine learning force-fields. We start with a description of the Behler and Parrinelo neural networks as implemented in the ÆNET package. Then we describe how to obtain accurate forces and stresses and we discuss the interpretability of the neural network force-fields. Afterwards, we present some example of force-fields and several applications: phonon dispersions, MD and melting temperatures, structure prediction and defects. We finish the chapter with an overview and outlook for the future. Part of the research presented here was published in Ref. [24].

4.1 Behler and Parrinelo neural networks in the ÆNET package

As mention before, one of our main interests pertains to global structure prediction and the study of materials. We hope to contribute to the search of materials that might satiate the electronics demands of this technological era or to the solution of the energetic problems we face.

However, as discussed in the previous chapters, this might require an uncountable number of calculations and the study of more than billions of materials. So, any hope to achieve this in our life-times, lies with the development of accurate, yet efficient methods to calculate energies, forces, and stresses, such as machine learning force-fields. And so it was, that our search for ways to speed up our global structure prediction calculations and other simulations with a huge number of atoms lead us to the recently published (at the time) work of Artrith *et al.* [212] describing an open-source implementation of the Behler and Parrinello approach (see section 3.5.2), the so called ÆNET package. This told us what we already knew about this rather successful approach: that it can yield fantastic accuracies in the reproduction of the PES of a system, with errors in the energy as low as a few meV/atom, that it can scale linearly with the number of atoms N in the unit cell, like classical force-fields, and that the analytical differentiation of the neural network function provides forces, needed for so many simulations and that was missing from some other approaches we found, such as cluster expansions. Furthermore, the ÆNET package allows for the usage of all symmetry functions proposed by Behler as features (see section 3.3), and trains its neural networks using the back-propagation algorithm.

Unfortunately, this particular implementation only optimized the neural networks with respect to the energy, and not forces or stresses. In fact, it even lacked a way to compute the stress tensor, which we so sorely need to optimize crystal structures. Meanwhile, forces are, sometimes, found wanting with this approach [212, 406, 449], with their errors remaining stubbornly high (above 100 meV/Å) and with directions that can differ from the reference ones by 100°. These problems in the forces occur whenever the training sets are not sufficiently rich (in size and variety) to capture the intricacies of the PES, when the features fail to capture both the similarities and dissimilarities between the structures in the training sets, or when the neural networks fail to retain all the information provided by the features. Then, subduing these errors requires considerable larger training sets, different descriptors, a new neural network structure and architecture, and/or the inclusion of force terms in the cost function.

4.2 Stress tensor

So, the first problem that we decided to solve concerns the calculation of the stress tensor. As seen already in section 1.1, the stress tensor is defined as the derivative of the total energy with respect to the infinitesimal strain ($\epsilon_{\alpha\beta}$) after the scaling of the space described in eq. (1.21). This includes the kinetic energy of the atoms. In this manner, the stress tensor can be divided into two parts: (i) A kinetic part issued from the derivative of the kinetic energy:

$$\sigma_{\alpha\beta}^{\text{kin}} = \frac{1}{\Omega}\sum_{k=1}^{N} m_k v_{k\alpha} v_{k\beta}, \tag{4.1}$$

where N is the number of atoms in the system, m_k the mass of atom k, and $v_{k\alpha}$ its velocity in the direction α. (ii) A static part obtained from the analytical differentiation of the neural network function

$$\sigma_{\alpha\beta}^{\text{static}} = -\frac{1}{\Omega}\frac{\partial E}{\partial \epsilon_{\alpha\beta}} = -\frac{1}{\Omega}\sum_{i}^{N}\sum_{\lambda}^{M_i}\sum_{\gamma}^{N}\frac{\partial E_i}{\partial o_{i\lambda}^{0}}\frac{\partial o_{i\lambda}^{0}}{\partial R_{\gamma\alpha}}\frac{\partial R_{\gamma\alpha}}{\partial \epsilon_{\alpha\beta}}, \tag{4.2}$$

where we used the same definitions as in section 3.4.1 and M_i represents the number of symmetry functions for atom i. Usually, we neglect the kinetic term, as we consider that the atoms are at rest (obviously, this is not the case for MD simulations). For simplicity, we will remove the static label from the rest of the discussion.

Similarly, the calculation of forces requires the differentiation of the neural network function with respect to the atomic positions

$$F_{\gamma\alpha} = -\frac{\partial E}{\partial R_{\gamma\alpha}} = -\sum_{i}^{N}\sum_{\lambda}^{M_i} \frac{\partial E_i}{\partial o_{i\lambda}^0}\frac{\partial o_{i\lambda}^0}{\partial R_{\gamma\alpha}}. \tag{4.3}$$

Here, $F_{\gamma\alpha}$ indicates the force acting on atom γ in the direction α. This shows that to obtain forces we just have to differentiate the neural network function with respect to the inputs $o_{i\lambda}^0$, and then differentiate them (the symmetry functions) with respect to the positions of the atom. While the computation of the stress tensor requires the same derivatives and the additional differentiation of the positions of the atoms with respect to the infinitesimal strains, which due to the scaling of the space eq. (1.21) is just

$$\frac{\partial R_{\gamma\alpha}}{\partial \epsilon_{\alpha\beta}} = R_{\gamma\beta}. \tag{4.4}$$

However, in order to compute the stress independently of the forces and for reasons that will become apparent in the next section (more precisely in section 4.3.2), we derived and implemented the derivatives of the symmetry functions with respect to the strains. For example, for the radial symmetry function we showed in section 3.3, we have

$$\frac{\partial g_{ij}^{\mathrm{r}}}{\partial R_{ij}} = \mathrm{e}^{-\eta(R_{ij}-R_{\mathrm{s}})^2}\Big(\frac{\partial f_{\mathrm{c}}(R_{ij})}{\partial R_{ij}} - 2\eta(R_{ij}-R_{\mathrm{s}})f_{\mathrm{c}}(R_{ij})\Big). \tag{4.5}$$

Then, the derivative with respect to the strains can be written as

$$\frac{\partial g_{ij}^{\mathrm{r}}}{\partial \epsilon_{\alpha\beta}} = \frac{\partial g_{ij}^{\mathrm{r}}}{\partial R_{i\alpha}}\frac{\partial R_{i\alpha}}{\partial \epsilon_{\alpha\beta}} + \frac{\partial g_{ij}^{\mathrm{r}}}{\partial R_{j\alpha}}\frac{\partial R_{j\alpha}}{\partial \epsilon_{\alpha\beta}} = \frac{\partial g_{ij}^{\mathrm{r}}}{\partial R_{ij}}\frac{R_{ij\alpha}}{R_{ij}}R_{ij\beta}, \tag{4.6}$$

where $R_{ij\alpha} = \mathbf{R}_j - \mathbf{R}_i$. For completion, we note that the derivatives with respect to the positions consist of

$$\frac{\partial g_{ij}^{\mathrm{r}}}{\partial R_{i\alpha}} = \frac{\partial g_{ij}^{\mathrm{r}}}{\partial R_{ij}}\frac{\partial R_{ij}}{\partial R_{i\alpha}} = -\frac{\partial g_{ij}^{\mathrm{r}}}{\partial R_{ij}}\frac{R_{ij\alpha}}{R_{ij}}, \tag{4.7}$$

and

$$\frac{\partial g_{ij}^{\mathrm{r}}}{\partial R_{j\alpha}} = \frac{\partial g_{ij}^{\mathrm{r}}}{\partial R_{ij}}\frac{\partial R_{ij}}{\partial R_{j\alpha}} = +\frac{\partial g_{ij}^{\mathrm{r}}}{\partial R_{ij}}\frac{R_{ij\alpha}}{R_{ij}}. \tag{4.8}$$

For the angular symmetry function showed in section 3.3, it is convenient to rearrange the terms as

$$\begin{aligned} G_{il}^{\mathrm{a}} &= \sum_{\substack{jk \\ i\neq j\neq k}}^{\text{neighbours}} 2\left(\frac{1+\lambda\cos\theta_{ijk}}{2}\right)^{\zeta} \mathrm{e}^{-\eta R_{ij}^2} f_{\mathrm{c}}(R_{ij}) \times \mathrm{e}^{-\eta R_{ik}^2} f_{\mathrm{c}}(R_{ik})\,\mathrm{e}^{-\eta R_{jk}^2} f_{\mathrm{c}}(R_{jk}) \\ &= \sum_{\substack{jk \\ i\neq j\neq k}}^{\text{neighbours}} g_{il}^{\mathrm{a}} = \sum_{\substack{jk \\ i\neq j\neq k}}^{\text{neighbours}} 2A_{ijk}B_{ij}B_{ik}B_{jk}, \end{aligned} \tag{4.9}$$

were $\cos\theta_{ijk} = \sum_\alpha \frac{R_{ij\alpha}}{R_{ij}}\frac{R_{ik\alpha}}{R_{ik}}$. Then for the derivatives with respect with the positions we get

$$\frac{\partial g^{\mathrm{a}}_{il}}{\partial R_{i\alpha}} = 2B_{jk}\Big[-\Big(\frac{\partial A_{ijk}}{\partial R_{j\alpha}} + \frac{\partial A_{ijk}}{\partial R_{k\alpha}}\Big)B_{ij}B_{ik} - \frac{R_{ij\alpha}}{R_{ij}}A_{ijk}\frac{\partial B_{ij}}{\partial R_{ij}}B_{ik} - \frac{R_{ik\alpha}}{R_{ik}}A_{ijk}B_{ij}\frac{\partial B_{ik}}{\partial R_{ik}}\Big] \tag{4.10}$$

for the derivative with respect to the positions of the central atoms $R_{i\alpha}$. For the derivatives with respect to the adjacent atoms we can write

$$\frac{\partial g^{\mathrm{a}}_{il}}{\partial R_{j\alpha}} = 2\Big[\frac{\partial A_{ijk}}{\partial R_{j\alpha}}B_{ij}B_{ik}B_{jk} + \frac{R_{ij\alpha}}{R_{ij}}A_{ijk}\frac{\partial B_{ij}}{\partial R_{ij}}B_{ik}B_{jk} - \frac{R_{jk\alpha}}{R_{jk}}A_{ijk}B_{ij}B_{ik}\frac{\partial B_{jk}}{\partial R_{jk}}\Big] \tag{4.11}$$

and

$$\frac{\partial g^{\mathrm{a}}_{il}}{\partial R_{k\alpha}} = 2\Big[\frac{\partial A_{ijk}}{\partial R_{k\alpha}}B_{ij}B_{ik}B_{jk} + \frac{R_{ik\alpha}}{R_{ik}}A_{ijk}B_{ij}\frac{\partial B_{ik}}{\partial R_{ik}}B_{jk} + \frac{R_{jk\alpha}}{R_{jk}}A_{ijk}B_{ij}B_{ik}\frac{\partial B_{jk}}{\partial R_{jk}}\Big] \tag{4.12}$$

Using these expressions, the derivatives with respect to the infinitesimal strains can be expressed as

$$\frac{\partial B_{ij}}{\partial \epsilon_{\alpha\beta}} = \frac{\partial B_{ij}}{\partial R_{i\alpha}}\frac{\partial R_{i\alpha}}{\partial \epsilon_{\alpha\beta}} + \frac{\partial B_{ij}}{\partial R_{j\alpha}}\frac{\partial R_{j\alpha}}{\partial \epsilon_{\alpha\beta}} = \frac{\partial B_{ij}}{\partial R_{ij}}\frac{R_{ij\alpha}}{R_{ij}}R_{ij\beta} \tag{4.13}$$

and

$$\begin{aligned}\frac{\partial A_{ijk}}{\partial \epsilon_{\alpha\beta}} &= \frac{\partial A_{ijk}}{\partial R_{j\alpha}}R_{j\beta} + \frac{\partial A_{ijk}}{\partial R_{k\alpha}}R_{k\beta} + \frac{\partial A_{ijk}}{\partial R_{i\alpha}}R_{i\beta} \\ &= \frac{\partial A_{ijk}}{\partial(\cos\theta_{ijk})}\Big[\Big(\frac{R_{ij\alpha}}{R_{ij}R_{ik}} - \cos\theta_{ijk}\frac{R_{ik\alpha}}{R^2_{ik}}\Big)R_{ij\beta} + \Big(\frac{R_{ij\alpha}}{R_{ij}R_{ik}} - \cos\theta_{ijk}\frac{R_{ik\alpha}}{R^2_{ik}}\Big)R_{ik\beta}\Big] \\ &= \frac{\partial A_{ijk}}{\partial R_{j\alpha}}R_{ij\beta} + \frac{\partial A_{ijk}}{\partial R_{k\alpha}}R_{ik\beta}\end{aligned} \tag{4.14}$$

Finally we just have to combine all these terms to obtain the derivative of this angular symmetry function with respect to the strains

$$\begin{aligned}\frac{\partial g^{\mathrm{a}}_{ij}}{\partial \epsilon_{\alpha\beta}} &= 2\Big[\frac{\partial A_{ijk}}{\partial \epsilon_{\alpha\beta}}B_{ij}B_{ik}B_{jk} + A_{ijk}\frac{\partial B_{ij}}{\partial \epsilon_{\alpha\beta}}B_{ik}B_{jk} + A_{ijk}B_{ij}\frac{\partial B_{ik}}{\partial \epsilon_{\alpha\beta}}B_{jk} + A_{ijk}B_{ij}B_{ik}\frac{\partial B_{jk}}{\partial \epsilon_{\alpha\beta}}\Big] \\ &= 2\Big[\Big(\frac{\partial A_{ijk}}{\partial R_{j\alpha}}R_{ij\beta} + \frac{\partial A_{ijk}}{\partial R_{k\alpha}}R_{ik\beta}\Big)B_{ij}B_{ik}B_{jk} + A_{ijk}\Big(\frac{\partial B_{ij}}{\partial R_{ij}}\frac{R_{ij\alpha}}{R_{ij}}R_{ij\beta}\Big)B_{ik}B_{jk} \\ &\quad + A_{ijk}B_{ij}\Big(\frac{\partial B_{ik}}{\partial R_{ik}}\frac{R_{ik\alpha}}{R_{ik}}R_{ik\beta}\Big)B_{jk} + A_{ijk}B_{ij}B_{ik}\Big(\frac{\partial B_{jk}}{\partial R_{jk}}\frac{R_{jk\alpha}}{R_{jk}}R_{jk\beta}\Big)\Big]\end{aligned} \tag{4.15}$$

Additionally, when the descriptors based on Chebyshev polynomials [177] were added to the ÆNET package, we also implemented their derivatives with respect to the strains.

To test our implementation, we strained a random structure of the TiO_2 example that comes with the ÆNET package, and then calculated the stress tensor using both our implementation of eq. (4.2) and a central 2-point rule. Figure 4.1 shows that, initially, the structure was not at a minimum of the energy, and that the stresses calculated with these two methods agree with each other. This was the case for all six independent components of the stress tensor. Therefore, we are confident in our implementation.

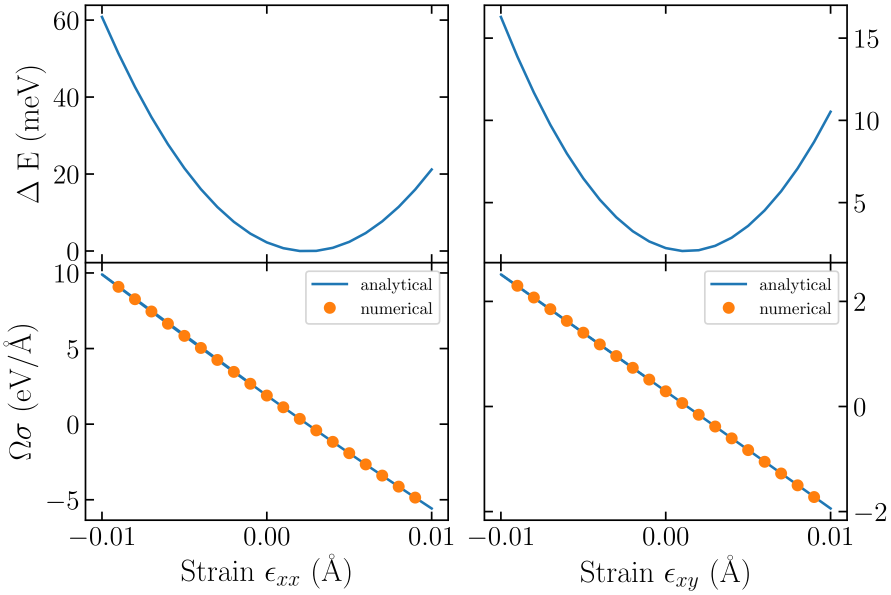


Figure 4.1: Energy and stress as a function of the strain applied to a random structure from the ÆNET package TiO_2 example. We multiplied the stress by the volume of the deformed structure and subtracted the energies by their minimum value. The strain was applied in the $xx-$direction in the left panels and in the $xy-$direction in the right panels. The stress was calculated analytically using eq. (4.2) and numerically with a central 2-point rule.

4.3 Training neural networks for forces and stresses

Whenever we perform a DFT calculation for a crystal structure, we obtain, among other properties, an energy, a set of $3N$ forces, and 6 independent components of the stress tensor. So, why not use all these values to reconstruct a potential energy surface? As mentioned already in section 3.5.2, this is equivalent to an increase of the training set size and to the addition of constrains to the learning of the PES: the minimization looks for the surface that contains a set of points, but no longer accepts any shape, only those that follow from a specific set of gradients. We should mention that other implementations of the Behler and Parrinello approach include this feature [204, 213, 417, 449], at least for forces, and that this is also quite common for other machine learning force-fields [187, 407, 427, 450].

The training of neural networks with information on the forces and stress requires a generalization of the cost function of the form

$$\epsilon = \sum_{\sigma} \left[\alpha \left(\sum_{i}^{N} E_i - E^{\text{ref}} \right)^2 + \frac{\beta}{3N} \sum_{\gamma}^{N} \sum_{\alpha}^{3} \left(F_{j\alpha} - F_{j\alpha}^{\text{ref}} \right)^2 + \frac{\gamma}{6} \sum_{k}^{6} \left(S_k - S_k^{\text{ref}} \right)^2 \right]_{\sigma} . \tag{4.16}$$

Here, α, β and γ represent parameters, that scale units and can be changed to increase the relative importance of different terms. Additionally, S_k stands for the k component of the 6 independent components of the stress tensor in the Voigt notation. We remind that the label "ref", for example in S_k^{ref}, identifies the value of the property calculated with the reference method. Then, the best values for the weights of the neural network come from the minimization of this cost function.

Our first approach to solve this problem involved a double optimization (of energy and forces) using a non-dominated sorting genetic algorithm (NSGA-II) [257, 451, 452]. At the time we did not consider the optimization of the stress tensor. While our second approach involved an extension of the back-propagation algorithm [417].

4.3.1 Genetic algorithms

The application of genetic algorithms to optimize the weights of a neural networks has been plentifully discussed in the past [453–465]. Recently, they have been revisited to train convolutional neural networks [466, 467] and to optimize their hyper-parameters [468, 469].

Here, we took advantage of the NSGA-II to optimize the RMSE of both energies and forces. Basically, this algorithm looks for candidate solutions in the Pareto front constrained by the objective function of each quantity. In this manner the fit can be greatly improved. While the forces improvements came only from the genetic algorithms, the optimization of the energies also involved the back-propagation algorithm.

The population was created from neural networks trained for energies using the TiO_2 example of the ÆNET package. Each of the 200 individuals consisted on a vector containing the weights of the neural networks. The neural networks of each element (Ti and O) contained 68 nodes on the input layer and two hidden layers with 10 nodes each. Without forgetting the bias and the output neuron, this corresponds to 1522 values to optimize. We ensured the variety of the initial population with the training (for energies) of 300 different potentials. Furthermore, we modified the weights of neural network potentials that provided similar predictions using uniformly distributed random numbers.

For the genetic algorithms, we used the implementation of the NSGA-II found in Ref. [470]. The mutation operator consisted on the increment of a random weight by a random factor, while for the recombination operator we used a simple two-point crossover. By this we mean that we cut the two parent's genotype vectors in two points and combined them as usual (see section 2.1.2) in order to form a different offspring. We did not construct a recombination operator that took into account the connections of each weight as discussed in Ref [454]. We tried different selection methods, but in the end we used the tournament selection.

Figure 4.2 shows the evolution of the Pareto fronts during the optimization. After 10 generations the neural networks still exhibit the behaviour of those trained only for energies: incredible small errors for the energies (below $\approx 3\,\mathrm{meV}$) and undesirable high errors for the forces (above $\approx 300\,\mathrm{meV/Å}$). After 60 generations the error in the forces improves, with a decrease of around 30 meV/Å for the minimum error. Meanwhile, the energy errors reduce to 0.5 meV. Finally, after 1000 generations, the energies become completely over-fitted, as can be expected by their errors below 2 meV. However the errors of the force predictions do not overcome the desired threshold of $\approx 100\,\mathrm{meV/Å}$. They still linger above 120 meV/Å. In fact, the desired threshold was only reached after tens of thousands of generations, which took a month in 200 cores of a supercomputer.

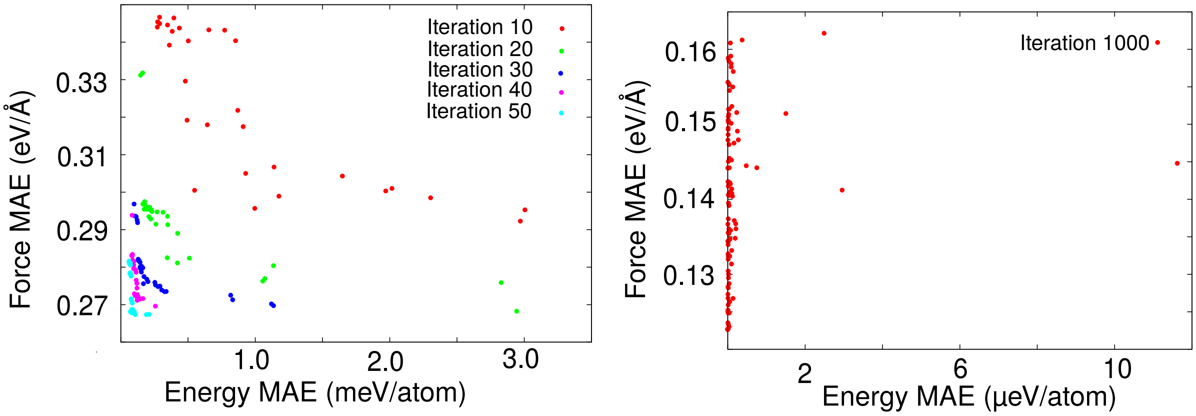


Figure 4.2: Pareto fronts for the optimization of the weights of the neural networks with respect to both energies and forces, after 10 generations in the GA, for the first 60 generations (left panel, only the best 30 individuals) and after 1000 iterations (right panel, only the 100 best individuals.

Our results agree with those present in the literature. While perfect for global optimizations, genetic algorithms suffer from a weakness in fine-tuned local search [456]. This reduces significantly their efficiency and its improvement requires more aggressive selection and mutation operators [455], more complex recombination operators [453–455], or their combination with a powerful local optimizer, such as the back-propagation algorithm [456].

Thus, we decided to follow another road to train the neural networks.

4.3.2 Back-propagation algorithm

In principle, the optimization of the cost function shown in eq. (4.16) through back-propagation can be achieved in two distinct ways: either by the extension of the neural network output (to also include forces and stresses) or by the generalization of the back-propagation algorithm. However, the former option will most probably lead to inconsistencies, as the relation between energy and its gradients is not imposed. Therefore, we chose the second option, which is much more robust, though it requires the derivatives of all three terms in the cost function with respect to the weights. The derivative of the first term (the energy term) has been shown already in section 3.4.1. For the others, consider the same definitions and that $\ddot{\varphi}^n$ represents the diagonal matrix that stores the second derivatives of the nodes of the n layer. Then, for the non trivial part of the derivative of the term containing the forces, we obtain

$$\begin{aligned}\frac{\partial F_{j\alpha}}{\partial w^n} = &\sum_i^N \Bigg(\left[\prod_{m>n}^{N_h+1} \dot{\varphi}^m w^m\right] \dot{\varphi}^n \otimes \xi^{n-1} \\ &+ \sum_{p=n}^{N+1} \left[\prod_{m>n}^{N_h+1} \dot{\varphi}^m w^m\right] \lambda^p \left[\prod_{q>1}^{p} w^q \dot{\varphi}^{q-1}\right] \otimes o^{n-1} \Bigg)_i ,\end{aligned} \tag{4.17}$$

where the products inside square brackets are ordered in decreasing order of layers, λ^p represents the diagonal matrix

$$\lambda^{\kappa}_{ij} = \left(\ddot{\varphi}^{\kappa}\ w^{\kappa}\ \xi^{\kappa}\right)_i \delta_{ij}, \tag{4.18}$$

and ξ^{κ} is defined through the recursion relation

$$\xi^{\kappa} = \dot{\varphi}^{\kappa}\ w^{\kappa}\ \xi^{\kappa-1}, \tag{4.19}$$

$$\xi^{0} = \frac{\partial o^{0}}{\partial R_{j\alpha}}. \tag{4.20}$$

We note that eq. (4.20) corresponds to the derivatives of the input of the neural network (or of the features) with respect to the position of the atom j which suffers the effect of the force $F_{j\alpha}$.

As an example, consider a neural network with two hidden layers such as the one shown in fig. 3.2. If the output of such a neural network is the contribution to the energy of a single atom, and the sum of all these contributions corresponds to the total energy of a structure, then the force can be calculated as

$$F_{\gamma\alpha} \ = \ -\sum_{i}^{N}\sum_{\kappa}\sum_{\zeta}\sum_{\lambda}\sum_{\tau}\varphi'(h^3_\kappa)w^3_{\zeta\kappa}\varphi'(h^2_\zeta)w^2_{\lambda\zeta}\varphi'(h^1_\lambda)w^1_{\tau\lambda}\frac{\partial o_{i\tau}}{\partial R_{\gamma\alpha}}. \tag{4.21}$$

Here we used $\varphi'(h^2_\zeta)$ to indicate the first derivative of the activation function at the point h^2_ζ. Then using the expressions above, we can calculate the derivatives of these forces with respect to the weights. For the weight matrix W^3, which contains the weights that connect the last hidden layer with the output layer we obtain

$$\begin{aligned}\frac{\partial F_{\gamma\alpha}}{\partial w^3_{\xi\phi}} = &-\sum_{i}^{N}\Bigg[\varphi'(h^3_\phi)\varphi'(h^2_\xi)\Big[\sum_{\lambda}\sum_{\tau}w^2_{\lambda\xi}\varphi'(h^1_\lambda)w^1_{\tau\lambda}\frac{\partial o_{i\tau}}{\partial R_{\gamma\alpha}}\Big]\\ &+\varphi''(h^3_\phi)o^2_{i\xi}\Big[\sum_{\zeta}\sum_{\lambda}\sum_{\tau}w^3_{\zeta\phi}\varphi'(h^2_\zeta)w^2_{\lambda\zeta}\varphi'(h^1_\lambda)w^1_{\tau\lambda}\frac{\partial o_{i\tau}}{\partial R_{\gamma\alpha}}\Big]\Bigg],\end{aligned} \tag{4.22}$$

where φ'' represents the second derivative of the activation function. For the weights W^2 that connect both hidden layers we get

$$\begin{aligned}\frac{\partial F_{\gamma\alpha}}{\partial w^2_{\theta\xi}} = &-\sum_{i}^{N}\sum_{\kappa}\Bigg[\varphi'(h^3_\kappa)w^3_{\xi\kappa}\Big(\varphi'(h^2_\xi)\varphi'(h^1_\theta)\Big[\sum_{\tau}w^1_{\tau\theta}\frac{\partial o_{i\tau}}{\partial R_{\gamma\alpha}}\Big]\\ &+\varphi''(h^2_\xi)o^1_{i\theta}\Big[\sum_{\lambda}\sum_{\tau}w^2_{\lambda\xi}\varphi'(h^1_\lambda)w^1_{\tau\lambda}\frac{\partial o_{i\tau}}{\partial R_{\gamma\alpha}}\Big]\Big)\\ &+\varphi''(h^3_\kappa)w^3_{\xi\kappa}\varphi'(h^2_\xi)o^1_{i\theta}\Big[\sum_{\zeta}\sum_{\lambda}\sum_{\tau}w^3_{\zeta\phi}\varphi'(h^2_\zeta)w^2_{\lambda\zeta}\varphi'(h^1_\lambda)w^1_{\tau\lambda}\frac{\partial o_{i\tau}}{\partial R_{\gamma\alpha}}\Big]\Bigg].\end{aligned} \tag{4.23}$$

Finally, for the weights W^1 that connect the input layer with the first hidden layer we can write

$$\begin{aligned}\frac{\partial F_{\gamma\alpha}}{\partial w^1_{\rho\theta}} = - \sum_i^N \sum_\kappa \sum_\zeta \Bigg[& \varphi'(h^3_\kappa) w^3_{\zeta\kappa} \Big\{ \varphi'(h^2_\zeta) w^2_{\theta\zeta} \Big(\varphi'(h^1_\theta) \frac{\partial o_{i\rho}}{\partial R_{\gamma\alpha}} \\ & + \varphi''(h^1_\theta) o_{i\rho} \Big[\sum_\tau w^1_{\tau\theta} \frac{\partial o_{i\tau}}{\partial R_{\gamma\alpha}} \Big] \Big) \\ & + \varphi''(h^2_\zeta) w^2_{\theta\zeta} \varphi'(h^1_\theta) o_{i\rho} \Big[\sum_\lambda \sum_\tau w^2_{\lambda\xi} \varphi'(h^1_\lambda) w^1_{\tau\lambda} \frac{\partial o_{i\tau}}{\partial R_{\gamma\alpha}} \Big] \Big\} \\ & + \varphi''(h^3_\kappa) w^3_{\zeta\kappa} \varphi'(h^2_\zeta) w^2_{\theta\zeta} \varphi'(h^1_\theta) o_{i\rho} \Big[\sum_\zeta \sum_\lambda \sum_\tau w^3_{\zeta\phi} \varphi'(h^2_\zeta) w^2_{\lambda\zeta} \varphi'(h^1_\lambda) w^1_{\tau\lambda} \frac{\partial o_{i\tau}}{\partial R_{\gamma\alpha}} \Big] \Bigg] .\end{aligned} \tag{4.24}$$

Finally, the derivative of the components of the stress tensor with respect to the weights of the neural network are acquired in a similar fashion. It just requires the replacement of $\frac{\partial o}{\partial R_{j\alpha}}$ by

$$\frac{1}{\Omega} \sum_j^N \frac{\partial o^0}{\partial R_{j\alpha}} \frac{\partial R_{j\alpha}}{\partial \epsilon_{\alpha\beta}} \tag{4.25}$$

in eqs. (4.17), (4.19) and (4.20), according to the definitions of forces and of the stress tensor (eqs. (4.2) and (4.3)).

We coded these expressions in the ÆNET package in a way that allows for the separate optimization of energy, forces, or stresses, or any of their combinations. We also implemented several activation functions and the derivatives necessary to satisfy eqs. (4.17), (4.19) and (4.20). Moreover, after testing the agreement between numerical and analytical derivatives for both the forces and stress derivatives, we used our implementation to construct several neural network force-fields, shown in the next section (section 4.4).

Yet, before proceeding, we would like to make some remarks concerning eqs. (4.17), (4.19) and (4.20). Activation functions with vanishing second derivatives, such as the ReLU, greatly simplify eq. (4.17), as only the first term remains. As a consequence, the equations above do not always provide an update for all weights of the neural network: the weight of the bias neuron will remain constant. Only training for the energy will update that neuron. Consequently, a double loop optimization procedure or a joint optimization is advised for these cases.

4.4 Example force-fields

We used the extension of the back-propagation algorithm to create force-fields for some selected semiconductors: elemental Si and Ge, the SiGe binary, and for some simple metals: Cu and Au. We chose these elements due to their comprehensive literature, which includes examples regarding classical force-fields [392, 393], tight-binding parameterizations [471, 472], and even machine learning force-fields [412, 473–475].

	Minima	Distorted	MD	2D	Total
Si	92	4999	13323	37	18451
Ge	94	3967	6435	37	10533
SiGe	671	15540	13561	0	29772
Cu	20	485	13191	0	13696
Au	27	1082	12259	0	13368

Table 4.1: Number of each kind of structures in our training sets. We note that the binary training dataset also included all the elemental minima structures.

As discussed in section 3.1, supervised learning, such as the regression we intend to perform using the neural networks to describe PES, requires five steps: the creation of a data set, that in this case contains information on crystal structures and the target properties, feature extraction, the selection of an algorithm, the training of the model, and the validation of the model to determine its predictability capabilities. We will proceed with the detailed discussion of each of these steps.

4.4.1 Data

The production of the data sets used in the training of the neural networks followed the strategies of Huran *et al.* [476] and Artrith *et al.* [212]. We start by the exploration of the chemical environment, and the selection of allowed crystal structures for each chemical composition using the MHM (see section 2.1.1). All the calculations were performed at the level of DFT with the PBE approximation for the exchange and correlation functional, as implemented in VASP code [477, 478]. More information on the calculations can be found in the appendix A. The output of this undertaking provided two sets of structures: one corresponding to the local minima of the PES, and the other to different steps of short MD simulations. Furthermore, we applied a series of geometrical distortions to the local-minima structure, namely volume-conserving orthorhombic and monoclinic strains (see fig. 4.3), and scaling of the lattice constants by up to ±10%. Lastly, we complemented the data sets of Si and Ge with two-dimensional minima structures obtained following the strategy of Borlido *et al.* [479, 480].

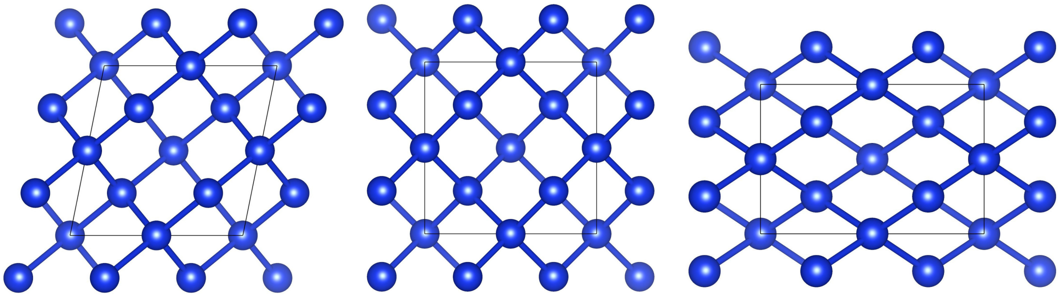


Figure 4.3: Cubic Si structure (middle panel) deformed by a volume-conserving monoclinic (left panel) and orthorhombic (right panel) strain.

Note that, contrary to parameterizations that include physical constrains, such as those for DFTB, and that for this reason might require a small number of training structures [476], machine learning algorithms are expected to identify the underlying patterns and physical relations only from the data. As such they normally require large training sets and, in fact, the accuracy of the model usually increases with increasing sizes of the training sets. To give an idea of the size of the data sets we will use Si as an example. Our data set for Si includes 26360 structures, out of which 131 correspond to minima, 54 to 2D minima, 7142 to distorted, and 19033 to MD structures. These structures were then divided into two sets: 70% for the training set and 30% for the test set. Table 4.1 displays the number corresponding to each type of structures in the training set for Si and for the other materials, that were divided in the same fashion. A quick glance at the table reveals that our data sets contain a rather low number of minima (both 3D and 2D), precisely the type of structures we are more interested in describing. For this reason, we increased the weight of these structures. Moreover, to insure a better description of the regions close to dynamical minima, we included a dimensionless weight factor u_σ in the objectives:

$$u_\sigma = \frac{0.2}{0.2 + \bar{F}_\sigma^2} \,. \tag{4.26}$$

where, $\bar{F}_\sigma$ corresponds to the average norm of forces acting on the atoms in the σ^{th} structure. The maximum of this function occurs for $\bar{F}_\sigma = 0$ and decreases monotonically with the increase of the forces.

On the other hand, we are mostly interested in energy differences and not in the absolute value of the total energy (which is meaningless for solids). An example of such difference is the formation energy. This quantity is defined as the difference between the energy of a structure and that of the ground-state of its elementary substances. Hence, we also found useful to increase the weight of the ground-state structures in the training set in order to improve the accuracy of their description with the neural network force-fields.

Table 4.2 exhibits the ranges of the energies, forces, and stresses in our data sets. These values will be important to understand the meaning of the errors calculated during the validation. The column for the norm of the forces reveals that no force in our training set exceeds a magnitude of 2.0 eV/Å. This is the case due to the filter we applied to clean the data when constructing the data sets. We removed duplicates and neglected structures with very high forces.

The distribution of the target properties in our data sets can be visualized, for example, in figs. 4.4 and 4.5 (for Cu and Si, respectively). It is evident that both data sets were constructed in a similar fashion. In fact, the differences between the distributions for Si and for Cu, come solely from the number of minima structures found for each of them. We note again that the structures identified as distorted come from distortions of these minima.

A comparison between the energy distributions reveals that the few lowest energy minima found for Cu are closer in energy than those found for Si. This is the reason why the energy distribution for Cu looks cut in fig. 4.4. Obviously, formation energies and energies follow the same distribution.

The forces distribution shows that our construction of the data sets focused on structures around the minima of the PES. Nevertheless, they also provide a rather complete description

	Formation Energy eV/atom	Energy eV/atom	Forces Component eV/Å	Forces Norm eV/Å	Stress kBar	Stress eV/Å^3
Si	[0.0,7.8]	[-5.4,2.3]	[-5.2,4.9]	[0.0,2.0]	[-1218,6827]	[-0.8,4.2]
Ge	[0.0,6.0]	[-4.6,1.4]	[-3.9,4.4]	[0.0,2.0]	[-549,3372]	[-0.3,2.1]
SiGe	[0.0,7.0]	[-5.4,1.7]	[-4.8,4.1]	[0.0,2.0]	[-1167,4633]	[-0.7,2.9]
Cu	[0.0,3.0]	[-4.1,-1.1]	[-3.3,3.3]	[0.0,2.0]	[-925,3098]	[-0.6,1.9]
Au	[0.0,3.3]	[-3.3,0.0]	[-4.2,4.2]	[0.0,2.0]	[-754,2726]	[-0.5,1.7]

Table 4.2: Range for the formation energy, energy, forces (norm and component) and stresses in our data sets (training and test).

of the forces space between 0 and 2 eV/Å. Both distributions look almost symmetrical around the maximum of 0.75 eV/Å (after neglecting the 0 eV/Å bar).

Lastly, the distribution of the stresses, for both Si and Cu, is the combination of 2 almost symmetric distributions, one around a maximum at 0 kBar and another around a maximum at -200 kBar. Most of the structures belong to the former, and the latter occurs since we performed some MHM runs at a pressure of 20 GPa. We included in the sets a few examples of high pressure structures (above 50 GPa) to increase the performance of the fit when the atoms are close together, during MD simulations or when the structures are compressed.

Furthermore, in fig. 4.6 we present the distributions for Cu separated by type of structure. As most of our structures come from MD simulations it is not surprising that the distributions for MD structures (panels g, h, and i) and the total distributions in our data sets look very similar. It is obvious that we mainly rely on this type of structures to obtain a fairly extensive description of all regions of the PES.

Regarding the minima structures (panels a, b, and c), we note that we do not have much control over their selection, they correspond to the structures found by the MHM. We can only increase the temperature to try to find higher energy structures or the pressure of the system to find structures subjected to some strain.

While the MD structures provide a rather broad set of forces, the distorted structures improve the description of the stresses in a broader region of the space (as observed in the panel f) corresponding to the stresses of the distorted structures). Meanwhile, the forces provided by the distorted set concern mainly the region in the proximity of the minima of the PES (see panel e).

Finally, we would like to mention that the distributions for the sets of the other materials resemble those depicted above for Cu and Si.

We anticipate that the force-fields, constructed from data sets generated in this manner, will provide an accurate description of structures close to dynamical equilibrium. Meanwhile, the inclusion of the MD and distorted structures assures the correct description of structures under different conditions of temperature and pressure, and with relatively large forces. Due to the Behler and Parrinelo approach and the extrapolation capabilities of neural networks, we do not expect force-fields created from these data sets to properly describe single atoms, molecules, or clusters. However, we believe them able to describe supercells that resemble locally the structures contained in the data sets, since the cut-off radius centered around

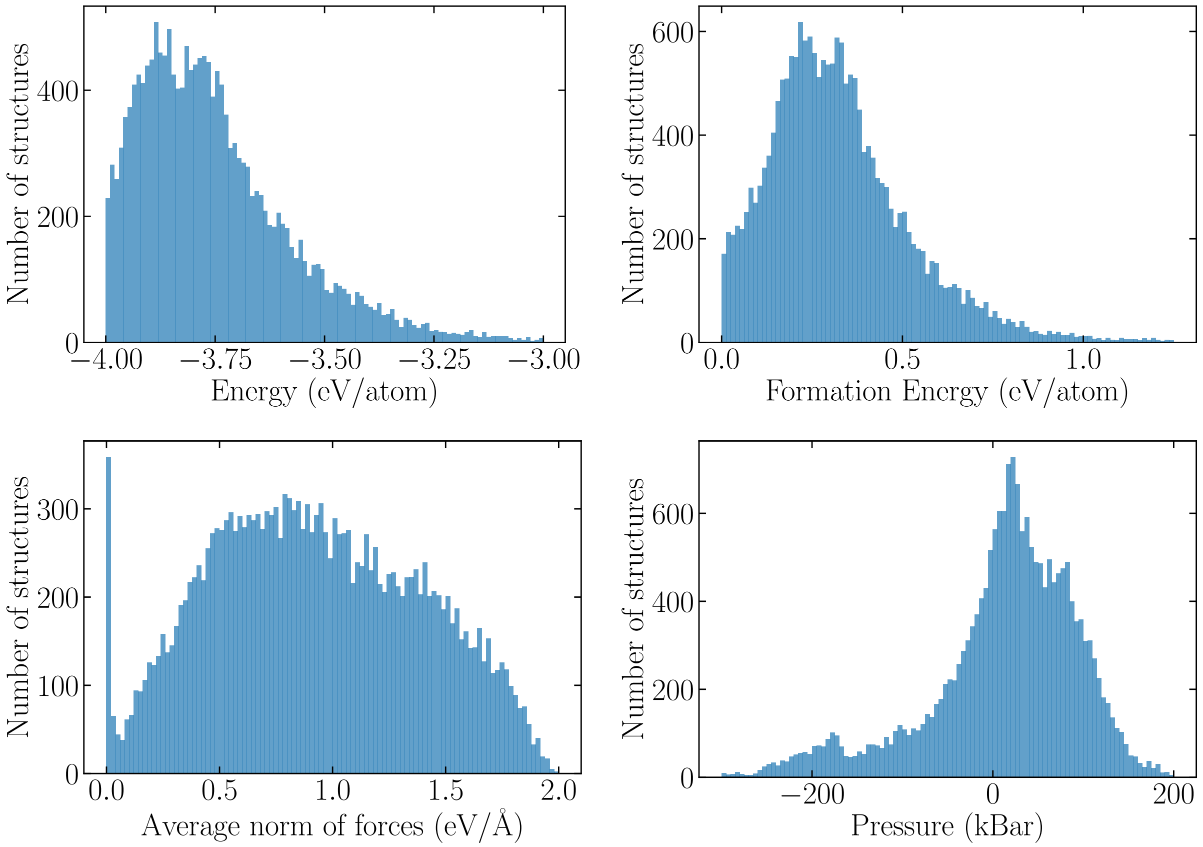


Figure 4.4: Distribution of the energies (top left panel), formation energies (top right panel), forces (bottom left panel), and stresses (bottom right panel) in our data set for Cu.

each atom encompasses the larger contributions to the energy.

4.4.2 Details on the features, algorithm and training

As features, we used a set of 8 radial symmetry functions and 18 angular symmetry functions for each elemental interaction. We present their definitions in eq. (3.3) and eq. (3.4), respectively (they correspond to types G^2 and G^4 in Ref. [213]). For the binary, these numbers increase to 16 radial and 54 angular functions. The values for these parameters can be found in Ref. [212] for titanium oxide. In principle, they should differ from element to element. For example, the value for η should be taken between the covalent radius and the cut-off radius. We resorted to pattern search techniques [481] in an attempt to optimize the values of the parameters of the symmetry functions, however this proved unsuccessful and we achieved no accuracy improvement.

Concerning the structure of the neural networks, we only used multilayer perceptron feed-forward neural networks. Nevertheless, we tried different architectures and, in section 4.4.3, we present results for the case of neural networks with three hidden layers and different number of nodes (5 and 50), trained using the Levenberg-Marquardt [482, 483] and the Broyden-Fletcher-Goldfarb-Shanno methods [484]. For the activation functions we chose the

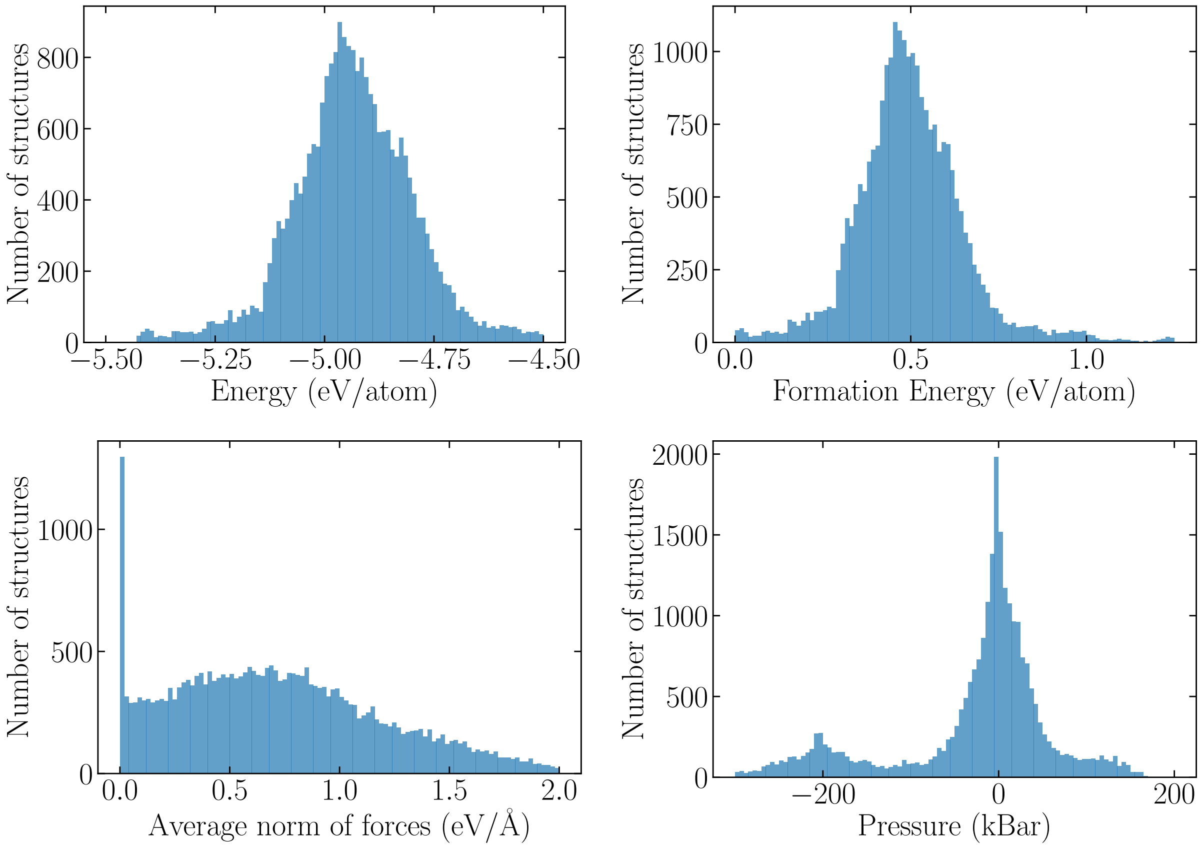


Figure 4.5: Distribution of the energies (top left panel), formation energies (top right panel), forces (bottom left panel), and stresses (bottom right panel) in our data set for Si.

softplus, except for the last layer, which is always linear in the ænet package.

Initially, we opted for the leaky ReLU, a popular activation function in many applications and that allows for a more efficient training: it remains the fastest activation function to evaluate, with the exception of the linear function, and permits a faster evaluation of the derivatives required for the back-propagation for forces and stresses, due to the vanishing terms. Nonetheless, usage of this activation function might result in an odd behaviour of several physical quantities, as shown in fig. 4.7. This results from the discontinuity of the first derivative of the (leaky) ReLU and, as such, this activation function should not be use in applications that require gradients of the energy, such as MD simulations.

4.4.3 Validation

The validation of our force-fields involved a series of tests, the visualization of both the reference target properties and their neural network counterparts, and the calculation of the (weighted) mean absolute error (wMAE) and root-mean square error (wRMSE) for the target properties: formation energies, forces and stresses.

Figure 4.8 shows the comparison between the formation energies computed by our potential and the DFT reference for Ge. The straight line $y = x$ represents the perfect fit.

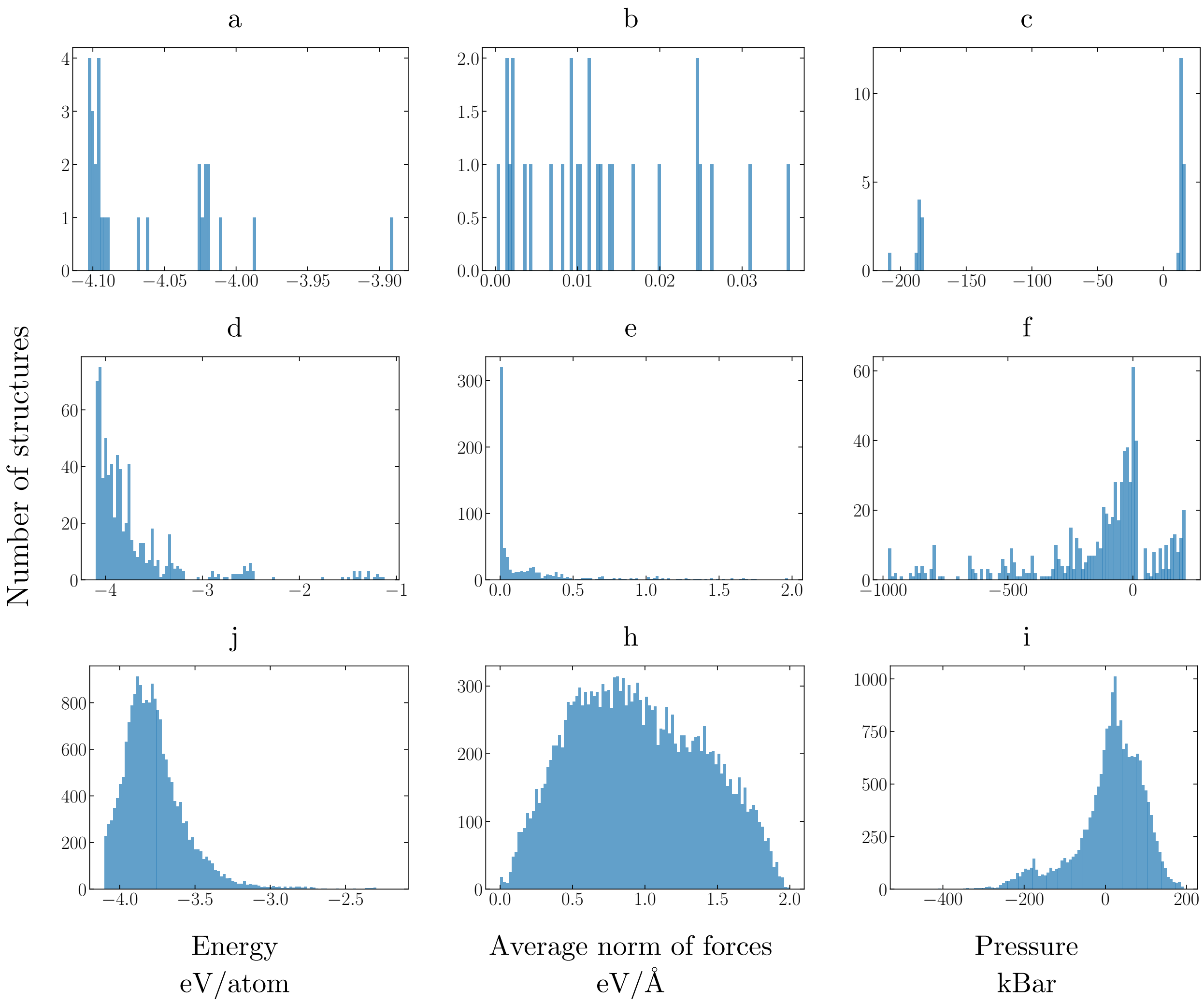


Figure 4.6: Distribution of the energies (left column), forces (middle column), and stresses (right column) in our data set for Cu separated by type of structures: minima (top row), distorted (middle row), and MD (bottom row).

The inset shows the region containing most of our structures: between 0 and 1 eV/atom. Evidently, the neural network predicts fairly accurately and consistently the energies across the whole range: the largest errors (around 100 meV) correspond to very unstable structures with high forces, to which we anyway attributed a small weight during training, while most structures exhibit errors below 25 meV. Meanwhile, fig. 4.9 displays a similar comparison for the forces (also for Ge). The left panel concerns a neural network trained only for the energy while the right panel is associated with a neural network trained for energies, forces, and stresses. The improvement of the results due to the joint training is astonishing. This indicates not only a decrease of the prediction errors, but also a correction over systematic errors. Furthermore, fig. 4.9 demonstrates plainly the importance of the addiction of the extra terms (in this case forces) in the cost function.

Additionally, we compared the error in the predictions made by our force-fields with other

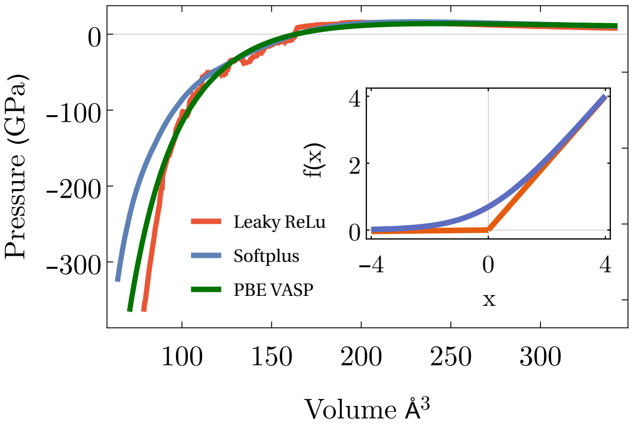


Figure 4.7: Change in pressure with the volume for the cubic silicon structure. The neural networks used to make this figure were only trained for energies and forces. The inset plot the leaky ReLu and the softplus activation function.

well known methods, namely classical force-fields and DFTB, and with a neural network force-field trained only for energies. In particular, we used Stillinger-Weber [393, 485] and the Tersoff potentials [486, 487] to compare with the Si and Ge results, and the Slater-Koster files from the (DFTB) parameters set pbc [472] and matsci [471] to compare with the results of Si and Cu. For the calculations, we used the LAMMPS code [488] for the classical force-fields and DFTB+ [489] for the DFTB calculations. Table 4.3 displays the results of these calculations. In general, our force-fields, trained with a joint optimization of energy, forces, and stresses, achieve errors in the formation energy below 50 meV/atom, forces smaller than 100 meV/Å, and stresses under 15 meV/Å^3.

The neural networks clearly outperform the other methods for all cases studied. We recall that neural networks exhibit a linear scaling with the number of atoms, such as classical force-fields, although classical force-fields provide more efficient calculations due to the smaller prefactor. Yet they scale considerably better and provide more efficient calculations than DFTB.

Noteworthy to mention are the small errors obtained for the simple metals, which are close to the expected errors of standard machine learning force-fields (around 5 meV/atom for the energy errors, depending on the material). Actually, the magnitude of these errors reveals that the joint training is not required to obtain accurate results, provided that the training sets are large enough and sufficiently complex.

On the other hand, the best energy RMSE error for Si that we show in table 4.3 is 38 meV/atom, which was obtained when the neural network was trained only for energies. In fact, training for energies, forces, and stresses constitutes a multi-objective optimization. Therefore its solution falls onto a Pareto curve, i.e., the optimization of one of the objectives can degrade the quality of others, and the degradation is larger, the further away the Pareto curve is from the origin. This is what we observe in table 4.3. In general, the multiple optimization improves the values of the forces and stresses, yet it also degrades the values for the energies. This degradation is more visible for the semiconductors because their errors

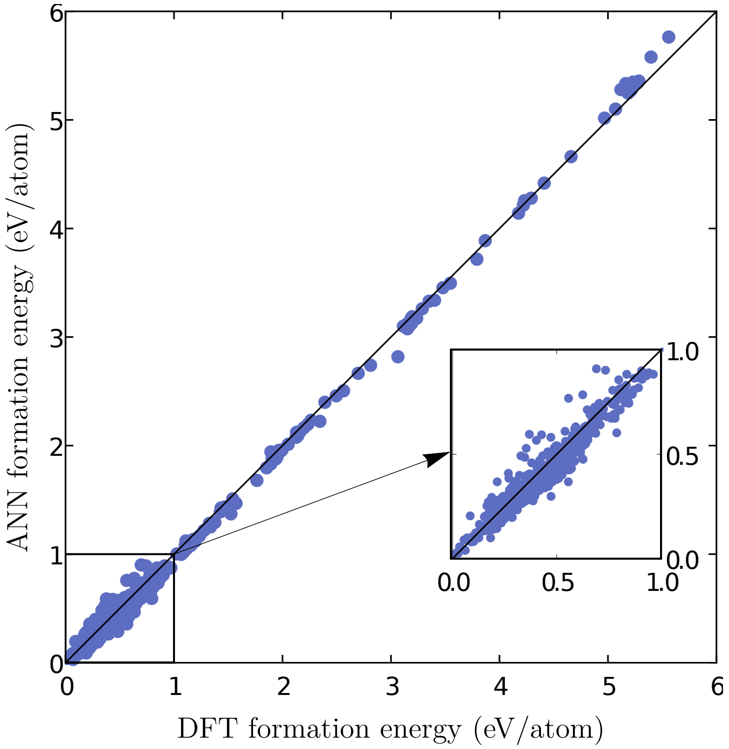


Figure 4.8: Comparison between formation energies calculated with DFT and with our force-field for Ge. The neural network used to model the PES had 3 hidden layers with 50 nodes each.

are higher. The reason for such values however remains a mystery. We used the same optimization algorithm for all the materials presented, so our optimization procedure should not be the reason for these high errors. It is possible that our optimization for silicon and for germanium is stopping in a local minimum though. In an attempt to resolve this problem, we tried to increase the number of symmetry functions, to optimize their parameters, to change the descriptor [177], to increase the number of layers of the neural network, and the training set size by addition of more MD structures. All of these proved unsuccessful. It is our conviction that this is due to the diversity of our data sets. And to subdue these errors would require more complex neural network structures, better input features, and significantly larger training sets. Maybe even training sets selectively increased using active or reinforcement learning.

We would like to direct attention to one of the main problems of machine learning methods nowadays: the lack of libraries and databases containing the different features, training sets, and methodologies proposed by different researchers, so that all of these could be benchmarked. This would provide a better comparison between different force-fields, than just a comparison between errors found with completely different data sets. As a small step to resolve this issue, we made available both our force-fields, our implementation, and our data sets.

Finally, we mentioned several tests at the beginning of this section. These consisted of small MHM runs, phonon calculations, and small simulations under different conditions to test the evolution of the target properties, such as the movement of two atoms in a box, the

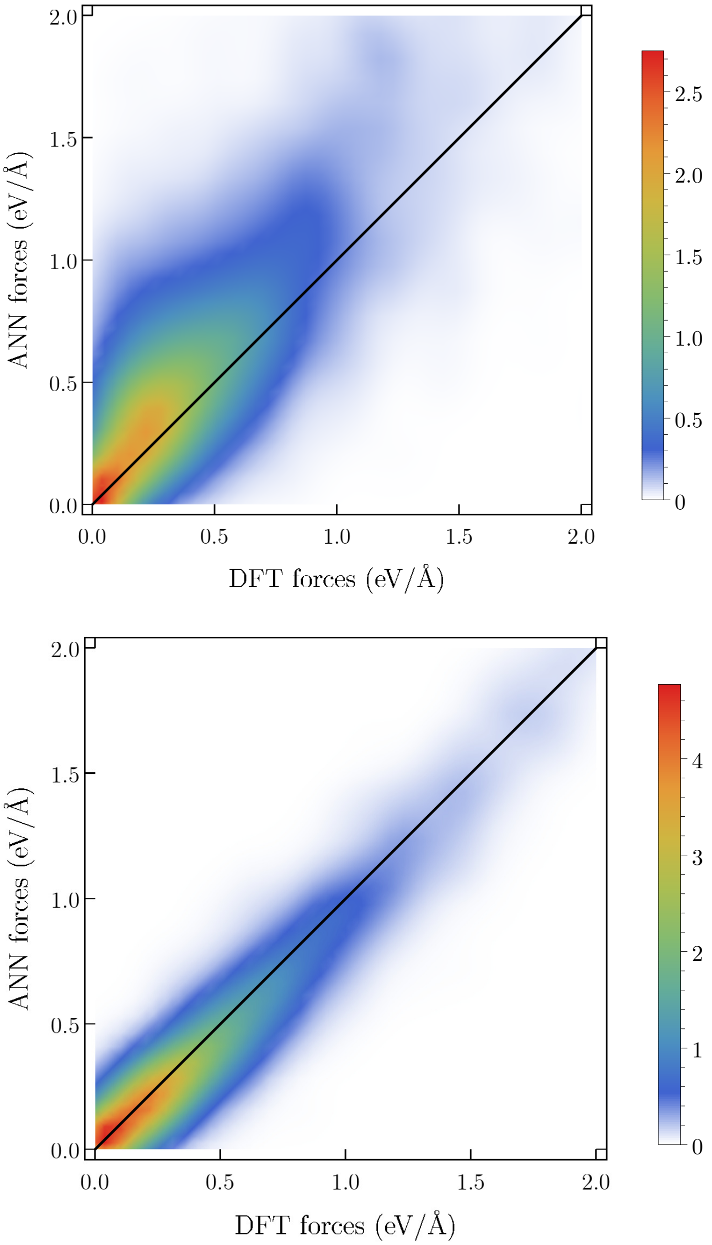


Figure 4.9: Comparison between forces calculated with DFT and with our force-field for Ge. The neural networks used to model the PES contained 3 hidden layers with 50 nodes each, and were optimized for energies (top panel) and with a joint training (bottom panel) of energies and its gradients. The probability density function (PDF), shown with the color gradient, displays the number of structures that can be found at each point (based on a smooth kernel density estimate).

movement of an atom inside a crystal structure, the gradual change of lattice parameters, and the gradual application of volume-conserving orthorhombic and monoclinic strains. These tests contributed to several improvements of our methodology. For example, the test with the lattice parameters revealed the problem with the ReLU in fig. 4.7. Furthermore, these

Table 4.3: Weighted mean absolute errors (MAE) and root mean square errors (RMSE) for formation energy (FE), forces and stresses calculated with different methods. All networks had 3 hidden layers with 5 neurons each. Results should be compared with the ranges in table 4.2.

		FE meV/atom		Forces meV/Å		Stress meV/Å^3	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
Si	this work	54	65	76	119	8	13
	e-training	29	38	162	298	11	16
	S.-Weber	784	1331	925	1594	231	490
	Tersoff	194	228	362	568	29	48
	pbc	299	386	190	282	57	106
	matsci	504	577	504	754	123	172
Ge	this work	34	40	46	75	5	9
	e-training	14	22	77	155	5	8
	S.-Weber	276	388	321	558	84	127
	Tersoff	434	559	448	759	81	122
SiGe	this work	78	89	87	127	6	10
	e-training	64	84	186	279	12	18
Cu	this work	4	6	13	18	5	8
	e-training	3	4	17	26	7	10
	matsci	228	343	656	937	133	228
Au	this work	8	11	25	37	2	3
	e-training	9	11	36	55	2	3

tests helped us to realize a problem with the description of the repulsion between atoms: if the training set lacks structures containing small inter-atomic distances, the neural network does not learn what to do for these instances. This can lead to unexpected behaviours during simulations, such as two atoms in the same position. Two easy solutions to this problem consist on the addition of such structures to the training set, or the inclusion of a repulsive term in the energy provided by the neural networks. We tried both. For the latter solution, we implemented a repulsion term similar to the repulsion part in the Lenard-Jones potential [490].

The neural network force-fields constructed here revealed an accuracy appropriate for the energies with respect to the training data. By appropriate accuracy we mean a similar definition as provided by Pople in his Nobel prize lecture [67]. Essentially, this corresponds to a global accuracy of 1 kcal/mol (roughly 43 meV/atom) for energies with respect to experimental values. Moreover, our force-fields provide an acceptable accuracy for forces and stresses, even when trained with rather humble data set sizes.

4.5 Applications

To further assess the predictability capabilities of our neural network force-fields and to demonstrate their usefulness, we calculated some properties which are not directly associated with the target properties we trained for. To be more precise, we calculated phonon dispersion curves for cubic Si and Cu, and the melting temperature of Cu and Au using our force-fields. Additionally, we combine them with the MHM, to investigate the formation of defects in large supercells of cubic silicon.

4.5.1 Phonon dispersion

We calculated the phonon dispersion for cubic silicon and copper, under the frozen-phonon technique using the PHONOPY package [491].

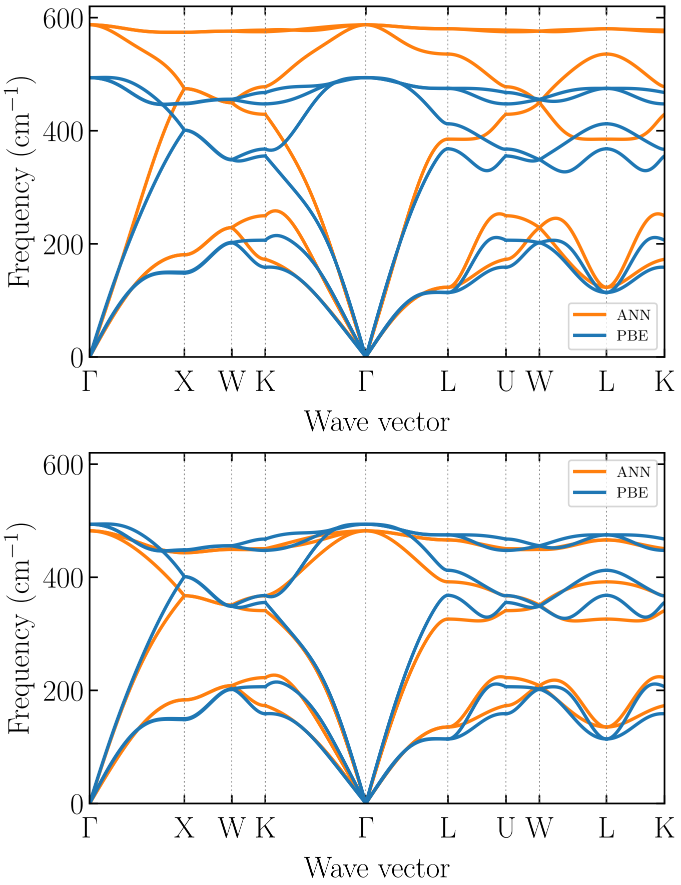


Figure 4.10: Phonon dispersion of cubic silicon calculated with the PBE functional and with a force-field only trained for energies (top panel) and another where we perform a joint training of energies, forces, and stresses (bottom panel).

Figure 4.10 displays the comparison between the phonon frequencies obtained with the our neural network force-fields for Si and the PBE functional. While the neural network used to calculate the phonons on the panel on the left was trained only for energies, the neural network used for the panel on the right was trained for energies, forces and stresses. When trained only for energies, the neural networks reproduce the PBE speed of sound, however they overestimate the optical phonons. Additionally, the acoustic branches deteriorate away from the Γ point. However, when optimized with a joint training, the neural networks reproduce remarkably the phonons across the entire Brillouin zone.

Moreover, we should note that the construction of our data sets is not focused on a particular kind of structure. In fact, we did not train a network with a training set based on cubic silicon structures and then calculated the phonons for it. Instead, we tried to represent a large region of the PES in an unbiased way. Therefore, we expect similar results for different kinds of structures. Then, this result is also quite general: for any structure, a better phonon dispersion curve can be obtained if the forces are improved.

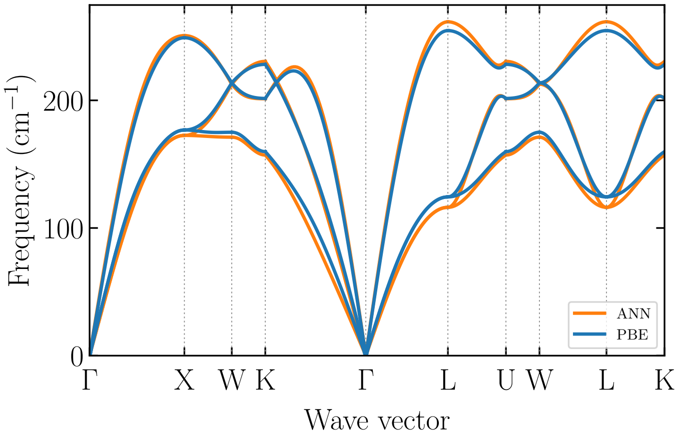


Figure 4.11: Phonon dispersion of cubic copper calculated with the PBE functional and with a force-field trained for energies, forces, and stresses.

On the other hand, the errors in the forces obtained by the neural networks in the description of Cu were rather low, independently of the target properties trained. So, we expect a good description of the phonons by the neural networks for both training methods. Figure 4.11 depicts the phonon dispersion of copper, and indeed the phonons frequencies are almost perfectly replicated over the whole Brilloin zone. Similar results can be found in the literature, for example for sodium in Ref. [409], for calcium fluoride in Ref. [492] or for copper, gold, and palladium in Ref. [421].

4.5.2 Molecular dynamics and melting temperature

We performed molecular dynamics simulations using the Berendsen thermostat and barostat [110] as implemented in the ASE package [493], to determine the melting temperature of face-centered cubic Cu and Au.

For the simulations of Cu, we chose a time step of 10 fs, and values of 500 fs and 1000 fs for the coupling constants of the thermostat and barostat, respectively. For Au, we took 400 fs for the thermostat and 500 fs for the barostat. We also used the isothermal compressibility of each element.

The methodology consisted on the thermalization of the system at 100 K and 1 bar, followed by a linear increase of the temperature of the thermostat using different constant heating rates. The melting temperature was then predicted from two distinct criteria: (i) From the maximum of the heat capacity, following the method of Qi, et. al [96]. (ii) From the maximum of the second derivative of $f_L = \sqrt{< u^2 >}/d$, where the numerator is the mean square atomic displacement and d is the inter-atomic distance. This corresponds to the Lindemann criterion [494]. However, instead of using the Lindemann constant, we analyzed the behaviour of the Lindemann function (f_L).

The change of the potential energy of Cu during the simulation can be seen in fig. 4.12 for different heating rates. The potential energy rises linearly with the increase of the temperature and leaps at the phase transition, after which it continues to rise linearly. For higher heating rates, the phase transition occurs so quickly, that there is no step, but instead a gradual increase of the potential energy. Moreover, the temperature of the melting point decreases with the decrease of the heating rate. Lastly, the increase of the number of atoms decreases the oscillations of the potential energy, as expected.

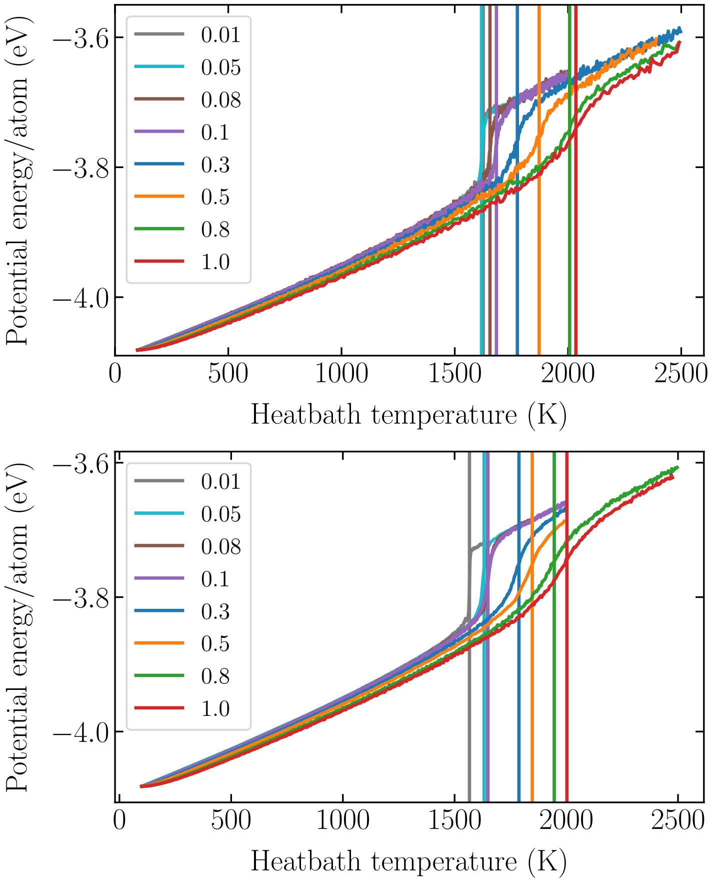


Figure 4.12: Variation of the potential energy of face-centered cubic Cu with the heathbath temperature for different heating rates, for a system with 500 (top panel) and 5000 (bottom panel) atoms. The vertical lines correspond to the transition temperature.

Similarly, fig. 4.13 shows the variation of Lindemann function for Cu atoms and different heating rates. We can clearly visualize that the melting temperature decreases with the decrease of the heating rate. Furthermore, the Lindemann function increases linearly in time, until a certain point and then the slope increases rapidly. The phase transition occurs at the point of maximum curvature of the Lindemann function, i.e., the maximum of its second derivative.

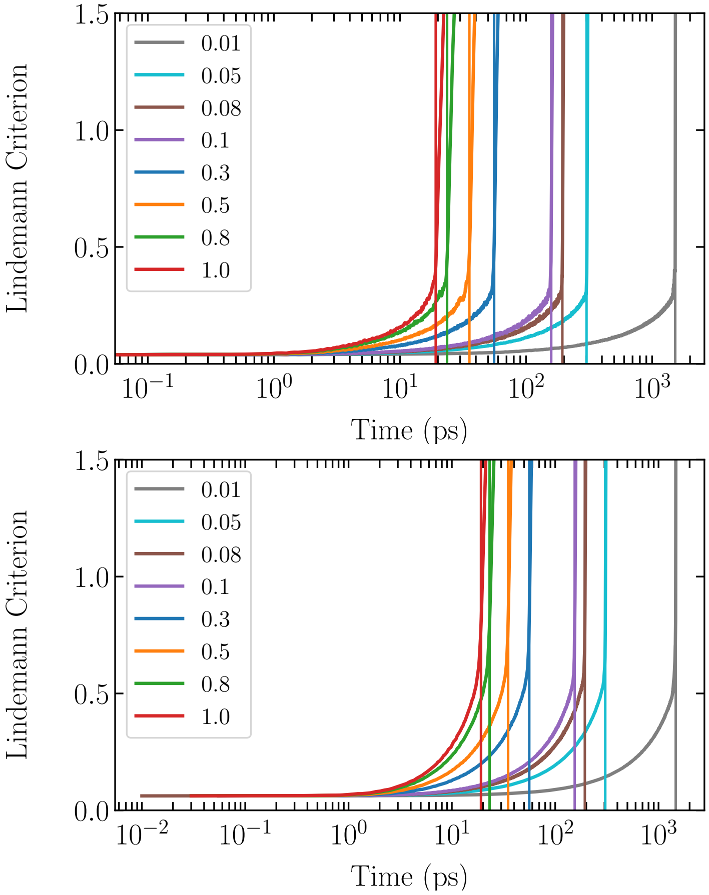


Figure 4.13: Variation of the Lindemann function face-centered cubic Cu with the simulation time for different heating rates, for a system with 500 (top panel) and 5000 (bottom panel) atoms. The vertical lines correspond to the transition temperature.

We condensed the results for Cu in fig. 4.14. Both methods provide similar results, and, as expected, the melting temperature converges with the decrease of the heating rate. Unexpectedly, the melting temperature seems to be converged with respect to the supercell size already for the simulation with just 500 atoms. In order to determine the melting temperature for an infinitesimal heating rate, we fitted a straight line to the points concerning heating rates smaller than 0.2 K/step. This resulted in a melting temperature of 1510 K, which slightly overestimates the experimental value of 1358 K [495].

In principle, we should compare this value with the melting temperature obtained by the PBE functional. This was the reference method to which we trained our neural networks, and shortcomings of the PBE functional might result in a different melting temperature. However, we found no DFT simulations related to this problem, probably due to the overwhelming

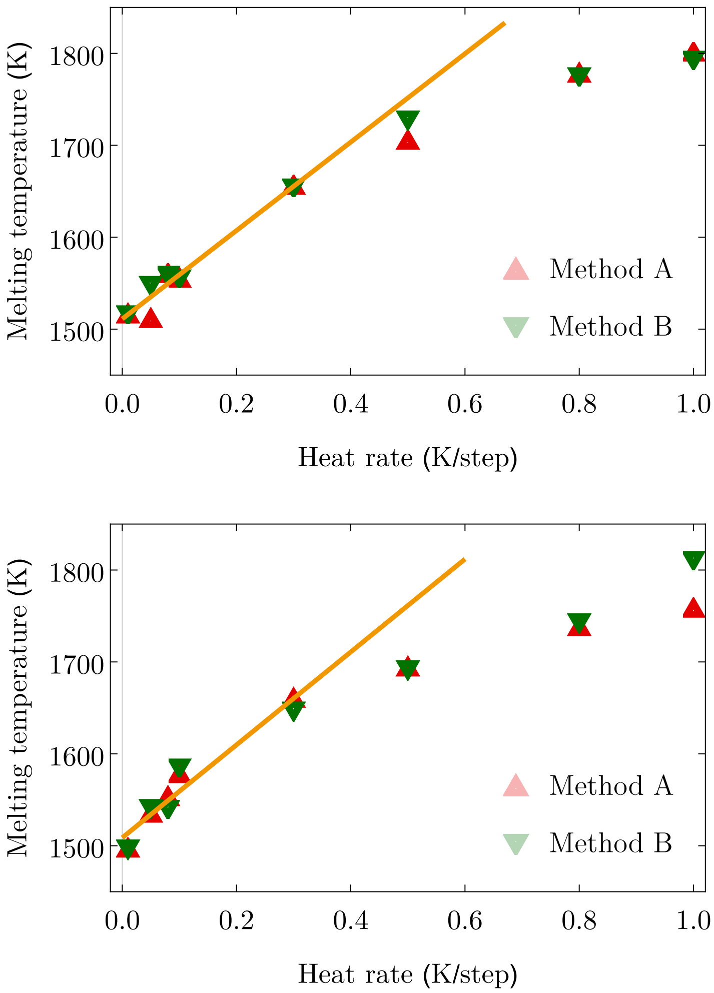


Figure 4.14: Variation of the melting temperature of Cu with the heating rate for a system with 500 atoms (top panel) and 5234 atoms (bottom panel). The straight lines are linear fits $(a + bx)$ to the points with heating rate below 0.4 K/step. Coefficients are $a = 1511.1 \pm 5.14$ and $b = 481.0 \pm 34.8$ (500 atoms) and $a = 1508.7 \pm 7.2$ and $b = 505.3 \pm 48.4$ (5324 atoms). The errors only concern the fit.

computational resources necessary. The literature contains some studies using embedded-atom method (EAM) potentials fitted to DFT energies, whose results are then corrected with DFT quantities. These studies delivered melting temperatures of 1176 ± 100 K [496] with the PW91 functional [497, 498], and 1251 ± 15 K [97] with the PBE functional. Moreover, another study with EAM potentials resulted in melting temperatures [98] of 1780 K and 1360 K. The former value came from the heat capacity method for a heating rate of 4 K/ps and a supercell with 500 atoms, while the latter was calculated from the extrapolation of melting temperatures of Cu clusters, to a cluster of infinite size.

We also used the same methodology to investigate the melting temperature of the face-centered cubic structure of gold, whose experimental value is 1338 K. Figure 4.15 shows our results for a supercell containing 500 atoms. We obtained a melting temperature of

1048 ± 12 K. A value which agrees perfectly with the melting temperature obtained by EAM potentials (1090 K) [499], and disagrees with the overestimated value of the ReaxFF [500] force-field (2125 ± 25 K).

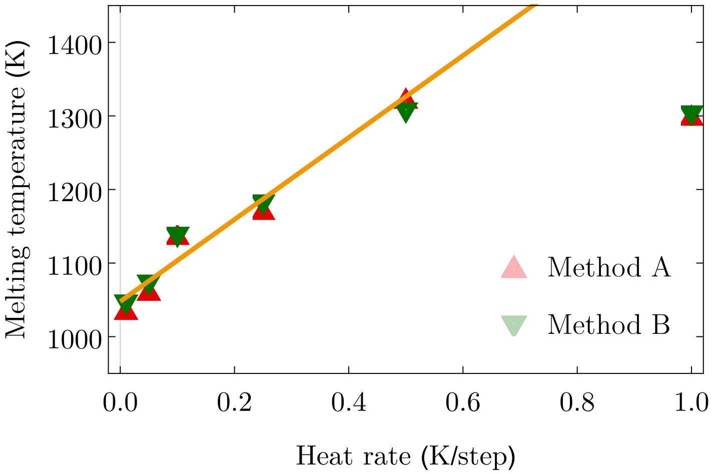


Figure 4.15: Variation of the melting temperature of Au with the heating rate for a system with 500 atoms. The straight line ($a + bx$ with $a = 1047.8 \pm 12.3$ and $b = 557.1 \pm 89.4$) was fitted for heating rates below 0.4 K/step. The errors only concern the fit.

The melting temperatures found resorting to our force-fields agree quite well with the experimental ones. We think that difference between them comes from limitations of the PBE approximation and not from the fit itself, which was rather good (see table 4.3.)

4.5.3 Structural prediction and defects

We studied the formation of defects in a supercell of cubic Si. In order to confirm our results with DFT we perform the calculation with cells containing only 216 atoms. Our methodology consisted in the exploration of the energy surface of Si using the MHM, coupled with our force-field for Si, while constraining three layers of silicon in every direction of the supercell. This ensures that the structures found by this global structure prediction method differ from the cubic silicon structure only by point defects (and deformations caused by them) inside the enclosing constrained volume. For example, the first local minima found by the MHM corresponds to the pristine structure. However, the second contained a vacancy and a interstitial. Furthermore, we repeated this methodology with supercells containing additional and fewer atoms, specifically 216 ± 3 atoms. We let the MHM run until around 400 minima were found for each supercell. We should note that performing these calculations with DFT would require months in a supercomputer, while with the neural network force-field this took about a week in a single core of a standard desktop computer. This is remarkable and shows the potential of our force-fields: indeed they allow to study complex large systems in an efficient, more systematic way.

After finding all these minima, we removed the constrains and relaxed all these structures with our force-field. Moreover we also relaxed them with DFTB, using the implementation

Defect	Energy (eV)	N	Spg
Split (X) interstitial	3.581	217	35
Tetrahedral (T) interstitial	3.647	217	215
Hexagonal (H) interstitial	3.625	217	160
Extended split (EX) interstitial	4.263	217	35/8
Vacancy	3.765	215	215
Vacancy (deformed)	3.664	215	111

Table 4.4: PBE formation energy of the most common point defects in Si that we found. N indicates the number of atoms and Spg the space group number.

of DFTB+ [489] and the parameterization of Ref. [476], and with DFT using the PBE approximation as explained in the appendix A.

We note that the neural network force-fields found fictitious minima that disappeared with the PBE relaxations. Furthermore, some of the defect formation energies were incorrectly estimated by our force-field. The DFTB parameterization also led to some of these problems, yet at a smaller scale. Regardless, we found most of the well-known low-energy point defects of silicon [501, 502] and listed them in table 4.4.

In particular, we found a slightly distorted version of the split (X) dumbbell interstitial that we depict in fig. 4.16 alongside the tetrahedral and the hexagonal interstitials. The DFT energies found agree with the values found in the literature, for example in Refs. [501, 502]. In fig. 4.17 we show the vacancies listed in table 4.4. While both correspond to the pristine diamond structure of silicon with a missing atom, in the latter the lattice is also slightly deformed. This leads to fewer symmetries and to a lower formation energy for the defect (akin to a Jahn-Teller deformation [503]). Additionally, the left panel of fig. 4.17 shows the extended split (EX) interstitial. Missing in table 4.4 is the fourfold coordinated defect [502] (FFCD) which has the lowest formation energy among all defects of diamond structured silicon. This defect is formed by a bond rotation followed by reconnecting all broken bonds [475]. Similarly to other force-fields [504], and even other machine learning

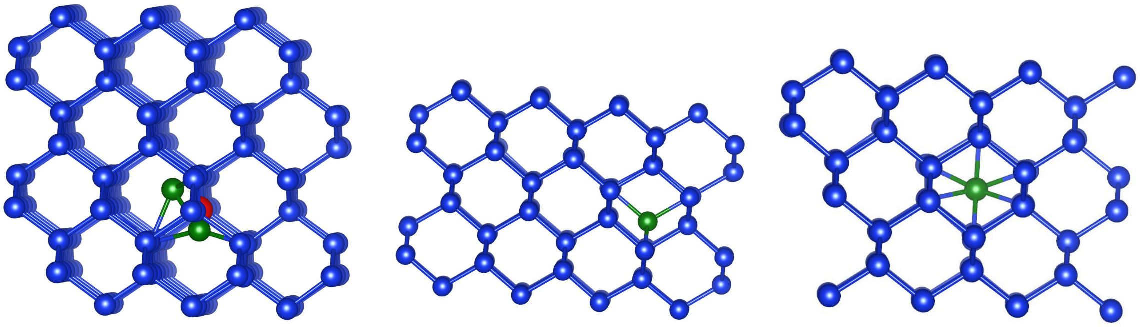


Figure 4.16: Example defects in Si: split (X) dumbbell (left), the tetrahedral (middle), and the hexagonal (right) interstitial. In blue we show Si atoms that are in the same positions as the atoms in the pristine structure, in red vacancies, in green the interstitial atoms, and in grey atoms that were displaced due to the defect.

Defect	Energy (eV)	N	Spg
Frenkel	4.732	216	35
Special FFCD pair	4.193	216	1
FFCD+vacancy	4.981	215	1
X+T	5.392	217	160
X+FFCD	4.362	217	1
Di-vacancy	5.441	214	12
W	5.214	218	1
complex	5.739	218	8

Table 4.5: PBE formation energy of the other low formation energy defects. N indicates the number of atoms and Spg the space group number.

force-fields [475], our neural network force-field could not stabilize this defect. In fact, relaxation of a structure containing this defect leads to the Si diamond structure. This happens because our data sets do not contain elements resembling this bond rotation process. So, we believe that the solution to this problem involves the extension of the training set with such structures and, maybe, the development of features (in this case symmetry functions) dependent on the torsion (or dihedral) angles present in the structures.

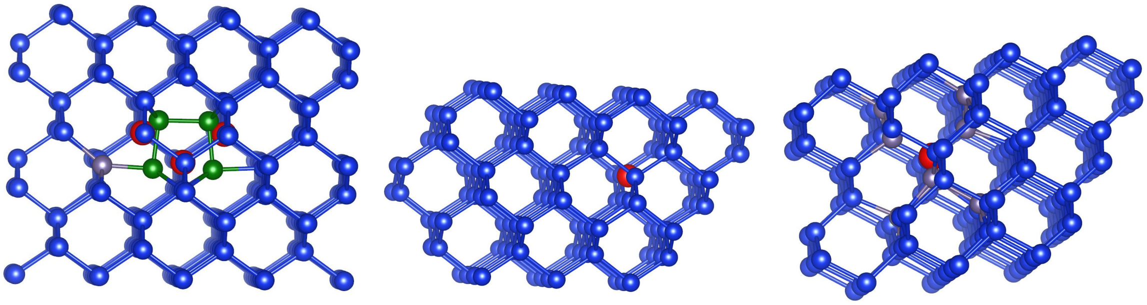


Figure 4.17: Extended split (EX) interstitial (left panel) and vacancy defects in Si without deformation (middle panel) and with deformation (right panel) of the neighbouring atoms. In blue we show Si atoms that are in the same positions as the atoms in the pristine structure, in red the vacancy, and in grey atoms that were displaced due to the defect.

Table 4.5 lists other interesting low energy defects found with our methodology. The first two defects in the table were found with a supercell of 216 atoms and can be visualized in fig. 4.18. The first of these corresponds to a Frenkel defect: one atom leaves its position in the lattice and becomes an interstitial. The second one is much more interesting, consisting of a special case of two pairs of FFCDs in which the pairs are rather close together. The distance between the closest atoms from each pair is just 2.33 Å, for comparison the distance between the interstitials in the FFCD is 2.25 Å. This is a remarkable finding, since our force-field does not stabilize the single FFCD but manages to stabilize the special case when two pairs of FFCD interact with each other. Since the pairs interact with each other, the formation energy of this defect (4.193 eV) is smaller than the combination of two FFCD (2.42 eV).

The following three defects from table 4.5 are depicted in fig. 4.19. The first of these was

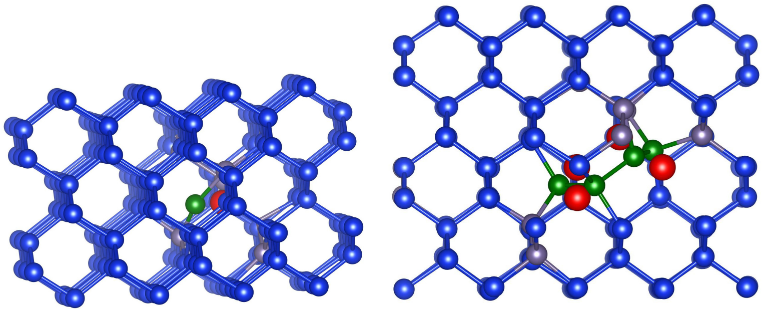


Figure 4.18: Frenkel defect (left panel) and special pair of two FFCDs (right panel). In blue we show Si atoms that are in the same positions as the atoms in the pristine structure, in red vacancies, in green the interstitial atoms, and in grey atoms that were displaced due to the defect.

discovered with a supercell containing 215 atoms and consists of a FFCD and a vacancy. The last two appeared from a supercell containing 217 atoms. The former consists of a combination of the X and the T interstitials. The latter involves a X interstitial and a FFCD. Once again, our force-field managed to find a composite of point defects involving the FFCD.

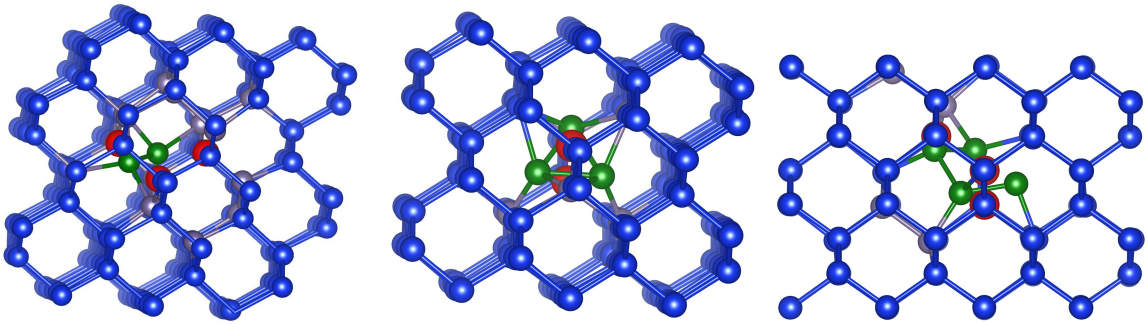


Figure 4.19: Composite defects: combination of a vacancy with the FFCD (left panel), XT di-interstitial (middle panel), and a combination of a X interstitial with the FFCD (right panel). In blue we show Si atoms that are in the same positions as the atoms in the pristine structure, in red vacancies, in green the interstitial atoms, and in grey atoms that were displaced due to the defect.

Finally, the last 3 defects of table 4.5 can be visualized in fig. 4.20. The first of these was found from a supercell containing 214 atoms and consists of a di-vacancy. In this defect two neighbouring atoms are missing from the lattice. Its formation energy agrees with other results present in the literature, for example in Ref. [501].

The last two defects appeared in the search involving a supercell containing 218 atoms. One is a W di-interstitial with a slight deformation of the lattice, while the other is a rather complex defect. It appears to be a FFCD combined with two X interstitials, with all interstitials so close together that they all interact and form a defect with a formation energy of 5.739 eV.

We found no low formation energy defects (below 6 eV) for supercells containg 213 and 219, i.e. supercells of diamond Si missing 3 atoms or containing 3 additional atoms. Moreover, here we discussed only the low-formation-energy defects that we found. Our force-fields

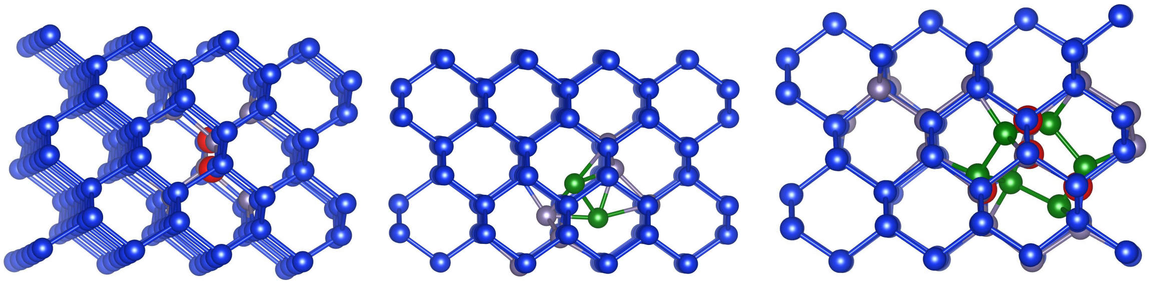


Figure 4.20: Three defects of diamond Si: Di-vacancy (left panel), W di-interstitial (middle panel), and a complex defect. In blue we show Si atoms that are in the same positions as the atoms in the pristine structure, in red vacancies, in green the interstitial atoms, and in grey atoms that were displaced due to the defect.

found other, such as di-insterstitials composed of 2 tetrahedral insterstitials and a combination of a tetrahedral and a hexagonal interstitial. The formation energy of these composite defects was close to the sum of the energies of the point defects.

To our knowledge, this was the first application of machine learning force-fields for the systematic exploration (with structure prediction methods) of defects in diamond structured silicon. Other machine learning studies involve only the relaxation of the defect structures with the force-fields.

Moreover, these results lead us to devise a methodology for future global structure predicting endeavours. Fast-to-evaluate neural network force-fields can be used to sample the energy surface and to find a set of local-minima structures. Afterwards, this set should be filtered with DFTB and finally refined using DFT calculations. A methodology such as this one can, in principle, reduce the amount of computer resources necessary to study large and complex systems and make the structure prediction search much more efficient.

4.6 Interpretability of the neural networks

Now, we would like to discuss the interpretability of neural networks. According to Lipton et al. [505] the abstract idea of interpretability of machine learning algorithms can be divided into 4 different concepts: simulatability, decomposability, algorithmic transparency and post hoc interpretability.

Simulatability concerns the ability of a human beings to follow the calculations that occur in the model. This is not the case for neural networks, as it is not the case for DFT calculations that we try to reproduce.

Decomposability pertains to the intuitive interpretation of the different parts of the model. The Behler symmetry functions that serve as our inputs can be seen as decaying pair wise or angular distribution functions and their parameters have been thoroughly described. The weights of the neural network, can be seen as fitting parameters. However, giving meaning to all of them stands as a seemingly impossible task, with the exception of bias neurons that shift the activation functions. It is difficult to give meaning to each calculation that occurs in the neural network and impossible to identify the contributions of each part to the energy, e.g., the part of the network that calculates the kinetic energy contribution to the energy. The calculation can however, be seen as an highly complex non-linear fit.

Algorithmic transparency on the other hand is related to the grasp of the error surface and the ability to predict the output of the model based on the inputs. In this regard, neural networks are frowned upon due to their non convexity. In fact, the same activation functions that are responsible for the success of neural networks due to their non linearity, are also responsible for the multiple local minima for which the neural network optimization tend to converge. This topic has been thoroughly discussed, for example in [506]. Nevertheless, even a local minima can provide more useful and accurate results than other methods. Furthermore, we believe that the training for forces and stresses is very helpful, as we are restricting the optimization of the neural network to a Pareto curve. We would also like to point out that these quantities are calculated in a consistent and physical manner in the neural networks force-fields that we presented: through analytical differentiation of the energy.

Finally, post hoc interpretability concerns the knowledge that can be gained from the model itself. One can for example study the importance of pair wise or three body interactions to the energy using the symmetry functions. Neural networks are often regarded as black box algorithms. However the picture that has been painted over the years might not be so grim. With every study, the understanding of neural networks increases and, at least, they can be seen as powerful mathematical tools that are capable of efficiently and accurately approximating functions.

So, what can we expect from neural network, or other machine learning, force-fields? We do not believe DFT will ever be replaced by machine learning force-fields. Yet, they will provide descriptions for regions of the PES that are not quite accessible within DFT, and by extent, other electronic structure methods. Moreover, these force-fields will allow for more accurate simulations (such as MD runs) than those provided by simpler fitting methods (such as classical force-fields), and will permit considerable faster samplings of the PES, which translates into a considerable speed-up for global structure prediction methods [414, 431, 434].

Chapter 5

Copper based materials and cluster expansions

> The world to me was a secret, which I desired to discover; to her it was a vacancy, which she sought to people with imaginations of her own.
>
> Mary W. Shelley
> Frankenstein

Discover... Our objective is to discover new materials and often that requires the development of different techniques and methodologies to study them. Usually a certain compound admits many crystal phases, however only a fraction of them are indeed stable, i.e., the compounds only crystallize in some of these phases. Moreover, some of these phases can even interact with each other and change the properties of the compound. For example, among copper based materials, a rather common occurrence is the stabilization of a compound due to copper vacancies.

In this chapter we present studies of copper based materials using cluster expansions. We start by explaining cluster expansions and then we discuss photovoltaics materials and transparent conducting semiconductors (TCSs), in particular CZTS and cuprous iodide, respectively. Our work with CZTS focus on a stability study using genetic algorithms, and on the transition between the kesterite and stannite structures under the incorporation of iron. Our application of cuprous iodine involves the formation of stable phases with copper vacancy complexes.

5.1 Cluster expansions

In the previous chapters, we discussed the construction of approximations for the potential energy surface using machine learning. A similar approach consists on the expansion of the energy of a system in terms of effective cluster interaction (ECI), which embody the energetic information of the underlying crystal structure. This approach is usually denoted as cluster expansion [507–509] and can be understood as a generalization of the Ising model Hamiltonian.

The definition of the cluster expansion starts with mapping of each site i in a parent lattice with a occupation variable σ_i. For the case of a binary allow, these variables mimic the spin and can take values of ±1 according to the type of atom that occupies the site. A specific arrangement of these occupation variables denotes a configuration and can be represented by a vector $\boldsymbol{\sigma}$ containing all the individual occupation variables. We note that the cluster expansion can be formally defined for arbitrary multi-component alloys [509]. However here we focus on cluster expansion for binary alloys as implemented in the MAPS code of the alloy theoretic automated toolkit [510] (ATAT), which we used to construct the cluster expansions presented in this chapter.

Continuing with the definition, the energy of an alloy can then be parameterized as the following polynomial of the occupation vector:

$$E(\boldsymbol{\sigma}) \; = \; \sum_{\alpha} J_\alpha m_\alpha \left\langle \prod_{i\in\alpha'} \sigma_i \right\rangle , \tag{5.1}$$

where the sum is taken over all the non-equivalent clusters α and the averaged product over all the equivalent clusters α'. By cluster we mean a set of sites i, and equivalent clusters means that they can not be transformed into another by a symmetry operation of the space group of the parent lattice. Furthermore, m_α denote the number of equivalent clusters, and J_α represent the coefficients of the expansion. In this formalism, they are usually designated as multiplicities and ECIs, respectively. The product between the multiplicities and the spin-products averaged over the entire lattice define the correlation matrix, which can be understood as the probability to find the cluster α in a configuration $\boldsymbol{\sigma}$. This quantity can be written more explicitly as

$$\overline{\Pi}_{k,n}(\boldsymbol{\sigma}) = m_\alpha \left\langle \prod_{i\in\alpha'} \sigma_i \right\rangle = \frac{1}{k}\cdot\frac{1}{\lambda}\cdot\frac{1}{m_\alpha}\sum_{I=1}^{\lambda}\sum_{I=1}^{m_\alpha} \sigma_{I_1}\sigma_{I_2}\cdots\sigma_{I_k} \tag{5.2}$$

where λ is the number of atomic sites i in the unit cell, and k the number of vertices in the cluster. If all the clusters α are included in the sum, then the cluster expansion can represent any function of the configuration (such as the energy $E(\boldsymbol{\sigma})$), provided we have appropriate ECIs. However, the expansion converges quickly in practise. So, usually only compact clusters are considered, such as small pairs and triplets.

The J_α remain as the only unknown variables and their determination follows from the Structure Inversion Method or the Collony-Williams method [511]. Basically, this method requires the calculation of the energy of a small number of configurations using first principles methods (in our case DFT as described in appendix A), and the calculation of the correlation matrix. Then, the J_α correspond to the least square solution of eq. (5.1), i.e., the solution of its normal equation [512].

To measure the predicative power of the cluster expansion, the MAPS code uses the cross-validation score defined as

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left(E_i^{\mathrm{ref}} - E_{(i)}\right)^2 , \tag{5.3}$$

where n represents the total number of structures, E_i^{ref} the energy of structure i obtained with the reference method, and $E_{(i)}$ the prediction of the cluster expansion obtained from the least-squares fit to the $n-1$ other structural energies. So, this quantity estimates the error made in the prediction of energies for structures not included in the fit [510].

Similarly to the neural networks force-fields from last chapter, the cluster expansion allows for a concise and computational efficient mapping of the configuration of an alloy and its energy. This allows the study of several thermodynamic properties and phase diagrams using statistical mechanical techniques, such as cluster variation method [513, 514], Monte Carlo simulations [515], and low and high temperature expansions [514, 516]. Often a well converged cluster expansion suited for this techniques requires fewer than 20 ECI and the calculation of the energy of 30 to 50 ordered structures. Contrary to the neural networks force-fields, the variation of the energy with respect to the positions (or the infinitesimal strains), can not be performed. Furthermore, this formalism does not allow for the calculation of displacements, as the lattice sites of each structure are fixed to those of the parent lattice. This means that forces and stresses can not be calculated using this formalism.

5.2 CZTS

Nowadays, one of the most promising and sustainable energy sources originates from photovoltaics, which converts solar energy into electricity. The market and the industry are centered around waver-based silicon solar cells due to silicon's optimal band gap, abundance in the earth's crust, non-toxicity, and also because chemical and semiconductor industries mastered its technology [517]. Here optimal band gap refers to the range [1.1,1.4] eV calculated according to the Shockley–Queisser detailed-balance efficiency limit [518, 519] for single-junction solar cells. Nevertheless, researchers undertook the challenge of developing more efficient solar cells. This can be achieved by improving the design of the devices or by employing different materials, in an attempt to optimize the properties that influence the efficiency of the cells. For example, the silicon band gap is close to optimal, yet it is indirect. As a result, the absorption coefficient is low and varies slowly around the gap, therefore a thicker wafer or film is required to absorb photons with energies above the gap. However, this leads to higher Auger recombination, which decreases the open circuit voltage of the cell [520]. Obviously, optimizing all of these influences the cost of the cell, which is another important factor in the development of the solar cells. Chemical and semiconductor industries are interested in mass production of these photovoltaic cells, and this requires cheap, easy to manufacture, non-toxic, efficient cells built from fairly available materials.

Presently, the most efficient device is a multi-junction solar cell [521–524], which consists, as the name implies, in combining several *p*-*n* junctions. Furthermore, using different materials, allows for a broader range of frequencies to be absorbed. An example of this type of device is the GaSb-based solar cell [525] that achieved a combined module efficiency of 41.2%. This would not, however, be possible without single *p*-*n* junctions, so in table 5.1, we show the world-record single junction solar cell efficiencies for different materials [520, 526]. When compared to the 25% efficiency of silicon solar cells from 1999 [527, 528], we see the advances in the field of photovoltaics. The maximum for silicon cells rose to 26.7%. From the table, we clearly see that some materials can be use as an alternative to silicon, basically those with efficiencies above 20%, such us CuInGa(Se,S) (CIGS) or GaAs (the most efficient material). Nonetheless, we note that these record efficiencies are constantly being surpassed.

Material	Efficiency (%)
Crystalline Si	26.7
Multicrystalline Si	22.3
Amorphous Si	10.2
Nanocrystalline Si	11.9
GaAs	29.1
GaInP	21.4
InP	24.2
CuInGa(Se,S)	22.9
CdTe	21.5
perovskite	23.7
CuZnSn(Se,S)	12.6
dye/TiO_2	11.9
organic	15.6
quantum dots	13.4

Table 5.1: Record efficiency for single-junction solar cells. More information can be found in Ref. [520, 531, 532]. Record efficiencies were taken from the references listed.

Although very promising, solar cells made out of some of these materials raised several concerns. For example arsenic is highly sought after, due to GaAs [529] applications (it is particular important in the most recent smartphones), yet it can be poisonous and has contaminated groundwater. Another example is indium, which is relatively rare (low abundance), toxic, and highly demanded, due to its use in screen displays. This caused problems for CIGS solar cells and the growing consensus was to replace CIGS by a cheaper, indium-free material with identical properties, such as the possibility to tune the band gap continuously in the range [1.04,1.65] eV, by varying the In/Ga ratio [530].

Examples of these indium free-materials are $Cu_2ZnSn(S,Se)_4$ [533–536] (which is usully designated as CZTS) and $Cu_2ZnGe(S,Se)_4$ [537, 538]. Although explored, the latter is not really an alternative, according to the above mentioned desires, as Ge is as expensive as Ga or In. On the other hand, the kesterite structured $Cu_2ZnSn(S,Se)_4$ has high absorption coefficients, direct energy band gap tuned in the range [1,1.5] eV by varying the S/Se ratio, and abundant, non-toxic, low-cost constituents [539]. Unfortunately, this material is more promising than efficient as we can see in table 5.1. The efficiency of 12.6%, which is below half of the theoretical S-Q detailed-balance efficiency limit, stems from the difficulty in preparing CTZS without the formation of secondary phases [540]. While benign secondary phases may exist, usually they are detrimental to cell efficiency, as they normally increase carrier recombination rates, leading to an increase of resistance and loss of open-circuit voltage. Furthermore, not only CZTS compounds exist in a rather narrow region of the phase space [541–543] but they can also decompose due to Mo back-contacts [544] and the evaporation of volatile S/Se and SnS/Se [545–547].

Additionally, CZTS can also crystallize in a stannite structure. Both the stannite and kesterite structures can be outlined as the same 1 × 1 × 2 zincblende supercell with the difference between them lying with the atom (Cu, Zn, or Sn) positioned at certain fcc lattice sites (see fig. 5.1). Even though the kesterite structure is the ground state structure, a mere 3 meV/atom difference separates both phases. This indicates that disorder in the cation sub-lattice can occur under standard growth condition, and this disorder can alter the band gap of the material by 0.15 eV [534]. Similarly, the incorporation of extrinsic impurities can also change the energy band gap and improve the performance of the material. The most discussed and already mention is the ratio between S/Se, nevertheless other replacements are certainly possible. For example the substitution of Zn by Fe can increase the band gap and allow for the application of the resulting material in Si-based tandem solar cells, as the lattice constant lies between those of CZTS and the stannite structured Cu_2FeSnS_4 (CFTS).

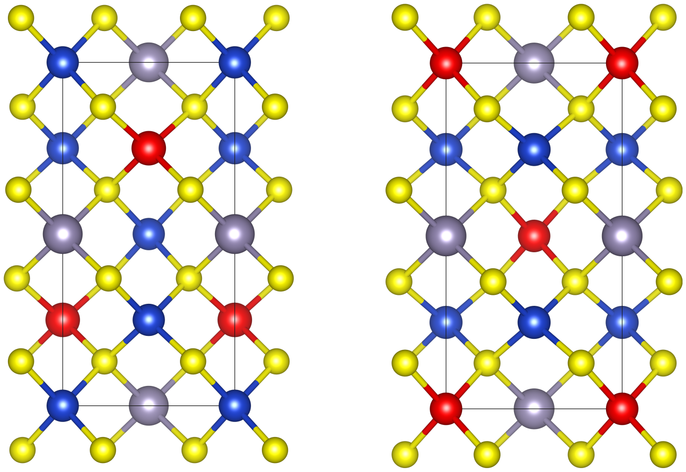


Figure 5.1: Kesterite (left panel) and stannite conventional tetragonal unit cells. Cu atoms are in blue, Se in yellow, Sn in gray, and Zn and Fe in red.

Also worth mentioning are the possible lattice defects that can change the properties of a material, such as conductivity, colour (i.e. absorbing or emitting light), and recombination processes. CTZS (a quaternary compound) allows for a plethora of possible lattice defects such as vacancies, antisites, and interstitials. The study of the formation energy [548, 549] of these defects reveal that the $p-$type conductivity and the difficulty in achieving $n-$type doping in CZTS comes from the lower formation energy of acceptor defects (with respect to donor defects). Moreover, the lowest energy defect is the Cu_{Zn} antisite and not the Cu vacancy like in other compounds (for example Cu−based chalcopyrites [550, 551]). However, this defect is detrimental to cell efficiency, so that Cu−poor and Zn−rich growth conditions are beneficial to improve the efficiency as shown experimentally [552].

This shows the importance of the description of secondary phases of CZTS, as well as the study of its defects, and of the incorporation of extrinsic impurities.

5.2.1 Genetic algorithms and stability

As mentioned above, one of the reasons for the low efficiency of CZTS, comes from the incapability to synthesise it without the formation of additional secondary phases. Thus, it becomes apparent how important it is to study the stable phases of this compound [540, 542, 543, 553, 554].

For example in Ref. [540], Schwartz *et al.* perform this study with atom probe tomography and DFT, and report the existence of two metastable phases with a distorted zincblende structure that might even be benign to the cell performance: $Cu_2Zn_5SnSe_8$ and $Cu_2Zn_6SnSe_9$. Their phases come from MHM runs and from a procedure that resembles its outer loop, which was employed for supercells containing 48 to 54 atoms. Furthermore, ZnSe pairs were added consecutively to $Cu_2ZnSn(S,Se)_4$ with the intent to find other iso-electronic compounds. This revealed that $Cu_2Zn_5SnSe_8$ and $Cu_2Zn_6SnSe_9$ are both stable with respect to the decomposition into binary and ternary compounds. Additionally, this showed that the most favorable decomposition is always the one that follows $Cu_2ZnSnSe_4$+ xZnSe, where x is a positive integer.

Here, we intended to validate these findings by performing structure prediction for supercells of Cu_2SnSe_3 + xZnSe compounds using genetic algorithms. However and due to the computational cost of such endeavour, we decided to first construct a cluster expansion for these compounds in the zincblende structure, and then use it for the global structure prediction search. This methodology had already been used in for example Ref. [555, 556]. Nevertheless, we had one additional objective, i.e., to evaluate the predictability power of the cluster expansion method.

We fitted a cluster expansion with ATAT [508] for these particular compounds and obtained a cross-validation error of 17 meV/atom. All the calculations required were performed at the level of DFT with the PBE approximation for the exchange and correlation functional, as implemented in VASP code [477, 478]. More information on the DFT calculations can be found in the appendix A.

For the genetic algorithms, we just considered the positions of the atoms in the zincblende structure and allowed for the exchange between Cu, Sn, and Zn atoms. Then, each individual of the population consisted on a vector containing the type of atom present at each position. We perform calculations for different number of atoms, from cells containing from 2 (for ZnSe) to 32 lattice sites. Most of the calculations took into account a population with 30 individuals but we also perform calculations with higher populations, up to 400 individuals.

The fitness function consisted on the evaluation of the energy of the structures using the cluster expansion. The recombination operator was just a 2-point crossover. Yet, special care was taken to keep the composition constant over the simulations. This means that a population that starts with 2 Cu atoms ends with 2 Cu atoms. The mutation consisted on the exchange of the positions of two atoms.

From the GA runs, we found that it was sufficient to perform simulations with 100 generations. The lowest energy minima found were then recomputed with the PBE approximation, and in these calculations the structures were allowed to relax. In this manner, only a small subset of the structures found were refined with DFT. Figure 5.2 depicts our results and those of Ref. [540].

Our methodology was able to find the lowest energy phase of $Cu_2ZnSnSe_4$ (found in the hull) and several phases for other concentrations. In particular, we found phases with similar energy to those found in Ref. [540] for $Cu_2Zn_5SnSe_8$ and $Cu_2Zn_6SnSe_9$. We should note that we found structures with 2 meV (for the former) and 4 meV (for the latter) lower formation energy. Additionally, we also used different references for the calculation of the formation energy, which explains the differences in this quantity in both studies. Regardless, most of the phases found possess a positive formation energy and are not placed in the convex hull of

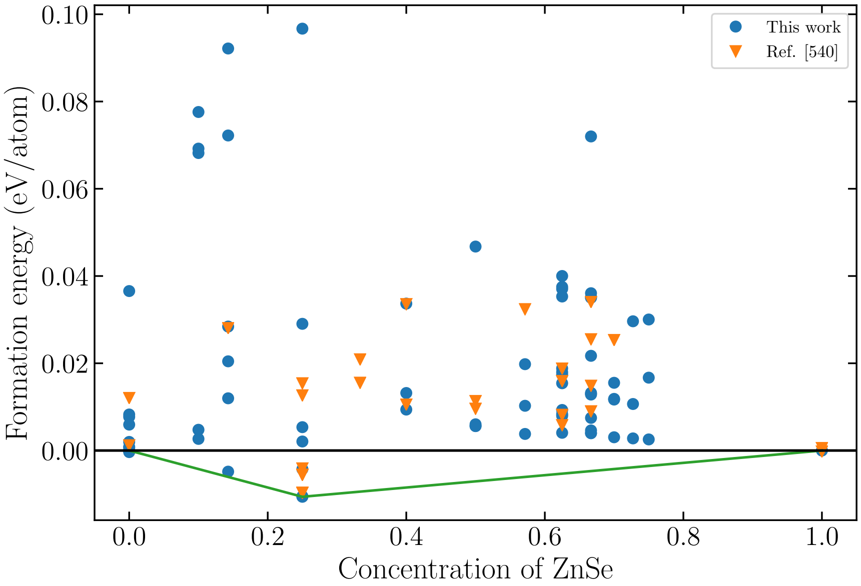


Figure 5.2: Phase diagram of $Cu_2Zn_xSnSe_{x+3}$ compounds. The formation energies presented are per atom. In green we show the convex hull of thermodynamic stability, with blue circles our results, and with orange triangles the results from Ref. [540].

stability, i.e. most of the structures found can decompose into ZnSe and other structures in the $Cu_2Zn_xSnSe_{x+3}$ compound. So, we conclude that these structures are thermodinamically unstable.

However, we found a lower energy structure for $Cu_4ZnSn_2Se_7$ (see fig. 5.3 for its depiction). This structure also appears to be a distortion of the zincblende structure, such as those found for $Cu_2Zn_5SnSe_8$ and $Cu_2Zn_6SnSe_9$. From fig. 5.2 we observe that this structure is almost at the convex hull of stability and the structure with highest probability to be stable, among those we found. This is an incredible result for this methodology that combines GA and cluster expansions.

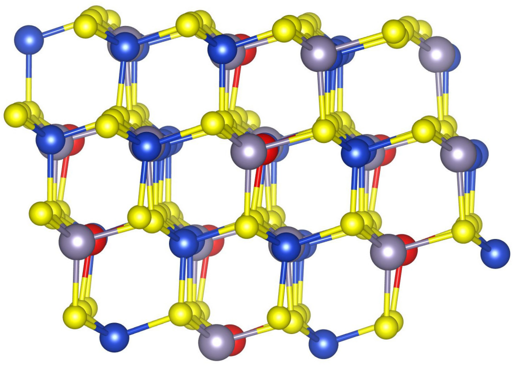


Figure 5.3: Crystal structure of $Cu_4ZnSn_2Se_7$. We use the same colour scheme as in fig. 5.1.

In this simple example, a cluster expansion proved to be a reliable method to predict DFT formation energies and construct phase diagrams. Instead of continuing with this study focused on the prediction of secondary phases of $Cu_2ZnSnSe_4$, which would involve the introduction of Cu vacancies and larger supercells, we decided to tackle the problem of the substitution of Zn by Fe, and investigate the stability of $Cu_2Zn_{1-x}SnFe_xSe_4$ compounds using this methodology.

5.2.2 Kesterite or stannite

We studied the stability of $Cu_2Zn_{1-x}SnFe_xSe_4$ compounds with a similar methodology to the one presented in the last section. Our objective was to find stable structures with lattice parameters close to those of Si. Thus, we constructed two cluster expansions, one for the kesterite structure and the other for the stannite structure, where the lattice sites corresponding to the Zn or Fe atoms (respectively) can be occupied by either of these elements. We obtained extremely accurate cluster expansions with cross-validation errors of 3.7 meV for the kesterite and 4.4 meV for the stannite structured compounds.

All the calculations required were performed at the level of DFT with the PBE approximation for the exchange and correlation functional, as implemented in VASP code [477, 478]. More information on these calculations can be found in the appendix A.

In fig. 5.2 we present some preliminary results, corresponding to the structures investigated while constructing the cluster expansion (with the blue circles and the orange squares). The next step would involve the prediction of the formation energy of supercells found with the genetic algorithms. However, before starting such study we found Ref. [557], where Shibuya *et al.* investigated the transition between the kesterite and stannite structures using the PBE approximation and supercells containing 64 atoms (5 concentrations of Fe for each type of crystal structure). Additionally the authors also calculated the band gaps of such phases. In fig. 5.2 we show the formation energies calculated for their structures with the green and the yellow symbols. Unfortunately, we did not manage to reproduce exactly their formation energies, as we limited the density of k-points to a maximum of 1000 per atom k-points, which usually yields a precision around 2 meV/atom in the total energy. This is a good approximation since, usually, entropic or Van der Waals effects (which we neglect) have a larger contribution to the formation energy. Yet, the differences in the formation energies for these structures is rather small, for example our phase diagram shows a maximum difference of around 10 meV.

Nevertheless, we observe the same behaviour as in Ref. [557]. The kesterite structure is more stable for concentrations of Fe in the range [0,0.5[while the stannite structure becomes more stable for concentrations of Fe above or equal to 0.5.

5.3 Cuprous iodide

Previously we discussed the importance of transparent conducting semiconductors (TCSs) in the development of solar cell devices. Additionally their high conductivity and transparency makes them suitable materials for many other applications, such as infrared reflective coatings and electrochromic displays, among other examples [558, 559].

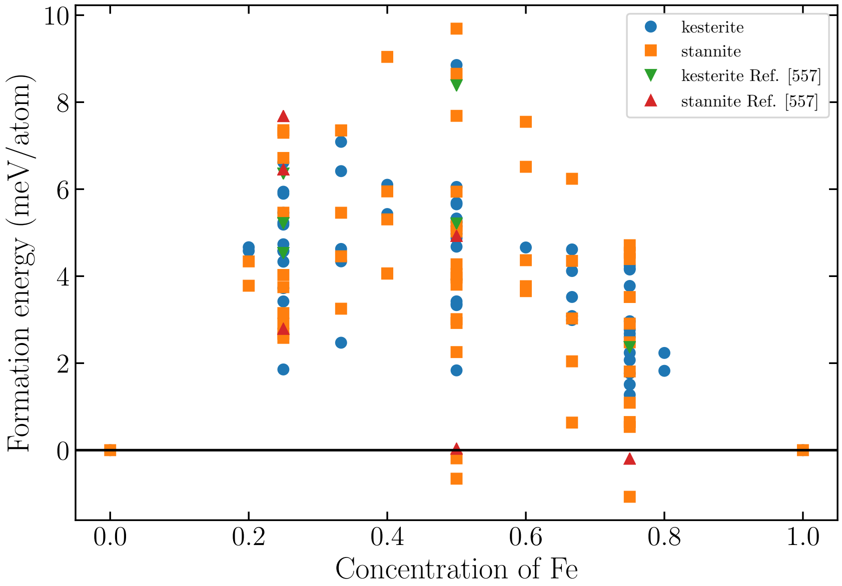


Figure 5.4: Phase Diagram of the $Cu_2Zn_{1-x}SnFe_xSe_4$ compounds. The formation energies presented are per atom. The blue circles and the orange squares represent our kesterite and stannite structures, respectively. The green triangles and the red triangles correspond to our attempt at reproducing the kesterite and stannite structures from Ref. [557]. We subtracted the formation energies of the kesterite structured Cu_2ZnSnS_4 and the stannite structured Cu_2FeSnS_4 from the corresponding phases.

As usual, these semiconductors can be divided based on their doping or percentage of major charge carriers. And while several n-type TCSs have been proposed in the recent decades and applied in industry, for example In_2O_3, SnO_2, ZnO, and GaN, not so many p-type TCS have been found [559]. The list of the known p-type TCS contains: NiO [560] which was the first to be found, the entire family of Cu oxides with the delafossite structure (such as $CuAlO_2$ [561]), $CuMO_2$compounds [562] where M is a trivalent cation and Mg-doped $CuCrO_2$, Mg-doped $CuCrO_2$ with a perovskite structure [563], and cuprous iodide (CuI) [564, 565]. Although discovered in 1907 [564], only recently did CuI appeared as the most promising p-type TCS [565].

Previously, the record for the highest conductivity among p-type TCS belonged to Mg-doped $CuCrO_2$ with 220 S/cm, yet it also displayed a low transmittance (around 30% for films with a thickness 250 nm) [562]. The $LaCrO_3$ perovskite doped with Sr exhibited better transmittance (42.3%), however it lacked conductivity (only 54 S/cm). Recently the record was taken by CuI thin films [566], which achieved a room-temperature hole conductivity of $\sigma > 280$ S/cm, while maintaining a transmittance over 70%. The growth of these thin films occurred in iodine-rich growth condition, which favors the generation of Cu vacancies. These vacancies manifest the lowest formation and ionization energies among all native defects (for both Cu-rich or I-rich equilibrium growth conditions [567, 568]) and constitute the dominant acceptors [569] in CuI. Consequently they are responsible for the p-type conduction of CuI thin films. This agrees with the findings of Maurer [570] and Bädeker [564].

An interesting observation, concerns the comparison between γ-CuI and other Cu compound, such as the already mentioned Cu(In,Ga)(S,Se)$_2$ (CIGS) and Cu$_2$ZnSn(S,Se)$_4$ (CZTS). In both of these direct band gap materials, the Cu vacancies occur in large quantities [571] and boast low formation energies. Moreover, the literature contains examples of neutral defect complexes that are energetically favorable in CIGS, such as the complex formed by two Cu vacancies and a Cu-{In,Ga} anti-site [572], which leads to the formation of ordered-defect phases that stabilized the stoichiometries CuIn$_5$Se$_8$, CuIn$_3$Se$_5$, Cu$_2$In$_4$Se$_7$ [573–576].

Furthermore, other studies of CuI revealed that this material can achieve a high optical transparency (90%) [577], hole concentrations in the range of 4.0×10^{16}–8.6×10^{19} cm^{-3}, with mobilities of 2–43.9 cm^{2}V^{-1}s^{-1} [565, 577], and effective masses of 0.30(1) m_e for the electrons, 2.4(3) m_e for heavy holes [578], and 0.2–0.25 m_e [579] for light holes. Concerning its (direct) band gap, measurements at $T = 80\,\mathrm{K}$ resulted in a value of 3.1 eV [580], while room temperatures measurements provided values in the range 2.93–3.03 eV [581–584].

Structurally, CuI is a very interesting material, that can exist in several polymorphs. At standard temperature and pressure, CuI crystallizes in the zincblende structure (usually designated as $\gamma-$CuI). Then, with the increase of temperature CuI takes a wurtzite structure between 643 and 673 K ($\beta-$CuI), and a rock salt structure above 673 K ($\alpha-$CuI). Meanwhile, with the increase of pressure the zincblende phase changes into a rhombohedral, a tetragonal, or a cubic phase [565, 585]. Additionally, CuI can also form some trigonal layered phases [586–589].

From a more technological point of view, these properties allowed for the creation of various opto-electronic devices, such as a hole transport layer in solid-state dye-sensitized [590], perovskite solar cells [591–593], hole-selective contacts in organic solar cells [594], light emitting diodes [595], and a transparent flexible thermoelectric material [596]. The application of CuI for the construction of these devices offers several advantages. First, this is an environment-friendly material: non-toxic and quite abundant. Moreover, the zincblende structure is ideal to match with those of the conventional semiconductors, while its direct band gap of $\sim$ 3 eV provides further benefits in the construction of p–n junctions [597]. Even from a chemistry point of view, this is an interesting compound since it exhibits a Cu$^\mathrm{I}$ oxidation state rather than Cu$^\mathrm{II}$ such in other halide salts of copper. This occurs due to the difference in ionic radii between Cu and I, and due to the powerful reducing capabilities of I$^-$, which can reduce spontaneously Cu$^\mathrm{II}$ to Cu$^\mathrm{I}$.

5.3.1 Copper vacancy complexes and stability

As mentioned above, the Cu vacancies exhibit the lowest formation energy among CuI defects and provide the greatest contribution to the $p-$type conductivity of CuI. However, the literature lacks information on how these Cu vacancies organize themselves in this material, do they form complexes like in CIGS semiconducting absorbers or do they spread all over the material as single point defects? Likewise, no exhaustive study answers questions related to their maximum and optimal concentrations. Here, we intend to provide an answer to these questions. This research was published in Ref. [25].

Our objective consisted in the study of stable phases of the Cu-I binary compound, with special focus on Cu vacancies in γ-CuI and their interaction with themselves. Furthermore, we tried to understand if the interaction between vacancies can lead to stable ordered-defect compounds. To accomplish this objective we explored exhaustively the phase diagram at zero temperature and pressure of the Cu–I system. We used the MHM to solve this global structural prediction problem. Moreover, we resorted to DFT with the PBE approximation for the exchange and correlation functional, as implemented in VASP code [477, 478] to perform the necessary calculations. More information on the calculations can be found in appendix A. The standard procedure to compare energies of multiple compounds is to use the highest k-point density and energy cutoff of all compounds to eliminate systematic errors [598]. The convergence tests for the zincblende CuI revealed that for a 1000/atom k-point mesh (which corresponds to a 8x8x8 mesh) the energy is converged to better than 1 meV/atom. Further tests on other structures (as we found them) revealed that a 1000/atom k-point mesh always ensured convergence to better than 2 meV/atom.

For efficiency reasons, we restricted the search to up to 6 atoms in the unit cell and we stopped each run after finding around 30–80 minima. Nevertheless, the result of the MHM was a rather complete view of the phase diagram of $Cu_{1-x}I_x$, with $0< x <1$. These results revealed that the Cu poor region of the phase diagram (between CuI and Cu_2I_3, which corresponds to 33% of Cu vacancies) contained the most stable structures. Furthermore, most of them corresponded to defected zincblende CuI. So, the next step involved the calculation of all possible crystal structures of γ-Cu, including a variable number of Cu vacancies. We took advantage of the software included in ATAT [508] to construct all possible supercells of γ-Cu, containing up to 14 sites (or 7 I atoms) and up to 50% of Cu vacancies (i.e. $1/2 < x < 2/3$). Note that this is already a very large number of vacancies, as we expect that a large concentration of vacancies will surely lead to a breakdown of the zincblende structure. Turns out that this results in 118 unique configurations, which we optimize with the PBE functional and added to the phase diagram. Finally, we constructed a cluster expansion using the ATAT [508] to predict the energy of all possible supercells containing up to 32 sites (16 I atoms). This was necessary since all possible combinations resulted in 30 849 geometries, which is a rather large number of structures with a substantial number of atoms to treat with DFT. The cluster expansion was fitted to the results of the small unit cells and achieved a cross-validation score of 33 meV/atom. The structures that the cluster expansion predicted to be closer to the convex hull of stability were then optimized with the PBE functional and added to the phase diagram. We should point out that it is useless to go beyond 30% of Cu vacancies ($x > 2/3$), since the breakdown of the zincblende structure will render the cluster expansion invalid.

In this manner, we obtained a rather complete phase diagram with the PBE approximation, an approximation known to incorrectly describe formation energies by more than 200 meV/atom [599–601] on average. So, in order to improve our results, we re-optimized all structures with the strongly constrained and appropriately normed [602] (SCAN) functional. This meta-GGA, that obeys 17 exact constrains of the exchange-correlation functional, is computational less efficient than the PBE but halves the average error of the formation energies for main group compounds [603, 604].

In the end, the phase diagram that we show in fig. 5.5 contains the SCAN formation energies of 623 unique $Cu_{1-x}I_x$ phases. The structures with lowest formation energy are, in descending order, Cu_3I_4 and Cu_4I_5. Both of these structures correspond to zincblende CuI with ordered lines of Cu vacancies. With the decrease of the Cu vacancies, comes a smooth

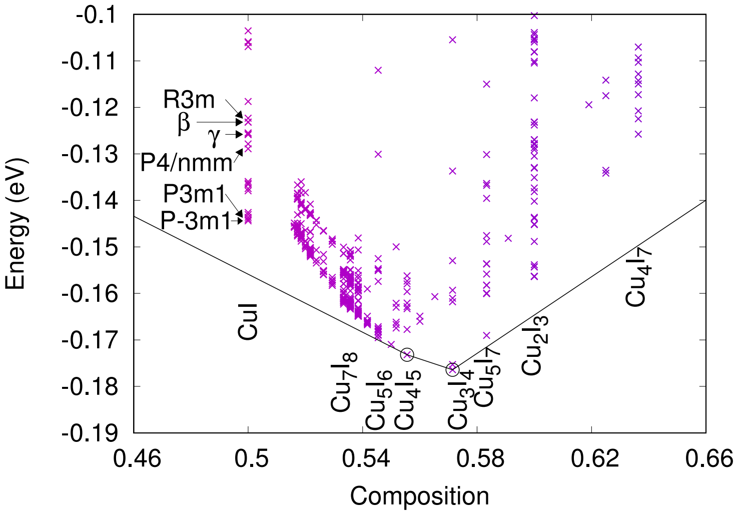


Figure 5.5: Binary phase diagram of $Cu_{1-x}I_x$ obtained with the SCAN functional [602]. We only show the relevant phases for p-type transparent conduction ($x \in [0.45, 0.65]$). Points that are strictly on the convex hull of thermodynamic stability (specifically, the lowest energy crystal structures Cu_4I_5 and Cu_3I_4) are represented by circles, while other phases are indicated by crosses. The chemical potentials of the elementary phases are set to zero, so the energy values indicated are in fact formation energies. The space groups identify some structures found in the materials project [70], and γ and β indicates the zincblende and wurzite structures, respectively.

increase of the formation energies until it reaches the energy per atom of γ-CuI. Moreover, the configurations with a high concentration of Cu vacancies would be further stabilized with the inclusion of the configurational entropy in the calculation.

In the insets of fig. 5.9 we display the geometry of the structures with the lowest energy, for which we also compute the density of electronic states (DOS). All of them exhibit lines of Cu vacancies in the [100] crystallographic direction (see fig. 5.6), though their distribution with respect to each other changes for each composition. This is remarkable since it indicates that the physics of p-type CuI might not come from isolated vacancies, but from ordered complexes of defects. Nevertheless, we should remark that although the lowest-energy structures exhibit these lines of Cu vacancies, we also found complexes with different patterns located just a few meV/atom higher in energy. This means that we should expect disorder vacancy configurations in real samples, due to entropic effects, and that the interaction between vacancies can indeed stabilize Cu-I binary systems.

Concerning the possibility to experimentally synthesize these ordered-defect structures (as it happens with CIGS [573–576]), we believe that it might be possible for compositions such as Cu_4I_5 or Cu_3I_4, due to the distance (in energy) between the ground state structure and the other polymorphs. For the other compositions the distance is too small (just a few meV/atom). Moreover, we note that there might be some disorder in the experimental samples, as the main difference between this structures is the distribution of the Cu lines of vacancies.

Similarly to other theoretical works [605], we confirm that the lowest formation energy structure of all polymorphs of stoichiometric CuI corresponds to the layered structure, which

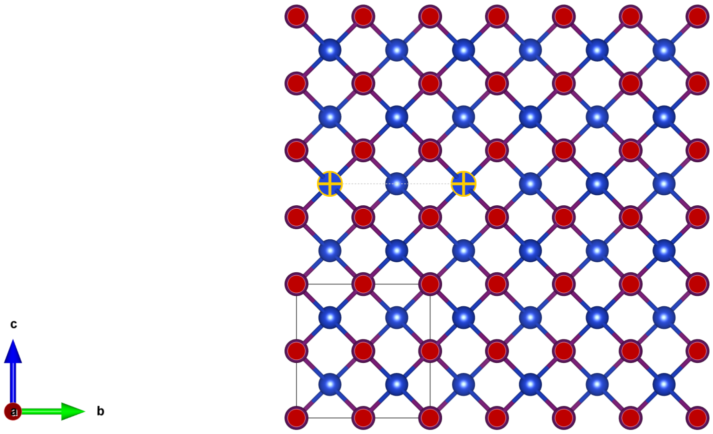


Figure 5.6: Zincblende CuI. Copper atoms are in blue, while iodine atoms are in red. Additionally, we show the [100] direction in red. The yellow crosses label two non-adjacent Cu atoms.

appears 18 meV/atom below the zincblende phase with the SCAN functional. Within this approximation, this structure is not thermodynamically stable (11 meV/atom above the convex hull), yet we believe that this results from the lack of van der Waals interactions in the SCAN approximation.

5.3.2 Phonon dispersion

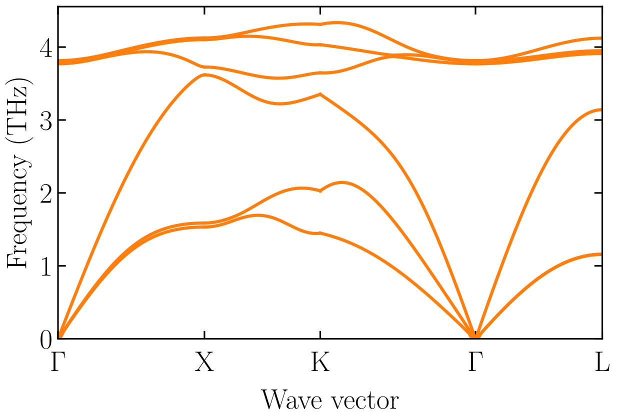
To further ascertain the stability of the lowest formation energy structures found, namely Cu_3I_4, and Cu_4I_5, we performed phonon calculations with PHONOPY [491] and VASP [477, 478], using the frozen-phonon technique. We approximated the exchange-correlation functional with the PBE functional [69] and used a cutoff energy of 720 eV, and a 10x10x12 and 10x10x8 k-point mesh for Cu_3I_4 and Cu_4I_5, respectively. The results and the additional phonon dispersion curve of γ-CuI can be found in fig. 5.7. No imaginary modes are present in any of the curves, which means that the structures are dynamically stable.

After gaining more confidence in the stability of these structures, we now proceed with the study of their electronic properties.

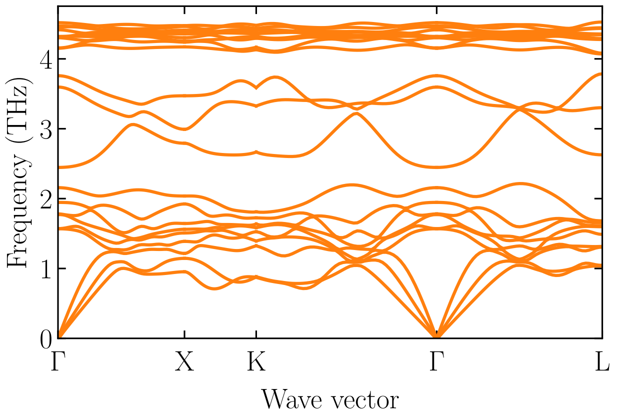
5.3.3 Density of states

Firstly, we switched here the exchange and correlation approximation to the PBE0 [607, 608] hybrid functional in order to describe acurately the band gap and the positioning of the *d*-states of these materials, which are incorrectly estimated by semi-local functionals, such as the PBE and the SCAN. After all, the PBE0 provides an excellent value for the gap of zincblende CuI [609].

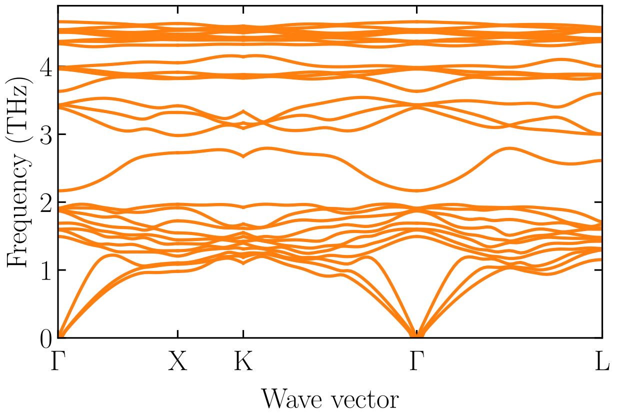
In fig. 5.9 we compare the DOS of the pristine zincblende CuI structure with the most



CuI



Cu_3I_4



Cu_4I_5

Figure 5.7: Calculated phonon band structures for CuI, Cu_3I_4, and Cu_4I_5.

relevant defect structures we found. For an easier comparison, the energy of each structure was subtracted by its Fermi energy, so that the curves are aligned. The DOS for all defect structures display similarities between themselves, and differ from the pristine DOS mainly due to some additional states that appear between the characteristic peaks of -3.7 and -3 eV and past the Fermi energy. The former states form a peak that emerges from the hybridiza-

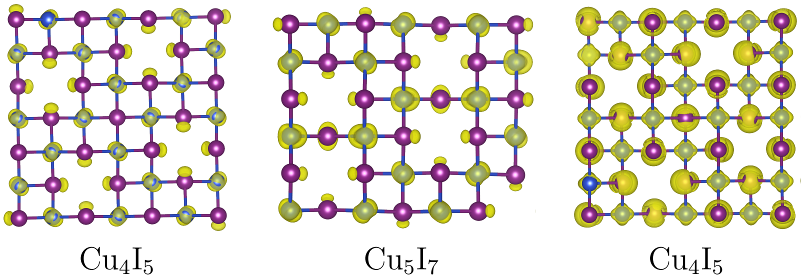


Figure 5.8: Partial charge density (yellow) integrated for the states around -3 eV for Cu_4I_5 (left) and Cu_5I_7 (middle) and for the conduction states of Cu_4I_5 (right) above the Fermi energy. This image was produced with VESTA [606].

tion of Cu *d*-states and I *p*-states, and we believe that it might serve as an experimental spectroscopy signature for these complexes of Cu vacancies. Figure 5.8 contains the depiction of these new states for Cu_4I_5 (left panel) and Cu_5I_7 (middle panel). Clearly, the *p*-states of I that contribute to the DOS are localized inside the lines of Cu vacancies.

Meanwhile, the additional states past the Fermi level, come primarily from the I *p*-states, and extend up to 0.7–0.9 eV, which is still below the visible range. This agrees with the experimental evidence that *p*-type CuI is transparent in this range. Furthermore, we show an example of these hole states in the right panel of fig. 5.8 (for Cu_4I_5). The partial charge densities of these states are very similar for the other compositions, and they all show that these states are rather delocalized as expected, due to the small hole mass of the CuI *p*-type conduction states.

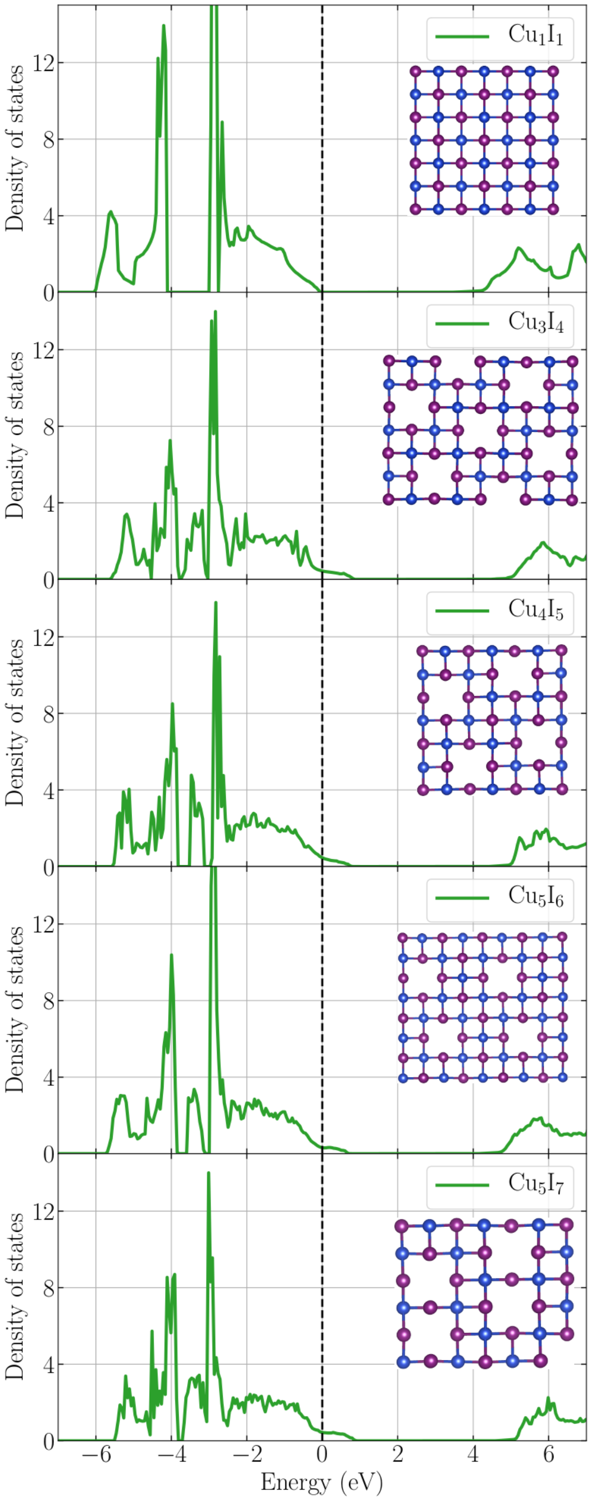


Figure 5.9: Density of electronic states for the lowest energy configuration of Cu_1I_1, Cu_3I_4, Cu_4I_5, Cu_5I_6, and Cu_5I_7, which are shown in the inset. Cu atoms are in blue, while I atoms are in purple. The lines of Cu vacancies run along the crystallographic [100] axis of the original zincblende structure. The insets were produced with VESTA [606]. The curves were calculated with the PBE0 hybrid exchange-correlation functional [607, 608], and were normalized to the number of I atoms in each structure.

Conclusions and outlook

> This lucid explanation of the phenomena we had witnessed appeared to me quite satisfactory. However great and mighty the marvels of nature may seem to us, they are always to be explained by physical reasons. Everything is subordinate to some great law of nature.
>
> Jules Verne
> Journey to the Center of the Earth

Explanation... In this thesis, we explained the problems of structure prediction and of complex and relatively large molecular dynamics simulations. Starting from the conception of electronic structure methods, we discussed some of the most recent developments in material science involving machine learning, and finished with our own humble contributions.

Concerning the neural network force-fields, we developed a methodology to construct accurate representations of the potential energy surface of solids using compact training sets. Firstly, we trained the neural networks with respect to the energy of each structure in the training set, the forces acting on each atom, and the stresses on the lattice. Furthermore, we implemented the extension of the back-propagation algorithm in the open-source ÆNET package [212]. Secondly, we constructed unbiased and varied data sets from minima of the potential energy surface, their distortions, structures visited during molecular dynamics simulations, and even from 2D minima (in some cases). In this manner, we represent many bonding configurations and we ensure an accurate description of the potential energy surface for many phases, such as supercells, structures under pressure or with varying temperatures. To determine the usefulness of our methodology, we trained several force-fields: namely for Si, Ge, SiGe, Cu and Au. The quality of the reproduction of all target quantities was satisfactory and enabled us to study relevant properties of these materials. In particular, we calculated phonon band-structures for Si and Cu with a fantastic agreement with density functional theory, melting temperatures for Cu and Au not that far from the experimental values, and we studied the defects of diamond structured Si. In the latter application, we found most of the known defects in silicon and devised a methodology to study large complex systems involving our neural network force-fields, density functional based tight binding, and density functional theory. To summarize, this methodology paves the way for high quality force-fields trained with rather small training sets. In the future, we intent to further improve this methodology with the optimization of the structure and the architecture of the neural networks, for example: with the addition of convolutional and pooling layers.

Regarding the cluster expansions and the copper based materials, we studied successfully supercells of CuI, $Cu_2ZnSnSe_4$, and $Cu_2(Zn,Fe)SnSe_4$ using cluster expansions. Using genetic algorithms, we found several stable phases of $Cu_2Zn_xSnSe_{x+3}$, in particular $Cu_2Zn_4SnSe_7$.

This is an important result, as the identification of all possible metastable secondary phases of of $Cu_2Zn_xSnSe_{x+3}$, may provide clues to optimize the efficiency of $Cu_2ZnSn(S,Se)_4$ solar cells. Moreover, we investigated the transition between kesterite and stannite structure in the $Cu_2Zn_{1-x}SnFe_xSe_4$ compound. According to our calculations, a concentration of Fe of 0.5 or above, makes the stannite structure more favorable. Moreover, we observed the expected increase of the lattice volume with the increase of the concentration of Fe, which might allow for the combination of these compounds with Si to form solar cells. We find that Cu vacancies in zincblende CuI can form complexes along the [100] crystallographic direction and that their interactions can further stabilize this binary compound. Furthermore, CuI admits a rather large concentration of these vacancies. From our calculations, compounds with 10–30% of Cu vacancies appear either in the convex hull of stability or in its proximity, which indicates that some compounds, such as Cu_4I_5 and Cu_3I_4, may be stable and able to form ordered defect compounds. As we neglect entropic effects in our calculations, this stabilization is only due to energetic effects. Actually, we expect even lower free energies for these compounds if the entropy is taken into account in the calculations. Furthermore, these results indicate that to proper understand the p-type conduction in CuI, researches should study not only isolated Cu vacancies but also complexes of Cu vacancies. Finally, these complexes of defects prompt the appearance of p-type conduction states, up to 0.7–0.9 eV above the Fermi energy, and are also responsible for the appearance of deeper lying electronic states, which may provide an experimental signature for these ordered defect compounds..

In this thesis we studied properties of six different materials and we developed efficient, yet accurate, methodologies capable of studying many other materials. Concerning its relevance to the field of materials science, our strategies allow to tackle problems usually left unsolved due to the high computational price required to simulate them.

Appendix A

DFT calculations with VASP

> It wasn't even a good note. 'If you are reading this I am probably dead.' What sort of a note is that?
>
> Patrick Rothfuss
> The Name of the Wind

Note... We use DFT with the PBE [69] approximation to the exchange-correlation functional as implemented in the Vienna Ab Initio Simulation Package (VASP) software [477, 478] to compute the total energies, forces, and stresses of the compounds that we study. The VASP code employs a plane-wave basis set. This basis set is orthonormal and the convergence of the calculations increases systematically with the number of plane-waves [16]. Moreover, the standard procedure to compare energies of multiple compounds is to use the highest k-point density and energy cutoff of all compounds to eliminate systematic errors [598]. For these reasons, we perform convergence tests of total energy with respect to k-point density and with respect to the number of plane-waves.

Often we rely on the projector augmented wave (PAW) method to model the core electrons with an energy cutoff of 520 eV, which normally assures an energy convergence below 1 meV/atom, and corresponds to 1.3 times the highest cutoff recommended among all the pseudopotentials we employ. This energy cutoff controls the number of plane-waves at a given k-point. We note that a convergence test in the total energy does not assure that all other properties are converged. So, when necessary we increase this value. For example, for some of the CuI phonon calculations (which involves second derivatives of the energy) in section 5.3.2, we increased it to 720 eV.

Concerning k-points, we usually start our convergence tests using only the Γ point and we keep increasing the k-point mesh up to 1000 per atom k-points. Usually this is sufficient to assure a precision of around 2 meV/atom in the total energy. In table A.1 we show an example of such a convergence. We should note that these are some of the structures discussed in section 5.3.1. Additionally, in Ref. [598] Jain *et al.* performed a convergence test of total energy with respect to k-point density and convergence energy difference for a subset of chemically diverse compounds and found that for a 500/atom k-point mesh, the numerical convergence for most compounds tested was within 5 meV/atom. Furthermore, 96% of compounds tested were converged to within 15 meV/atom.

Finally, most of our calculations are performed at 0 K and 0 kBar.

k-points		energy per	atom (eV)
mesh	number	CuI	Cu_3I_4
1x1x1	1		-2.66669404286
2x2x2	8	-2.511989175	-2.80401045857
3x3x3	27	-2.87548646	-2.79831699429
4x4x4	64	-2.935506285	-2.80029404571
5x5x5	125	-2.94796096	**-2.79950188428**
6x6x6	216	-2.951144695	
7x7x7	343	-2.952851745	
8x8x8	512	**-2.95228803**	-2.80003013857
9x9x9	729	-2.952937085	
10x10x10	1000		-2.799941054
12x12x12	1728	-2.95276668	

Table A.1: Convergence test of total energy with respect to k-point density for CuI and Cu_3I_4. A 1000/atom k-point mesh corresponds to 8x8x8 for CuI and 5x5x5 for Cu_3I_4 (in bold).

Bibliography

[1] G. B. Olson, "Designing a new material world," Science **288**, 993–998 (2000).

[2] A. R. Oganov, ed., *Modern methods of crystal structure prediction* (Wiley-VCH Verlag GmbH & Co. KGaA, 2010).

[3] A. R. Oganov and C. W. Glass, "Crystal structure prediction using ab initio evolutionary techniques: Principles and applications," J. Chem. Phys. **124**, 244704 (2006).

[4] R. E. Newnham, *Properties of materials: anisotropy, symmetry, structure* (Oxford University Press, 2005).

[5] F. H. Allen, G. Gergerhoff, and R. Sievers, eds., *Crystallographic databases* (International Union of Crystallography, Chester, 1987).

[6] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, "The Cambridge structural database," Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater. **72**, 171–179 (2016).

[7] "CAS REGISTRY - the gold standard for chemical substance information," `https://www.cas.org/support/documentation/cas-databases` (accessed in 2019).

[8] A. Walsh, "The quest for new functionality," Nat. Chem. **7**, 274–275 (2015).

[9] S. Goedecker, "Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems," J. Chem. Phys. **120**, 9911–9917 (2004).

[10] M. Amsler and S. Goedecker, "Crystal structure prediction using the minima hopping method," J. Chem. Phys. **133**, 224104 (2010).

[11] C. M. Freeman and C. R. A. Catlow, "Structure predictions in inorganic solids," J. Chem. Soc., Chem. Commun. , 89–91 (1992).

[12] B. P. van Eijck and J. Kroon, "Structure predictions allowing more than one molecule in the asymmetric unit," Acta Crystallogr. B **56**, 535–542 (2000).

[13] T. S. Bush, C. R. A. Catlow, and P. D. Battle, "Evolutionary programming techniques for predicting inorganic crystal structures," J. Mater. Chem. **5**, 1269–1272 (1995).

[14] S. M. Woodley and R. Catlow, "Crystal structure prediction from first principles," Nat. Mater. **7**, 937–946 (2008).

[15] R. M. Dreizler, E. K. U. Gross, and K. U. Eberhard, *Density Functional Theory: An approach to the Quantum Many-Body Problem* (Springer-Verlag Berlin Heidelberg, 1990).

[16] C. Fiolhais, F. Nogueira, and M. A. Marques, eds., *A Primer in Density Functional Theory* (Springer-Verlag Berlin Heidelberg, 2003).

[17] R. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, Inc., 1989).

[18] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, 2017).

[19] S. Marsland, *Machine Learning* (CRC Press, Taylor & Francis Inc, 2014).

[20] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods* (Cambridge University Press, 2008).

[21] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques (The Morgan Kaufmann Series in Data Management Systems)* (Morgan Kaufmann, 2011).

[22] J. P. M. de Sá, *Pattern Recognition: Concepts, Methods and Applications* (Springer Berlin Heidelberg, 2001).

[23] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," npj Comput. Mater. **5**, 83 (2019).

[24] M. R. G. Marques, J. Wolff, C. Steigemann, and M. A. L. Marques, "Neural network force fields for simple metals and semiconductors: construction and application to the calculation of phonons and melting temperatures," Phys. Chem. Chem. Phys. **21**, 6506–6516 (2019).

[25] S. Jaschik, M. R. G. Marques, M. Seifert, C. Rödl, S. Botti, and M. A. L. Marques, "Stable ordered phases of cuprous iodide with complexes of copper vacancies," Chem. Mater. **31**, 7877–7882 (2019).

[26] A. Szabo and N. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory* (Dover Books on Chemistry Series. Dover Publications, 1996).

[27] M. R. G. Marques, *Optical and Magnetical Properties of Endohedral Silicon Cages*, Master's thesis, University of Coimbra, FCTUC, Coimbra, Portugal (2015).

[28] P. Ehrenfest, "Bemerkung über die angenäherte gültigkeit der klassischen mechanik innerhalb der quantenmechanik," Z. Phys. **45**, 455–457 (1927).

[29] M. Born and V. Fock, "Beweis des adiabatensatzes," Z. Phys. **51**, 165–180 (1928).

[30] W. Pauli, *General Principles of Quantum Mechanics* (Springer Berlin Heidelberg, 1980).

[31] D. Andrae, *Hans Hellmann: Einführung in die Quantenchemie* (Springer Berlin Heidelberg, 2015).

[32] J. C. Slater, *Solid State and Molecular Theory: A Scientific Biography* (Wiley, 1975).

[33] R. P. Feynman, "Forces in molecules," Phys. Rev. **56**, 340–343 (1939).

[34] O. H. Nielsen and R. M. Martin, "Stresses in semiconductors: Ab initio calculations on si, ge, and gaas," Phys. Rev. B **32**, 3792–3805 (1985).

[35] O. H. Nielsen and R. M. Martin, "First-principles calculation of stress," Phys. Rev. Lett. **50**, 697–700 (1983).

[36] O. H. Nielsen and R. M. Martin, "Quantum-mechanical theory of stress and force," Phys. Rev. B **32**, 3780–3791 (1985).

[37] H. Bethe, *Quantentheorie* (Springer Berlin Heidelberg Imprint Springer, Berlin, Heidelberg, 1933).

[38] M. Born, W. Heisenberg, and P. Jordan, "Zur Quantenmechanik. II." Zeitschrift für Physik **35**, 557–615 (1926).

[39] J. C. Slater, "The virial and molecular structure," J. Chem. Phys. **1**, 687–691 (1933).

[40] V. Fock, "Näherungsmethode zur lösung des quantenmechanischen mehrkörperproblems," Z. Phys. **61**, 126–148 (1930).

[41] M. A. L. Marques, N. T. Maitra, F. Nogueira, E. K. U. Gross, and A. Rubio, eds., *Fundamentals of time-dependent density functional theory* (Lecture Notes in Physics, Vol. 837, Springer, Berlin, 2012).

[42] L. H. Thomas, "The calculation of atomic fields," Math. Proc. Camb. Philos. Soc. **23**, 542–548 (1927).

[43] E. Fermi, "Un metodo statistico per la determinazione di alcune priorieta dell'atome," Rend. Accad. Naz. Lincei **6**, 602–607 (1927).

[44] P. A. M. Dirac, "Note on exchange phenomena in the thomas atom," Math. Proc. Camb. Philos. Soc. **26**, 376–385 (1930).

[45] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," Phys. Rev. **136**, B864–B871 (1964).

[46] A. A. Abrikosov, *Methods of Quantum Field Theory in Statistical Physics* (Dover Publications Inc., 1976).

[47] A. L. Fetter and J. D. Walecka, *Quantum theory of many-particle systems* (Dover Publications Inc., 2003).

[48] M. E. Peskin and D. V. Schroeder, *An Introduction To Quantum Field Theory* (Taylor & Francis Inc, 1995).

[49] G. Stefanucci and R. van Leeuwen, *Nonequilibrium Many-Body Theory of Quantum Systems* (Cambridge University Press, 2015).

[50] L. Hedin, "New method for calculating the one-particle green's function with application to the electron-gas problem," Phys. Rev. **139**, A796–A823 (1965).

[51] E. E. Salpeter and H. A. Bethe, "A relativistic equation for bound-state problems," Phys. Rev. **84**, 1232–1242 (1951).

[52] K. Capelle, "A bird's-eye view of density-functional theory," arXiv:cond-mat/0211443 (2002).

[53] W. Kohn, "Nobel lecture: Electronic structure of matter-wave functions and density functionals," Rev. Mod. Phys. **71**, 1253–1266 (1999).

[54] M. Levy, "Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem," Proc. Natl. Acad. Sci. U.S.A **76**, 6062–6065 (1979).

[55] M. Levy, "Electron densities in search of Hamiltonians," Phys. Rev. A **26**, 1200–1208 (1982).

[56] R. M. Dreizler and J. Providência, *Density Functional Methods In Physics* (Springer, Boston, MA, 1985).

[57] A. Shimony and H. Feshbach, *Physics as natural philosophy : essays in honor of Laszlo Tisza on his seventy-fifth birthday* (MIT Press, Cambridge, Mass, 1982).

[58] E. H. Lieb, "Density functionals for Coulomb systems," Int. J. Quantum Chem. **24**, 243–277 (1983).

[59] J. P. Perdew and A. Zunger, "Self-interaction correction to density-functional approximations for many-electron systems," Phys. Rev. B **23**, 5048–5079 (1981).

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

[61] R. van Leeuwen, "Density functional approach to the many-body problem: Key concepts and exact functionals," (Academic Press, 2003) pp. 25 – 94.

[62] A. D. Becke, "Perspective: Fifty years of density-functional theory in chemical physics," J. Chem. Phys. **140**, 18A301 (2014).

[63] N. Mardirossian and M. Head-Gordon, "Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals," Mol. Phys. **115**, 2315–2372 (2017).

[64] M. A. Marques, M. J. Oliveira, and T. Burnus, "Libxc: A library of exchange and correlation functionals for density functional theory," Comput. Phys. Commun. **183**, 2272 – 2281 (2012).

[65] S. Lehtola, C. Steigemann, M. J. Oliveira, and M. A. Marques, "Recent developments in libxc — a comprehensive library of functionals for density functional theory," SoftwareX **7**, 1 – 5 (2018).

[66] J. P. Perdew and K. Schmidt, "Jacob's ladder of density functional approximations for the exchange-correlation energy," AIP Conference Proceedings **577**, 1–20 (2001).

[67] J. A. Pople, "Nobel lecture: Quantum chemical models," Rev. Mod. Phys. **71**, 1267–1274 (1999).

[68] Y. Zhao, B. J. Lynch, and D. G. Truhlar, "Doubly hybrid meta DFT: new multi-coefficient correlation and density functional methods for thermochemistry and thermochemical kinetics," J. Phys. Chem. A **108**, 4786–4791 (2004).

[69] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865–3868 (1996).

[70] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).

[71] X. Xu and W. A. Goddard, "The extended perdew-burke-ernzerhof functional with improved accuracy for thermodynamic and electronic properties of molecular systems," J. Chem. Phys. **121**, 4068–4082 (2004).

[72] J. Maddox, "Crystals from first principles," Nature **335**, 201–201 (1988).

[73] E. Mooser and W. B. Pearson, "On the crystal chemistry of normal valence compounds," Acta Crystallogr. **12**, 1015–1022 (1959).

[74] J. K. Burdett, G. D. Price, and S. L. Price, "Factors influencing solid-state structure—an analysis using pseudopotential radii structural maps," Phys. Rev. B **24**, 2903–2912 (1981).

[75] D. Pettifor, "A chemical scale for crystal-structure maps," Solid State Commun. **51**, 31–34 (1984).

[76] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, "Predicting crystal structures with data mining of quantum calculations," Phys. Rev. Lett. **91**, 135503 (2003).

[77] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, "Predicting crystal structure by merging data mining with quantum mechanics," Nat. Mater. **5**, 641–646 (2006).

[78] C. J. Pickard and R. J. Needs, "High-pressure phases of silane," Phys. Rev. Lett. **97**, 045504 (2006).

[79] C. J. Pickard and R. J. Needs, "Ab initio random structure searching," J. Phys.: Condens. Matter **23**, 053201 (2011).

[80] J. Pannetier, J. Bassas-Alsina, J. Rodriguez-Carvajal, and V. Caignaert, "Prediction of crystal structures from crystal chemistry rules by simulated annealing," Nature **346**, 343–345 (1990).

[81] J. C. Schön and M. Jansen, "First step towards planning of syntheses in solid-state chemistry: Determination of promising structure candidates by global optimization," Angew. Chem. Int. Ed. **35**, 1286–1304 (1996).

[82] K. Doll, J. C. Schön, and M. Jansen, "Structure prediction based on ab initio simulated annealing," J. Phys.: Conf. Ser. **117**, 012014 (2008).

[83] R. Martoňák, A. Laio, and M. Parrinello, "Predicting crystal structures: The Parrinello-Rahman method revisited," Phys. Rev. Lett. **90**, 075503 (2003).

[84] D. Gottwald, G. Kahl, and C. N. Likos, "Predicting equilibrium structures in freezing processes," J. Chem. Phys. **122**, 204503 (2005).

[85] W. Paszkowicz, "Genetic algorithms, a nature-inspired tool: Survey of applications in materials science and related fields," Mater. Manuf. Process. **24**, 174–197 (2009).

[86] C. W. Glass, A. R. Oganov, and N. Hansen, "USPEX–evolutionary crystal structure prediction," Comput. Phys. Commun. **175**, 713 – 720 (2006).

[87] Y. Wang, J. Lv, L. Zhu, and Y. Ma, "Crystal structure prediction via particle-swarm optimization," Phys. Rev. B **82**, 094116 (2010).

[88] Y. Wang, J. Lv, L. Zhu, and Y. Ma, "CALYPSO: A method for crystal structure prediction," Comput. Phys. Commun. **183**, 2063 – 2070 (2012).

[89] J. E. S. Agoston E. Eiben, *Introduction to Evolutionary Computing* (Springer Berlin Heidelberg, 2010).

[90] J. Brownlee, *Clever Algorithms: Nature-Inspired Programming Recipes* (LULU PR, 2012).

[91] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence* (MIT Press, Cambridge, MA, USA, 1992).

[92] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley Professional, 1989).

[93] A. De Jong, *Analysis of the behavior of a class of genetic adaptive systems*, Ph.D. thesis, University of Michigan, Michigan, USA (1975).

[94] S. G. Volz and G. Chen, "Molecular-dynamics simulation of thermal conductivity of silicon crystals," Phys. Rev. B **61**, 2651–2656 (2000).

[95] R. N. Salaway and L. V. Zhigilei, "Molecular dynamics simulations of thermal conductivity of carbon nanotubes: Resolving the effects of computational parameters," Int. J. Heat Mass Transf. **70**, 954 – 964 (2014).

[96] Y. Qi, T. Çağin, W. L. Johnson, and W. A. Goddard, "Melting and crystallization in Ni nanoclusters: The mesoscale regime," J. Chem. Phys. **115**, 385–394 (2001).

[97] L.-F. Zhu, B. Grabowski, and J. Neugebauer, "Efficient approach to compute melting properties fully from ab initio with application to Cu," Phys. Rev. B **96**, 224202 (2017).

[98] L. Wang, Y. Zhang, X. Bian, and Y. Chen, "Melting of Cu nanoclusters by molecular dynamics simulation," Phys. Lett. A **310**, 197 – 202 (2003).

[99] V. S. Pande, I. Baker, J. Chapman, S. P. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C. D. Snow, E. J. Sorin, and B. Zagrovic, "Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing," Biopolymers **68**, 91–109 (2003).

[100] T. D. Newport, M. S. Sansom, and P. J. Stansfeld, "The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions," Nucleic Acids Res. **47**, D390–D397 (2018).

[101] A. J. Wirth, Y. Liu, M. B. Prigozhin, K. Schulten, and M. Gruebele, "Comparing fast pressure jump and temperature jump protein folding experiments and simulations," J. Am. Chem. Soc. **137**, 7152–7159 (2015).

[102] Y. Li, C. Ji, W. Xu, and J. Z. Zhang, "Dynamical stability and assembly cooperativity of β-sheet amyloid oligomers – effect of polarization," J. Phys. Chem. B **116**, 13368–13373 (2012).

[103] B. J. Alder and T. E. Wainwright, "Studies in molecular dynamics. I. General method," J. Chem. Phys. **31**, 459–466 (1959).

[104] B. J. Alder and T. E. Wainwright, "Phase transition for a hard sphere system," J. Chem. Phys. **27**, 1208–1209 (1957).

[105] L. Verlet, "Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules," Phys. Rev. **159**, 98–103 (1967).

[106] H. Störmer, "Zufällige Überdeckungen auf dem Kreis," J. Appl. Math. Mech. **51**, 91–96 (1971).

[107] W. G. Hoover, A. J. C. Ladd, and B. Moran, "High-strain-rate plastic flow studied via nonequilibrium molecular dynamics," Phys. Rev. Lett. **48**, 1818–1820 (1982).

[108] D. J. Evans and B. L. Holian, "The Nose–Hoover thermostat," J. Chem. Phys. **83**, 4069–4074 (1985).

[109] H. C. Andersen, "Molecular dynamics simulations at constant pressure and/or temperature," J. Chem. Phys. **72**, 2384–2393 (1980).

[110] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," J. Chem. Phys. **81**, 3684–3690 (1984).

[111] M. Parrinello and A. Rahman, "Crystal structure and pair potentials: A molecular-dynamics study," Phys. Rev. Lett. **45**, 1196–1199 (1980).

[112] M. Parrinello and A. Rahman, "Crystal structure and pair potentials: A molecular-dynamics study," Phys. Rev. Lett. **45**, 1196–1199 (1980).

[113] H. Berendsen, D. van der Spoel, and R. van Drunen, "Gromacs: A message-passing parallel molecular dynamics implementation," Comput. Phys. Commun. **91**, 43 – 56 (1995).

[114] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *2014 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2014).

[115] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2015).

[116] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the 30th International Conference on Machine Learning (ICML), Deep Learning for Audio, Speech and Language Processing* , (2013).

[117] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)* (2015) pp. 1026–1034.

[118] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2015).

[119] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," arXiv:1604.07316 (2016).

[120] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," Nature **518**, 529–533 (2015).

[121] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," Nature **529**, 484–489 (2016).

[122] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," arXiv:1712.01815 (2017).

[123] "Alphastar: Mastering the real-time strategy game starcraft ii," `https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii` (accessed in 2019).

[124] "Openai five," `https://openai.com/blog/openai-five/` (accessed in 2019).

[125] S.-S. Liu and Y.-T. Tian, "Facial expression recognition method based on gabor wavelet features and fractional power polynomial kernel PCA," in *Advances in Neural Networks - ISNN 2010* (Springer Berlin Heidelberg, 2010) pp. 144–151.

[126] A. Waibel and K.-F. Lee, eds., *Readings in Speech Recognition* (Morgan Kaufmann, 1990).

[127] C.-L. Huang, M.-C. Chen, and C.-J. Wang, "Credit scoring with a data mining approach based on support vector machines," Expert Syst. Appl. **33**, 847–856 (2007).

[128] P. K. Chan and S. J. Stolfo, "Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection," in *KDD'98 Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (AAAI Press, New York, NY, 1998) pp. 164–168.

[129] M. Pazzani and D. Billsus, "Learning and revising user profiles: The identification of interesting web sites," Mach. Learn. **27**, 313–331 (1997).

[130] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to spam filtering," Expert Syst. Appl. **36**, 10206–10222 (2009).

[131] P. Baldi and S. Brunak, *Bioinformatics: the machine learning approach* (The MIT Press, 2001).

[132] J. H. Noordik, *Cheminformatics Developments: History, Reviews and Current Research* (IOS Press, 2004).

[133] K. Rajan, "Materials informatics," Mater. Today **8**, 38–45 (2005).

[134] K. Rajan, "Materials informatics: The materials gene and big data," Annu. Rev. Mater. Res. **45**, 153–169 (2015).

[135] T. Mueller, A. G. Kusne, and R. Ramprasad, "Machine learning in materials science," in *Reviews in Computational Chemistry* (John Wiley & Sons, Inc, 2016) pp. 186–273.

[136] J.-P. Correa-Baena, K. Hippalgaonkar, J. van Duren, S. Jaffer, V. R. Chandrasekhar, V. Stevanovic, C. Wadia, S. Guha, and T. Buonassisi, "Accelerating materials development via automation, machine learning, and high-performance computing," Joule **2**, 1410–1420 (2018).

[137] Y. Liu, T. Zhao, W. Ju, and S. Shi, "Materials discovery and design using machine learning," J. Materiomics **3**, 159–177 (2017).

[138] L. Ward, M. Aykol, B. Blaiszik, I. Foster, B. Meredig, J. Saal, and S. Suram, "Strategies for accelerating the adoption of materials informatics," MRS Bull. **43**, 683–689 (2018).

[139] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature **559**, 547–555 (2018).

[140] K. T. Butler, J. M. Frost, J. M. Skelton, K. L. Svane, and A. Walsh, "Computational materials design of crystalline solids," Chem. Soc. Rev. **45**, 6138–6146 (2016).

[141] S. Shi, J. Gao, Y. Liu, Y. Zhao, Q. Wu, W. Ju, C. Ouyang, and R. Xiao, "Multi-scale computation methods: Their applications in lithium-ion battery research and development," Chin. Phys. B **25**, 018212 (2016).

[142] L. Ward and C. Wolverton, "Atomistic calculations and materials informatics: A review," Curr. Opin. Solid State Mater. Sci. **21**, 167–176 (2017).

[143] "A beginner's guide to deep reinforcement learning," `https://skymind.ai/wiki/deep-reinforcement-learning` (accessed in 2019).

[144] R. S. Sutton and A. G. Barto, *Reinforcement Learning* (The MIT Press, 2018).

[145] H. Nguyen, S.-i. Maeda, and K. Oono, "Semi-supervised learning of hierarchical representations of molecules using neural message passing," arXiv:1711.10168 (2017).

[146] R. R. Picard and R. D. Cook, "Cross-validation of regression models," J. Am. Stat. Assoc. **79**, 575–583 (1984).

[147] B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta, and L. Ward, "Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery," Mol. Syst. Des. Eng. **3**, 819–825 (2018).

[148] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," Neural Comput. **4**, 1–58 (1992).

[149] C. Sammut and G. I. Webb, *Encyclopedia of machine learning and data mining* (Springer Publishing Company, Incorporated, 2017).

[150] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR guiding principles for scientific data management and stewardship," Sci. Data **3**, 160018 (2016).

[151] C. Draxl and M. Scheffler, "NOMAD: The FAIR concept for big data-driven materials science," MRS Bull. **43**, 676–682 (2018).

[152] "Materials genome initiative," `https://www.mgi.gov/` (accessed in 2019).

[153] "The NOMAD archive," `https://metainfo.nomad-coe.eu/nomadmetainfo_public/archive.html` (accessed in 2019).

[154] "Supercon, national institute of materials science," `http://supercon.nims.go.jp/index_en.html` (2011).

[155] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD)," JOM **65**, 1501–1509 (2013).

[156] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, "The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies," npj Comput. Mater. **1**, 15010 (2015).

[157] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, and A. Aspuru-Guzik, "The Harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid," J. Phys. Chem. Lett. **2**, 2241–2251 (2011).

[158] B. Puchala, G. Tarcea, E. A. Marquis, M. Hedstrom, H. V. Jagadish, and J. E. Allison, "The materials commons: A collaboration platform and information repository for the global materials community," JOM **68**, 2035–2044 (2016).

[159] R. Mullin, "Citrine informatics," C&EN Global Enterprise **95**, 34–34 (2017).

[160] M. de Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. K. Ande, S. van der Zwaag, J. J. Plata, C. Toher, S. Curtarolo, G. Ceder, K. A. Persson, and M. Asta, "Charting the complete elastic properties of inorganic crystalline compounds," Sci. Data **2**, 150009 (2015).

[161] A. Zakutayev, N. Wunder, M. Schwarting, J. D. Perkins, R. White, K. Munch, W. Tumas, and C. Phillips, "An open experimental database for exploring inorganic materials," Sci. Data **5**, 180053 (2018).

[162] P. Villars, H. Okamoto, and K. Cenzual, *ASM alloy phase diagrams database* (ASM International, Materials Park, OH, USA, 2006).

[163] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moeck, R. T. Downs, and A. L. Bail, "Crystallography open database (COD): an open-access collection of crystal structures and platform for world-wide collaboration," Nucleic Acids Res. **40**, D420–D427 (2011).

[164] P. Villars, M. Berndt, K. Brandenburg, K. Cenzual, J. Daams, F. Hulliger, T. Massalski, H. Okamoto, K. Osaki, A. Prince, H. Putz, and S. Iwata, "The Pauling file, binaries edition," J. Alloy. Comp. **367**, 293–297 (2004).

[165] P. Gorai, D. Gao, B. Ortiz, S. Miller, S. A. Barnett, T. Mason, Q. Lv, V. Stevanović, and E. S. Toberer, "TE design lab: A virtual laboratory for thermoelectric material design," Comput. Mater. Sci. **112**, 368–376 (2016).

[166] S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, J. Gath, K. W. Jacobsen, J. J. Mortensen, T. Olsen, and K. S. Thygesen, "The Computational 2D Materials Database: high-throughput modeling and discovery of atomically thin crystals," 2D Mater. **5**, 042002 (2018).

[167] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist, "Machine-learning-assisted materials discovery using failed experiments," Nature **533**, 73–76 (2016).

[168] K. Ryan, J. Lengyel, and M. Shatruk, "Crystal structure prediction via deep learning," J. Am. Chem. Soc. **140**, 10158–10168 (2018).

[169] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: Critical role of the descriptor," Phys. Rev. Lett. **114**, 105503 (2015).

[170] R. E. Bellman, *Adaptive Control Processes: A Guided Tour* (Princeton university press, 2015).

[171] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, and M. A. L. Marques, "Predicting the thermodynamic stability of solids combining density functional theory and machine learning," Chem. Mater. **29**, 5090–5103 (2017).

[172] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, "Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals," Phys. Rev. Lett. **117**, 135502 (2016).

[173] C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli, and M. Scheffler, "New tolerance factor to predict the stability of perovskite oxides and halides," arXiv:1801.07700 (2018).

[174] W. Ye, C. Chen, Z. Wang, I.-H. Chu, and S. P. Ong, "Deep neural networks for accurate predictions of crystal stability," Nat. Commun. **9**, 3800 (2018).

[175] W. Li, R. Jacobs, and D. Morgan, "Predicting the thermodynamic stability of perovskite oxides using machine learning models," Comput. Mater. Sci. **150**, 454–463 (2018).

[176] A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, and R. Ramprasad, "Machine learning strategy for accelerated design of polymer dielectrics," Sci. Rep. **6**, 20952 (2016).

[177] N. Artrith, A. Urban, and G. Ceder, "Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species," Phys. Rev. B **96**, 014112 (2017).

[178] T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," Phys. Rev. Lett. **120**, 145301 (2018).

[179] Y. Zhang and C. Ling, "A strategy to apply machine learning to small datasets in materials science," npj Comput. Mater. **4**, 25 (2018).

[180] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798–1828 (2013).

[181] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**, 184115 (2013).

[182] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett. **108**, 058301 (2012).

[183] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, "Crystal structure representations for machine learning models of formation energies," Int. J. Quantum Chem. **115**, 1094–1101 (2015).

[184] 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," Phys. Rev. B **89**, 205118 (2014).

[185] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).

[186] A. Seko, A. Takahashi, and I. Tanaka, "Sparse representation for a potential energy surface," Phys. Rev. B **90**, 024101 (2014).

[187] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett. **104**, 136403 (2010).

[188] E. Sanville, A. Bholoa, R. Smith, and S. D. Kenny, "Silicon potentials investigated using density functional theory fitted neural networks," J. Phys.: Condens. Matter **20**, 285219 (2008).

[189] V. E. Kuz'min, A. G. Artemenko, P. G. Polischuk, E. N. Muratov, A. I. Hromov, A. V. Liahovskiy, S. A. Andronati, and S. Y. Makan, "Hierarchic system of QSAR models (1D–4D) on the base of simplex representation of molecular structure," J. Mol. Model. **11**, 457–467 (2005).

[190] V. E. Kuz'min, A. G. Artemenko, and E. N. Muratov, "Hierarchical QSAR technology based on the simplex representation of molecular structure," J. Comput. Aid. Mol. Des. **22**, 403–421 (2008).

[191] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, "Universal fragment descriptors for predicting properties of inorganic crystals," Nat. Commun. **8**, 15679 (2017).

[192] T. L. Pham, H. Kino, K. Terakura, T. Miyake, K. Tsuda, I. Takigawa, and H. C. Dam, "Machine learning reveals orbital interaction in materials," Sci. Technol. Adv. Mat. **18**, 756–765 (2017).

[193] P. B. Jørgensen, K. W. Jacobsen, and M. N. Schmidt, "Neural message passing with edge updates for predicting properties of molecules and materials," arXiv:1806.03146 (2018).

[194] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, "Predicting crystal structure by merging data mining with quantum mechanics," Nat. Mater. **5**, 641–646 (2006).

[195] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, "Graph networks as a universal machine learning framework for molecules and crystals," Chem. Mater. **31**, 3564–3572 (2019).

[196] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated Graph Sequence Neural Networks," arXiv:1511.05493 (2015).

[197] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," Nat. Commun. **8** (2017), 10.1038/ncomms13890.

[198] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," arXiv:1609.02907 (2016).

[199] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: moving beyond fingerprints," J. Comput. Aided Mol. Des. **30**, 595–608 (2016).

[200] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral Networks and Locally Connected Networks on Graphs," arXiv:1312.6203 (2013).

[201] P. W. Battaglia, R. Pascanu, M. Lai, D. Rezende, and K. Kavukcuoglu, "Interaction Networks for Learning about Objects, Relations and Physics," arXiv:1612.00222 (2016).

[202] M. Defferrard, X. Bresson, and P. Vand ergheynst, "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering," arXiv:1606.09375 (2016).

[203] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in Neural Information Processing Systems 28*, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Curran Associates, Inc., 2015) pp. 2224–2232.

[204] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "Schnet – a deep learning architecture for molecules and materials," J. Chem. Phys. **148**, 241722 (2018).

[205] L. Raff, R. Komanduri, and M. Hagan, *Neural Networks in Chemical Reaction Dynamics* (Oxford Univrsity Press, 2012).

[206] B. J. Braams and J. M. Bowman, "Permutationally invariant potential energy surfaces in high dimensionality," Int. Rev. Phys. Chem. **28**, 577–606 (2009).

[207] S. J. Swamidass, J. Chen, J. Bruand, P. Phung, L. Ralaivola, and P. Baldi, "Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity," Bioinformatics **21**, i359–i368 (2005).

[208] H. Weyl, *The Classical Groups: Their Invariants and Representations* (Princeton University Press, 1997).

[209] F. Jensen, *Introduction to Computational Chemistry* (Wiley, 2013).

[210] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 70, edited by D. Precup and Y. W. Teh (PMLR, International Convention Centre, Sydney, Australia, 2017) pp. 1263–1272.

[211] J. Behler, "Perspective: Machine learning potentials for atomistic simulations," J. Chem. Phys. **145**, 170901 (2016).

[212] N. Artrith and A. Urban, "An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for TiO_2," Comput. Mater. Sci. **114**, 135–150 (2016).

[213] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," J. Chem. Phys. **134**, 074106 (2011).

[214] L. Himanen, M. O. Jäger, E. V. Morooka, F. F. Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, "Dscribe: Library of descriptors for machine learning in materials science," arXiv:1904.08875 (2019).

[215] L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, and A. Jain, "Matminer: An open source toolkit for materials data mining," Comput. Mater. Sci. **152**, 60–69 (2018).

[216] K. Yao, J. E. Herr, D. Toth, R. Mckintyre, and J. Parkhill, "The tensormol-0.1 model chemistry: a neural network augmented with long-range physics," Chem. Sci. **9**, 2261–2269 (2018).

[217] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, "Python materials genomics (pymatgen): A robust, open-source python library for materials analysis," Comput. Mater. Sci. **68**, 314–319 (2013).

[218] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller, "SchNetPack: A deep learning toolbox for atomistic systems," J. Chem. Theory Comput. **15**, 448–455 (2018).

[219] M. O. J. Jäger, E. V. Morooka, F. F. Canova, L. Himanen, and A. S. Foster, "Machine learning hydrogen adsorption on nanoclusters through structural descriptors," npj Comput. Mater. **4**, 37 (2018).

[220] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood, and S. P. Ong, "A performance and cost assessment of machine learning interatomic potentials," arXiv:1906.08888 (2019).

[221] A. Goncharsky, V. V. Stepanov, A. N. Tikhonov, and A. G. Yagola, *Numerical Methods for the Solution of Ill-Posed Problems* (Springer Netherlands, 1995).

[222] L. Devroye, L. Györfi, and G. Lugosi, "Vapnik-Chervonenkis theory," in *A Probabilistic Theory of Pattern Recognition* (Springer New York, 1996) pp. 187–213.

[223] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (MIT Press Ltd, 2005).

[224] T. Evgeniou, C. A. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," J. Mach. Learn. Res. **6**, 615–637 (2005).

[225] M. A. Álvarez, L. Rosasco, and N. D. Lawrence, "Kernels for vector-valued functions: A review," Found. Trends Mach. Learn. **4**, 195–266 (2012).

[226] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory - COLT'92* (ACM Press, 1992).

[227] F. Santosa and W. W. Symes, "Linear inversion of band-limited reflection seismograms," SIAM J. Sci. Stat. Comp. **7**, 1307–1330 (1986).

[228] R. Tibshirani, "Regression shrinkage and selection via the LASSO," J. Roy. Statist. Soc. Ser. B (1996).

[229] L. M. Ghiringhelli, J. Vybiral, E. Ahmetcik, R. Ouyang, S. V. Levchenko, C. Draxl, and M. Scheffler, "Learning physical descriptors for materials science by compressed sensing," New J. Phys. **19**, 023017 (2017).

[230] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L. M. Ghiringhelli, "SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates," Phys. Rev. Mater. **2**, 083802 (2018).

[231] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," Philos. Mag. **2**, 559–572 (1901).

[232] I. Jolliffe, *Principal Component Analysis* (Springer-Verlag, 2002).

[233] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow* (O'Reilly UK Ltd., 2017).

[234] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (The MIT Press, 2017).

[235] K. P. Murphy, *Machine Learning: A Probabilistic Perspective* (MIT Press Ltd, 2012).

[236] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag New York Inc., 2006).

[237] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies* (MIT Press Ltd, 2015).

[238] "The mostly complete chart of neural networks, explained," `https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464` (accessed in 2019).

[239] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," Psychol. Rev. **65**, 386 (1958).

[240] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for Boltzmann machines," Cognitive Sci. **9**, 147–169 (1985).

[241] P. Smolensky, "Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1," (MIT Press, Cambridge, MA, USA, 1986) Chap. Information Processing in Dynamical Systems: Foundations of Harmony Theory, pp. 194–281.

[242] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput. **9**, 1735–1780 (1997).

[243] D. C. Plaut and G. E. Hinton, "Learning sets of filters using back-propagation," Computer Speech & Language **2**, 35–61 (1987).

[244] G. E. Hinton, "Reducing the dimensionality of data with neural networks," Science **313**, 504–507 (2006).

[245] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," arXiv:1312.6114 (2013).

[246] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., 2014).

[247] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten Digit Recognition with a Back-Propagation Network," in *Advances in Neural Information Processing Systems 2* (1990) pp. 396–404.

[248] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436–444 (2015).

[249] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," Bull. Math. Sci. **5**, 115–133 (1943).

[250] J. F. Kolen and S. C. Kremer, "Gradient flow in recurrent nets: The difficulty of learning long-term dependencies," in *A Field Guide to Dynamical Recurrent Networks* (Wiley-IEEE Press, 2001).

[251] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *ICML'10 Proceedings of the 27th International Conference on International Conference on Machine Learning* (Omnipress, 2010) pp. 807–814.

[252] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 15, edited by G. Gordon, D. Dunson, and M. Dudík (PMLR, 2011).

[253] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," arXiv:1511.07289 (2015).

[254] Z. Ren and Y. J. Lee, "Cross-domain self-supervised multi-task feature learning using synthetic imagery," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2018).

[255] M. Nielsen, "Neural networks and deep learning," `http://neuralnetworksanddeeplearning.com/index.html` (accessed in 2019).

[256] K. Hornik, "Approximation capabilities of multilayer feedforward networks," Neural Networks **4**, 251 – 257 (1991).

[257] J. Brownlee, *Clever algorithms : nature-inspired programming recipes* (Lulu, Melbourne, 2011).

[258] R. Rojas, *Neural Networks* (Springer Berlin Heidelberg, 1996).

[259] P. V. Balachandran, D. Xue, and T. Lookman, "Structure–Curie temperature relationships in $BaTiO_3$-based ferroelectric perovskites: Anomalous behavior of $(Ba,Cd)TiO_3$ from DFT, statistical inference, and experiments," Phys. Rev. B **93**, 144111 (2016).

[260] S. Sanvito, C. Oses, J. Xue, A. Tiwari, M. Zic, T. Archer, P. Tozman, M. Venkatesan, M. Coey, and S. Curtarolo, "Accelerated discovery of new magnets in the Heusler alloy family," Sci. Adv. **3**, e1602241 (2017).

[261] P. V. Balachandran, B. Kowalski, A. Sehirlioglu, and T. Lookman, "Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning," Nat. Commun. **9**, 1668 (2018).

[262] X. Zhai, M. Chen, and W. Lu, "Accelerated search for perovskite materials with higher Curie temperature based on the machine learning methods," Comput. Mater. Sci. **151**, 41–48 (2018).

[263] H. C. Dam, V. C. Nguyen, T. L. Pham, A. T. Nguyen, K. Terakura, T. Miyake, and H. Kino, "Important descriptors and descriptor groups of Curie temperatures of rare-earth transition-metal binary alloys," J. Phys. Soc. Jpn. **87**, 113801 (2018).

[264] B. Zhang, X.-Q. Zheng, T.-Y. Zhao, F.-X. Hu, J.-R. Sun, and B.-G. Shen, "Machine learning technique for prediction of magnetocaloric effect in $La(Fe,Si/Al)_{13}$-based materials," Chin. Phys. B **27**, 067503 (2018).

[265] F. Legrain, J. Carrete, A. van Roekeghem, S. Curtarolo, and N. Mingo, "How chemical composition alone can predict vibrational free energies and entropies of solids," Chem. Mater. **29**, 6220–6227 (2017).

[266] Y. Zhuo, A. M. Tehrani, and J. Brgoch, "Predicting the band gaps of inorganic solids by machine learning," J. Phys. Chem. Lett. **9**, 1668–1673 (2018).

[267] P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon, and K. Rajan, "Informatics-aided bandgap engineering for solar materials," Comput. Mater. Sci. **83**, 185–195 (2014).

[268] J. Lee, A. Seko, K. Shitara, K. Nakayama, and I. Tanaka, "Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques," Phys. Rev. B **93**, 115104 (2016).

[269] G. Pilania, J. Gubernatis, and T. Lookman, "Multi-fidelity machine learning models for accurate bandgap predictions of solids," Comput. Mater. Sci. **129**, 156–163 (2017).

[270] A. C. Rajan, A. Mishra, S. Satsangi, R. Vaish, H. Mizuseki, K.-R. Lee, and A. K. Singh, "Machine-learning-assisted accurate band gap predictions of functionalized MXene," Chem. Mater. **30**, 4031–4038 (2018).

[271] T. Sparks, S. Kauwe, and T. Welker, "Extracting knowledge from DFT: Experimental band gap predictions through ensemble learning," ChemRxiv:10.26434/chemrxiv.7236029.v1 (2018).

[272] L. Weston and C. Stampfl, "Machine learning the band gap properties of kesterite $I_2-II-IV-V_4$ quaternary compounds for photovoltaics applications," Phys. Rev. Mater. **2**, 085407 (2018).

[273] T. Gu, W. Lu, X. Bao, and N. Chen, "Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors," Solid State Sci. **8**, 129–136 (2006).

[274] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, and T. Lookman, “Machine learning bandgaps of double perovskites,” Sci. Rep. **6**, 19375 (2016).

[275] W. Setyawan, R. M. Gaume, S. Lam, R. S. Feigelson, and S. Curtarolo, “High-throughput combinatorial database of electronic band structures for inorganic scintillator materials,” ACS Comb. Sci. **13**, 382–390 (2011).

[276] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, and J. Wang, “Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning,” Nat. Commun. **9**, 3405 (2018).

[277] G. Pilania and X.-Y. Liu, “Machine learning properties of binary wurtzite superlattices,” J. Mater. Sci. **53**, 6652–6664 (2018).

[278] C. Kim, G. Pilania, and R. Ramprasad, “From organized high-throughput data to phenomenological theory using machine learning: The example of dielectric breakdown,” Chem. Mater. **28**, 1304–1311 (2016).

[279] C. Kim, G. Pilania, and R. Ramprasad, “Machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX_3 perovskites,” J. Phys. Chem. C **120**, 14575–14580 (2016).

[280] F. Yuan and T. Mueller, “Identifying models of dielectric breakdown strength from high-throughput data via genetic programming,” Sci. Rep. **7**, 17594 (2017).

[281] A. Furmanchuk, A. Agrawal, and A. Choudhary, “Predictive analytics for crystalline materials: bulk modulus,” RSC Advances **6**, 95246–95251 (2016).

[282] S. K. Kauwe, J. Graser, A. Vazquez, and T. D. Sparks, “Machine learning prediction of heat capacity for solid inorganics,” Integr. Mater. Manuf. Innov. **7**, 43–51 (2018).

[283] D. R. Cassar, A. C. de Carvalho, and E. D. Zanotto, “Predicting glass transition temperatures using neural networks,” Acta Mater. **159**, 249–256 (2018).

[284] Y. Liu, T. Zhao, G. Yang, W. Ju, and S. Shi, “The onset temperature (Tg) of As_xSe_{1-x} glasses transition prediction: A comparison of topological and regression analysis methods,” Comput. Mater. Sci. **140**, 315–321 (2017).

[285] T. Zhan, L. Fang, and Y. Xu, “Prediction of thermal boundary resistance by the machine learning method,” Sci. Rep. **7**, 7109 (2017).

[286] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka, “Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization,” Phys. Rev. Lett. **115**, 205901 (2015).

[287] J. Carrete, W. Li, N. Mingo, S. Wang, and S. Curtarolo, “Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling,” Phys. Rev. X **4**, 011019 (2014).

[288] A. van Roekeghem, J. Carrete, C. Oses, S. Curtarolo, and N. Mingo, “High-throughput computation of thermal conductivity of high-temperature solid phases: The case of oxide and fluoride perovskites,” Phys. Rev. X **6**, 041061 (2016).

[289] M. W. Gaultois, A. O. Oliynyk, A. Mar, T. D. Sparks, G. J. Mulholland, and B. Meredig, “Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties,” APL Mater. **4**, 053213 (2016).

[290] S. Ju, T. Shiga, L. Feng, Z. Hou, K. Tsuda, and J. Shiomi, “Designing nanostructures for phonon transport via Bayesian optimization,” Phys. Rev. X **7**, 021024 (2017).

[291] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, and I. Tanaka, “Representation of compounds for machine-learning prediction of physical properties,” Phys. Rev. B **95**, 144110 (2017).

[292] G. C. Sosso, V. L. Deringer, S. R. Elliott, and G. Csányi, "Understanding the thermal properties of amorphous solids using machine-learning-based interatomic potentials," Mol. Simulat. **44**, 866–880 (2018).

[293] H. Wei, S. Zhao, Q. Rong, and H. Bao, "Predicting the effective thermal conductivities of composite materials and porous media by machine learning methods," Int. J. Heat Mass Tran. **127**, 908–916 (2018).

[294] Y.-J. Wu, M. Sasaki, M. Goto, L. Fang, and Y. Xu, "Electrically conductive thermally insulating Bi–Si nanocomposites by interface design for thermal management," ACS Appl. Nano Mater. **1**, 3355–3363 (2018).

[295] T.-L. Pham, N.-D. Nguyen, V.-D. Nguyen, H. Kino, T. Miyake, and H.-C. Dam, "Learning structure-property relationship in crystalline materials: A study of lanthanide–transition metal alloys," J. Chem. Phys. **148**, 204106 (2018).

[296] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, "Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids," Phys. Rev. B **89**, 054303 (2014).

[297] G. Pilania, J. E. Gubernatis, and T. Lookman, "Structure classification and melting temperature prediction in octet AB solids via machine learning," Phys. Rev. B **91**, 214302 (2015).

[298] S. Kikuchi, H. Oda, S. Kiyohara, and T. Mizoguchi, "Bayesian optimization for efficient determination of metal oxide grain boundary structures," Physica B **532**, 24–28 (2018).

[299] S. Kiyohara, H. Oda, T. Miyata, and T. Mizoguchi, "Prediction of interface structures and energies via virtual screening," Sci. Adv. **2**, e1600746 (2016).

[300] Q. Zhu, A. Samanta, B. Li, R. E. Rudd, and T. Frolov, "Predicting phase behavior of grain boundaries with evolutionary search and machine learning," Nat. Commun. **9**, 467 (2018).

[301] S. Kiyohara, H. Oda, K. Tsuda, and T. Mizoguchi, "Acceleration of stable interface structure searching using a kriging approach," Jpn. J. Appl. Phys. **55**, 045502 (2016).

[302] C. W. Rosenbrock, E. R. Homer, G. Csányi, and G. L. W. Hart, "Discovering the building blocks of atomic systems using machine learning: application to grain boundaries," npj Comput. Mater. **3**, 29 (2017).

[303] A. Furmanchuk, J. E. Saal, J. W. Doak, G. B. Olson, A. Choudhary, and A. Agrawal, "Prediction of Seebeck coefficient for compounds without restriction to fixed stoichiometry: A machine learning approach," J. Comput. Chem. **39**, 191–202 (2017).

[304] M. Abdellahi, M. Bahmanpour, and M. Bahmanpour, "Modeling Seebeck coefficient of $\mathrm{Ca}_{3-x}\mathrm{M}_x\mathrm{Co}_4\mathrm{O}_9$ (M = Sr, Pr, Ga, Ca, Ba, La, Ag) thermoelectric ceramics," Ceram. Int. **41**, 345–352 (2015).

[305] J. Carrete, N. Mingo, S. Wang, and S. Curtarolo, "Nanograined half-Heusler semiconductors as advanced thermoelectrics: An ab initio high-throughput statistical study," Adv. Funct. Mater. **24**, 7427–7432 (2014).

[306] M. de Jong, W. Chen, R. Notestine, K. Persson, G. Ceder, A. Jain, M. Asta, and A. Gamst, "A statistical learning framework for materials science: Application to elastic moduli of k-nary inorganic polycrystalline compounds," Sci. Rep. **6**, 34256 (2016).

[307] J. D. Evans and F.-X. Coudert, "Predicting the mechanical properties of zeolite frameworks by machine learning," Chem. Mater. **29**, 7833–7839 (2017).

[308] A. M. Tehrani, A. O. Oliynyk, M. Parry, Z. Rizvi, S. Couper, F. Lin, L. Miyagi, T. D. Sparks, and J. Brgoch, "Machine learning directed search for ultraincompressible, superhard materials," J. Am. Chem. Soc. **140**, 9844–9853 (2018).

[309] B. C. Yeo, D. Kim, C. Kim, and S. S. Han, "Pattern learning electronic density of states," arXiv:1808.03383 (2018).

[310] S. R. Broderick, H. Aourag, and K. Rajan, "Classification of oxide compounds through data-mining density of states spectra," J. Am. Ceram. Soc. **94**, 2974–2980 (2011).

[311] B. Meredig and C. Wolverton, "Dissolving the periodic table in cubic zirconia: Data mining to discover chemical trends," Chem. Mater. **26**, 1985–1991 (2014).

[312] Y. Zhang and E.-A. Kim, "Quantum loop topography for machine learning," Phys. Rev. Lett. **118**, 216401 (2017).

[313] P. Zhang, H. Shen, and H. Zhai, "Machine learning topological invariants with neural networks," Phys. Rev. Lett. **120**, 066401 (2018).

[314] D.-L. Deng, X. Li, and S. D. Sarma, "Machine learning topological states," Phys. Rev. B **96**, 195145 (2017).

[315] N. Sun, J. Yi, P. Zhang, H. Shen, and H. Zhai, "Deep learning topological invariants of band insulators," Phys. Rev. B **98**, 085402 (2018).

[316] M. J. S. Beach, A. Golubeva, and R. G. Melko, "Machine learning vortices at the Kosterlitz-Thouless transition," Phys. Rev. B **97**, 045207 (2018).

[317] L. Pilozzi, F. A. Farrelly, G. Marcucci, and C. Conti, "Machine learning inverse problem for topological photonics," Commun. Phys. **1**, 57 (2018).

[318] J. Carrasquilla and R. G. Melko, "Machine learning phases of matter," Nat. Phys. **13**, 431–434 (2017).

[319] T. O. Owolabi, K. O. Akande, and S. O. Olatunji, "Prediction of superconducting transition temperatures for Fe-based superconductors using support vector machine," Adv. Phys. Theor. Appl. **35**, 12–26 (2014).

[320] T. O. Owolabi, K. O. Akande, and S. O. Olatunji, "Estimation of superconducting transition temperature T_C for superconductors of the doped MgB_2 system from the crystal lattice parameters using support vector regression," J. Supercond. Nov. Magn. **28**, 75–81 (2014).

[321] V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, and I. Takeuchi, "Machine learning modeling of superconducting critical temperature," npj Comput. Mater. **4**, 29 (2018).

[322] O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, and S. Curtarolo, "Materials cartography: Representing and mining materials space using structural and electronic fingerprints," Chem. Mater. **27**, 735–743 (2015).

[323] J. Ling, M. Hutchinson, E. Antono, S. Paradiso, and B. Meredig, "High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates," Integr. Mater. Manuf. Innov. **6**, 207–217 (2017).

[324] A. D. Sendek, E. D. Cubuk, E. R. Antoniuk, G. Cheon, Y. Cui, and E. J. Reed, "Machine learning-assisted discovery of solid Li-ion conducting materials," Chem. Mater. **31**, 342–352 (2019).

[325] W. Waag, C. Fleischer, and D. U. Sauer, "Critical review of the methods for monitoring of lithium-ion batteries in electric and hybrid vehicles," J. Power Sources **258**, 321 – 339 (2014).

[326] D. J. Tozer, V. E. Ingamells, and N. C. Handy, "Exchange-correlation potentials," J. Chem. Phys. **105**, 9200–9213 (1996).

[327] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, "Bypassing the Kohn-Sham equations with machine learning," Nat. Commun. **8**, 872 (2017).

[328] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, "Finding density functionals with machine learning," Phys. Rev. Lett. **108**, 253002 (2012).

[329] J. C. Snyder, M. Rupp, K. Hansen, L. Blooston, K.-R. Müller, and K. Burke, "Orbital-free bond breaking via machine learning," J. Chem. Phys. **139**, 224104 (2013).

[330] Q. Liu, J. Wang, P. Du, L. Hu, X. Zheng, and G. Chen, "Improving the performance of long-range-corrected exchange-correlation functional with an embedded neural network," J. Phys. Chem. A **121**, 7273–7281 (2017).

[331] R. Nagai, R. Akashi, S. Sasaki, and S. Tsuneyuki, "Neural-network Kohn-Sham exchange-correlation potential and its out-of-training transferability," J. Chem. Phys. **148**, 241737 (2018).

[332] J. Schmidt, C. L. Benavides-Riveros, and M. A. L. Marques, "Machine learning the physical nonlocal exchange–correlation functional of density-functional theory," J. Phys. Chem. Lett. **10**, 6425–6431 (2019).

[333] V. M. Goldschmidt, "Die gesetze der krystallochemie," Die Naturwissenschaften **14**, 477–485 (1926).

[334] G. Pilania, P. V. Balachandran, J. E. Gubernatis, and T. Lookman, "Classification of ABO_3 perovskite solids: a machine learning study," Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater. **71**, 507–513 (2015).

[335] P. V. Balachandran, A. A. Emery, J. E. Gubernatis, T. Lookman, C. Wolverton, and A. Zunger, "Predictions of new ABO_3 perovskite compounds by combining machine learning and density functional theory," Phys. Rev. Mater. **2**, 043802 (2018).

[336] J. Schmidt, L. Chen, S. Botti, and M. A. L. Marques, "Predicting the stability of ternary intermetallics with density functional theory and machine learning," J. Chem. Phys. **148**, 241728 (2018).

[337] K. Kim, L. Ward, J. He, A. Krishna, A. Agrawal, and C. Wolverton, "Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Heusler compounds," Phys. Rev. Mater. **2**, 123801 (2018).

[338] A. O. Oliynyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig, and A. Mar, "High-throughput machine-learning-driven synthesis of full-Heusler compounds," Chem. Mater. **28**, 7324–7331 (2016).

[339] X. Zheng, P. Zheng, and R.-Z. Zhang, "Machine learning material properties from the periodic table using convolutional neural networks," Chem. Sci. **9**, 8426–8432 (2018).

[340] L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, and C. Wolverton, "Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations," Phys. Rev. B **96**, 024104 (2017).

[341] D. G. Pettifor, "The structures of binary compounds. I. Phenomenological structure maps," J. Phys. C: Solid State Phys. **19**, 285–313 (1986).

[342] D. G. Pettifor, "Structure maps for pseudobinary and ternary phases," Mater. Sci. Tech. **4**, 675–691 (1988).

[343] H. Glawe, A. Sanna, E. K. U. Gross, and M. A. L. Marques, "The optimal one dimensional periodic table: a modified pettifor chemical scale from data mining," New J. Phys. **18**, 093011 (2016).

[344] T. Morita, "Cluster variation method of cooperative phenomena and its generalization II. quantum statistics," J. Phys. Soc. Jpn. **12**, 1060–1063 (1957).

[345] N. A. Sinkov and J. J. Harynuk, "Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling," Talanta **83**, 1079–1087 (2011).

[346] J. Graser, S. K. Kauwe, and T. D. Sparks, "Machine learning and energy minimization approaches for crystal structure predictions: A review and new horizons," Chem. Mater. **30**, 3601–3612 (2018).

[347] A. Nouira, N. Sokolovska, and J.-C. Crivello, "Crystalgan: Learning to discover crystallographic structures with generative adversarial networks," arXiv:1810.11203 (2018).

[348] X. Li, Y. Zhang, H. Zhao, C. Burkhart, L. C. Brinson, and W. Chen, "A transfer learning approach for microstructure reconstruction and structure-property predictions," Sci. Rep. **8**, 13461 (2018).

[349] X. Li, Z. Yang, L. C. Brinson, A. Choudhary, A. Agrawal, and W. Chen, "A deep adversarial learning methodology for designing microstructural material systems," in *ASME. International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Volume 2B: 44th Design Automation Conference* (ASME, 2018).

[350] W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin, and K.-S. Sohn, "Classification of crystal structure using a convolutional neural network," IUCrJ **4**, 486–494 (2017).

[351] A. Ziletti, D. Kumar, M. Scheffler, and L. M. Ghiringhelli, "Insightful classification of crystal structures using deep learning," Nat. Commun. **9**, 2775 (2018).

[352] D. M. Dimiduk, E. A. Holm, and S. R. Niezgoda, "Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes, and structures engineering," Integr. Mater. Manuf. Innov. **7**, 157–172 (2018).

[353] S. V. Kalinin, B. G. Sumpter, and R. K. Archibald, "Big–deep–smart data in imaging for guiding materials design," Nat. Mater. **14**, 973–980 (2015).

[354] Z. Liu, T. Bicer, R. Kettimuthu, D. Gursoy, F. De Carlo, and I. Foster, "Tomogan: Low-dose x-ray tomography with generative adversarial networks," arXiv:1902.07582 (2019).

[355] R. Liu, A. Agrawal, W. Liao, A. Choudhary, and M. De Graef, "Materials discovery: Understanding polycrystals from large-scale electron patterns," in *2016 IEEE International Conference on Big Data (Big Data)* (2016) pp. 2261–2269.

[356] B. Wang, Z. Guan, S. Yao, H. Qin, M. H. Nguyen, K. Yager, and D. Yu, "Deep learning for analysing synchrotron data streams," in *2016 New York Scientific Data Summit (NYSDS)* (2016) pp. 1–5.

[357] B. L. DeCost, H. Jain, A. D. Rollett, and E. A. Holm, "Computer vision and machine learning for autonomous characterization of am powder feedstocks," JOM **69**, 456–465 (2017).

[358] A.-T. Nguyen, S. Reiter, and P. Rigo, "A review on simulation-based optimization methods applied to building performance analysis," Appl. Energ. **113**, 1043–1058 (2014).

[359] A. I. Forrester and A. J. Keane, "Recent advances in surrogate-based optimization," Prog. Aerosp. Sci. **45**, 50–79 (2009).

[360] P. V. Balachandran, D. Xue, J. Theiler, J. Hogden, and T. Lookman, "Adaptive strategies for materials design using uncertainties," Sci. Rep. **6**, 19660 (2016).

[361] G. Schneider, M. Hartenfeller, M. Reutlinger, Y. Tanrikulu, E. Proschak, and P. Schneider, "Voyages to the (un)known: adaptive design of bioactive compounds," Trends Biotechnol. **27**, 18–26 (2009).

[362] J. Bajorath, L. Peltason, M. Wawer, R. Guha, M. S. Lajiness, and J. H. V. Drie, "Navigating structure–activity landscapes," Drug Discov. Today **14**, 698–705 (2009).

[363] S. R. Johnson, "The trouble with QSAR (or how I learned to stop worrying and embrace fallacy)," J. Chem. Inf. Model. **48**, 25–26 (2008).

[364] G. M. Maggiora, "On outliers and activity cliffs – why QSAR often disappoints," J. Chem. Inf. Model. **46**, 1535–1535 (2006).

[365] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, "Accelerated search for materials with targeted properties by adaptive design," Nat. Commun. **7**, 11241 (2016).

[366] D. Xue, P. V. Balachandran, R. Yuan, T. Hu, X. Qian, E. R. Dougherty, and T. Look-

man, “Accelerated search for $BaTiO_3$-based piezoelectrics with vertical morphotropic phase boundary using Bayesian learning,” Proc. Natl. Acad. Sci. U. S. A. **113**, 13301–13306 (2016).

[367] A. Solomou, G. Zhao, S. Boluki, J. K. Joy, X. Qian, I. Karaman, R. Arróyave, and D. C. Lagoudas, “Multi-objective Bayesian materials discovery: Application on the discovery of precipitation strengthened NiTi shape memory alloys through micromechanical modeling,” Mater. Des. **160**, 810–827 (2018).

[368] A. Talapatra, S. Boluki, T. Duong, X. Qian, E. Dougherty, and R. Arróyave, “Autonomous efficient experiment design for materials discovery with Bayesian model averaging,” Phys. Rev. Mat. **2**, 113803 (2018).

[369] A. M. Gopakumar, P. V. Balachandran, D. Xue, J. E. Gubernatis, and T. Lookman, “Multi-objective optimization for materials discovery via adaptive design,” Sci. Rep. **8**, 3738 (2018).

[370] P. V. Balachandran, D. Xue, J. Theiler, J. Hogden, J. E. Gubernatis, and T. Lookman, “Importance of feature selection in machine learning and adaptive design for materials,” in *Materials Discovery and Design* (Springer International Publishing, 2018) pp. 59–79.

[371] M. Hutchinson, S. Paradiso, and L. Ward, “Citrine informatics lolo,” `https://github.com/CitrineInformatics/lolo` (2016).

[372] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton, “A survey of Monte Carlo tree search methods,” IEEE T. Comp. Intel. AI **4**, 1–43 (2012).

[373] T. M. Dieb, S. Ju, K. Yoshizoe, Z. Hou, J. Shiomi, and K. Tsuda, “MDTS: automatic complex materials design using Monte Carlo tree search,” Sci. Technol. Adv. Mat. **18**, 498–503 (2017).

[374] S. Kiyohara and T. Mizoguchi, “Searching the stable segregation configuration at the grain boundary by a Monte Carlo tree search,” J. Chem. Phys. **148**, 241741 (2018).

[375] T. M. Dieb, Z. Hou, and K. Tsuda, “Structure prediction of boron-doped graphene by machine learning,” J. Chem. Phys. **148**, 241716 (2018).

[376] D. D. Johnson, “Evolutionary algorithms applied to electronic-structure informatics,” in *Informatics for Materials Science and Engineering* (Elsevier, 2013) pp. 349–364.

[377] R. Sawada, Y. Iwasaki, and M. Ishida, “Boosting material modeling using game tree search,” Phys. Rev. Mat. **2**, 103802 (2018).

[378] R. Dehghannasiri, D. Xue, P. V. Balachandran, M. R. Yousefi, L. A. Dalton, T. Lookman, and E. R. Dougherty, “Optimal experimental design for materials discovery,” Comput. Mater. Sci. **129**, 311–322 (2017).

[379] Y. Wang, K. G. Reyes, K. A. Brown, C. A. Mirkin, and W. B. Powell, “Nested-batch-mode learning and stochastic optimization with an application to sequential multi-stage testing in materials science,” SIAM J. Sci. Comput. **37**, B361–B381 (2015).

[380] S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, and D. Morgan, “AFLOW: An automatic framework for high-throughput materials discovery,” Comput. Mater. Sci. **58**, 218 – 226 (2012).

[381] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson, “The Materials Project: A materials genome approach to accelerating materials innovation,” APL Mater. **1**, 011002 (2013).

[382] G. Ceder, “Opportunities and challenges for first-principles materials design and applications to Li battery materials,” MRS Bull. **35**, 693–701 (2010).

[383] T. J. Goncalves, U. Arnold, P. N. Plessow, and F. Studt, “Theoretical investigation of the acid catalyzed formation of oxymethylene dimethyl ethers from trioxane and dimethoxymethane,” ACS Catal. **7**, 3615–3621 (2017).

[384] R. Sarmiento-Pérez, T. F. T. Cerqueira, S. Körbel, S. Botti, and M. A. L. Marques, "Prediction of stable nitride perovskites," Chem. Mater. **27**, 5957–5963 (2015).

[385] S. Körbel, M. A. L. Marques, and S. Botti, "Stable hybrid organic–inorganic halide perovskites for photovoltaics from ab initio high-throughput calculations," J. Mater. Chem. A **6**, 6463–6475 (2018).

[386] N. Drebov, A. Martinez-Limia, L. Kunz, A. Gola, T. Shigematsu, T. Eckl, P. Gumbsch, and C. Elsässer, "Ab initio screening methodology applied to the search for new permanent magnetic materials," New J. Phys. **15**, 125023 (2013).

[387] D. R. Bowler and T. Miyazaki, "Calculations for millions of atoms with density functional theory: linear scaling shows its potential," J. Phys. Condens. Matter. **22**, 074207 (2010).

[388] A. Nakata, Y. Futamura, T. Sakurai, D. R. Bowler, and T. Miyazaki, "Efficient calculation of electronic structure using O(N) density functional theory," J. Chem. Theory Comput. **13**, 4146–4153 (2017).

[389] L. E. Ratcliff, S. Mohr, G. Huhs, T. Deutsch, M. Masella, and L. Genovese, "Challenges in large scale quantum mechanical calculations," Wiley Interdiscip. Rev. Comput. Mol. Sci. **7**, e1290 (2017).

[390] T. D. Huan, M. Amsler, M. A. L. Marques, S. Botti, A. Willand, and S. Goedecker, "Low-energy polymeric phases of alanates," Phys. Rev. Lett. **110**, 135502 (2013).

[391] H. D. Tran, M. Amsler, S. Botti, M. A. L. Marques, and S. Goedecker, "First-principles predicted low-energy structures of $NaSc(BH_4)_4$," J. Chem. Phys. **140**, 124708 (2014).

[392] J. Tersoff, "New empirical model for the structural properties of silicon," Phys. Rev. Lett. **56**, 632–635 (1986).

[393] F. H. Stillinger and T. A. Weber, "Computer simulation of local order in condensed phases of silicon," Phys. Rev. B **31**, 5262–5271 (1985).

[394] A. C. T. van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard, "ReaxFF: a reactive force field for hydrocarbons," J. Phys. Chem. A **105**, 9396–9409 (2001).

[395] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins," J. Phys. Chem. B **102**, 3586–3616 (1998).

[396] M. S. Daw and M. I. Baskes, "Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals," Phys. Rev. B **29**, 6443–6453 (1984).

[397] M. S. Daw, S. M. Foiles, and M. I. Baskes, "The embedded-atom method: a review of theory and applications," Mater. Sci. Rep. **9**, 251–310 (1993).

[398] C. A. Becker, F. Tavazza, Z. T. Trautt, and R. A. B. de Macedo, "Considerations for choosing and using force fields and interatomic potentials in materials science and engineering," Curr. Opin. Solid State Mater. Sci. **17**, 277–283 (2013).

[399] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, and R. Kaschner, "Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon," Phys. Rev. B **51**, 12947–12957 (1995).

[400] G. Seifert and J.-O. Joswig, "Density-functional tight binding – an approximate density-functional theory method," Wiley Interdisciplinary Reviews: Computational Molecular Science **2**, 456–465 (2012).

[401] P. Koskinen and V. Mäkinen, "Density-functional tight-binding for beginners," Comput. Mater. Sci. **47**, 237–253 (2009).

[402] B. G. Sumpter and D. W. Noid, "Potential energy surfaces for macromolecules. a neural network technique," Chem. Phys. Lett. **192**, 455–462 (1992).

[403] T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, "Neural network models of potential energy surfaces," J. Chem. Phys. **103**, 4129–4137 (1995).

[404] C. M. Handley and P. L. A. Popelier, "Potential energy surfaces fitted by artificial neural networks," J. Phys. Chem. A **114**, 3371–3383 (2010).

[405] J. Behler, "First principles neural network potentials for reactive simulations of large molecular and condensed systems," Angew. Chem., Int. Ed. **56**, 12828–12840 (2017).

[406] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).

[407] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, "Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials," J. Comput. Phys. **285**, 316–330 (2015).

[408] R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, and M. Parrinello, "Nucleation mechanism for the direct graphite-to-diamond phase transition," Nat. Mater. **10**, 693–697 (2011).

[409] H. Eshet, R. Z. Khaliullin, T. D. Kühne, J. Behler, and M. Parrinello, "Ab initio quality neural-network potential for sodium," Phys. Rev. B **81**, 184107 (2010).

[410] N. Artrith, T. Morawietz, and J. Behler, "High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide," Phys. Rev. B **83**, 153101 (2011).

[411] G. C. Sosso, G. Miceli, S. Caravati, J. Behler, and M. Bernasconi, "Neural network interatomic potential for the phase change material GeTe," Phys. Rev. B **85**, 174103 (2012).

[412] N. Artrith and J. Behler, "High-dimensional neural network potentials for metal surfaces: A prototype study for copper," Phys. Rev. B **85**, 045439 (2012).

[413] J. R. Boes, M. C. Groenenboom, J. A. Keith, and J. R. Kitchin, "Neural network and ReaxFF comparison for Au properties," Int. J. Quantum Chem. **116**, 979–987 (2016).

[414] N. Artrith, A. Urban, and G. Ceder, "Constructing first-principles phase diagrams of amorphous $\mathrm{Li}_x\mathrm{Si}$ using machine-learning-assisted sampling with an evolutionary algorithm," J. Chem. Phys. **148**, 241711 (2018).

[415] R. Kobayashi, D. Giofré, T. Junge, M. Ceriotti, and W. A. Curtin, "Neural network potential for Al-Mg-Si alloys," Phys. Rev. Mater. **1**, 053604 (2017).

[416] J. B. Witkoskie and D. J. Doren, "Neural network models of potential energy surfaces: prototypical examples," J. Chem. Theory Comput. **1**, 14–23 (2005).

[417] A. Pukrittayakamee, M. Hagan, L. Raff, S. Bukkapatnam, and R. Komanduri, "Fitting a function and its derivative," in *Intelligent Engineering Systems Through Artificial Neural Networks: Smart Systems Engineering Computational Intelligence in Architecting Complex Engineering Systems, Volume 17* (ASME Press, 2007) pp. 469–474.

[418] A. Pukrittayakamee, M. Malshe, M. Hagan, L. M. Raff, R. Narulkar, S. Bukkapatnum, and R. Komanduri, "Simultaneous fitting of a potential-energy surface and its corresponding force fields using feedforward neural networks," J. Chem. Phys. **130**, 134101 (2009).

[419] S. A. Ghasemi, A. Hofstetter, S. Saha, and S. Goedecker, "Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network," Phys. Rev. B **92**, 045131 (2015).

[420] S. Faraji, S. A. Ghasemi, S. Rostami, R. Rasoulkhani, B. Schaefer, S. Goedecker, and M. Amsler, "High accuracy and transferability of a neural network potential through charge equilibration for calcium fluoride," Phys. Rev. B **95**, 104105 (2017).

[421] S. Hajinazar, J. Shao, and A. N. Kolmogorov, "Stratified construction of neural network based interatomic models for multicomponent materials," Phys. Rev. B **95**, 014114 (2017).

[422] W. J. Szlachta, A. P. Bartók, and G. Csányi, "Accuracy and transferability of Gaussian approximation potential models for tungsten," Phys. Rev. B **90**, 104108 (2014).

[423] V. L. Deringer and G. Csányi, "Machine learning based interatomic potential for amorphous carbon," Phys. Rev. B **95**, 094203 (2017).

[424] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," Phys. Chem. Chem. Phys. **18**, 13754–13769 (2016).

[425] V. Vítek, "Intrinsic stacking faults in body-centred cubic crystals," Philos. Mag. **18**, 773–786 (1968).

[426] D. Dragoni, T. D. Daff, G. Csányi, and N. Marzari, "Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron," Phys. Rev. Mater. **2**, 013808 (2018).

[427] P. Rowe, G. Csányi, D. Alfè, and A. Michaelides, "Development of a machine learning potential for graphene," Phys. Rev. B **97**, 054303 (2018).

[428] A. Kamath, R. A. Vargas-Hernández, R. V. Krems, T. Carrington, and S. Manzhos, "Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy," J. Chem. Phys. **148**, 241702 (2018).

[429] G. Schmitz and O. Christiansen, "Gaussian process regression to accelerate geometry optimizations relying on numerical differentiation," J. Chem. Phys. **148**, 241704 (2018).

[430] V. L. Deringer, C. J. Pickard, and G. Csányi, "Data-driven learning of total and local energies in elemental boron," Phys. Rev. Lett. **120**, 156001 (2018).

[431] R. Jinnouchi, F. Karsai, and G. Kresse, "On-the-fly machine learning force field generation: Application to melting points," Phys. Rev. B **100**, 014105 (2019).

[432] M. A. Wood and A. P. Thompson, "Extending the accuracy of the SNAP interatomic potential form," J. Chem. Phys. **148**, 241721 (2018).

[433] X.-G. Li, C. Hu, C. Chen, Z. Deng, J. Luo, and S. P. Ong, "Quantum-accurate spectral neighbor analysis potential models for Ni–Mo binary alloys and fcc metals," Phys. Rev. B **98**, 094104 (2018).

[434] T. Jacobsen, M. Jørgensen, and B. Hammer, "On-the-fly machine learning of atomic potential in density functional theory structure optimization," Phys. Rev. Lett. **120**, 026102 (2018).

[435] Z. Li, J. R. Kermode, and A. D. Vita, "Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces," Phys. Rev. Lett. **114**, 096405 (2015).

[436] I. Kruglov, O. Sergeev, A. Yanilkin, and A. R. Oganov, "Energy-free machine learning force field for aluminum," Sci. Rep. **7**, 8512 (2017).

[437] A. Glielmo, P. Sollich, and A. D. Vita, "Accurate interatomic force fields via machine learning with covariant kernels," Phys. Rev. B **95**, 214302 (2017).

[438] A. Glielmo, C. Zeni, and A. D. Vita, "Efficient nonparametric n-body force fields from machine learning," Phys. Rev. B **97**, 184307 (2018).

[439] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," J. Roy. Statist. Soc. Ser. B **67**, 301–320 (2005).

[440] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer New York, 2009).

[441] J. Han, L. Zhang, R. Car, and E. Weinan, "Deep potential: A general representation of a many-body potential energy surface," Commun. Comput. Phys. **23**, 629 (2018).

[442] L. Zhang, J. Han, H. Wang, R. Car, and W. E, "Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics," Phys. Rev. Lett. **120**, 143001 (2018).

[443] A. V. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials," Multiscale Model Simul. **14**, 1153–1173 (2016).

[444] J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: an extensible neural network potential with dft accuracy at force field computational cost," Chem. Sci. **8**, 3192–3203 (2017).

[445] J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules," Sci. Data **4** (2017), 10.1038/sdata.2017.193.

[446] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. Roitberg, "Outsmarting quantum chemistry through transfer learning," ChemRxiv:10.26434/chemrxiv.6744440.v1 (2018).

[447] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is more: Sampling chemical space with active learning," J. Chem. Phys. **148**, 241733 (2018).

[448] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," Nat. Commun. **8** (2017), 10.1038/ncomms13890.

[449] A. Khorshidi and A. A. Peterson, "Amp : A modular approach to machine learning in atomistic simulations," Comput. Phys. Commun. **207**, 310–324 (2016).

[450] V. L. Deringer, N. Bernstein, A. P. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, and G. Csányi, "Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics," J. Phys. Chem. Lett. **9**, 2879–2885 (2018).

[451] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," IEEE Transactions on Evolutionary Computation **6**, 182–197 (2002).

[452] "esa/pagmo2: pagmo 2.11.4," `http://doi.org/10.5281/zenodo.3464510` (accessed in 2019).

[453] D. J. Montana and L. Davis, "Training feedforward neural networks using genetic algorithms," in *Proceedings of the 11th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'89 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1989) pp. 762–767.

[454] J. Branke, "Evolutionary algorithms for neural network design and training," in *In Proceedings of the first Nordic workshop on genetic algorithms and its applications* (1995) pp. 145–163.

[455] D. Thierens, J. Suykens, J. Vandewalle, and B. De Moor, "Genetic weight optimization of a feedforward neural network controller," in *Artificial Neural Nets and Genetic Algorithms*, edited by R. F. Albrecht, C. R. Reeves, and N. C. Steele (Springer Vienna, Vienna, 1993) pp. 658–663.

[456] H. Kitano, "Empirical studies on the speed of convergence of neural network training using genetic algorithms," in *AAAI-90 Proccedings* (1990) pp. 789–795.

[457] B. Yoon, D. J. Holmes, G. Langholz, and A. Kandel, "Efficient genetic algorithms for training layered feedforward neural networks," Information Sciences **76**, 67 – 85 (1994).

[458] H. Kitano, "Designing neural networks using genetic algorithms with graph generation system," Complex Syst. **4**, 461–476 (1990).

[459] B.-T. Zhang and H. Muhlenbein, "Evolving optimal neural networks using genetic algorithms with occam's razor," Complex Syst. **7**, 199–220 (1993).

[460] H. Kitano, "Neurogenetic learning: an integrated method of designing and training neural networks using genetic algorithms," Physica D **75**, 225 – 238 (1994).

[461] J. D. Schaffer, D. Whitley, and L. J. Eshelman, "Combinations of genetic algorithms and neural networks: a survey of the state of the art," in *[Proceedings] COGANN-92: International Workshop on Combinations of Genetic Algorithms and Neural Networks* (1992) pp. 1–37.

[462] D. E. Goldberg and J. H. Holland, “Genetic algorithms and machine learning,” Machine Learning **3**, 95–99 (1988).

[463] J. N. Gupta and R. S. Sexton, “Comparing backpropagation with a genetic algorithm for neural network training,” Omega **27**, 679 – 684 (1999).

[464] M. N. H. Siddique and M. O. Tokhi, “Training neural networks: backpropagation vs. genetic algorithms,” in *IJCNN’01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, Vol. 4 (2001) pp. 2673–2678 vol.4.

[465] G. F. Miller, P. M. Todd, and S. U. Hegde, “Designing neural networks using genetic algorithms,” in *Proceedings of the Third International Conference on Genetic Algorithms* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1989) pp. 379–384.

[466] O. E. David and I. Greental, “Genetic algorithms for evolving deep neural networks,” Proceedings of the 2014 conference companion on Genetic and evolutionary computation companion - GECCO Comp ’14 (2014), 10.1145/2598394.2602287.

[467] F. P. Such, V. Madhavan, E. Conti, J. Lehman, K. O. Stanley, and J. Clune, “Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning,” arXiv:1712.06567 (2017).

[468] Y. Sun, B. Xue, M. Zhang, and G. G. Yen, “Automatically designing CNN architectures using genetic algorithm for image classification,” arXiv:1808.03818 (2018).

[469] H. Chung and K.-S. Shin, “Genetic algorithm-optimized multi-channel convolutional neural network for stock market prediction,” Neural Comput. and Applic. (2019), 10.1007/s00521-019-04236-3.

[470] “NSGA-II-python,” `https://code.google.com/archive/p/nsga-ii-python/` (accessed in 2019).

[471] L. Guimarães, A. N. Enyashin, J. Frenzel, T. Heine, H. A. Duarte, and G. Seifert, “Imogolite nanotubes: Stability, electronic, and mechanical properties,” ACS Nano **1**, 362–368 (2007), pMID: 19206688.

[472] A. Sieck, T. Frauenheim, and K. A. Jackson, “Shape transition of medium-sized neutral silicon clusters,” Phys. Status Solidi B **240**, 537–548 (2003).

[473] J. Behler, R. Martoňák, D. Donadio, and M. Parrinello, “Pressure-induced phase transitions in silicon studied by neural network-based metadynamics simulations,” Phys. Status Solidi B **245**, 2618–2629 (2008).

[474] G. C. Sosso, G. Miceli, S. Caravati, J. Behler, and M. Bernasconi, “Neural network interatomic potential for the phase change material GeTe,” Phys. Rev. B **85**, 174103 (2012).

[475] A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, “Machine learning a general-purpose interatomic potential for silicon,” Phys. Rev. X **8**, 041048 (2018).

[476] A. W. Huran, C. Steigemann, T. Frauenheim, B. Aradi, and M. A. L. Marques, “Efficient automatized density-functional tight-binding parametrizations: Application to group IV elements,” J. Chem. Theory Comput. **14**, 2947–2954 (2018).

[477] G. Kresse and J. Furthmüller, “Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set,” Phys. Rev. B **54**, 11169–11186 (1996).

[478] G. Kresse and J. Furthmüller, “Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set,” Comput. Mater. Sci. **6**, 15 – 50 (1996).

[479] P. Borlido, C. Steigemann, N. N. Lathiotakis, M. A. L. Marques, and S. Botti, “Structural prediction of two-dimensional materials under strain,” 2D Mater. **4**, 045009 (2017).

[480] P. Borlido, C. Rödl, M. A. L. Marques, and S. Botti, “The ground state of two-dimensional silicon,” 2D Mater. **5**, 035010 (2018).

[481] B. Adams, L. Bauman, W. Bohnhoff, K. Dalbey, M. Ebeida, J. Eddy, M. Eldred, P. Hough, K. Hu, J. Jakeman, J. Stephens, L. Swiler, D. Vigil, and T. Wildey, *Dakota, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 6.0 User's Manual* (Sandia Technical Report SAND2014-4633,(Version 6.3), 2015).

[482] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," Quart. Appl. Math. **2**, 164–168 (1944).

[483] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," SIAM J. Appl. Math. **11**, 431–441 (1963).

[484] R. H. Byrd and all, "A limited memory algorithm for bound constrained optimization," SIAM J. Sci. Stat. Comp. **16**, 1190–1208 (1995).

[485] Z. Jian, Z. Kaiming, and X. Xide, "Modification of Stillinger-Weber potentials for Si and Ge," Phys. Rev. B **41**, 12915–12918 (1990).

[486] J. Tersoff, "New empirical approach for the structure and energy of covalent systems," Phys. Rev. B **37**, 6991–7000 (1988).

[487] J. Tersoff, "Modeling solid-state chemistry: Interatomic potentials for multicomponent systems," Phys. Rev. B **39**, 5566–5568 (1989).

[488] S. Plimpton, "Fast parallel algorithms for short-range molecular dynamics," J. Comput. Phys. **117**, 1 – 19 (1995).

[489] B. Aradi, B. Hourahine, and T. Frauenheim, "DFTB+, a sparse matrix-based implementation of the DFTB method," J. Phys. Chem. A **111**, 5678–5684 (2007), pMID: 17567110.

[490] J. E. Jones, "On the determination of molecular fields. II. from the equation of state of a gas," Proc. Royal Soc. A **106**, 463–477 (1924).

[491] A. Togo and I. Tanaka, "First principles phonon calculations in materials science," Scr. Mater. **108**, 1–5 (2015).

[492] S. Faraji, S. A. Ghasemi, S. Rostami, R. Rasoulkhani, B. Schaefer, S. Goedecker, and M. Amsler, "High accuracy and transferability of a neural network potential through charge equilibration for calcium fluoride," Phys. Rev. B **95**, 104105 (2017).

[493] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, "The atomic simulation environment—a python library for working with atoms," J. Phys.: Condens. Matter **29**, 273002 (2017).

[494] J. M. Ziman, *Principles of the Theory of Solids*, 2nd ed. (Cambridge University Press, 1972).

[495] V. Y. Chekhovskoi, V. D. Tarasov, and Y. V. Gusev, "Calorific properties of liquid copper," High Temp. **38**, 394–399 (2000).

[496] L. Vočadlo, D. Alfè, G. D. Price, and M. J. Gillan, "Ab initio melting curve of copper by the phase coexistence approach," J. Chem. Phys. **120**, 2872–2878 (2004).

[497] Y. Wang and J. P. Perdew, "Correlation hole of the spin-polarized electron gas, with exact small-wave-vector and high-density scaling," Phys. Rev. B **44**, 13298–13307 (1991).

[498] J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais, "Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation," Phys. Rev. B **46**, 6671–6687 (1992).

[499] L. J. Lewis, P. Jensen, and J.-L. Barrat, "Melting, freezing, and coalescence of gold nanoclusters," Phys. Rev. B **56**, 2248–2257 (1997).

[500] T. T. Järvi, A. Kuronen, M. Hakala, K. Nordlund, A. C. van Duin, W. A. Goddard, and T. Jacob, "Development of a reaxff description for gold," Eur. Phys. J. B **66**, 75–79 (2008).

[501] M. G. Ganchenkova, I. A. Supryadkina, K. K. Abgaryan, D. I. Bazhanov, I. V. Mutigullin, and V. A. Borodin, "Influence of the ab-initio calculation parameters on prediction of energy of point defects in silicon," Mod. Electron. Mater. **1**, 103 – 108 (2015).

[502] S. Goedecker, T. Deutsch, and L. Billard, "A fourfold coordinated point defect in silicon," Phys. Rev. Lett. **88**, 235501 (2002).

[503] N. Ashcroft and N. Mermin, *Solid State Physics* (Brooks/Cole, 1976).

[504] S. A. Ghasemi, M. Amsler, R. G. Hennig, S. Roy, S. Goedecker, T. J. Lenosky, C. J. Umrigar, L. Genovese, T. Morishita, and K. Nishio, "Energy landscape of silicon systems and its description by force fields, tight binding schemes, density functional methods, and quantum monte carlo methods," Phys. Rev. B **81**, 214107 (2010).

[505] Z. C. Lipton, "The mythos of model interpretability," Queue **16**, 30:31–30:57 (2018).

[506] B. Rister and D. L. Rubin, "Piecewise convexity of artificial neural networks," Neural Networks **94**, 34 – 45 (2017).

[507] A. van de Walle, M. D. Asta, and G. Ceder, "The Alloy Theoretic Automated Toolkit: A user guide," Calphad **26**, 539–553 (2002).

[508] A. van de Walle, "Multicomponent multisublattice alloys, nonconfigurational entropy and other additions to the Alloy Theoretic Automated Toolkit," Calphad **33**, 266–278 (2009).

[509] J. Sanchez, F. Ducastelle, and D. Gratias, "Generalized cluster description of multicomponent systems," Physica A **128**, 334–350 (1984).

[510] A. van de Walle and G. Ceder, "Automating first-principles phase diagram calculations," J. Phase Equilib. **23**, 348–359 (2002).

[511] J. W. D. Connolly and A. R. Williams, "Density-functional theory applied to phase transformations in transition-metal alloys," Phys. Rev. B **27**, 5169–5172 (1983).

[512] "Normal equation," `http://mathworld.wolfram.com/NormalEquation.html` (accessed in 2019).

[513] R. Kikuchi, "A theory of cooperative phenomena," Phys. Rev. **81**, 988–1003 (1951).

[514] F. Ducastelle, "Order and phase stability in alloys," Interatomic Potential and Structural Stability , 133–142 (1993).

[515] K. Binder and D. W. Heermann, *Monte Carlo Simulation in Statistical Physics* (Springer Berlin Heidelberg, 2010).

[516] A. Kohan, P. Tepesch, G. Ceder, and C. Wolverton, "Computation of alloy phase diagrams at low temperatures," Comput. Mater. Sci. **9**, 389 – 396 (1998).

[517] L. C. Andreani, A. Bozzola, P. Kowalczewski, M. Liscidini, and L. Redorici, "Silicon solar cells: toward the efficiency limits," Adv. Phys. X **4**, 1548305 (2018).

[518] W. Shockley and H. J. Queisser, "Detailed balance limit of efficiency of p-n junction solar cells," J. Appl. Phys. **32**, 510–519 (1961).

[519] T. Tiedje, E. Yablonovitch, G. Cody, and B. Brooks, "Limiting efficiency of silicon solar cells," IEEE Transactions on Electron Devices **31**, 711–716 (1984).

[520] A. Polman, M. Knight, E. C. Garnett, B. Ehrler, and W. C. Sinke, "Photovoltaic materials: Present efficiencies and future challenges," Science **352**, aad4424 (2016).

[521] R. King, "Multijunction cells: Record breakers," Nat. Photonics **2**, 284–286 (2008).

[522] M. Steiner, G. Siefer, T. Schmidt, M. Wiesenfarth, F. Dimroth, and A. W. Bett, "43% sunlight to electricity conversion efficiency using CPV," IEEE J. Photovolt. **6**, 1020–1024 (2016).

[523] M. A. Green, Y. Hishikawa, E. D. Dunlop, D. H. Levi, J. Hohl-Ebinger, and A. W. Ho-Baillie, "Solar cell efficiency tables (version 51)," Prog Photovolt. **26**, 3–12 (2018).

[524] M. Green, *Third Generation Photovoltaics Advanced Solar Energy Conversion* (Springer-Verlag Berlin Heidelberg, 2006).

[525] M. P. Lumb, S. Mack, K. J. Schmieder, M. González, M. F. Bennett, D. Scheiman, M. Meitl, B. Fisher, S. Burroughs, K.-T. Lee, and et al., "GaSb-based solar cells for full solar spectrum energy harvesting," Adv. Energy Mater. **7**, 1700345 (2017).

[526] J. Praveen and V. VijayaRamaraju, "Materials for optimizing efficiencies of solar photovoltaic panels," Mater. Today: Proc. **4**, 5233 – 5238 (2017), 6th International Conference of Materials Processing and Characterization (ICMPC 2016), 5-7 December 2016.

[527] J. Zhao, A. Wang, and M. A. Green, "24·5% efficiency silicon pert cells on mcz substrates and 24·7% efficiency perl cells on fz substrates," Prog Photovolt. **7**, 471–474 (1999).

[528] M. A. Green, "The path to 25% silicon solar cell efficiency: History of silicon cell evolution," Prog Photovolt. **17**, 183–189 (2009).

[529] R. Moskalyk, "Gallium: the backbone of the electronics industry," Miner. Eng. **16**, 921 – 929 (2003).

[530] M. I. Alonso, K. Wakita, J. Pascual, M. Garriga, and N. Yamamoto, "Optical functions and electronic structure of $CuInSe_2$, $CuGaSe_2$, $CuInS_2$, and $CuGaS_2$," Phys. Rev. B **63**, 075203 (2001).

[531] "Light management in new photovoltaic materials," `http://www.lmpv.nl/sq/` (accessed in 2019).

[532] A. Zakutayev, "Brief review of emerging photovoltaic absorbers," Curr. Opin. Green Sustain. Chem. **4**, 8 – 15 (2017).

[533] K. Ito and T. Nakazawa, "Electrical and optical properties of stannite-type quaternary semiconductor thin films," Jpn. J. Appl. Phys. **27**, 2094–2097 (1988).

[534] A. Walsh, S. Chen, S.-H. Wei, and X.-G. Gong, "Kesterite thin-film solar cells: Advances in materials modelling of Cu_2ZnSnS_4," Adv. Energy Mater. **2**, 400–409 (2012).

[535] M. Ravindiran and C. Praveenkumar, "Status review and the future prospects of CZTS based solar cell – a novel approach on the device structure and material modeling for CZTS based photovoltaic device," Renew. Sustain. Energy Rev. **94**, 317 – 329 (2018).

[536] F. A. Jhuma, M. Z. Shaily, and M. J. Rashid, "Towards high-efficiency czts solar cell through buffer layer optimization," Mater. Renew. Sustain. Energy **8**, 6 (2019).

[537] S. Chen, X. G. Gong, A. Walsh, and S.-H. Wei, "Electronic structure and stability of quaternary chalcogenide semiconductors derived from cation cross-substitution of ii-vi and i-iii-vi_2 compounds," Phys. Rev. B **79**, 165211 (2009).

[538] L. Shi, P. Yin, H. Zhu, and Q. Li, "Synthesis and photoelectric properties of Cu_2ZnGeS_4 and $Cu_2ZnGeSe_4$ single-crystalline nanowire arrays," Langmuir **29**, 8713–8717 (2013).

[539] S. Körbel, D. Kammerlander, R. Sarmiento-Pérez, C. Attaccalite, M. A. L. Marques, and S. Botti, "Optical properties of Cu-chalcogenide photovoltaic absorbers from self-consistent *GW* and the bethe-salpeter equation," Phys. Rev. B **91**, 075134 (2015).

[540] T. Schwarz, M. A. L. Marques, S. Botti, M. Mousel, A. Redinger, S. Siebentritt, O. Cojocaru-Mirédin, D. Raabe, and P.-P. Choi, "Detection of $Cu_2Zn_5SnSe_8$ and $Cu_2Zn_6SnSe_9$ phases in co-evaporated $Cu_2ZnSnSe_4$ thin-films," Appl. Phys. Lett. **107**, 172102 (2015).

[541] T. Maeda, S. Nakamura, and T. Wada, "First-principles calculations of vacancy formation in In-free photovoltaic semiconductor $Cu_2ZnSnSe_4$," Thin Solid Films **519**, 7513 – 7516 (2011), proceedings of the EMRS 2010 Spring Meeting Symposium M: Thin Film Chalcogenide Photovoltaic Materials.

[542] I. Dudchak and L. Piskach, "Phase equilibria in the Cu_2SnSe_3–$SnSe_2$–ZnSe system," J. Alloys Compd. **351**, 145 – 150 (2003).

[543] I. Olekseyuk, I. Dudchak, and L. Piskach, "Phase equilibria in the Cu_2S–ZnS–SnS_2 system," J. Alloys Compd. **368**, 135 – 143 (2004).

[544] J. J. Scragg, T. Ericson, T. Kubart, M. Edoff, and C. Platzer-Björkman, "Chemical insights into the instability of $Cu_2ZnSnSe_4$ films during annealing," Chem. Mater. **23**, 4625–4633 (2011).

[545] J. J. Scragg, P. J. Dale, D. Colombara, and L. M. Peter, "Thermodynamic aspects of the synthesis of thin-film materials for solar cells," ChemPhysChem **13**, 3035–3046 (2012).

[546] A. Redinger and S. Siebentritt, "Coevaporation of cu2znsnse4 thin films," Appl. Phys. Lett. **97**, 092111 (2010).

[547] A. Redinger, D. M. Berg, P. J. Dale, and S. Siebentritt, "The consequences of kesterite equilibria for efficient solar cells," J. Am. Chem. Soc. **133**, 3320–3323 (2011).

[548] S. Chen, X. G. Gong, A. Walsh, and S.-H. Wei, "Defect physics of the kesterite thin-film solar cell absorber Cu_2ZnSnS_4," Appl. Phys. Lett. **96**, 021902 (2010).

[549] S. Chen, J.-H. Yang, X. G. Gong, A. Walsh, and S.-H. Wei, "Intrinsic point defects and complexes in the quaternary kesterite semiconductor $Cu_2ZnSnSe_4$," Phys. Rev. B **81**, 245204 (2010).

[550] S. B. Zhang, S.-H. Wei, A. Zunger, and H. Katayama-Yoshida, "Defect physics of the $CuInSe_2$ chalcopyrite semiconductor," Phys. Rev. B **57**, 9642–9656 (1998).

[551] S. Siebentritt, M. Igalson, C. Persson, and S. Lany, "The electronic structure of chalcopyrites—bands, point defects and grain boundaries," Prog. Photovolt: Res. Appl. **18**, 390–410 (2010).

[552] T. K. Todorov, K. B. Reuter, and D. B. Mitzi, "High-efficiency solar cell with earth-abundant liquid-processed absorber," Adv. Mater. **22**, E156–E159 (2010).

[553] D. J. Chakrabarti and D. E. Laughlin, "The Cu-S (Copper-Sulfur) system," Bull. Alloy Phase Diagrams **4**, 254 (1983).

[554] V. M. Glazov, A. S. Pashinkin, and V. A. Fedorov, "Phase equilibria in the Cu-Se system," Inorg. Mater. **36**, 641–652 (2000).

[555] V. Blum, G. L. W. Hart, M. J. Walorski, and A. Zunger, "Using genetic algorithms to map first-principles results to model Hamiltonians: Application to the generalized Ising model for alloys," Phys. Rev. B **72**, 165113 (2005).

[556] M. d'Avezac and A. Zunger, "Identifying the minimum-energy atomic configuration on a lattice: Lamarckian twist on Darwinian evolution," Phys. Rev. B **78**, 064102 (2008).

[557] T. Shibuya, Y. Goto, Y. Kamihara, M. Matoba, K. Yasuoka, L. A. Burton, and A. Walsh, "From kesterite to stannite photovoltaics: Stability and band gaps of the Cu_2ZnSnS_4 alloy," Appl. Phys. Lett. **104**, 021912 (2014).

[558] T. Minami, "New n-type transparent conducting oxides," MRS Bull. **25**, 38–44 (2000).

[559] R. A. Afre, N. Sharma, M. Sharon, and M. Sharon, "Transparent conducting oxide films for various applications: A review," Rev. Adv. Mater. Sci **53**, 79–89 (2018).

[560] H. Sato, T. Minami, S. Takata, and T. Yamada, "Transparent conducting p-type NiO thin films prepared by magnetron sputtering," Thin Solid Films **236**, 27 – 31 (1993).

[561] H. Kawazoe, M. Yasukawa, H. Hyodo, M. Kurita, H. Yanagi, and H. Hosono, "P-type electrical conduction in transparent thin films of $CuAlO_2$," Nature **389**, 939–942 (1997).

[562] R. Nagarajan, A. D. Draeseke, A. W. Sleight, and J. Tate, "P-type conductivity in $CuCr_{1-x}Mg_xO_2$ films and powders," J. Appl. Phys. **89**, 8022–8025 (2001).

[563] K. H. L. Zhang, Y. Du, A. Papadogianni, O. Bierwagen, S. Sallis, L. F. J. Piper, M. E. Bowden, V. Shutthanandan, P. V. Sushko, and S. A. Chambers, "Perovskite Sr-doped $LaCrO_3$ as a new p-type transparent conducting oxide," Adv. Mater. **27**, 5191–5195 (2015).

[564] K. Baedeker, "Über die elektrische leitfähigkeit und die thermoelektrische kraft einiger schwermetallverbindungen," Annalen der Physik **327**, 749–766 (1907).

[565] M. Grundmann, F.-L. Schein, M. Lorenz, T. Böntgen, J. Lenzner, and H. von Wenckstern, "Cuprous iodide – a p-type transparent semiconductor: history and novel applications," Phys. Status Solidi A **210**, 1671–1703 (2013).

[566] C. Yang, M. Kneiß, M. Lorenz, and M. Grundmann, "Room-temperature synthesized copper iodide thin film as degenerate p-type transparent conductor with a boosted figure of merit," Proc. Natl. Acad. Sci. U.S.A. **113**, 12929–12933 (2016).

[567] J. Wang, J. Li, and S.-S. Li, "Native p-type transparent conductive CuI via intrinsic defects," J. Appl. Phys. **110**, 054907 (2011).

[568] M. Graužinytė, S. Botti, M. A. L. Marques, S. Goedecker, and J. A. Flores-Livas, "Computational acceleration of prospective dopant discovery in cuprous iodide," Phys. Chem. Chem. Phys. , Advance Article, DOI:10.1039/C9CP02711D (2019).

[569] Y. Kokubun, H. Watanabe, and M. Wada, "Electrical properties of CuI thin films," Jpn. J. Appl. Phys. **10**, 864–867 (1971).

[570] R. J. Maurer, "Deviations from stoichiometric proportions in cuprous iodide," J. Chem. Phys. **13**, 321–326 (1945).

[571] L. L. Baranowski, P. Zawadzki, S. Lany, E. S. Toberer, and A. Zakutayev, "A review of defects and disorder in multinary tetrahedrally bonded semiconductors," Semicond. Sci. Technol. **31**, 123004 (2016).

[572] J.-F. Guillemoles, L. Kronik, D. Cahen, U. Rau, A. Jasenek, and H.-W. Schock, "Stability issues of Cu(In,Ga)Se_2-based solar cells," J. Phys. Chem. B **104**, 4849–4862 (2000).

[573] H. Mönig, C.-H. Fischer, A. Grimm, B. Johnson, C. A. Kaufmann, R. Caballero, I. Lauermann, and M. C. Lux-Steiner, "Surface Cu-depletion of Cu(In,Ga)Se_2 thin films: Further experimental evidence for a defect-induced surface reconstruction," J. Appl. Phys. **107**, 113540 (2010).

[574] D. Schmid, M. Ruckh, F. Grunwald, and H. W. Schock, "Chalcopyrite/defect chalcopyrite heterojunctions on the basis of $CuInSe_2$," J. Appl. Phys. **73**, 2902–2909 (1993).

[575] H. Zhao, M. Kumar, and C. Persson, "Density functional theory study of ordered defect Cu-(In,Ga)-Se compounds," Phys. Status Solidi C **9**, 1600–1603 (2012).

[576] S. B. Zhang, S.-H. Wei, and A. Zunger, "Stabilization of ternary compounds via ordered arrays of defect pairs," Phys. Rev. Lett. **78**, 4059–4062 (1997).

[577] N. Yamada, R. Ino, and Y. Ninomiya, "Truly transparent p-type γ-cui thin films with high hole mobility," Chem. Mater. **28**, 4971–4981 (2016).

[578] B. Hönerlage, C. Klingshirn, and J. Grun, "Spontaneous emission due to exciton—electron scattering in semiconductors," Phys. Status Solidi (b) **78**, 599–608 (1976).

[579] D. Huang, Y.-J. Zhao, S. Li, C.-S. Li, J.-J. Nie, X.-H. Cai, and C.-M. Yao, "First-principles study of γ-CuI for p-type transparent conducting materials," J. Phys. D **45**, 145102 (2012).

[580] O. Gogolin, J. Deiss, and E. Tsitsichvili, "The piezobirefringence in copper halides," Il Nuovo Cimento D **11**, 1525–1534 (1989).

[581] T. Chaudhuri, P. Basu, A. Patra, R. Saraswat, and H. Acharya, "A chemical method for preparing copper iodide thin films," Jpn. J. Appl. Phys. **29**, L352 (1990).

[582] D. Chen, Y. Wang, Z. Lin, J. Huang, X. Chen, D. Pan, and F. Huang, "Growth strategy

and physical properties of the high mobility p-type cui crystal," Cryst. Growth Des. **10**, 2057–2060 (2010).

[583] S. Ves, D. Glötzel, M. Cardona, and H. Overhof, "Pressure dependence of the optical properties and the band structure of the copper and silver halides," Phys. Rev. B **24**, 3073 (1981).

[584] A. Gruzintsev and W. Zagorodnev, "Temperature-dependent conductivity and photoconductivity of p-cui crystals," Semiconductors **46**, 35–40 (2012).

[585] S. Miyake, S. Hoshino, and T. Takenaka, "On the phase transition in cuprous iodide," J. Phys. Soc. Jpn. **7**, 19–24 (1952).

[586] R. Kurdyumova and R. Baranova, "An electron diffraction study of thin films of cuprous iodide," Kristallografiya **6**, 402–405 (1961).

[587] T. Sakuma, "Crystal structure of β-CuI," J. Phys. Soc. Jpn. **57**, 565–569 (1988).

[588] P. Villars and K. Cenzual, eds., *Landolt-Börnstein - Group III Condensed Matter: Crystal Structures of Inorganic Compounds* (Springer-Verlag Berlin Heidelberg, Heidelberg, 2009).

[589] I. K. Akopyan, V. V. Golubkov, O. A. Dyatlova, A. N. Mamaev, B. V. Novikov, and A. N. Tsagan-Mandzhiev, "Specific features of the CuI nanocrystal structure in photochromic glasses," Phys. Solid State **52**, 805–809 (2010).

[590] M. N. Amalina and M. Rusop, "The properties of p-type copper (I) iodide (CuI) as a hole conductor for solid-state dye sensitized solar cells," in *RSM 2013 IEEE Regional Symposium on Micro and Nanoelectronics* (2013) pp. 300–303.

[591] J. A. Christians, R. C. M. Fung, and P. V. Kamat, "An inorganic hole conductor for organo-lead halide perovskite solar cells. improved hole conductivity with copper iodide," J. Am. Chem. Soc. **136**, 758–764 (2014).

[592] G. A. Sepalage, S. Meyer, A. Pascoe, A. D. Scully, F. Huang, U. Bach, Y.-B. Cheng, and L. Spiccia, "Copper(I) iodide as hole-conductor in planar perovskite solar cells: Probing the origin of J–V hysteresis," Adv. Funct. Mater. **25**, 5650–5661 (2015).

[593] Z. Yu and L. Sun, "Recent progress on hole-transporting materials for emerging organometal halide perovskite solar cells," Adv. Energy Mater. **5**, 1500213 (2015).

[594] S. A. Mohamed, J. Gasiorowski, K. Hingerl, D. R. Zahn, M. C. Scharber, S. S. Obayya, M. K. El-Mansy, N. S. Sariciftci, D. A. Egbe, and P. Stadler, "CuI as versatile hole-selective contact for organic solar cell based on anthracene-containing PPE–PPV," Sol. Energy Mater. Sol. Cells **143**, 369 – 374 (2015).

[595] M. Shan, H. Jiang, Y. Guan, D. Sun, Y. Wang, J. Hua, and J. Wang, "Enhanced hole injection in organic light-emitting diodes utilizing a copper iodide-doped hole injection layer," RSC Adv. **7**, 13584–13589 (2017).

[596] C. Yang, D. Souchay, M. Kneiß, M. Bogner, H. M. Wei, M. Lorenz, O. Oeckler, G. Benstetter, Y. Q. Fu, and M. Grundmann, "Transparent flexible thermoelectric material based on non-toxic earth-abundant p-type copper iodide thin film," Nat. Commun. **8**, 16076 (2017).

[597] K. G. Godinho, J. J. Carey, B. J. Morgan, D. O. Scanlon, and G. W. Watson, "Understanding conductivity in $SrCu_2O_2$: stability, geometry and electronic structure of intrinsic defects from first principles," J. Mater. Chem. **20**, 1086–1096 (2010).

[598] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, "A high-throughput infrastructure for density functional theory calculations," Comput. Mater. Sci. **50**, 2295 – 2310 (2011).

[599] R. Sarmiento-Pérez, S. Botti, and M. A. L. Marques, "Optimized exchange and correlation semilocal functional for the calculation of energies of formation," J. Chem. Theory Comput. **11**, 3844–3850 (2015).

[600] F. Tran, J. Stelzl, and P. Blaha, "Rungs 1 to 4 of dft jacob's ladder: Extensive test on the lattice constant, bulk modulus, and cohesive energy of solids," J. Chem. Phys. **144**, 204120 (2016).

[601] V. Stevanović, S. Lany, X. Zhang, and A. Zunger, "Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies," Phys. Rev. B **85**, 115104 (2012).

[602] J. Sun, A. Ruzsinszky, and J. P. Perdew, "Strongly constrained and appropriately normed semilocal density functional," Phys. Rev. Lett. **115**, 036402 (2015).

[603] E. B. Isaacs and C. Wolverton, "Performance of the strongly constrained and appropriately normed density functional for solid-state materials," Phys. Rev. Materials **2**, 063801 (2018).

[604] Y. Zhang, D. A. Kitchaev, J. Yang, T. Chen, S. T. Dacek, R. A. Sarmiento-Pérez, M. A. Marques, H. Peng, G. Ceder, J. P. Perdew, and J. Sun, "Efficient first-principles prediction of solid stability: Towards chemical accuracy," npj Comput. Mater. **4**, 9 (2018).

[605] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).

[606] K. Momma and F. Izumi, "*VESTA3* for three-dimensional visualization of crystal, volumetric and morphology data," J. Appl. Crystallogr. **44**, 1272–1276 (2011).

[607] C. Adamo and V. Barone, "Toward reliable density functional methods without adjustable parameters: The PBE0 model," J. Chem. Phys. **110**, 6158–6170 (1999).

[608] M. Ernzerhof and G. E. Scuseria, "Assessment of the Perdew–Burke–Ernzerhof exchange-correlation functional," J. Chem. Phys. **110**, 5029–5036 (1999).

[609] S. Koyasu, N. Umezawa, A. Yamaguchi, and M. Miyauchi, "Optical properties of single crystalline copper iodide with native defects: Experimental and density functional theoretical investigation," J. Appl. Phys. **125**, 115101 (2019).

List of Publications

1. M. R. G. Marques, J. Wolff, C. Steigemann, and M. A. L. Marques, “Neural network force fields for simple metals and semiconductors: construction and application to the calculation of phonons and melting temperatures,” Phys. Chem. Chem. Phys. 21, 6506–6516 (2019).

2. J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, “Recent advances and applications of machine learning in solid-state materials science,” npj Comput. Mater. 5, 83 (2019).

3. S. Jaschik, M. R. G. Marques, M. Seifert, C. Röodl, S. Botti, and M. A. L. Marques, “Stable ordered phases of cuprous iodide with complexes of copper vacancies,” Chem. Mater. 31, 7877–7882 (2019).

Acknowledgements

At the end of this arduous quest, I would like to acknowledge all those that contributed and supported me while writing this thesis.

First of all, I would like to express my gratitude to Prof. Dr. Miguel Marques, my supervisor, for this fantastic opportunity. None of this would be possible if not for him. He not only taught me about physics, but also music, responsibility, and even cooking! Moreover he always had time (and patience) for my foolish questions. He constantly had some scientific problem to peek our curiosity and motivate us in this structural prediction quest.

Furthermore, I would like to thank everyone in the group here in Halle: Conrad, Carlos, Katja, Jonathan, Ahmad, Matheus, Haichen, Tomás, Thorsten, Tugce, Fernanda, Stefan, and Jakob. We had a fantastic, friendly working environment on all the occasions! We did amazing work and I will remember these years fondly.

I would like thank Prof. Dr. Silvana Botti and her group in Jena: Borlido, Claudia, Sun Lin, Michael, Sofia, and Tiago. We had constructive discussions and a lot of fun while working together.

I would like thank Prof. Dr. Fernando Nogueira and Dr. Micael Oliveira, for all the help and for the introduction to electronic structure methods. They put me on this path.

I would like thank Prof. Dr. Jamal Berakdar and his group in Halle: Jonas (my gym master), Michael, Stefan, Alex, Dominik, Mikheil, Levan, Anna, and Carlos. Our discussions were very productive and will for sure help me in the upcoming defense!

I would like to thank my parents, my mother's stories developed my imagination and my father's page skipping my wit. Moreover, I would like to thank my favourite brother.

I would like to thank all my friends in Portugal and Germany, in particular Tiago, Ana, Rafael, Pedro, João, and Vandeta.

Finally, I would like to thank Vika for her support and love.

Lebenslauf / Curriculum Vitae

Persönliche Daten / Personal data

Name	Mário Rui Gonçalves Marques
Geburtsdatum / Date of birth	24.10.1992
Geburstort / Place of birth	Coimbra, Portugal
Staatsangehörigkeit / Nationality	Portugiesisch / Portuguese
Adresse / Address	Heidealle 4, 06120 Halle (Saale), Germany
E-mail	mario.goncalves-marques@physik.uni-halle.de

Bildung / Education

10.2015 - Present	PhD in Physics Martin-Luther University Halle-Wittenberg, Germany Supervisor: Miguel A. L. Marques
09.2013 - 07.2015	Master's degree in Physics specialization in Computational Modelling and Simulation University of Coimbra, Portugal Master thesis: "Optical and Magnetical Properties of Endohedral Silicon Cages" Supervisors: Fernando M. S. Nogueira, Micael J. T. Oliveira
09.2010 - 07.2013	Bachelor's degree in Physics University of Coimbra, Portugal
09.2007 - 07.2010	High School Escola Secundária José Falcão, Coimbra, Portugal

Date, Place

Mário Rui Gonçalves Marques

Eidesstattliche Erklärung / Statutory statement

Hiermit erkläre ich, gemäß §5 der Promotionsordnung der Naturwissenschaftlichen Fakultät II - Chemie, Physik und Mathematik der Martin-Luther-Universität Halle-Wittenberg vom 13.06.2012, dass die vorliegende Arbeit

"The structure and dynamics of materials using machine learning"

selbständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Ich erkläre, die Angaben wahrheitsgemäß gemacht und keine Dissertation an einer anderen wissenschaftlichen Einrichtung zur Erlangung eines akademischen Grades eingereicht zu habe

Date, Place	Mário Rui Gonçalves Marques