# Artificial Intelligence for Crystal Structure Prediction

vorgelegt von
M. Sc.
Emre Ahmetcik

an der Fakultät II - Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
Dr. rer. nat.
genehmigte Dissertation

Promotionsausschuss:

| | |
|---|---|
| Vorsitzender: | Prof. Dr. Wolf-Christian Müller, Technische Universität Berlin |
| Gutachter: | Prof. Dr. Andreas Knorr, Technische Universität Berlin |
| Gutachter: | Prof. Dr. Matthias Scheffler, Fritz-Haber-Institut der Max-Planck-Gesellschaft |
| Gutachter: | Dr. Matthias Rupp, Universität Konstanz |

Tag der wissenschaftlichen Aussprache: 4. Mai 2022

Berlin 2022

# Contents

# Abstract

Predicting the ground-state and metastable crystal structures of materials from just knowing their composition is a formidable challenge in computational materials discovery. Recent studies that were published in the group of M. Scheffler have investigated how the relative stability of compounds between two crystal-structure types can be predicted from the properties of their atomic constituents within the framework of symbolic regression. By using a novel compressed-sensing-based method, the sure independence screening and sparsifying operator (SISSO), the descriptor that best captured the structural stability was identified from billions of candidates. A descriptor is a vector of analytical formulas built from simple physical quantities.

In the first part of the thesis, a multi-task-learning extension of SISSO (MT-SISSO) that enables the treatment of the structural stability of compounds among multiple structure types is introduced. We show how the multi-task method that identifies a single descriptor for all structure types enables the prediction of a well-defined structural stability and, therefore, the design of a crystal-structure map. Moreover, we present how MT-SISSO determines accurate, predictive models even when trained with largely incomplete databases.

A different artificial-intelligence approach proposed for tackling the crystal-structure-prediction challenge is based on approximating the Born-Oppenheimer potential-energy surface (PES). In particular, Gaussian Approximation Potentials that are typically composed of a combination of two-, three-, and many-body potentials and fitted to elemental systems have attracted attention in recent years. First examples that were published in the group of G. Csanyi have demonstrated how the ground-state and metastable phases could correctly be identified for Si, C, P, and B, by exploring the PES that was predicted by such *machine-learning potentials* (ML potentials). However, the ML potentials introduced so far show limited transferability, i.e. their accuracy rapidly decreases in regions of the PES that are distant from the training data. As a consequence, these ML potentials are usually fitted to large training databases. Moreover, such training data needs to be constructed for every new material (more precisely, tuple of species types) that was not in the initial training database. For instance, the chemical-species information does not enter the ML potentials in the form of a variable.

The second part of the thesis introduces a neural-network-based scheme to make ML potentials, specifically two- and three-body potentials, explicitly chemical-species-type dependent. We call the models *chemical transferable potentials* (CTP). The methodology enables the prediction of materials not included in the training data. As a showcase example, we consider a set of binary materials. The thesis tackles two challenges at the same time: a) the prediction of the PES of a material not contained in the training data and b) constructing robust models from a limited set of crystal structures. In particular, our tests examine to which extent the ML potentials that were trained on such sparse data allow an accurate prediction of regions of the PES that are far from the training data (in the structural space) but are sampled in a global crystal-structure

search. When performing both constrained structure searches among a set of considered crystal-structure prototypes and an unbiased global structure search, we find that missing data in those regions does not hinder our models from identifying the ground-state phases of materials, even if the materials are not in the training data. Moreover, we compare our method to two state-of-the-art ML methods that, similarly to CTP, are capable of predicting the potential energies of materials not included in the training data. These are the extension of the smooth overlap of atomic positions by an alchemical similarity kernel (ASOAP) introduced in the group of M. Ceriotti, and the crystal graph convolutional neural networks (CGCNN) introduced in the group of J. C. Grossman. In the literature so far, the ASOAP and CGCNN have been benchmarked on single-point energy calculations but have not been investigated in combination with global, unbiased structure-search scenarios. We include the ASOAP and CGCNN in our structure-search tests. Our analysis reveals that, unlike CTP, these two approaches learn unphysical shapes of the PES in regions that surround the training data which are typically sampled in a structure-search application. This shortcoming is particularly evident in the unbiased global-search scenario.

# Zusammenfassung

Die Vorhersage der Grundzustands- und metastabilen Kristallstrukturen von Materialien anhand der Kenntnis ihrer Zusammensetzung ist in der computergestützten Materialwissenschaft eine Herausforderung. In neueren Studien der Forschungsgruppe M. Schefflers wurde untersucht, wie die Energiedifferenz zwischen zwei Kristallstrukturtypen der gleichen chemischen Zusammensetzung anhand der Eigenschaften ihrer atomaren Bestandteile im Rahmen der symbolischen Regression vorhergesagt werden kann. Mithilfe der Verwendung einer neuartigen Compressed-Sensing-basierten Methode, des Sure Independence Screening and Sparsifying Operator (SISSO), wurde aus Milliarden von Kandidaten der Deskriptor identifiziert, der die strukturelle Stabilität am besten erfasst. Ein Deskriptor ist ein Vektor aus analytischen Formeln, die sich aus einfachen physikalischen Größen zusammensetzen.

Im ersten Teil der Arbeit wird eine Multi-Task-Learning-Erweiterung von SISSO (MT-SISSO) vorgestellt, die das Behandeln von Energiedifferenzen zwischen mehreren Kristallstrukturtypen des gleichen Materials ermöglicht. Wir demonstrieren, wie die Multi-Task-Methode, die einen einzigen Deskriptor für alle Strukturtypen identifiziert, die Vorhersage einer wohldefinierten strukturellen Stabilität und damit das Erstellen einer Kristallstrukturkarte ermöglicht. Darüber hinaus zeigen wir, wie MT-SISSO genaue Vorhersagemodelle bildet, selbst wenn die Modelle mit weitgehend unvollständigen Daten trainiert werden.

Ein weiterer bekannter Ansatz zur Bewältigung der Herausforderung der Kristallstrukturvorhersage mit künstlicher Intelligenz basiert auf der Approximation der Born-Oppenheimer-Potentialenergieoberfläche (PEO). Insbesondere haben Gaussian Approximation Potentials, die in der Regel aus einer Kombination von Zwei-, Drei- und Vielteilchenpotentialen bestehen und an Materialien, die aus einem chemischen Element bestehen, gefittet werden, in den letzten Jahren Aufmerksamkeit erregt. Erste Beispiele, die in der Gruppe von G. Csanyi veröffentlicht wurden, haben gezeigt, wie die Grundzustands- und metastabilen Kristallstrukturen von Si, C, P und B korrekt identifiziert werden können. Dabei wurde die PEO erkundet, die durch die Gaussian Approximation Potentials - oder allgemeiner *Machine-Learning-Potentials* (ML-Potentials) - vorhergesagt wurde. Die Transferierbarkeit der bisher bekannten ML-Potentials ist allerdings begrenzt, d. h. ihre Genauigkeit nimmt in Bereichen der PEO, die weit entfernt von den Trainingsdaten liegen, rapide ab. Folglich werden diese ML-Potentiale an große Trainingsdatenbanken gefittet. Des Weiteren müssen solche Trainingsdaten für jedes neue Material (genauer gesagt, Tupel von chemischen Elementen), das nicht in der aktuellen Trainingsdatenbank enthalten ist, konstruiert werden. Beispielsweise fehlt in den ML-Potentials eine Beschreibung der Eigenschaften der chemischen Elemente der Materialien in Form einer Variable.

Im zweiten Teil der Arbeit wird eine auf Neuronalen-Netzen-basierende Methode entwickelt, die eine explizite Abhängigkeit der ML-Potentials, insbesondere Zwei- und Drei-Teilchen-Potentiale, von den chemischen Elementen des Materials erlaubt. Wir nennen die Modelle *Chemical Transferable Potentials* (CTP). Die Methodik ermöglicht

die Vorhersage von Materialien, die nicht in den Trainingsdaten enthalten sind. Als Vorzeigebeispiel betrachten wir eine Reihe von binären Materialien. Die Arbeit befasst sich mit zwei Herausforderungen zur gleichen Zeit: a) der Vorhersage der PEO eines Materials, das nicht in den Trainingsdaten enthalten ist, und b) das Bilden robuster Modelle aus einer begrenzten Anzahl an Kristallstrukturen. In unseren Untersuchungen wird insbesondere evaluiert, inwieweit die auf solch spärlichen Daten trainierten ML-Potentiale eine genaue Vorhersage von Regionen der PEO ermöglichen, die zwar weit von den Trainingsdaten (im Kristallstrukturraum) entfernt liegen, aber in einer globalen Kristallstruktursuche mit abgetastet werden. Sowohl bei eingeschränkten Kristallstruktursuchen unter einer Reihe von betrachteten Kristallstrukturprototypen als auch bei einer uneingeschränkten globalen Kristallstruktursuche stellen wir fest, dass fehlende Daten in diesen Kristallstrukturregionen unsere Modelle nicht daran hindern, die Grundzustandskristallstrukturen von Materialien zu identifizieren, selbst wenn die Materialien nicht in den Trainingsdaten enthalten sind. Darüber hinaus vergleichen wir unsere Methode mit zwei modernen ML-Methoden, die ähnlich wie die CTP in der Lage sind, die potentielle Energie von Materialien vorherzusagen, die nicht in den Trainingsdaten enthalten sind. Die eine Methode basiert auf einer Erweiterung des Smooth Overlap of Atomic Positions um einen *alchemical* Ähnlichkeitsmaß (ASOAP), welche in der Gruppe von M. Ceriotti entwickelt wurde. Die zweite Methode heißt Crystal Graph Convolutional Neural Networks (CGCNN) und wurde in der Gruppe von J. C. Grossman eingeführt. Bisher wurden ASOAP und CGCNN in der Literatur anhand von Einzelpunkt-Energieberechnungen validiert, aber nicht im Rahmen globaler uneingeschränkter Kristallstruktursuchen. Wir wenden unsere Kristallstruktursuchtests ebenso auf ASOAP und CGCNN an. Unsere Untersuchungen zeigen, dass die beiden Methoden im Gegensatz zu den CTP unphysikalische Formen der PEO in Regionen lernen, die weit von den Trainingsdaten entfernt liegen, aber in einer Kristallstruktursuche üblicherweise abgetastet werden. Diese Limitation kommt besonders im uneingeschränkten und globalen Suchszenario zur Geltung.

## Acknowledgments

I want to thank my wife for the seamless support throughout my time at the Fritz Haber Institute of the Max Planck Society. Having her as a motivator allowed me to go beyond myself and achieve record working efficiencies.
Next, I would like to thank Luca M. Ghiringhelli for the outstanding supervision. I did not only enjoy the deep technical discussions. I found someone supporting my passion of developing mathematical modelling approaches from scratch. This made the guy out of me that I am today, a data-analysis nerd.
Moreover, I would like to thank Matthias Scheffler for giving me the opportunity to work in a world-class research group on cutting-edge science.
Last but not least, I want to thank Hagen-Henrik Kowalski. He has been available literally 24 h a day for any kind of discussions. Having a childhood friend as a colleague during my PhD has been a great asset.

# 1 Introduction

Designing a new material for a specific technological application can be a slow and uncertain process [1]. For instance, while the elements of the periodic table provide an enourmous number of combinations (pairs, triplets, ...) to construct chemical compounds, only 16% of ternary, 0.6% of quaternary, 0.03% of quinternary and probably a much smaller fraction of systems involving more elements have been experimentally investigated [2]. Facing moreover the fact that it takes typically 15-20 years from the invention of a new material in the laboratory to its widespread adoption in the market [3,4], a progress (actually a revolution) of the presently slow materials-discovery process is required to determine promising materials early.
Facilitated by advances in both computational power and first-principles techniques in the last decades, modern computational approaches based on a quantum-mechanical description of matter have proven to be an appropriate step in the materials design process [5] simulating several materials properties and overcoming different experimental drawbacks, e.g. the need to synthesize the material first. Still, the number of unknown materials is pratically infinite and the application of first-principles based techniques even to a fraction of the theoretical materials space stays computationally infeasible. In recent years, the growth of materials databases [6–9] together with intelligent data analysis techniques has led to the development of a data-driven and informatics-based methodology [10] to alleviate the need of solving quantum-mechanical equations from scratch repetitively. Machine learning (ML) has become the most promising technique to accelerate and systematically facilitate insights into computational materials science. Many ML-based methods have been introduced for the description of technologically important materials properties [11] where the models allow for a fast prediction of the properties of up to millions of compounds [12]. However, only a small number of promising compounds published were hitherto validated by experiments or quantum-mechanical simulations and the number of "unsolved" materials might increase with the exponential growth of ML-based research in materials science [13]. Crucially, when property predictions with machine-learning models go beyond a reduced set of materials that are already known to exist, the question if a hypothetical material is stable in nature is often not adressed. Many candidates might not be stable. Thus, a reliable and efficient ML-based methodology that predicts what crystal structures a compound will form is required to determine the "realistic" materials in the theoretical materials space. Such a methodology is currently missing.

Predicting the ground-state and metastable crystal structures of a compound from its

chemical composition is one of the most fundamental challenges in materials science because exploring the complex, non-convex potential-energy surface (PES) is a combinatorial problem. In 1988, John Maddox stated [14]: "One of the continuing scandals in the physical sciences is that it remains impossible to predict the structure of even the simplest crystalline solids from a knowledge of their composition." In 1982, pioneering steps in predicting atomic structures and the relative structural stability of solids with *ab initio* methods were demonstrated [15]: the equations of state for several phases of Si and Ge were predicted. Since then, the field of crystal-structure prediction has undergone significant progress and methods combining *ab initio* approaches with structure-search algorithms [16–21] have asserted themselves as their application led to the discovery of several new materials [22, 23]. The dilemma of the computational cost of the underlying *ab initio* method is, nonetheless, present as ever and the current status of crystal-structure prediction is becoming shaped by the development of so-called *ML potentials* [24, 25].

The ground-state solution of the electronic Schrödinger equation provides the Born-Oppenheimer PES [26]. The potential energy depends on the species types of the atoms in the compound and their structural arrangement[1]. The PES describes the space of potential energies in dependence of the crystal structure. Usually, a ML potential approximates specific regions of a PES for a considered set of a few atomic species types (typically not more than four species). The restriction to this set of species types dos not necessarily hinder the treatment of different compositions of the same species types. For example, given two species types $A$ and $B$, a ML potential may be fitted to $AB$ structures, however, possibly also to compounds with chemical formulas $AB_2$ and $A_2B_3$ as well as elemental compounds of the constituents, i.e. $A$ and $B$. In contrast, the prediction of the PES of a compound that includes a different species type $C$ is not possible because the chemical-species information does not enter the ML potentials in the form of a variable. We note that when referring to the the mentioned restriction of a ML potential, we will, nonetheless, use the term *composition* instead of set of atomic species types to say that new sets of species types cannot be modeled. In this regard, we may define that a ML potential is a function $f_{\mathrm{ML}}$ that maps the atomic positions in a material onto the potential energy $E$ of a composition:

$$E_{\mathrm{composition}} = f_{\mathrm{ML}}(\mathrm{structure}). \tag{1.1}$$

ML potentials are fitted to data obtained from *ab inito* calculations. Once the learning step has been performed, the computational cost of predicting the PES with a ML potential is orders of magnitudes lower than the one of the reference *ab initio* methods. However, obtaining the data for the learning may well be very elaborate.
Recent examples have demonstrated how the PES predicted by a ML potential could be

[1] A full description is given by adding the number of electrons. Throughout this work, the considered systems are uncharged. Therefore, the sum of the atomic numbers equals the number of electrons in the system.

explored using a structure-search algorithm to identify the ground-state and metastable phases of C [27, 28], Si [29], B [28, 30], P [31], and Na [28]. Among these examples, Ref. [27, 30, 31] used a hierarichal combination of a two-, three-, and many-body potential and Ref. [29] one of a two- and and many-body potential within the framework of Gaussian Approximation Potentials [32]. Other works have considered clusters and nanoparticles [33–36] as well as surface reconstructions [37] and sheets [38]. However, our studies will focus on three-dimensional crystal structures only. The consideration of systems with defects is out of the scope of this thesis as well. A key component of ML potentials introduced so far has been the available flexibility of the functional form. This choice aims at minimizing the human effort in determining the functional form while allowing for a high fitting accuracy when tackling new training data sets. Typically, the high accuracy is limited to regions of the PES that are close to the training data. Accordingly, such methodology comes with the need for reliable data design techniques to adapt the domain of applicability of the model to regions in the structure space relevant for the targeted application. In Ref. [27, 29], the selection of training structures has relied on past experience. Attempts to autonomize the creation of the training database were studied in Ref. [28, 30, 31, 39]. However, the outcome of both approaches has been the same: a large training database. Crucially, it is unclear to which extent the creation of such databases challenges the speed-up of sampling the PES with a ML potential. Moreover, new training data needs to be constructed for every new composition that was not in the initial training data base. Thus, the acceleration of first-principles crystal-structure prediction through the help of ML is currently uncertain.

In recent years, a conceptually different artificial-intelligence approach to tackle materials-science problems has attracted the attention: a combination of symbolic regression and compressed sensing [40]. In contrast to predifined functional forms that allow for a high flexibility, the expressions are being determined by descriptive parameters based on the physics captured from the data. By using compressed sensing, a short vector of (typically not more than five) analytical formulas is determined where the formulas are identified out of a large pool of candidates. This vector is termed *descriptor*. However, only the recently introduced sure independence screening and sparsifying operator (SISSO) [41] enabled tackling spaces of billions of candidate analytical formulas. SISSO outperforms state-of-the-art compressed-sensing methods if the restriction of the model to a few analytical formulas is required.

The symbolic-regression- and compressed-sensing-based scheme is well suited for the description of a wide range of materials properties and it was first demonstrated for a crystal-structure prediction problem predicting the relative stability of octet binary compounds between two crystal-structure types from descriptors based on chemical information of the atoms only. A descriptor that is based on atomic properties allows for the prediction across the composition space. However, due to the absence of a structural description of the material, the approximation of the PES through a model based on such descriptor is

reduced to selected points on the PES, e.g. equilibria of crystal-structure types instead of whole regions. Accordingly, the compressed-sensing based models map the chemical composition onto the energy of one fixed structure type (or the energy difference of two structure types):

$$E_{\text{structure}} = f_{\text{ML}}(\text{composition}). \tag{1.2}$$

An extension of the approach to model across the structure space as well could be realized by introducing a structural input quantity. However, the symbolic-regression and compressed-sensing based scheme is limited to scalar input features. As good (flexible) structural descriptors are often vectorial quantities, it is currently not clear how the scheme can be applied to the crystal-structure-prediction challenge.

Both models, the ML potential in Eq. 1.1 and the model in Eq. 1.2 that results from the specific application [40] of the symbolic-regression- and compressed-sensing-based scheme, are limited, i.e. predictions are possible across either structure space only or composition space only. As a consequence, the description of the entire materials space can only be realized by separate independent models that are specific to only one composition or structure (pair of structures). The consideration of new compositions or phases that are not contained (or only sparsely contained) in the training database involves new quantum-mechanical calculations.
For example, assume we use a model in the form of Eq. 1.2, however, to predict the relative stability of compositions between two space groups (structure types) instead of the energy of one structure. Then, at least 229 different independent models of that form need to be determined to cover the structural stability among all possible 230 space groups. Accordingly, for every model a sufficiently large and *good* training database needs to be constructed in order to determine reliable models. It would be of great value if an artificial-intelligence model could connect the 229 different optimization problems such that information about the relative stability of a certain composition among a subset of space groups can be used to predict the relative stability of that composition among a different set of space groups. As a result, the number of known compositions for some or all space groups could be reduced.
Analogously, in order to describe the entire composition space with ML potentials, for every composition a sufficiently large and *good* training database of structures needs to be present. The ability to model additionally across chemical composition space, could allow for predictions of the PES of new compositions or reduce the number of calculated data for them.

In this thesis, we show how both approaches can be extended by introducing a so-far missing structural/chemical component. The work is split into two studies. Both studies focus on the prediction of the crystal-structure stability of octet binary compounds among multiple polymorphs considering sparse data sets. Note that the methodologies developed

in this thesis are not restricted to binary systems. The approaches can be further extended to multispecies systems beyond binaries in future projects.
In the first part of the thesis, we extend the symbolic-regression- and compressed-sensing-based scheme to a multi-task learning approach, specifically for SISSO. More in general, if a specific input can be categorized into a set of tasks (here structure types), multi-task SISSO provides an elegant solution by decoupling the models such that the information of this input (structural information) enters only the fitting coefficients parametrically. This enables the treatment of multiple structure types without the need for a definition of a structural quantity. We demonstrate how multi-task SISSO, which identifies one single descriptor capturing different materials properties (structure-type energies), stabilizes the predictions and outperforms the ones of the independent single-task models on sparse and incomplete data sets. Furthermore, we present how the dependence of the models on only one descriptor allows the projection of the multidimensional model onto a ground-state structure map. Contrary to typical black-box ML models, the methodology provides a promising way to visualize the crystal-structure stability of materials and provide insights into the physical mechanisms that drive the stability. Moreover, the visualization of the distribution of the training data in the descriptor space, e.g. in a materials map, reveals *unknown* regions of a model. The ability to locate those regions could facilitate the development of model error estimators. Apart from the crystal-structure-prediction problem that we tackle in this work, our approach is applicable to problems involving other materials properties. The work is, in general, a first showcase of a multi-task-learning application in materials science.
The second part of the thesis introduces a novel class of ML potentials that are able to predict across the chemical composition space, namley *chemical-transferable potentials* (CTP). The dependence on the atomic species types is realized by making the regression coefficients of $n$-body potentials functions of the chemical composition using a neural network. We demonstrate how the additional interpolation across the chemical compound space stabilizes the potentials on the prediction across the crystal-structure space. Furthermore, we tackle a task not investigated yet: the prediction of the PES of a composition that is not included in the training database. In particular, we investigate to which extent models that are built on sparse training data are able to predict the PES of a composition with an accuracy that allows to identify the ground-state and metastable phases in a crystal-structure search. We perform both constrained structure searches among a set of considered crystal-structure prototypes and an unbiased global structure search. Our investigations involve critical factors important for a reliable crystal-structure search, e.g. if a structure appears as a minimum in the predicted PES or spurious minima hinder the identification of the most stable structure(s). We highlight possible challenges in fulfilling these factors and show why other state-of-the-art ML methods fail in determining reliable models. The state-of-the-art methods to which we compare our models are the crystal graph convolutional neural networks [42] and the alchemical smooth overlap of atomic positions [43]. Beyond the specific class of potentials that we introduce, our analysis yields a first step towards reliable next-generation models

that describe the unified structure and chemical composition space. The models aim at a coarse grained prediction of the PES, so that the number of quantum-mechanical calculations in a structure search is greatly reduced, i.e. the ML model is meant as an accelerator, but the reference is meant to be always the considered quantum-mechanical method.

The thesis is divided into six chapters. In Chapter 2, an introduction to the statistical methods behind the ML methods used in this work is given. The mathematical framework of the ML potentials is introduced in Chapter 4. The main results of this work are presented in Chapter 3 and Chapter 5. The former considers the multi-task extension of the symbolic-regression- and compressed-sensing-based approach, the latter the chemical-transferable potentials. Chapter 6 summarizes and concludes the work. The thesis is complemented by a rich appendix where more details and analyses that would have broken the flow of the main story are reported.

# 2 Statistical learning

This chapter deals with the theoretical background behind the ML algorithms used or extended in this thesis. In particular, it focuses on different statistical-learning approaches from the literature, independent of their application to a physical problem. In contrast, Chapter 4 introduces a theoretical framework that implements the concepts of this chapter within a physical methodology that considers a certain problem (i.e. fitting the potential-energy surface).
The different ML methods discussed in this chapter are divided into linear (Sec. 2.1) and kernel (Sec. 2.2) ridge regression, artificial neural networks (Sec. 2.3), and compressed-sensing methods (Sec. 2.4). Note that Sec. 2.4.3 includes the recently introduced sure independence screening and sparsifying operator (SISSO) [41] while its multi-task extension (MT-SISSO) which is introduced in this work will be presented in Chapter 3.

At the heart of the studies presented in this thesis lies the development of new machine-learning (ML) based approaches for identifying correlations and trends in materials databases in order to yield physical insights and allow for the prediction of properties of materials not contained in the reference databases. Typically, the goal is finding a function $f$ from a function space $\mathcal{F}$ that approximates a materials property $P$ in order to accelerate or enable its determination, e.g. surrogating (intensive) quantum-mechanical calculations or reducing the need for expensive and wasteful experimental scan of materials spaces. The functions are learned by a set of $N$ input-output pairs $\{(\boldsymbol{d}_1, P_1), ..., (\boldsymbol{d}_N, P_N)\}$. $P_i$ represents the value of a targeted property of the data point $i$ to be predicted from a vectorial *descriptor* $\boldsymbol{d}_i \in \mathbb{R}^M$ of the data point. A general formulation of the ML optimization problem is given by [44]

$$\arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} L(f(\boldsymbol{d}_i), P_i) + \lambda r(f), \tag{2.1}$$

where $L$ is a loss function and $r(f)$ is a measurement of the *complexity* of $f$. The parameter $\lambda \geq 0$ regulates the compromise between the fitting accuracy and a low complexity of the model. One reason for regularization is to avoid *overfitting* (e.g. fitting to noise) in order to increase the prediction accuracy on data points out of the training set. The choice of $\mathcal{F}$, $L$ and $r$ determines the ML method. All ML methods presented in this chapter are based on the squared error loss $L(f(\boldsymbol{d}_i), P_i) = (f(\boldsymbol{d}_i) - P_i)^2$.

In the following sections, we will write many equations in matrix form and represent the $N$ data points as a property vector $\boldsymbol{P} \in \mathbb{R}^N$ and an input matrix $\boldsymbol{D} \in \mathbb{R}^{N \times M}$.

## 2.1 Linear regression

The most basic ML method is linear regression. One reason is that many linear-regression problems can be solved by linear algebra and convex optimization. The least-squares problem

$$\underset{\boldsymbol{c} \in \mathbb{R}^M}{\arg\min} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|^2 \tag{2.2}$$

formulates the fundament of linear regression analysis determining the regression coefficients $\boldsymbol{c}$ of a linear model $f(\boldsymbol{d}) = \boldsymbol{d}\boldsymbol{c}$. Its solution is given by the closed-form expression $\boldsymbol{c} = (\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})^{-1}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{P}$. Geometrically, the linear model is given by the orthogonal projection of the target vector $\boldsymbol{P}$ onto the column space of the descriptor matrix $\boldsymbol{D}$. It minimizes the Euclidean distance $\|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|$ between the target $\boldsymbol{P}$ and itself, i.e $\boldsymbol{D}\boldsymbol{c}$.

In order to avoid overfitting, the problem 2.2 is often extended by an $\ell_2$ penalty $\lambda\|c\|_2^2 = \lambda\|c\|^2$ (linear ridge regression):

$$\underset{\boldsymbol{c} \in \mathbb{R}^M}{\arg\min} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|^2 + \lambda\|\boldsymbol{c}\|^2. \tag{2.3}$$

The solution to this problem is given by $\boldsymbol{c} = (\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D} + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{P}$, where $\boldsymbol{I} \in \mathbb{R}^{M \times M}$ is the identity matrix.

## 2.2 Kernel ridge regression

The kernel ridge regression is a generalization of the linear ridge regression 2.3 towards considering nonlinar models $f$. Let us consider the optimization problem in Eq. 2.1 using the squared error loss and $r(f) = f^2$. Then, according to the representer theorem [45], the solution to the optimization problem has the form

$$f(\boldsymbol{d}) = \sum_{i=1}^{N} \alpha_i k(\boldsymbol{d}, \boldsymbol{d}_i) \tag{2.4}$$

where $k : X \times X \to \mathbb{R}$ is the associated positive-definite real-valued kernel on our input (descriptor) space $X$. It can be shown that the coefficients $\{\alpha_i\}$ are a closed-form solution $\boldsymbol{\alpha} = (\boldsymbol{K} + \lambda\boldsymbol{I})^{-1}\boldsymbol{P}$ to the optimization problem (kernel ridge regression)

$$\underset{\boldsymbol{\alpha} \in \mathbb{R}^N}{\arg\min} \|\boldsymbol{P} - \boldsymbol{K}\boldsymbol{\alpha}\|^2 + \lambda\boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha}, \tag{2.5}$$

where $\boldsymbol{K} \in \mathbb{R}^{N\times N}$ represents the matrix with elements $K_{ij} = k(\boldsymbol{d}_i, \boldsymbol{d}_j)$. For a linear kernel $k(\boldsymbol{d}_i, \boldsymbol{d}_j) = \boldsymbol{d}_i \cdot \boldsymbol{d}_j$ the model 2.4 becomes a linear function and equals the linear model that results from the optimization problem 2.3 of the linear ridge regression. For instance, the kernel ridge regression can alternatively be derived by kernelizing the linear ridge regression and, moreover, introducing a potentially nonlinear feature map $\phi$ such that $k(\boldsymbol{d}_i, \boldsymbol{d}_j) =< \phi(\boldsymbol{d}_i), \phi(\boldsymbol{d}_j) >$. However, $\phi$ does not need to be known and rather the kernel is specified directly. A typical (nonlinear) kernel is given by the Gaussian kernel $k(\boldsymbol{d}_i, \boldsymbol{d}_j) = \exp(-\frac{\|\boldsymbol{d}_i - \boldsymbol{d}_j\|^2}{2\sigma^2})$. Note that such a kernel function is, generally, viewed as a similarity measurements between two data points.

## 2.3 Artificial neural networks

Artificial neural networks, or simply *neural networks*, are (nonlinear) ML models whose design are (vaguely) inspired by the nervous system of animals. The neural network consists of (partially) connected neurons. Each neuron that receives signals processes them and sends a single output signal to other neurons connected to it. The processing of a signal is modeled by

$$y_{\text{output}} = \phi(\sum_i w_i x_{\text{input},i} + b), \tag{2.6}$$

where $y_{\text{output}}$ represents the output signal, $x_{\text{input},i}$ the input signal received from a neuron $i$ that is weighted by $w_i$, $b$ a bias term, and $\phi$ a so-called *activation function*, typically nonlinear. One popular example for an activation function is given by the rectified linear unit (ReLU) $\phi(x) = \max(0, x)$. In this thesis, we consider feed-forward neural network for regression. If such a neural network consist of one hidden layer, it is written:

$$f(\boldsymbol{d}) = \sum_h^{N_{\text{neurons}}} w_h^{(2)} \phi \left[ \sum_j^M w_{h,j}^{(1)} d_j + b_j^{(1)} \right] + b^{(2)}. \tag{2.7}$$

A typical way to train the weights and biases of neural networks is based on the backpropagation algorithm in combination with a squared error loss [46, 47].

## 2.4 Compressed-sensing based methods

Often, it is desirable to find a linear model $f(\boldsymbol{d}) = \boldsymbol{d}\boldsymbol{c}$ with a sparse solution $\boldsymbol{c}$. We measure the sparsity by the number of non-zero coordinates of the coefficient vector, the $\ell_0$ norm:

$$\|\boldsymbol{c}\|_0 = \#\{j : c_j \neq 0\}. \tag{2.8}$$

A rigorous reformulation of the least-squares problem 2.2 to promote and control sparsity is given by adding an $\ell_0$ penalty:

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\arg\min} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|^2 + \lambda\|\boldsymbol{c}\|_0. \tag{2.9}$$

However, Eq. 2.9 is not a convex but a combinatorial problem and computationally infeasible if $m$ is large. The naive way of determining the solution to it is to solve the least-squares problem

$$\underset{\tilde{\boldsymbol{c}}\in\mathbb{R}^S}{\arg\min} \|\boldsymbol{P} - \tilde{\boldsymbol{D}}\tilde{\boldsymbol{c}}\| \tag{2.10}$$

for all submatrices $\tilde{\boldsymbol{D}} \in \mathbb{R}^{N\times S}$ that consist of a specified number $S = \|\boldsymbol{c}\|_0$ of columns from $\boldsymbol{D}$. The solution $\tilde{\boldsymbol{c}}$ for the submatrix that yields the smallest least-squares error $\|\boldsymbol{P} - \tilde{\boldsymbol{D}}\tilde{\boldsymbol{c}}\|^2$ determines the (only) non-zero elements of the solution $\boldsymbol{c}$ at specified $S$, where the non-zero vector indices of $\boldsymbol{c}$ correspond to the columns of $\tilde{\boldsymbol{D}}$. Varying $S$ corresponds to varying $\lambda$.

### 2.4.1 Least absolute shrinkage and selection operator

A convex reformulation of the $\ell_0$ problem 2.9 is realized by the least absolute shrinkage and selection operator (LASSO) [48], which replaces the $\ell_0$ norm by the $\ell_1$ norm:

$$\|\boldsymbol{c}\|_1 = \sum_i^M |c_i|. \tag{2.11}$$

The resulting optimization problem is given by:

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\arg\min} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|^2 + \lambda\|\boldsymbol{c}\|_1. \tag{2.12}$$

The $\ell_1$ problem 2.12 is an approximation of the $\ell_0$ problem 2.9 and the conditions on $\boldsymbol{P}$ and $\boldsymbol{D}$ that guarantee that the solutions of the two problems coincide, or at least do not differ much, are formulated in the theory of compressed sensing.

### 2.4.2 Orthogonal matching pursuit

The orthogonal matching pursuit (OMP) is a greedy algorithm that iteratively expands the set of columns $\{\boldsymbol{d}^j\}$ of $\boldsymbol{D}$ to which non-zero coefficients are assigned:

1. Initialize the residual $\boldsymbol{R}_0 = \boldsymbol{P}$ and the set $S = \emptyset$ of saved columns $\{\boldsymbol{d}^j\}$. Chose a target number $\Omega$ of columns for the final linear model and let the iteration counter $k = 1$.

2. Extend $S$ by the column $\boldsymbol{d}^j$ with the maximum projection score:

$$\max_j |\boldsymbol{d}^j \cdot \boldsymbol{R}_k|. \tag{2.13}$$

3. Build the submatrix $\tilde{\boldsymbol{D}}$ out of all $\boldsymbol{d}^j \in S$. Calculate $c^* = \arg\min_{\boldsymbol{c}\in\mathbb{R}^M} \|\boldsymbol{P} - \tilde{\boldsymbol{D}}\boldsymbol{c}\|^2$ and let $\boldsymbol{R}_{k+1} = \boldsymbol{P} - \tilde{\boldsymbol{D}}\boldsymbol{c}^*$.

4. If $k = \Omega$, stop. Otherwise set $k = k + 1$ and return to the 2. step.

Note that, Eq. 2.13 in the 2. step selects the column $\boldsymbol{d}^j$ of $\boldsymbol{D}$ that is the closest one to $\boldsymbol{R}_k$, measured by the Euclidean distance. The OMP is characterized by its low computational complexity (relative to the LASSO) because Eq. 2.13 is evaluated by the matrix multiplication $\boldsymbol{v} = \boldsymbol{D}^T\boldsymbol{R}_k$ and the search for the index of $\boldsymbol{v}$ with the highest absolute value.

### 2.4.3 Sure independence screening and sparsifying operator

The $\ell_1$ penalty in Eq. 2.9 (LASSO) has been established as an alternative regularizer for linear regression besides the $\ell_2$ problem 2.3. Often the regularization parameter $\lambda$ is optimized with the aim of yielding the best prediction performance. A different reason for searching for sparse solutions is based on a descriptive methodolgy, i.e. understand on which features (columns) $\boldsymbol{d}^j$ the target $\boldsymbol{P}$ depends. For example, studies that used the compressed-sensing-based methodology introduced in Ref. [40] for materials science had the motivation of identifying the best few, typically not more than five, features in order to provide insights into the physics behind the target property. The iterative scheme in Ref. [40] that combined $\ell_1$ and $\ell_0$ regularization improves the results of a method solely based on $\ell_1$ regularization [49]. However, the combined method is still limited by the $\ell_1$ part particularly when dealing with large and correlated feature spaces. For instance, the goal of the compressed-sensing based methodology introduced in in Ref. [40] is to generate a feature space of billions of candidates which may be highly correlated.
The sure independence screening and sparsifying operator (SISSO) [41] was specifically designed to identify the best few descriptors out of a space of billions of candidates. It makes use of a step of lower computational complexity similar to the second step of the OMP algorithm 2.4.2, however it improves the results of the OMP by enforcing for *stricter* searches for accurate low-dimensional descriptors. Note that in this work we consider only the $\ell_0$ regularization as the sparsifying operator of the SISSO algorithm. While the SISSO method might be considered as independent of how the descriptor matrix $\boldsymbol{D}$ is built, processing the descriptor matrix is an important prestep in the total framework of a symbolic-regression- and compressed-sensing-based scheme introduced in Ref. [40].

The construction of the feature space is an iterative procedure. The basis is given by the set of primary features $\Phi_0$, e.g. the column labels or indices of the input matrix $\boldsymbol{D}$, and

a set of unary and binary operators (such as $+, -, \exp, \sqrt{}, \ldots$). At each iteration $q$ a new feature space $\Phi_q$ is constructed by combining every feature (pair of features) of $\Phi_{q-1}$ with each unary (binary) operator. Sums and differences are constrained to be taken only among homogeneous quantities.
The SISSO algorithm searches for a model based on the best few descriptors out of the final $\Phi_k$. The targeted model is linear in the identified descriptors of $\Phi_k$ but nonlinear in the original features $\Phi_0$. The algorithm is given by:

1. Initialize the residual $\boldsymbol{R}_0 = \boldsymbol{P}$ and the (subspace) set $S = \emptyset$ of saved columns $\{\boldsymbol{d}^j\}$. Chose a target number $\Omega$ of columns for the final linear model and a number $N_S$ of columns added to $S$ at each iteration $k$. Furthermore, let the iteration counter $k = 1$.

2. SIS step: Extend $S$ by the $N_S$ columns $\{\boldsymbol{d}^j\}$ having the largest linear correlations $|\boldsymbol{d}^j \cdot \boldsymbol{R}_k|$ with $\boldsymbol{R}_k$.

3. SO step: Build the submatrix $\tilde{\boldsymbol{D}}$ out of all $\boldsymbol{d}^j \in S$ and solve the $\ell_0$ problem 2.9 for $\tilde{\boldsymbol{D}}$ and $k$ non-zero coefficients. Let $\boldsymbol{R}_{k+1} = \boldsymbol{P} - \tilde{\boldsymbol{D}}\boldsymbol{c}^*$ where $\boldsymbol{c}^*$ is the solution of the $\ell_0$ problem.

4. If $k = \Omega$, stop. Otherwise set $k = k + 1$ and return to the 2. step.

For $N_S = 1$ the SISSO algorithm becomes the same as the OMP. Note that when $k = 1$, the $\ell_0$ problem in the 3. step is not performed because its solution is the model based on the descriptor with the maximum linear correlation with $\boldsymbol{R}_0 = \boldsymbol{P}$, found in the SIS step.
For extending SISSO to classification problems, we refer to the reference [41].

# 3 Multi-task SISSO for crystal structure prediction on incomplete databases

## 3.1 Introduction

The wealth of available data in materials databases [6–8] has opened the era of the data-driven materials science [50, 51]. A critical goal of the field is the acceleration of materials discovery through the establishment of artificial-intelligence tools in order to find patterns and trends in the databases and allow to draw conclusions for properties of (theoretical) materials not contained in the databases. Clearly, the interest in machine-learning (ML) aided research is rapidly increasing: The number of materials and chemistry publications referencing ML has grown exponentially, even as a fraction of total research [13]. When the prediction of a materials property is targeted, the formulation of the ML problem is, typically, composed of a) the definition of a quantity that represents the materials property to be predicted, b) the identification of a materials representation (or descriptor) from which the target-property representation needs to be predicted, and c) the choice or development of the ML algorithm that determines a model mapping the quantity of a) to the one of b). In that sense, a) and b) yield the output and input of a ML model.
Choosing the most convenient quantity for task a) may be not trivial. Let us consider the example of this work. The original property that we want to predict is the energetically lowest crystal structure (ground-state crystal structure) on the potential-energy surface of a material from its chemical formula. A crystal structure is defined by a multi-component quantity given by the atomic numbers, the coordinates of the atomic positions in the crystal and, in case of a solid, the ones of the lattice vectors. Probably, the most naive choice for the output of the ML model would be these crystal coordinates of the ground-state crystal structure. However, the fact that the output is given by a multi-component quantity does not only limit the number of ML algorithms applicable, the high dimensionality of the output space might lead to a poor learning performance of the models. To overcome this problem there are different ways to define the ML problem for the crystal-structure-prediction challenge and, in this thesis, we consider two approaches. One is presented in Chapter 5 and we refer to it for a precise description of the problem. The one of this chapter follows the strategy of dividing first the crystal-structure space into categories that are characterized by crystal symmetries such as space groups (crystal-structure types). As a consequence, the infinite crystal-structure space is split into a set

of a *few* crystal-structure types, e.g. 230 space groups[1]. The most direct ML approach to learn the ground-state crystal-structure prototype can be realized by a classification task. For example, the input of the ML model is given by the chemical formula and the output by the ground-state structure type. Nevertheless, the uncertainty if the structure of a material in the training data is correctly labeled as the ground state may introduce biases to the model. For instance, the ground states can only be determined with certainty if the energy of every other possible crystal-structure type is present in the training data for every chemical formula, a requirement that is not fulfilled in current databases and would, moreover, lead to an immense number of *ab initio* calculations, which is a task that we want to avoid with the approach of this chapter. Instead of considering a classification task, we choose to present the output of the ML model by the potential energy of a structure. More precisely: A data point is specified by the tuple (chemical formula, crystal-structure type), and the output is given by the potential energy of the energetically lowest crystal structure within the crystal symmetry that corresponds to the considered crystal-structure type. Furthermore, we will show that learning energy differences between structure types instead of potential energies directly will simplify learning the crystal-structure stability.

The discussion above of choosing the quantity to be predicted highlights that task a) is indeed not necessarily trivial. Note that we will consider only five different crystal-structure types in this chapter as a showcase for our approach. For a global crystal-structure prediction challenge, the work needs to be extended to include every space group at least for a few chemical formulas in the training data.

Although the $T = 0$ K properties of a material are fully described by the many-body Hamiltonian which is uniquely identified by the atomic positions and numbers $\{\boldsymbol{R}_I, Z_I\}$ in the crystal[2], the connection between $\{\boldsymbol{R}_I, Z_I\}$ and the materials properties is too complicated and indirect. As a result, the description of processes ruling materials properties and functions requires to add as much domain knowledge to the ML problem as available. The general approach is to incorporate this knowledge into b), the input of the ML model or simply *descriptor*, while c) is typically given by well-known ML approaches such as kernel-based methods or neural networks. The “critical role of the descriptor” in "big data of materials science" was highlited by Ghiringhelli *et al.* in 2015 [40]. The work has, moreover, introduced a novel artificial-intelligence based approach to systematically search for a materials descriptor within the framework of symbolic regression applying a two-step procedure: the construction of a large space of candidate descriptors (feature space) and the application of a compressed-sensing-based method that selects the best few descriptors out of the constructed feature space. However, only the introduction of the sure independence screening and sparsifying operator (SISSO) in 2018 [41] enabled to tackle feature spaces of billions of descriptor

[1] Note that this simplification of the ML problem comes at the cost of the need for a search, e.g. with the reference *ab initio* method, for the exact crystal coordinates substantially after the ground-state crystal-structure type was predicted

[2] A full description is given by adding the number of electrons. However, in this work we consider only uncharged system.

candidates, outperforming state-of-the-art compressed sensing methods, if the target number of descriptors is restricted to be low (e.g. $< 6$), and replaced the LASSO (least absolute shrinkage and selection operator) [48] based approach of the reference work[3].

SISSO has been already successfully applied to identifying descriptors for relevant materials-science properties [12, 41, 52, 53]. Still, its extension for two types of problems is a challenge: I) if an input feature is multi-component, e.g. a vector with many coordinates, and needs to appear in a descriptor with all or many of its components and II) if the target property depends strongly on a certain input feature but the data is sparse in that input feature. Note that the collection of input (or *primary*) features is determined by all quantities that are hypothesized to be relevant for describing the target property and, in the feature-space-construction step, non-linear combinations out of them are built using arithmetic operations in an iterative procedure. I) is prevented by the fact that the compressed-sensing-based scheme is conceptually based on scalar features and the only way to include a multi-component quantity considering all its elements is to contract it to a scalar first, e.g. by taking the average of the components. If, alternatively, not all elements of the multi-component feature need to appear in the descriptor, each component might be considered as an independent scalar feature. The candidate descriptors of the constructed feature space would, then, each contain a few or none components. Nevertheless, the applicability of SISSO is computationally limited by the number of input features (often $< 30$), as the construction of the space of candidate descriptors is a combinatorial process. An example for a multi-component feature is a descriptor that represents a crystal structure in a vector and the one of Chapter 5 for building ML potentials based on two- and three-body terms contains 786 elements. Now, let us give an example for II). Consider a training data which is set up of materials whose target property is given by a thermodynamic quantity (depending on the temperature) and was calculated at only a few (e.g. three) different temperatures. To distinguish the properties with respect to their corresponding temperatures one would expect that a descriptor needs to depend on the temperature. However, given the small number of different temperatures in the training set, such treatment carries the risk of capturing the dependence of the target on the temperature wrongly[4]. In such a case, it might be desirable to determine one single model that describes the whole data set, however, does not explicitly depend on the temperature.
We show how the problems I) and II) can be solved by extending the compressed-sensing based method of the reference works to a multi-task (MT) compressed-sensing approach[5],

[3]Note that in Ref. [41], furthermore, a powerful extension of the compressed-sensing based scheme to classification tasks was introduced.

[4]Still, a relationship between the property and the temperature that is simple, i.e. linear, might be captured correctly by the descriptor which is typically a nonlinear function.

[5]More precisely, we will focus only on problem I) but it will be clear that the MT approach provides the needed concept for II).

a framework that belongs to the wider class of learning schemes known as MT learning [54–61]. A *task* for a learning algorithm is the learning of one target property from a single input source (set of features). The learning of multiple tasks (or MT learning) is an umbrella term that refers to [60] the learning of multiple target properties using a single input source, the joint learning of a single target property using multiple input sources, or a mixture of both. The key aspect is the parallel learning of multiple tasks, with the (sometimes implicit) assumption that the shared information among different tasks can lead to better learning performance if all the tasks are learned jointly, as compared to learning them independently. In other words, MT learning assumes that the learning of one task can improve the learning of the other tasks [60].
We demonstrate the MT approach specifically for the SISSO algorithm, i.e. MT-SISSO. More general than the problems I) and II), if an input feature can be categorized into a set of classes (tasks), e.g. structure types or a set of temperatures, MT-SISSO, provides an elegant solution by decoupling the models such that the information about the categorized feature $f_{\text{task}}$ enters only the linear fitting coefficients $c_i$, parametrically, instead of the descriptors $d_i$ (as in the case of all other input features):

$$P = \sum_{i}^{\Omega} c_i(\{f_{\text{task}}\}) d_i(F_{\text{rest},i}). \tag{3.1}$$

Here, $P$ denotes the modeled target property, $F_{\text{rest},i}$ a subset (with index $i$) of all features but $f_{\text{task}}$, $\Omega$ the number of descriptors to be selected by SISSO from the space of candidate descriptors, and the brackets $\{\}$ highlight the parametrical dependence of the coefficients on the tasks. While, in principle, it is possible, to determine alternatively an independent model for each task, the result can be a different descriptor for each task and a loss of information in each separate learning process as the properties of all data points are described by the same physical mechanisms. In fact, the central result of this chapter is that the unified learning process, realized by MT-SISSO, that identifies a single set of descriptors for all tasks improves the prediction performance of the models on incomplete databases compared to using separate single-task (ST) models which, typically, all depend on different sets of descriptors. MT-SISSO can learn accurate predictive models also with high levels of incompleteness, e.g., when 50% or more of the information is randomly missing. We demonstrate this by considering the crystal structure stability of 82 octet binary compounds among five different crystal structure types. Furthermore, we show how the fact that one single descriptor is found for all structures types, or equivalently for every compound,

$$P = \sum_{i}^{\Omega} c_i(\{\text{structure type}\}) d_i(\text{compound}), \tag{3.2}$$

allows to visualize the structural stability and materials distribution in the descriptor space in a ground-state structure map. In general, the work is a first showcase of a MT-learning application in materials science.

This chapter presents my contributions to the publication in Ref. [62]. The contributions include the formulation of theoretical aspects of the methodology for continuous properties, the development of the 2D-structure-map approach as well as the design, implementation, and analysis of all tests for continuous properties.

## 3.2 Theoretical framework

As in the reference works [40, 41], the compressed-sensing-based scheme consists of two steps: the construction of a large space of candidate descriptors (feature space) given a set of input (*primary*) features and the application of a compressed-sensing-based algorithm that identifies the best few descriptors out of the constructed feature space to fit the target property $P$. The initial data set is given by $N$ input-output pairs $\{(\boldsymbol{d}_1^{\text{primary}}, P_1), ..., (\boldsymbol{d}_N^{\text{primary}}, P_N)\}$. $\boldsymbol{d}_i^{\text{primary}}$ represents the vector of input primary features of a data point $i$. We write the values of the target properties $P_i$ into a vector $\boldsymbol{P} \in \mathbb{R}^N$.

The construction of the feature space is an iterative procedure. The basis is given by the set of primary features $\Phi_0$ and a set of unary and binary operators (such as $+, -, \exp, \sqrt{}, \dots$). At each iteration $l$, a new feature space $\Phi_l$ is constructed by combining every feature (pair of features) of $\Phi_{l-1}$ with every unary (binary) operator. Typically sums and differences are constrained to be taken only among homogeneous quantities, i.e. quantities which are expressed by the same units. The SISSO algorithm searches for a model based on the best few descriptors out of the final $\Phi_l$. The targeted model is linear in the identified descriptors of $\Phi_l$ but nonlinear in the primary features $\Phi_0$ if $l > 0$. As a result of the feature-space-construction step, every data point $i$ is represented by a possibly huge descriptor vector $\boldsymbol{d}$ of size $M$. We write the descriptor vectors of all data points down as the rows of a matrix $\boldsymbol{D} \in \mathbb{R}^{N \times M}$.

We split the data, $\boldsymbol{P}$ and $\boldsymbol{D}$, into $Q$ tasks $\{(\boldsymbol{D}^1, \boldsymbol{P}^1), ..., (\boldsymbol{D}^Q, \boldsymbol{P}^Q)\}$. Note that each task may have a different number of samples $N_q$, while the number $M$ of columns of all matrices $\boldsymbol{D}^q$ is the same. In order to make the feature vectors (columns $D_j^q$) comparable (within each task), we standardize each of them to have zero mean and a variance of one.

The goal of single-task (ST) compressed-sensing methods is to approximate the solution of the $\ell_0$ problem (defined in Eq. 2.9). We write the MT analogy of the $\ell_0$ problem

$$\underset{\boldsymbol{C} \in \mathbb{R}^{M \times Q}}{\arg\min} \sum_{q=1}^{Q} \frac{1}{N_q} \left\| \boldsymbol{P}^q - \boldsymbol{D}^q \boldsymbol{C}^q \right\|_2^2 + \lambda \left\| \boldsymbol{C} \right\|_0 , \tag{3.3}$$

where $\boldsymbol{C}$ is the coefficient matrix, with $M$ rows and $Q$ columns, i.e. its $q$-th column $\boldsymbol{C}^q$ is the vector of coefficients projecting $\boldsymbol{D}^q$ onto $\boldsymbol{P}^q$. The $\ell_0$ norm of the matrix $\boldsymbol{C}$ counts the number of *rows* that have at least one nonzero element. The regularization imposes that

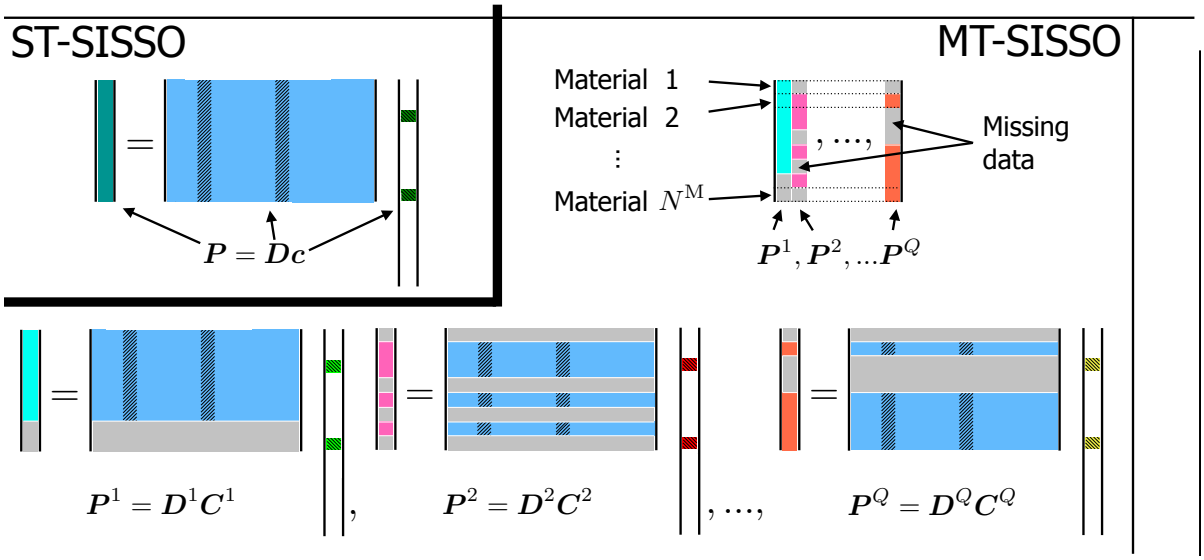



**Figure 3.1:** Schematic representation of single-task (ST) SISSO vs multi-task (MT) SISSO. The white regions in the (sparse) fitting vectors $\boldsymbol{c}$ and $\boldsymbol{C}^q$ represent zero-valued elements, while the colored-hatched squares represent the nonzero-valued elements, which *select* the hatched columns in the descriptor matrices $\boldsymbol{D}$ and $\boldsymbol{D}^q$. The gray areas in the target property vectors $\boldsymbol{P}^q$ - and correspondingly in the descriptor matrices - represent missing/unknown data in the training database, i.e. for some materials (each material occupies always the same position in the property vector and is related to the same row in the descriptor matrix) some property vectors are not known. Crucially, the non-zero values of the fitting vectors $\boldsymbol{C}^q$ are always in the same position (they have the same indices) for the same MT-SISSO learning, and correspondingly the selected columns are the same in all descriptor matrices $\boldsymbol{D}^q$.

when a feature $\boldsymbol{D}_j^q$ is selected (i.e. it has nonzero coefficient $C_j^q$) for one task $q$, then it is selected for all tasks. While the $\ell_0$ minimization in Eq. 3.3 yields exactly what we want (identify the same best few descriptors for all tasks), it is computationally infeasible if $M$ is large as, in practice, we perform the minimization in a combinatorial way:
For a target number $\Omega$ of descriptors to be selected, the set $S_\Omega$ of all column-index $\Omega$-combinations of $\boldsymbol{D}$ is defined. For every $\Omega$-tuple out of $S_\Omega$ the submatrices $\boldsymbol{D}_\Omega^q \in \mathbb{R}^{N_q \times \Omega}$ are built and the (MT-extended) least-squares problem is solved:

$$\underset{\boldsymbol{C}_\Omega \in \mathbb{R}^{\Omega \times Q}}{\arg\min} \sum_{q=1}^{Q} \frac{1}{N_q} \left\| \boldsymbol{P}^q - \boldsymbol{D}_\Omega^q \boldsymbol{C}_\Omega^q \right\|_2^2 . \tag{3.4}$$

Then, the solution of the $\ell_0$ problem is the $\Omega$-tuple that exhibits the lowest least-squares error[6]. We say that the features determined by the best $\Omega$-tuple are the identified best $\Omega$ descriptors (or $\Omega$-dimensional descriptor). The corresponding $Q$ models are given by $\boldsymbol{P}_{\mathrm{fit}}^q = \boldsymbol{D}_\Omega^q \boldsymbol{C}_\Omega^{q*}$ with the single least-squares solutions $\boldsymbol{C}_\Omega^{q*} = (\boldsymbol{D}_\Omega^{q\mathrm{T}} \boldsymbol{D}_\Omega^q)^{-1} \boldsymbol{D}_\Omega^{q\mathrm{T}} \boldsymbol{P}^q$. Varying the target number $\Omega$ of descriptors corresponds to varying $\lambda$.

[6] More precisely, the solution is the coefficient vector $\boldsymbol{C}$ with all entries zero except that the rows that correspond to the $\Omega$-tuple are filled with the non-zero coefficients of the solution to Eq. 3.4

The problem solved by MT-SISSO is an approximation to the $\ell_0$ problem in Eq. 3.3. An important feature of the MT-learning concept that we have introduced is the ability to handle tasks of different sample sizes. For instance, an incomplete database (one where for every data point not every task is known) typically consists of tasks with different sample sizes. Figure 3.1 shows graphically the setup for MT-SISSO, in particular in terms of the possibility to deal with incomplete data. Note that, in this work, we consider SISSO only using the $\ell_0$ optimization as the SO step. The MT-SISSO algorithm is given by:

1. Initialize $Q$ residuals with $\boldsymbol{R}_0^q = \boldsymbol{P}^q$ and the (subspace) set $S = \emptyset$ of saved column (descriptor) indices. Chose the target dimension $\Omega$ of the descriptor for the final linear models and a number $N_S$ of columns added to $S$ at each iteration $k$. Furthermore, let the iteration counter $k = 1$.

2. SIS step: Extend $S$ by the $N_S$ column indices $\{j\}$ having the largest linear correlation score with the residuals:

$$\theta_j = \sqrt{\sum_{q=1}^{Q} < \boldsymbol{D}_j^q, \boldsymbol{R}^q >^2 / N_q}. \tag{3.5}$$

3. SO step: Build for each task the submatrix $\tilde{\boldsymbol{D}}^q$ out of all columns $\boldsymbol{D}_j^q$ with saved indices $j \in S$ and solve the MT-$\ell_0$ problem 3.3 for the set of $Q$ submatrices $\tilde{\boldsymbol{D}}^q$ and a target dimension $k$ of the descriptor. Let every $\boldsymbol{R}_{k+1}^q = \boldsymbol{P} - \tilde{\boldsymbol{D}}^q \boldsymbol{C}^{q*}$ where $\boldsymbol{C}^{q*}$ are the solutions of the $\ell_0$ problem.

4. If $k = \Omega$, stop. Otherwise set $k = k + 1$ and return to the 2. step.

For $Q = 1$, MT-SISSO becomes ST-SISSO (described in Sec. 2.4.3). Note that when $k = 1$, the $\ell_0$ problem in the 3. step is, in practice, not performed because its solution is known already in the 2. step, i.e. the models based on the descriptor $j$ with maximum $\theta_j$. A derivation of $\theta_j$ is given in App. A.

The number of iterations $l$ in the construction of the feature space $\Phi_l$ and the dimension $\Omega$ of the descriptor are (hyper-)parameters of the SISSO method, to be optimized with respect to the prediction error on a validation set, typically via cross validation (CV). The size $\Omega \cdot N_S$ of the subspace selected by SIS is also a parameter, but not a hyperparameter to be optimized. In fact, ideally it is large enough to include the solution of the $\ell_0$ problem 3.3. In practice, we invoke the relationship that the compressed-sensing theory establishes between the size $\Omega \cdot N_S$ of the feature space $S$ of selected descriptors by all SIS steps, dimensionality of the solution $\Omega$, and the number of data points $N$: $\Omega N_S = \exp\left(N/(\kappa \cdot \Omega)\right)$, where $\kappa$ is a dimensionless constant that the compressed-sensing theory locates between 1 and 10.

## 3.3 Data set

The data set of our example consists of 82 octet binary compounds, each in five crystal-structure types, including rock-salt (RS) and zinc-blende (ZB) of the reference works [40, 41] based on ST learning, and three further phases: the CsCl, NiAs, and CrB prototypes. Every compound was optimized in each of the five different crystal-structure types by fully relaxing all degrees of freedom compatible with the corresponding crystal symmetry (1 degree of freedom for RS, ZB, and CsCl, 2 degrees of freedom for NiAs, and 5 for CrB). This results in $82 \cdot 5$ data points, e.g. (structure, potential energy)-tuples. The data was calculated in Ref. [63], using DFT within the local-spin-density approximation, and downloaded from the NOMAD Repository [6].

Modeling four independent energy differences, each between two crystal structures for every compound, is sufficient for describing the relative stability among the five crystal-structure types, i.e. using only one structure as reference and learning the energy differences to that structure. However, learning four energy differences may lead to large errors for the relative stability of any two other phases whose difference was not considered as a target to be learned. In contrast, the simultaneous learning of all ten possible energy differences limits the prediction error of the relative stability between all phases. Therefore, the MT problem that we consider is given by ten tasks made of all possible energy differences. The distribution of the ten energy differences is shown in Fig. 3.2.
Note that considering energy differences as the targets simplifies modeling the structural stability compared to using cohesive energies (DFT energy of a structure minus the DFT energy of the gas-phase atomic constituents). For instance, the root mean square error (RMSE) on predicting all ten energy differences with a MT model that was fitted to the five cohesive energies is 0.15 eV/atom[7], while it is only 0.07 eV/atom if the energy differences were fitted directly. For both models the parameters chosen in the next section were used.

## 3.4 Choice of the parameters

For the descriptor identification, we use atomic properties as input features: the ionization potential ($IP$), electron affinity ($EA$), number of valence electrons $n_{\text{val}}$, the group number $G$ in the periodic table, and the radii $r_{s,p,d}$ where the radial probability density of the valence $s$, $p$, and $d$ orbitals are maximal. Furthermore, equilibrium distances $d_{ij}$ of homonuclear $AA$ and $BB$, and $AB$ dimers are included.
We set the parameter $\kappa$ that determines the sizes of the SIS subspaces (see theory in Sec. 3.2) to 3.3. With $N = 82$, the subspace sizes $\Omega \cdot N_S$ are approximately $2 \cdot 10^5$

[7] This means that first five cohesive energies per compound were predicted from a model fitted to the five cohesive energies. Then, ten energy differences were derived out of the five predicted cohesive energies.

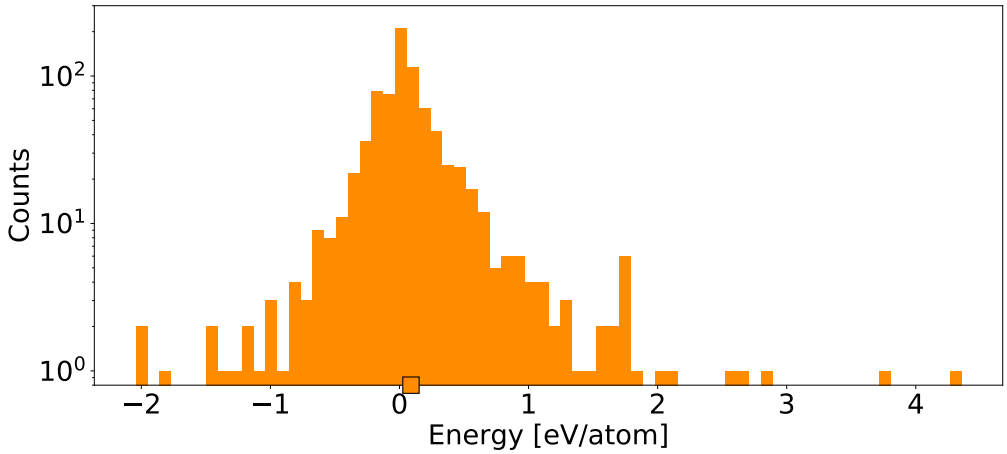



**Figure 3.2:** Distribution of reference energy differences (10 pairs of structure types) for all 82 octet binaries. The square marks the average value of the distribution.

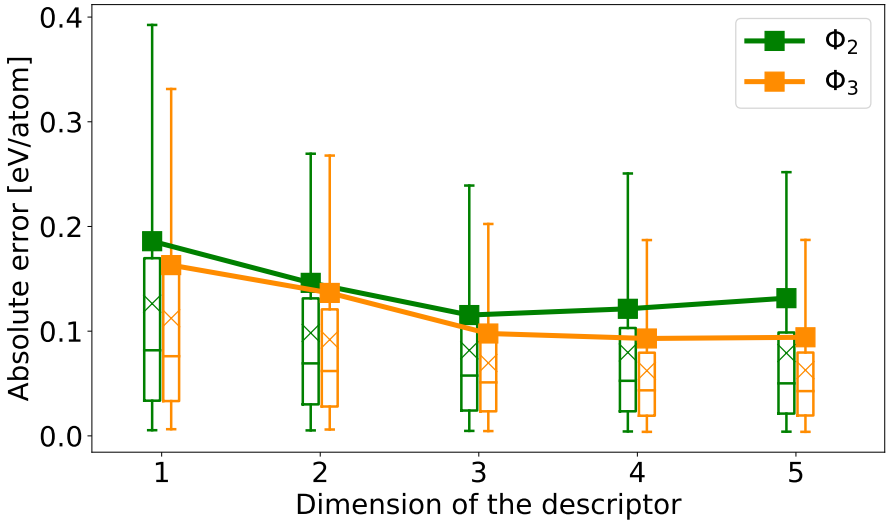



**Figure 3.3:** Prediction errors of MT-models, as a function of the dimension $\Omega$ of the descriptor, for the feature space $\Phi_2$ and $\Phi_3$. All errors are averaged over 30 repetitions of a leave-10%-out CV (MT-SISSO is trained over 90% of randomly selected data and tested on the remaining 10%). The "box plots" mark the 25th and 75th percentiles (extrema of the rectangle), the 5th and 95th percentiles (extrema of the "whiskers"), and the median (horizontal line inside the rectangle). Shown are also the mean absolute error (MAE, cross) and the root mean square error (RMSE, solid square).

| | $\Omega$ | RMSE | Median | $p_{75}$ | $p_{95}$ | MaxAE |
|---|---|---|---|---|---|---|
| $\boldsymbol{\Phi}_2$ | 1 | 0.186 | 0.082 | 0.170 | 0.393 | 1.098 |
| | 2 | 0.146 | 0.069 | 0.131 | 0.272 | 1.055 |
| | 3 | 0.115 | 0.058 | 0.112 | 0.240 | 0.649 |
| | 4 | 0.121 | 0.053 | 0.103 | 0.252 | 0.968 |
| | 5 | 0.132 | 0.050 | 0.099 | 0.252 | 1.385 |
| $\boldsymbol{\Phi}_3$ | 1 | 0.163 | 0.076 | 0.158 | 0.332 | 1.056 |
| | 2 | 0.137 | 0.062 | 0.121 | 0.268 | 0.973 |
| | 3 | 0.098 | 0.051 | 0.090 | 0.205 | 0.548 |
| | 4 | 0.093 | 0.043 | 0.079 | 0.187 | 0.742 |
| | 5 | 0.094 | 0.043 | 0.080 | 0.189 | 0.709 |

**Table 3.1:** Tabulated values from Figure 3.3. $p_{75}$ and $p_{95}$ are the 75th and 95th percentiles, respectively, RMSE is the root mean square error, and MaxAE is the maximum absolute error. All quantities are given in [eV/atom].

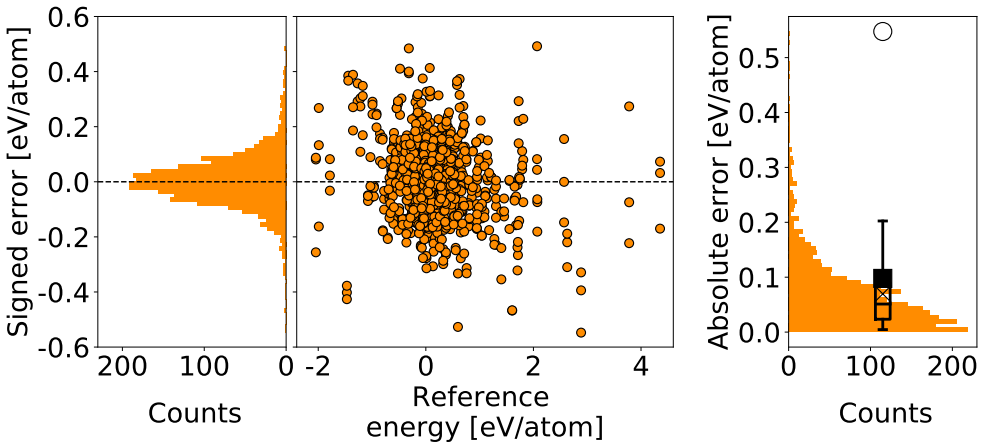



**Figure 3.4:** Prediction errors of MT-SISSO models (using $\Phi_3$, $\Omega = 3$) within a leave-10%-out cross validation. Central panel: Prediction errors vs. reference energies (energy differences). Left panel: distribution of the prediction errors. Right panel: distribution of the absolute values of the prediction errors and corresponding box plot. The box plot and symbols are consistent with Fig. 3.3, except that the maximum absolute error (MaxAE, circle) was added, here.

, $4 \cdot 10^3$, $5 \cdot 10^2$, and $10^2$ for $\Omega = 2, 3, 4, 5$[8]. These values are kept fixed through all our numerical tests, e.g. also when the sample size $N$ is decreased in the cross-validation (CV) tests. For the routine application of ST and MT-SISSO, we note that the sizes $N_S$ of the feature subspaces used in this work are rather large. We checked that even for $\kappa = 4$, the same descriptors are always found at $\Omega = 2$, while for $\Omega = 3$ even $\kappa = 5$ is small enough to yield the same descriptor as for $\kappa = 3.3$

In Fig. 3.3 (the corresponding numerical values are tabulated in Tab. 3.1), results of a leave-10%-out CV test are shown, performed in order to assess the two hyperparameters of MT-SISSO: the (size of the) constructed feature space $\Phi_l$ and the dimensionality $\Omega$ of the descriptor. The leave-10%-out CV is performed in the following way: A random set of 10% of the 82 materials is left out of the training set, the MT-SISSO model is trained on the remaining 90% of the materials, and the errors are measured for the left-out materials. This random selection of training and validation sets was repeated 30 times, which we found sufficient to converge the validation RMSE to 0.01 eV. Note that all the ten target properties of a material are excluded from the training set when it is left out.
Analysis of Fig. 3.3 reveals that models trained by using the larger feature space $\Phi_3$ (containing $\sim 2 \cdot 10^{10}$ features) are consistently better performing (in terms of prediction errors) than models trained on $\Phi_2$ (containing $\sim 2.4 \cdot 10^5$ features), for all dimensions. RMSE and mean absolute errors (MAE) are only marginally better when going from $\Phi_2$ to $\Phi_3$, but we notice that the largest percentiles (75th and 95th) improve significantly, especially for $3 \leq \Omega \leq 5$. The overall best model is $(\Phi_3, \Omega = 5)$, but we also notice that, for $\Phi_3$, the improvement of all error indicators when going from $\Omega = 3$ to $\Omega = 5$ is only marginal. Therefore, in view of the significantly smaller computational time needed to train $\Omega = 3$ vs $\Omega = 5$, in the tests of the next section, we focus on $(\Phi_3, \Omega = 3)$. The detailed analysis of the signed and absolute errors for the latter setting is shown in Fig. 3.4.

## 3.5 Stabilization of the SISSO models through multi-task learning

Besides conceptual advantages of MT vs ST learning (see next section), the shared learning across different tasks realized by MT learning stabilizes (reduces the risk of high prediction errors of) the models if only incomplete data is available. We will show this by performing two tests.

In the first test, we selected left-out sets in this way: One material and one crystal structure are randomly selected and all the energy differences involving the selected structure are eliminated from the training set for the selected material. The procedure is repeated until

[8] Recall that for $\Omega = 1$, no subspace is considered as the best descriptor is determined already by the SIS step on all features.

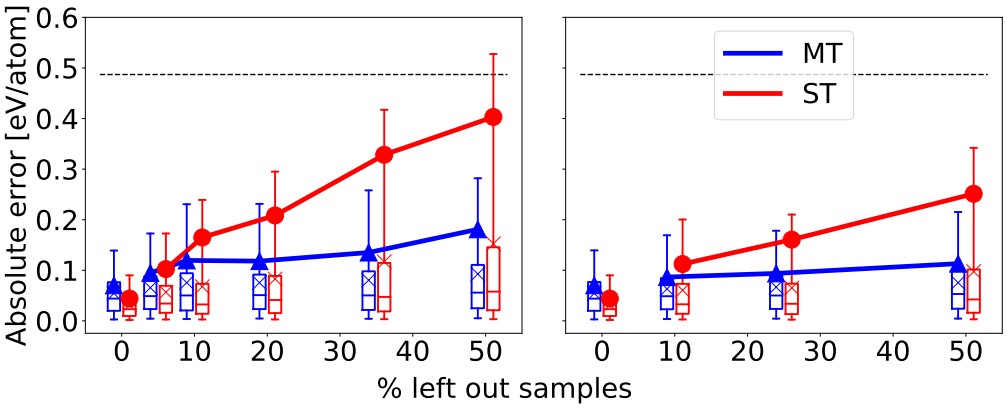



**Figure 3.5:** Prediction errors of MT-SISSO vs (average) ST-SISSO for (left panel) "leave $x$% of (materials, structure) data out" and (right panel) "leave $y$% of data for one crystal structure out". The symbol convention is the same as in Fig. 3.3. The errors at 0% data out in both panels are training errors. The horizontal line at 0.49 eV/atom represents the *basline* (standard deviation of the reference energy differences) against which the root mean square errors (solid markers) are benchmarked.

a prefixed $x$% of pairs (material, structure) are eliminated. We recall, the total number of such pairs is $82 \times 5 = 410$. This test simulates the training over a materials database where for some (or many) materials the information for only some crystal structures is available. It would be of great value if from such dishomogenous database, one could predict the missing information. For a meaningful test, we added the following two constraints in the simulated elimination of database fields: For each material, the energy of at least two crystal structures (e.g. one energy difference) is known and for each of the ten tasks (energy differences) there are at least four materials carrying the information, in order to have enough data to train the four fitting coefficients of the $\Omega = 3$ model[9]. For each $x$% selected value, we train one MT-SISSO model and 10 independent ST-SISSO models (one for each task of MT-SISSO). We then look at the prediction errors on the missing data. The left panel of Fig. 3.5 shows the outcome of the test. With abuse of notation, the values at 0% refer to the training error on all data points. As one should expect, ST-SISSO yields lower training error due to higher flexibility (for each task, a different descriptor can be chosen). However, as soon as data are missing, MT-SISSO rules with lower RMSE and, crucially, with lower largest errors. Interestingly, the quality of MT-SISSO stays pretty unchanged, for all error indicators, over a wide range of amount of missing data.

In the second test, we select, first, one crystal structure and then we remove the energy values for a given $y$% of materials. Removing the energy value of one structure implies the removal of four energy differences from the (material, energy differences) database. One MT-SISSO model and four ST-SISSO models are trained and the errors for the selected structures are evaluated on the missing materials. This test simulates the case of a new crystal structure being identified for only few materials in the database and one wants

[9]Three coefficients are assigned to the three descriptors and one determines the constant/intercept.

to learn with the fewest possible data the predicted energy in such new crystal structure for all materials. We repeat this test for leaving each of the five structures once out and report the prediction errors averaged over all five tests.
The right panel of Fig. 3.5 shows the performance of the MT-SISSO models vs the one of the ST-SISSO models. Again the training error (at 0%) favors ST-SISSO and again MT-SISSO's performance remains impressively constant over a wide range of amount of missing data.

These two tests show numerically what should be expected from a physical point of view: It is reasonable to assume that the energy of different crystal structures depend on the same mechanism encoded in the properties of the gas-phase atoms used as primary features. Therefore MT-SISSO uses at best the (possibly scarce) information scattered over all crystal structures to identify such mechanism. In this way the prediction on the scarcely known materials and/or crystal structures is more reliable than a model that uses information from only one crystal structure (or, one pair of crystal structures, as in the presented case) to identify the descriptor.

## 3.6 Design of a crystal-structure map

The identification of a single descriptor for multiple tasks (crystal structures), enabled only by the MT extension of the compressed-sensing-based scheme, allows to draw a phase-diagram (crystal-structure map) whose axes are given by the found descriptor. Let us consider the ($\Omega = 2$) MT-SISSO model trained over all data points. The linear models of the different tasks can be represented as planes in a 3D space, where the coordinates $(x, y)$ are the components of the descriptor and coordinate $z$ is the predicted energy. A subtle implication of the MT-SISSO learning energy differences is that the models maintain an *internal consistency* with respect to a common energy zero, which allows for the unambiguous determination of the predicted lowest-energy structure for each coordinate $(x, y)$. For example, for any three structures $\alpha$, $\beta$, $\gamma$, the difference in energy $E(\alpha) - E(\gamma)$ is by construction equal to $(E(\alpha) - E(\beta)) - (E(\gamma) - E(\beta))$. This is not (necessarily) true if the three energy differences are learned with separate, independent models. For instance, the internal consistency obtained through the MT-SISSO models is a result of the fact that all models depend on the same descriptor (see derivation in App. B).

The left panel of Figure 3.6 represents the structure map for the octet binaries. The colored areas refer to the predictions and the colored squares are the reference data. A color is associated with any specific crystal structure and assigned to a square (pixel) $(\delta x, \delta y)$ centered on $(x, y)$ when the corresponding structure is the lowest in energy at $(x, y)$. The white color marks areas where the energy difference between the lowest-energy and the second lowest-energy structures differs by less than 0.03 eV/atom. In order to give

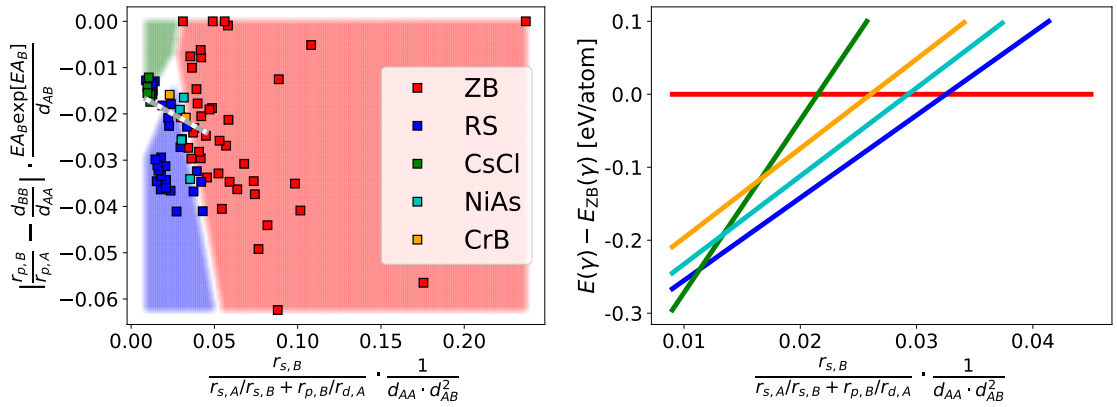



**Figure 3.6:** Left: MT-SISSO-learned ground-state structure map for the octet binaries. The colored areas represent the predicted stability region for the structure with the same color in the legend. The squares are colored according to the reference lowest-energy structure. The white color marks areas where the difference between the energy of the lowest-energy and the second lowest-energy structure differs by less or equal 0.03 eV/atom. Right: cut of the phase-diagram along the dashed line shown in the left panel. The lines are the traces of the planes representing the predicted energy difference from the baseline (ZB structure).

an insight into the 3D visualization of the structure map, we show in the right panel of Fig. 3.6, a cut along the gray-white dotted line marked in the left panel of Fig. 3.6. This shows that some crystal structures are predicted to be very close in energy for certain values of the descriptors. In a realistic application, one may conclude that the actual ground state in the neighborhood of those values of the descriptor may be any of the low-energy structures, while those that are predicted to be very high in energy can be safely discarded as candidate ground states. To gauge the trustfulness of the presented phase diagram, we mention that the largest prediction error for a structure that appears "misclassified" (the color of its symbol does not match the predicted color of the background) is 0.09 eV/atom.

## 3.7 Outlook

While the extension of the approach to considering a more global description of the structure space, e.g. modeling all 230 space groups, is conceptually possible, there are some practical challenges.

Clearly, for each of the 230 space groups at least a few data points need to be present in the training data. If no data is available for a space group, a task cannot be assigned to it. Furthermore, the number of present data points in a space group needs to be at least as high as the number of fitting coefficients to determine the model corresponding to the task (there are $\Omega + 1$ fitting coefficients in a model based on a $\Omega$-dimensional descriptor). As discussed in Sec. 3.3, predicting energy differences instead of cohesive energies is

advantageous in terms of accuracy, if not even essential. However, following the spirit of this work that it is beneficial to consider for the tasks any energy difference possible among a set of given structures, the extension of our approach to 230 structure types is computationally impractical, as the set up would need to consist of 26 335 tasks. Attempts to *sparsify* the number of tasks need to be discussed in a future work. Moreover, out of a scope of this thesis, the consideration of formation energies (DFT energy of structure minus the DFT energy of elemental solids built by the atomic constituents) instead of energy differences or cohesive energies, might be testes as an alternative, as the assignation of one energy to a structure type would result in maximally 230 tasks. To give a feeling on the computational effort necessary to find the MT-SISSO descriptor and model, we report that the learning for the settings ($\boldsymbol{\Phi}_3, \Omega = 3$) and $82 \cdot 10$ data points (materials $\times$ energy differences) was run on an Intel Xeon E5-2698 v3 node with 2 CPUs per node (16 cores/CPU @ 2.3 GHz) and it took 5 h. On a 4-cores laptop, it would take a couple of days of runtime. We remind that the size of the features space is huge: $2 \times 10^{10}$ features.

## 3.8 Conclusions

In conclusion, we have introduced an extension of the symbolic-regression- and compressed-sensing-based scheme of Ref. [40, 41] to a MT-learning approach, considering specifically the SISSO algorithm. If an input feature can be categorized into classes, MT-SISSO yields an elegant way to take account of this input feature. This is needed if the feature is high dimensional, such as structural information, because SISSO is limited to scalar features. We have shown the applicability of MT-SISSO to the challenge of crystal-structure prediction and demonstrated how only the MT extension of the symbolic-regression- and compressed-sensing-based scheme enabled the description of the structural stability in a well-defined phase diagram and ground-state structure map. Furthermore, we highlighted how MT learning stabilized the models on incomplete data sets. As opposed to ST-SISSO, MT-SISSO was able to learn accurate predictive models also with high levels of incompleteness (e.g. when 50% or more of the information was randomly missing).

# 4 Machine-learning Potentials

## 4.1 Definition of interatomic potential

An **interatomic potential** is a closed-form expression that calculates the potential energy of a system of atoms as a function of their arrangement. The design of an interatomic potential is often based on a quantum-mechanical description of matter. The ground-state solution of the electronic Schrödinger equation provides the Born-Oppenheimer PES, which is typically the reference for an accurate treatment of the interactions between nuclei. Interatomic potentials approximate specific regions of the Born-Oppenheimer PES and DFT-based calculations are a possible source of data[1]. Traditionally, the functional form of potentials was determined based on chemical intuition [64–66]. Such potentials are referred to as analytical, empirical, semiempirical, or classical potentials. With machine-learning (ML) potentials, a new class of interatomic potentials was introduced that aim at the determination of accurate models at much less human intervention. Their main difference to analytical potentials is a more flexible functional form that has less physical background built in. However, the higher flexibility comes with the need for new data-design techniques because the accuracy of ML potentials may decrease even more rapidly in regions of the PES that are distant from the training data than the classical approach.
In this chapter, we will present the theoretical framework of the ML potentials developed in this work. We will introduce the chemical-transferable potentials (CTP) and summarize two other state-of-the-art methods to which the CTP will be compared in Chapter 5. The two other methods are the extension of the smooth overlap of atomic positions by an alchemical kernel [43, 67] and the crystal graph convolutional neural networks [42].
The chemical-transferable potentials are based on $n$-body potentials. We note that our terms and mathematical definitions might deviate from the ones of other works which treat $n$-body potentials. The definitions mainly serve to decompose and illustrate the mathematical framework behind our methods. Basically, many of our concepts follow the framework of $n$-body potentials [27, 68] that were introduced within the framework of Gaussian Approximation Potentials [32] and the respective implementation in the QUIP code [69]. However, our implementations are similar but not equivalent [70] to the ones of the Gaussian Approximation Potentials of Ref. [32].

[1] More precisely, the calculations are based on an approximation to DFT and provide only an approximation to the Born-Oppenheimer PES.

Usually, there are two fundamental properties of interatomic potentials and a possible definition is given in the following [32].

**Definition 4.1.** Let $(\boldsymbol{S}, \boldsymbol{Z})$ be a system of atoms with positions $\boldsymbol{S} = (\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_{N_{\text{atoms}}})$[2] and atomic numbers $\boldsymbol{Z} = (Z_1, Z_2, ..., Z_{N_{\text{atoms}}})$. Furthermore, let $A \subset \{1, 2, ..., 118\}$ be a set of atomic numbers. A **local interatomic potential** is a function

$$\begin{aligned} E : \mathbb{R}^{3N_{\text{atoms}}} \times A^{N_{\text{atoms}}} &\to \mathbb{R} \\ (\boldsymbol{S}, \boldsymbol{Z}) &\mapsto E(\boldsymbol{S}, \boldsymbol{Z}) \end{aligned}$$

that approximates the Born-Oppenheimer PES and has the properties:

a) The energy is a sum of atomic contributions $\epsilon_i$,

$$E = \sum_{i}^{N_{\text{atoms}}} \epsilon_i.$$

b) Each atomic contribution $\epsilon_i$ to the total energy depends on the positions (and chemical species) of atoms $j$ that are within a cutoff region (neighbourhood) from atom $i$.

All ML potentials introduced in this chapter and investigated in Chapter 5 are such interatomic potentials. A typical choice of the neighbourhood is a spherical cutoff region $\rho(\boldsymbol{r}_i) = \{\boldsymbol{r}_j \in S \mid \|\boldsymbol{r}_i - \boldsymbol{r}_j\| < r_{\text{cut}}\}$ [32,71]. Note that, generally, ML interatomic potentials treat $\boldsymbol{Z}$ only parametrically, i.e. the information about the species types enters only the fitting coefficients.
An interatomic potential defined through Def. 4.1 exhibits practical advantages. Property a) ensures size-consistency, e.g. doubling the unit cell of a crystal to a supercell doubles also the potential energy per crystal, and its computational cost scales only linearly with the number of atoms in the unit cell. Property b) simplifies the atomic environment to a small number of interactions and the size of the cutoff is a major factor for computational speed. It puts, however, a constraint on the accuracy of the model [32] as interactions with atoms outside of the cutoff region are, in the best case, mapped only effectively into the neigbourhood.

Given Def. 4.1, the effort in designing a potential lies in finding an appropriate function for $\epsilon$. In the case of ML potentials, this task is separated into two components, namely finding an appropriate representation $\boldsymbol{q}$ of the crystal system and a function that maps $\boldsymbol{q}$ onto $\epsilon$ (or onto a *part* of $\epsilon$).

[2] In case of periodic systems, we assume that also the unit cell is given and the atom index runs only over atoms in the unit cell.

## 4.2 $n$-body descriptors

In the field of data-driven materials science, the search for the materials representation is considered as the major challenge while the function mapping onto the materials property is determined by an ML algorithm, typically a kernel-based method or a neural network. Fundamental symmetries, such as invariance for translation, rotation, reflection of the crystal, or permutation of the same species types, are chosen to be incorporated into the crystal representation. Representation design has become an attractive field [72–75] that has led to many new descriptors [76–82] or, in the case of general materials properties, novel approaches [41, 48]. However, in interatomic potentials, representations like explicit pairwise or three-body descriptors often stay a key term [27, 29, 39, 83]. A possible definition of an $n$-body descriptor is given below.

**Definition 4.2.** Let $P = (\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_n) \subset S$ be positions of a set of atoms. Then the function $q$

$$\begin{aligned} q : \mathbb{R}^{n\times 3} &\to \mathbb{R}^m \\ P &\mapsto \boldsymbol{q} \end{aligned}$$

with

$$\frac{\partial q}{\partial \boldsymbol{r}_i} \neq 0 \; \forall \; \boldsymbol{r}_i \in P$$

is called an ($m$-dimensional) ***n*-body descriptor**.

The two-body descriptor is given by

$$q_{ij} = r_{ij} = \|\boldsymbol{r}_i - \boldsymbol{r}_j\| \tag{4.1}$$

with $j \in \rho_i$. $\rho_i = \{j \mid \|\boldsymbol{r}_i - \boldsymbol{r}_j\| < r_{\text{cut}}, \; \boldsymbol{r}_j \in S\}$ is the set of indices of atoms in the neighbourhood of atom $i$. The two-body descriptor often builds the basis of more complex descriptors such as many-body descriptors (e.g. the smooth overlap of atomic positions in Sec. 4.5) or the following three-body descriptor

$$\boldsymbol{q}_{ijk} = (r_{ij}, r_{ik}, r_{jk}) \tag{4.2}$$

with $j, k \in \rho_i$. In this form, the three-body descriptor is ordered, and symmetry in exchanging the neighbour atoms $j$ and $k$ needs to be imposed either by the potential function or by assigning to each triplet of atoms a second three-body descriptor with $j$ and $k$ exchanged. In this work, we implement the latter symmetrization technique. Note that, in the following sections, we will often denote the pairwise distance in

a vector form $\boldsymbol{q}_{ij}$ to keep consistent with the vectorial form of the three-body descriptor.

## 4.3 $n$-body potentials

A common approach to design potentials is the so-called *many-body expansion* in which the local energies $\epsilon_i$ in Def. 4.1 are written in linear combinations of $n$-body contributions [64, 65].

**Definition 4.3.** Let $E = \sum_i^{N_\text{atoms}} \epsilon_i$ be an interatomic potential and $\{\boldsymbol{q}_{ij}, \boldsymbol{q}_{ijk}, ...\}$ $n$-body descriptors. The **many-body expansion** expresses the local energy $\epsilon$ formally through

$$\epsilon_i = \epsilon_\text{1b} + \sum_{j \in \rho_i} \epsilon_\text{2b}(\boldsymbol{q}_{ij}) + \sum_{j,k \in \rho_i} \epsilon_\text{3b}(\boldsymbol{q}_{ijk}) + ... \tag{4.3}$$

and $\epsilon_{n\text{b}}$ is called an ***n*-body potential**.

The general modelling strategy is based on truncating the expansion after a rather small number $n$, e.g. $n < 5$, due to the combinatorial growth of the number of $n$-tuples. Sometimes, the expansion is complemented with a term based on a many-body descriptor [27, 29, 39] like the smooth overlap of atomic positions (Sec. 4.5).
Note that while $n \geq$ 2-body terms are interaction-based models, $\epsilon_\text{1b}$ is typically a quantity that depends on the composition of the material, e.g. a quantity based on the energies of the atomic constituents.

### 4.3.1 $n$-body potentials for multiple species types

Considering a sum of a two- and three-body potential for an elemental solid, e.g. Si, the energy of the crystal can be separated into a two- and three-body contribution

$$E = E_\text{SiSi} + E_\text{SiSiSi} \tag{4.4}$$

with $E_\text{SiSi} = \sum_i^{N_\text{atoms}} \sum_{j \in \rho_i} \epsilon_\text{2b}(\boldsymbol{q}_{ij})$ and $E_\text{SiSiSi} = \sum_i^{N_\text{atoms}} \sum_{j,k \in \rho_i} \epsilon_\text{3b}(\boldsymbol{q}_{ijk})$. One way to model a multi-species systems, e.g. a binary AB, is to separate the interactions of the different species combinations into a linear combination of contributions:

$$\begin{aligned} E = {} & E_{AA} + E_{AB} + E_{BB} \\ & + E_{AAA} + E_{AAB} + E_{ABB} \\ & + E_{BAA} + E_{BAB} + E_{BBB}. \end{aligned} \tag{4.5}$$

Correspondingly, nine functions $\epsilon_{nb,I}$ are specified for nine types of interactions $I$ ($AA$, $AB$,... $BAB$, $BBB$) which leads to a higher flexibility of the multi-species potential compared to the single-species one. The indices $i$, $j$ ($k$) in $E_{AB}$ ($E_{ABC}$) run only over atoms of species types $A$, $B$ ($C$). Note that the first element $A$ of a three-body interaction $ABC$ specifies the species type of the central atom $i$ of a triplet represented by the descriptor $(r_{ij}, r_{ik}, r_{jk})$. Its neighbours $B$ and $C$ correspond to $j$ and $k$ in the descriptor.

### 4.3.2 Functional forms of machine-learning $n$-body potentials

The task of determining functions $\epsilon_{n\mathrm{b}}(\boldsymbol{q}_{n\mathrm{b}})$ starts with the specification of the function space from which a corresponding ML algorithm selects an optimal element. A possible result of the demand for efficient optimization are functions that are linear in regression coefficients $\alpha_\mu$ with basis functions $b_\mu$ nonlinear in the input $\boldsymbol{q}_{n\mathrm{b}}$:

$$\epsilon_{n\mathrm{b}}(\boldsymbol{q}_{n\mathrm{b}}) = \sum_{\mu}^{N_{\mathrm{basis}}} \alpha_\mu b_{n\mathrm{b},\mu}(\boldsymbol{q}_{n\mathrm{b}}). \tag{4.6}$$

Then also the total energy of the system can be written in a linear form (considering again a two- and three-body potential)

$$E = \sum_{\mu}^{N_{\mathrm{basis,2b}}} \alpha_{2\mathrm{b},\mu} B_{2\mathrm{b},\mu} + \sum_{\mu}^{N_{\mathrm{basis,3b}}} \alpha_{3\mathrm{b},\mu} B_{3\mathrm{b},\mu} \tag{4.7}$$

with

$$B_{n\mathrm{b},\mu} = \sum_{i}^{N_{\mathrm{atoms}}} \sum_{j,k\ldots\in\rho_i} b_{n\mathrm{b},\mu}(\boldsymbol{q}_{ijk\ldots}). \tag{4.8}$$

For a data set of multiple structure-energy tuples the equations become a linear system

$$\boldsymbol{E} = \boldsymbol{B}\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{B}_{2\mathrm{b}} & \boldsymbol{B}_{3\mathrm{b}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_{2\mathrm{b}} \\ \boldsymbol{\alpha}_{3\mathrm{b}} \end{pmatrix}. \tag{4.9}$$

The coefficients $\boldsymbol{\alpha}_{2\mathrm{b}}$ and $\boldsymbol{\alpha}_{3\mathrm{b}}$ are determined by (regularized) linear regression (i.e. Eq. 2.3). Note that the closed-form expression $\boldsymbol{\alpha} = (\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} + \lambda\boldsymbol{I})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{E}$ is equivalent to the one that is obtained in the sparsification process [68] of the Gaussian Approximation Potentials [32] if the same two- and three-body potential (with the same basis functions) is used. The rows of $\boldsymbol{B}$ can be considered as structural descriptors that could alternatively be used as a vectorial input of a nonlinear ML model. The columns of $\boldsymbol{B}$ represent the

basis functions (or their indices $\mu$).
Following the works [27,32], a Gaussian basis set is used

$$\epsilon_{n\mathrm{b}}(\boldsymbol{q}) = \sum_{\mu}^{N_{\mathrm{basis}}} \alpha_\mu \exp(-\frac{(\boldsymbol{q}_{n\mathrm{b}} - \boldsymbol{q}_\mu)^2}{2\sigma_\mu^2}), \tag{4.10}$$

where the collection of Gaussian centers $\boldsymbol{q}_\mu$ and widths $\sigma_\mu$ is a set of hyperparameters.
In order to guarantee smooth decay of energies and forces to zero at the cutoff $r_c$, a cutoff function is multiplied with the (radial) basis functions:

$$\tilde{b}_{2\mathrm{b},\mu}(r_{ij}) = b_{2\mathrm{b},\mu}(r_{ij}) \cdot f_{\mathrm{cut,2b}}(r_{ij}). \tag{4.11}$$

The following cutoff function was used [68]:

$$f_{\mathrm{cut,2b}}(r_{ij}) = \begin{cases} 1, & r \leq r_c - r_w \\ \frac{1}{2}(\cos\left[\pi \frac{(r_{ij} - r_c + r_w)}{r_w}\right] + 1), & r > r_c - r_w. \end{cases} \tag{4.12}$$

Here, $r_w$ specifies the width of the smooth decay region. It is a hyperparameter of the ML model. For the three-body terms we implement:

$$\tilde{b}_{3\mathrm{b},\mu}(r_{ij}, r_{ik}, r_{jk}) = b_{3\mathrm{b},\mu}(r_{ij}, r_{ik}, r_{jk}) \cdot f_{\mathrm{cut,2b}}(r_{ij}) \cdot f_{\mathrm{cut,2b}}(r_{ik}). \tag{4.13}$$

## 4.4 Chemical-transferable potentials

Normally, ML potentials do not take the chemical-species types explicitly as variables but treat them parametrically, e.g. by letting information on the chemical species enter the fitting coefficients. For instance, ML potentials have often been fitted to data of a single chemical formula (or systems with not more than three species types) which did not require an explicit dependence of the potentials on the species type in form of a variable. ML models that take the chemical species types as variables do exist [42, 43, 78, 84], for example crystal graph convolutional neural networks (Sec. 4.6) or a kernel-based model depending on the alchemical SOAP (Sec. 4.5.1). However, a methodology that makes explicitly $n$-body potentials (defined in the form of Def. 4.3) species dependent was, to the best of our knowledge, not investigated yet and is introduced in this work. A definition of a chemical-transferable $n$-body potential is given in the following.

**Definition 4.4.** Let $\mathcal{Z} = (1, 2, 3, ...) = (Z_{\mathrm{H}}, Z_{\mathrm{He}}, Z_{\mathrm{Li}}, ...)$ be the set of all atomic numbers (species types), $\boldsymbol{Z}$ a vector of $n$ atomic numbers $Z_i \in \mathcal{Z}$ , and $\epsilon_{n\mathrm{b}}(\boldsymbol{q}_{n\mathrm{b}}) = \sum_{\mu}^{N_{\mathrm{basis}}} \alpha_\mu b_\mu(\boldsymbol{q}_{n\mathrm{b}})$

an $n$-body potential as defined in Def. 4.3. Then a **chemical-transferable $n$-body potential** is written

$$\epsilon_{n\mathrm{b}}(\boldsymbol{q}_{n\mathrm{b}}, \boldsymbol{Z}) = \sum_{\mu}^{N_{\mathrm{basis}}} \alpha_\mu(\boldsymbol{Z}) b_\mu(\boldsymbol{q}_{n\mathrm{b}}). \tag{4.14}$$

where $\alpha_\mu$ are functions of the species types $\boldsymbol{Z}$.

In this way, the chemical information about the species types enters only the coefficients, while the basis functions stay solely a structural description of the material. Note that the $\alpha_\mu$ are explicit functions of species tuples and not of the chemical formula of a compound, e.g. the compounds MgS, MgO, and $Mg_2O$ are described by the same Mg-Mg and Mg-Mg-Mg coefficients, $\alpha_\mu([Z_{\mathrm{Mg}}, Z_{\mathrm{Mg}}])$ in the two-body and $\alpha_\mu([Z_{\mathrm{Mg}}, Z_{\mathrm{Mg}}, Z_{\mathrm{Mg}}])$ in the three-body case.

The determination of the relationship between the species types and coefficients $\alpha_\mu$ follows the approach of first representing the input $\boldsymbol{Z}$ with an appropriate descriptor vector $\boldsymbol{d}$ to be mapped with a ML model to the coefficients. This means we replace $\boldsymbol{Z}$ by $\boldsymbol{d}$. Atomic properties as used in [40] are one example for building $\boldsymbol{d}$, e.g ionization potential, electron affinity, or orbital-based radii. To ensure the invariance of the potential energy with respect to interchanging atoms of different species types, we symmetrize the descriptors. In case of a two-body potential for an $AB$ compound and using an atomic property $d$, we implement:

$$\boldsymbol{d} = \begin{pmatrix} d(A) + d(B) \\ |d(A) - d(B)| \end{pmatrix}. \tag{4.15}$$

Our symmetrized descriptor for the three-body potential is defined by:

$$\boldsymbol{d} = \begin{pmatrix} d(A) + d(B) + d(A) + d(C) \\ |d(A) - d(B)| + |d(A) - d(C)| \end{pmatrix}. \tag{4.16}$$

This choice of the three-body symmetrization takes account of the symmetry of our three-body potential, i.e. that $E_{ABC} = E_{ACB}$ (see paragraph below Eq. 4.6). In case multiple atomic properties $d_1, d_2, ...$ are used, we write the symmetrized elements of all properties in one vector,

$$\boldsymbol{d} = \begin{pmatrix} d_1(A) + d_1(B) \\ |d_1(A) - d_1(B)| \\ d_2(A) + d_2(B) \\ |d_2(A) - d_2(B)| \\ \vdots \end{pmatrix}, \tag{4.17}$$

if considering, for example, the two-body case.

Our experience is that a linear relationship between the chemical descriptors $\boldsymbol{d}$ and the coefficients $\alpha_\mu$ does not yield sufficiently accurate potentials. One way to obtain a nonlinear relationship could be a symbolic type of regression in which the input descriptors are combined by arithmetic combinations to be selected by a compressed-sensing-based technique, see Chapter 3. In fact, this is possible because the linearity of the symbolic regression models in their regression coefficients allows to write the total energy in a linear combination of products between combined descriptors and structural basis functions. However, in this work we used a neural network as described in the next section and the symbolic-regression-based approach could be investigated in a future work. Note that a neural network that determines parameters of a predefined expression (here $n$-body potential) was already introduced in a recent work by S. A. Ghasemi *et al.* [85]. The neural network mapped atomic environments onto atomic electronegativities that were components of an explicit potential.

### 4.4.1 Chemical-transferability via neural networks

In order to realize a nonlinear connection between the chemical descriptors $\boldsymbol{d}$ and the coefficients $\alpha_\mu$ of the potentials, a neural network is used. An example with one hidden layer is written

$$\alpha_\mu(\boldsymbol{Z}) = \sum_{h}^{N_{\text{neurons}}} W^{(2)}_{\mu,h} \phi \left[ \sum_{j}^{N_d} W^{(1)}_{h,j} d_j(\boldsymbol{Z}) + b^{(1)}_j \right] + b^{(2)}_\mu \tag{4.18}$$

or in matrix form

$$\boldsymbol{\alpha}(\boldsymbol{Z}) = \boldsymbol{W}^{(2)} \phi \left[ \boldsymbol{W}^{(1)} \boldsymbol{d}(\boldsymbol{Z}) + \boldsymbol{b}^{(1)} \right] + \boldsymbol{b}^{(2)} \tag{4.19}$$

where the activation function $\phi$ is applied to a vector elementwise and $\boldsymbol{W}^{(i)}$ and $\boldsymbol{b}^{(i)}$ are weights and biases, respectively.

Consider the example of a two- and three-body chemical-transferable potential for a binary compound. The energy of a crystal is decomposed into species-interaction contributions (as already described in Sec. 4.3.1):

$$\begin{aligned} E = {} & E_{AA} + E_{AB} + E_{BB} \\ & + E_{AAA} + E_{AAB} + E_{ABB} \\ & + E_{BAA} + E_{BAB} + E_{BBB}. \end{aligned} \tag{4.20}$$

Then we use two neural networks, one for the two- and one for the three-body potential. If implementing one-hidden-layer neural networks and using Eq. 4.8 to sum the potential basis functions over the atomic environments in the crystal into the elements of a matrix $\boldsymbol{B}$, the energy of a crystal is written:

$$
\begin{aligned}
E = \; & \boldsymbol{B}_{AA}^T \left( \boldsymbol{W}_{2\mathrm{b}}^{(2)} \phi \left[ \boldsymbol{W}_{2\mathrm{b}}^{(1)} \boldsymbol{d}_{AA} + \boldsymbol{b}_{2\mathrm{b}}^{(1)} \right] + \boldsymbol{b}_{2\mathrm{b}}^{(2)} \right) \\
& + \boldsymbol{B}_{AB}^T \left( \boldsymbol{W}_{2\mathrm{b}}^{(2)} \phi \left[ \boldsymbol{W}_{2\mathrm{b}}^{(1)} \boldsymbol{d}_{AB} + \boldsymbol{b}_{2\mathrm{b}}^{(1)} \right] + \boldsymbol{b}_{2\mathrm{b}}^{(2)} \right) \\
& + \boldsymbol{B}_{BB}^T \left( \boldsymbol{W}_{2\mathrm{b}}^{(2)} \phi \left[ \boldsymbol{W}_{2\mathrm{b}}^{(1)} \boldsymbol{d}_{BB} + \boldsymbol{b}_{2\mathrm{b}}^{(1)} \right] + \boldsymbol{b}_{2\mathrm{b}}^{(2)} \right) \\
& + \boldsymbol{B}_{AAA}^T \left( \boldsymbol{W}_{3\mathrm{b}}^{(2)} \phi \left[ \boldsymbol{W}_{3\mathrm{b}}^{(1)} \boldsymbol{d}_{AAA} + \boldsymbol{b}_{3\mathrm{b}}^{(1)} \right] + \boldsymbol{b}_{3\mathrm{b}}^{(2)} \right) \\
& + \boldsymbol{B}_{AAB}^T \left( \boldsymbol{W}_{3\mathrm{b}}^{(2)} \phi \left[ \boldsymbol{W}_{3\mathrm{b}}^{(1)} \boldsymbol{d}_{AAB} + \boldsymbol{b}_{3\mathrm{b}}^{(1)} \right] + \boldsymbol{b}_{3\mathrm{b}}^{(2)} \right) \\
& + \boldsymbol{B}_{ABB}^T \left( \boldsymbol{W}_{3\mathrm{b}}^{(2)} \phi \left[ \boldsymbol{W}_{3\mathrm{b}}^{(1)} \boldsymbol{d}_{ABB} + \boldsymbol{b}_{3\mathrm{b}}^{(1)} \right] + \boldsymbol{b}_{3\mathrm{b}}^{(2)} \right) \\
& + \boldsymbol{B}_{BAA}^T \left( \boldsymbol{W}_{3\mathrm{b}}^{(2)} \phi \left[ \boldsymbol{W}_{3\mathrm{b}}^{(1)} \boldsymbol{d}_{BAA} + \boldsymbol{b}_{3\mathrm{b}}^{(1)} \right] + \boldsymbol{b}_{3\mathrm{b}}^{(2)} \right) \\
& + \boldsymbol{B}_{BAB}^T \left( \boldsymbol{W}_{3\mathrm{b}}^{(2)} \phi \left[ \boldsymbol{W}_{3\mathrm{b}}^{(1)} \boldsymbol{d}_{BAB} + \boldsymbol{b}_{3\mathrm{b}}^{(1)} \right] + \boldsymbol{b}_{3\mathrm{b}}^{(2)} \right) \\
& + \boldsymbol{B}_{BBB}^T \left( \boldsymbol{W}_{3\mathrm{b}}^{(2)} \phi \left[ \boldsymbol{W}_{3\mathrm{b}}^{(1)} \boldsymbol{d}_{BBB} + \boldsymbol{b}_{3\mathrm{b}}^{(1)} \right] + \boldsymbol{b}_{3\mathrm{b}}^{(2)} \right) .
\end{aligned}
\tag{4.21}
$$

### 4.4.2 Optimization of the neural networks

The weights of the two neural networks in Eq. 4.21 are determined by solving the optimization problem

$$
\underset{\boldsymbol{W}_{2\mathrm{b}}, \boldsymbol{b}_{2\mathrm{b}}, \boldsymbol{W}_{3\mathrm{b}}, \boldsymbol{b}_{3\mathrm{b}}}{\arg\min} \sum_{i=1}^{N_{\mathrm{data}}} \left[ E_i^{\mathrm{ref}} - E_i^{\mathrm{CTP}}(\boldsymbol{W}_{2\mathrm{b}}, \boldsymbol{b}_{2\mathrm{b}}, \boldsymbol{W}_{3\mathrm{b}}, \boldsymbol{b}_{3\mathrm{b}}) \right]^2 \tag{4.22}
$$

using backpropagation and stochastic gradient descent. In this work we regularize the neural networks by *early stopping* at an epoch that optimizes the prediction error on a validation test, similar to other works in materials science that use neural networks such as the crystal graph convolutional neural networks [42] or SchNet [84].

It might be desirable to make also the parameters of the basis functions $b_\mu$ dependent on the compounds or species tuples in the form of variables, i.e. the widths $\sigma_\mu$ and centers $\boldsymbol{r}_\mu$ in a basis set of Gaussians $b_\mu = \exp(-\frac{(\boldsymbol{q}-\boldsymbol{q}_\mu)^2}{2\sigma_\mu^2})$. However, the fact that, in this work, the basis function parameters do not enter the optimization problem 4.22 as variables has a practical advantage: the basis functions $b_\mu$ can be summed together into a matrix $\boldsymbol{B}$ (see Eq. 4.8) that is treated as a constant factor in the optimization problem. In contrast, an additional optimization of the basis function parameters would require to calculate the gradient along every basis function $b_\mu$ resulting in a significantly higher computational expense in the optimization process. An investigation of the computational limits of an algorithm that optimizes also the basis function parameters is a possible future work.

### 4.4.3 Constrained two-body potentials

By using zero-centered Gaussians $b_\mu(r_{ij}) = \exp(-\frac{(r_{ij}-0)^2}{2\sigma_\mu^2})$ as basis functions, the two-body potential (Eq. 5.3) becomes:

$$\epsilon_{2\mathrm{b}}(r_{ij}) = \sum_{\mu}^{N_{\mathrm{basis}}} \alpha_\mu \exp(-\frac{(r_{ij}-0)^2}{2\sigma_\mu^2}). \tag{4.23}$$

A Gaussian is, then, defined only by its width $\sigma_\mu$. One of the most fundamental relationships that is typically incorporated into an interatomic potential, is a pairwise interaction that consists of a repulsion behaviour ($\frac{\partial \epsilon_{2\mathrm{b}}}{\partial r_{ij}} < 0$) at small distances and an attraction part ($\frac{\partial \epsilon_{2\mathrm{b}}}{\partial r_{ij}} > 0$) at larger distances. We found that the choice of zero-centered Gaussians allows to assign specific roles to the different basis functions, i.e. ones with relatively small widths are suited to model a (strong) repulsion, while the ones with larger widths are proper to fit an attraction part. To control the roles of the basis functions when using the CTP, we constrain the coefficients $\alpha_\mu$ of the Gaussians with smaller widths to be positive and the ones that correspond to Gaussians with larger widths to be negative, see Fig. 4.1. This is realized by squaring the outputs of the neural network in Eq. 4.18 and multiplying them with a $s_\mu = 1$ or $s_\mu = -1$ depending on the basis function. Eq. 4.18, then, becomes for the coefficients of the two-body potentials:

$$\alpha_\mu(\boldsymbol{Z}) = s_\mu \left( \sum_{h}^{N_{\mathrm{neurons}}} W_{\mu,h}^{(2)} \phi \left[ \sum_{j}^{N_d} W_{h,j}^{(1)} d_j(\boldsymbol{Z}) + b_j^{(1)} \right] + b_\mu^{(2)} \right)^2 . \tag{4.24}$$

Note that an alternative way to control the signs of the the coefficients $\alpha_\mu$ is to apply ReLU activation functions to the original neural-network outputs instead of squaring them. The ReLU activation function is given by $f(x) = \max(0, x)$. When using ReLUs, we found that some coefficients $\alpha_\mu$ converged against zero during the training process in our tests. Accordingly, we observed a missing repulsion in some $AB$ pairwise potentials when coefficients defined to be positive became zero.

## 4.5 The smooth overlap of atomic positions

In Chapter 5, we will compare the chemical-transferable potentials (Sec. 4.4) introduced in this thesis to two state-of-the-art ML methods. The two methods are the extension of the smooth overlap of atomic positions (SOAP) by an alchemical kernel (this section) and the crystal graph convolutional neural networks (Sec. 4.6).

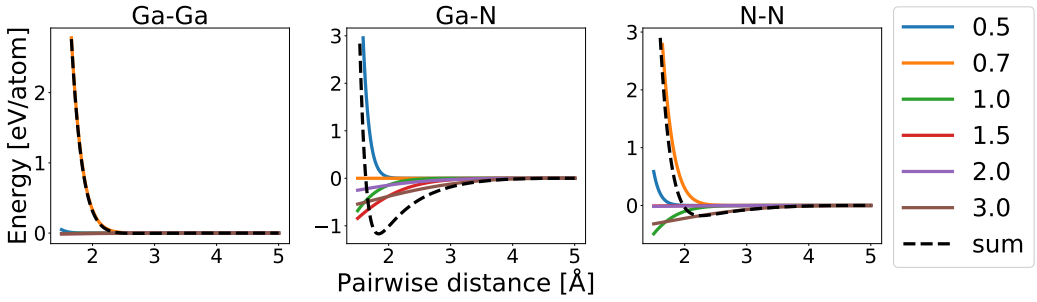



**Figure 4.1:** Two-body potentials decomposed into contributions of their basis functions. The potentials are obtained from a chemical-transferable-potential (CTP) model trained on structures of 77 octet binary compounds, GaN not included. The CTP was fitted within the leave-one-compound-out cross validation in Sec. 5.4.6. The three panels show three two-body potentials modeling the Ga-Ga, Ga-N and N-N interaction. The black dashed line represents the two-body potential $\epsilon_{2b}(r_{ij}) = \sum_{\mu}^{N_{\text{basis}}} \alpha_\mu \exp(-\frac{(r_{ij}-0)^2}{2\sigma_\mu^2})$, i.e. the sum of basis-function contributions $\alpha_\mu \exp(-\frac{(r_{ij}-0)^2}{2\sigma_\mu^2})$. Each basis-function contribution is shown by a solid line and represented in the legend by its Gaussian width $\sigma_\mu$. The coefficients $\alpha_\mu$ of the Gaussians with the two smallest widths 0.5 and 0.7 were constrained to be positive, the remaining coefficients negative. We checked that with this choice, the energy-vs-distance curves of dimers in the gas phase from elements of the 78 octet binary compounds could accurately be fitted. The positive coefficients were able to describe the repulsive part of the energy-vs-distance curves of the dimers.

The SOAP kernel [76]

$$k(U_i, U_j) = \left( \frac{\tilde{k}(U_i, U_j)}{\sqrt{\tilde{k}(U_i, U_i) k(U_j, U_j)}} \right)^{\zeta} \tag{4.25}$$

$$\tilde{k}(U_i, U_j) = \int dR \left| \int \rho_{U_i}(\boldsymbol{r}) \rho_{U_j}(\boldsymbol{r}) (R\boldsymbol{r}) d\boldsymbol{r} \right|^2 \tag{4.26}$$

measures the similarity (overlap) between the local atomic neighbour densities

$$\rho_{U_i}(\boldsymbol{r}) = \sum_{k \in U_i} \exp(\frac{(\boldsymbol{r}_k - \boldsymbol{r})^2}{2\sigma^2}) \tag{4.27}$$

of atom $i$ and $j$, with their neighbourhoods $U_i$ and $U_j$. The integration is performed over all three-dimensional rotations $R$. Furthermore, $\zeta$ is a positive integer. Expanding

$$\rho_{U_i}(\boldsymbol{r}) = \sum_{nlm} c_{nlm} g_n(r) Y_{lm}(\hat{\boldsymbol{r}}) \tag{4.28}$$

in terms of spherical harmonics $Y_{lm}$ and orthonormal basis functions $g_n(r)$ the (not normalized) SOAP kernel becomes

$$\tilde{k}(U_i, U_j) = \boldsymbol{p}_i \cdot \boldsymbol{p}_j = \sum_{n_1 n_2 l} p_{i,n_1 n_2 l} p_{j,n_1 n_2 l} \tag{4.29}$$

with the power spectrum

$$p_{n_1 n_2 l} = \sum_m c_{n_1 l m} (c_{n_2 l m})^\dagger. \tag{4.30}$$

### 4.5.1 The alchemical smooth overlap of atomic positions

One way to distinguish between different species types is to split the neighbor densities into different parts

$$\rho^\alpha_{U_i}(\boldsymbol{r}) = \sum_{k \in U_i^\alpha} \exp(\frac{(\boldsymbol{r}_k - \boldsymbol{r})^2}{2\sigma^2}), \tag{4.31}$$

where the index $k$ runs only over atom types $\alpha$, and extend the SOAP kernel (Sec. 4.5) to:

$$\tilde{k}(U_i, U_j) = \int dR \left| \sum_\alpha \int \rho^\alpha_{U_i}(\boldsymbol{r}) \rho^\alpha_{U_j}(\boldsymbol{r})(R\boldsymbol{r}) d\boldsymbol{r} \right|^2. \tag{4.32}$$

By introducing, furthermore, an alchemical kernel $\kappa_{\alpha\beta}$, also the similarity between atomic species types can be taken into account [43]:

$$\tilde{k}(U_i, U_j) = \int dR \left| \sum_{\alpha\beta} \kappa_{\alpha\beta} \int \rho^\alpha_{U_i}(\boldsymbol{r}) \rho^\beta_{U_j}(\boldsymbol{r})(R\boldsymbol{r}) d\boldsymbol{r} \right|^2. \tag{4.33}$$

Setting the alchemical kernel to the Kronecker-delta $\kappa_{\alpha\beta} = \delta_{\alpha\beta}$ recovers the original expression Eq. 4.32. An explicit chemical dependence is realized, for example, by assigning nonzero similarities to a pair of different species types, e.g. with a Gaussian kernel and some quantity that represents the atomic species types such as the atomic number $Z$: $\kappa_{\alpha\beta} = \exp\left[-(Z_\alpha - Z_\beta)^2/(2\sigma^2)\right]$.

### 4.5.2 The smooth overlap of atomic positions for crystals

The SOAP kernel as defined in Eq. 4.26 or 4.33 measures the similarity between atomic environments. In order to compare two structures, the SOAP kernel is evaluated for all pairs of atoms from structure $X_1$ and $X_2$, resulting in a covariance matrix:

$$C_{ij}(X_1, X_2) = \tilde{k}(U_i, U_j) \ , i \in X_1, j \in X_2. \tag{4.34}$$

Following the framework of decomposing structural energies into sums of local contributions, a kernel for comparing structures can be written as an average of atomic pairwise comparisons [43]

$$\tilde{K}(X_1, X_2) = \frac{1}{N_A N_B} \sum_{i \in X_1, j \in X_2} \tilde{k}(U_i, U_j), \tag{4.35}$$

if the target energy is averaged over the number of atoms in the structure. Using the kernel 4.35 to compare data points (structures), the target energy is fitted via kernel ridge regression (Sec. 2.2).

## 4.6 Crystal graph convolutional neural networks

In the crystal graph convolutional neural network [42] introduced by T. Xie and J. C. Grossman the local energies are given by a linear transformation

$$\epsilon_i = \boldsymbol{\nu}_i \boldsymbol{W_l} + b_l \tag{4.36}$$

of a vector $\boldsymbol{\nu}_i$, which represents both the species type and structural environment of an atom. The representation is learned in the training process by a neural network as described in the following.
First $\boldsymbol{\nu}_i$ is initialized by a random vector $\boldsymbol{\nu}_i^{(0)}$ which is species-type specific [86]. Such a random initialization of the atomic descriptor was implemented by Schütt *et al.* in another deep-learning method applied to fitting PESs as well [84]. However, note that this choice precludes transferability to compositions with unseen combinations of species types.
Following the work in Ref. [42], for the structural description of the neighbourhood, only the twelve nearest neighbours are taken into account with vectors

$$\boldsymbol{u}_{ij}[\mu] = \exp(-\frac{(r_{ij} - r_\mu)^2}{\sigma^2}). \tag{4.37}$$

In each iteration $k$, a concatenated vector

$$\boldsymbol{z}_{ij}^{(k)} = \boldsymbol{\nu}_i^{(k-1)} \oplus \boldsymbol{\nu}_j^{(k-1)} \oplus \boldsymbol{u}_{ij}^{(k-1)} \tag{4.38}$$

is built and the representation according to

$$\boldsymbol{\nu}_i^{(k)} = \boldsymbol{\nu}_i^{(k-1)} + \sum_j \left[\sigma(\boldsymbol{z}_{ij}^{(k-1)} \boldsymbol{W}_f^{(k-1)} + \boldsymbol{b}_f^{(k-1)}) \odot \phi(\boldsymbol{z}_{ij}^{(k-1)} \boldsymbol{W}_s^{(k-1)} + \boldsymbol{b}_s^{(k-1)})\right] \tag{4.39}$$

transformed. $\odot$ denotes element-wise multiplication, $\boldsymbol{W}$ and $\boldsymbol{b}$ are weights and biases, and $\sigma$ denotes a sigmoid and $\phi$ a nonlinear (here unspecified) activation function.

## 4.7 Comparison of the extrapolation of chemical information for the chemical-transferable potentials, alchemical smooth overlap of atomic positions, and crystal graph convolutional neural networks

This section presents a short theoretical analysis of what chemical information is *assumed* to be known by a model when predicting a compound not seen in the training set, i.e. how chemical similarities between compounds are incorporated into the analytical forms of the models. As an example we consider the energy of a GaN compound to be predicted.

One crucial property of the chemical-transferable potentials (CTP, Sec. 4.4) is the fact that the total energy $E$ of the GaN compound can be decomposed into addends independent of each other that yield energy contributions of Ga-Ga (atoms of type Ga surrounded by neighbours of type Ga), N-N, and Ga-N (a mixture of Ga and N atoms) interactions:

$$E = E_{\text{Ga-Ga}} + E_{\text{Ga-N}} + E_{\text{N-N}}. \tag{4.40}$$

In case of a two- and three-body CTP, $E_{\text{Ga-Ga}} = E_{\text{GaGa}} + E_{\text{GaGaGa}}$ depends on the two- and three-body contribution $E_{\text{GaGa}}$ and $E_{\text{GaGaGa}}$, respectively (see Sec. 4.3.1). Besides the independence of the three terms in Eq. 4.40, a further consequence of the CTP implementation is that for any compound that contains the specific atom type Ga, e.g. GaP, GaAs, or $Ga_2P_4$, the determination of $E_{\text{Ga-Ga}}$ is based on the same chemical relationships in the model. More precisely, the two-body potential $\epsilon_{\text{2b}}(\boldsymbol{q}_{\text{2b}}, (Z_{\text{Ga}}, Z_{\text{Ga}})) = \sum_{\mu}^{N_{\text{basis}}} \alpha_{\mu}(Z_{\text{Ga}}, Z_{\text{Ga}}) b_{\mu}(\boldsymbol{q}_{\text{2b}})$, see Def. 4.4, is based on the same regression coefficients $\alpha_{\mu}$ independent of what other species is present in the compound. The same rule applies to the three-body potential $\epsilon_{\text{3b}}(\boldsymbol{q}_{\text{3b}}, (Z_{\text{Ga}}, Z_{\text{Ga}}, Z_{\text{Ga}}))$. Crucially, if for example a GaN compound is not in the training set, the model *assumes* that it already *knows* the Ga-Ga interaction of GaN if compounds like GaP or GaAs are included in the training set. Moreover, if also compounds containing N atoms (BN, AlN, etc.) are contained in the training set, the only *unknown* (chemically interpolated or extrapolated) part is the interaction between Ga and N atoms. Accordingly, if neither compounds consisting of Ga atoms nor ones with N atoms are included in the training set, the prediction of all three interactions relies on the interpolation/extrapolation along the chemical space. The fact that interactions are only species-tuple and not directly compound dependent is a choice of implementation in this work. Its benefits or limitations are, however, not clear and might be investigated in a future work. An example for a different choice of implementation are, if considering again a GaN compound, coefficients $\alpha_{\mu}$ that are assigned to Ga-Ga interactions however depend on chemical descriptors of both Ga and N.

The alchemical smooth overlap of atomic positions (ASOAP, Sec. 4.5.1) [43] captures chemical information in a conceptually different way. The total energy of a structure modeled by the ASOAP can be decomposed into atomic energies, which each again can be decomposed into measured similarities between the atomic environments of the structure and reference atomic environments in the training set. Considering again the case, that we want to predict the energy of a GaN compound using a training set that contains GaP, one term in the energy of the GaN compound will be given by a similarity measurement between the atomic environment $U_{\mathrm{Ga,GaN}}$ of a Ga atom in GaN and the environment $U_{\mathrm{Ga,GaP}}$ of a Ga atom in GaP. For the sake of clarity, we will denote the environments of the two Ga atoms by $U_{\mathrm{GaN}}$ and $U_{\mathrm{GaP}}$. The ASOAP similarity (Eq. 4.33) of the two Ga atoms is given by:

$$\tilde{k}(U_{\mathrm{GaN}}, U_{\mathrm{GaP}}) = \int dR \Bigg| \int \kappa_{\mathrm{GaGa}} \rho^{\mathrm{Ga}}_{\mathrm{GaN}}(\boldsymbol{r}) \rho^{\mathrm{Ga}}_{\mathrm{GaP}}(\boldsymbol{r})(R\boldsymbol{r}) d\boldsymbol{r} \tag{4.41}$$

$$+ \int \kappa_{\mathrm{GaP}} \rho^{\mathrm{Ga}}_{\mathrm{GaN}}(\boldsymbol{r}) \rho^{\mathrm{P}}_{\mathrm{GaP}}(\boldsymbol{r})(R\boldsymbol{r}) d\boldsymbol{r} \tag{4.42}$$

$$+ \int \kappa_{\mathrm{NGa}} \rho^{\mathrm{N}}_{\mathrm{GaN}}(\boldsymbol{r}) \rho^{\mathrm{Ga}}_{\mathrm{GaP}}(\boldsymbol{r})(R\boldsymbol{r}) d\boldsymbol{r} \tag{4.43}$$

$$+ \int \kappa_{\mathrm{NP}} \rho^{\mathrm{N}}_{\mathrm{GaN}}(\boldsymbol{r}) \rho^{\mathrm{P}}_{\mathrm{GaP}}(\boldsymbol{r})(R\boldsymbol{r}) d\boldsymbol{r} \Bigg|^2 . \tag{4.44}$$

Note that the superscript $\alpha$ in the neighbour densities $\rho^{\alpha}_{U_i}(\boldsymbol{r}) = \sum_{k \in U_i^{\alpha}} \exp(\frac{(\boldsymbol{r}_k - \boldsymbol{r})^2}{2\sigma^2})$ denotes the species type of the neighbours around the (central) Ga atom, e.g. the index $k$ runs only over neighbours of type $\alpha$. Given that the pure Ga-Ga comparison in line 4.41 does not rely on an appropriate choice of the chemical kernel $\kappa_{\alpha\beta}$ (except that $\kappa_{\alpha\beta} = 1$ for $\alpha = \beta$), interactions between Ga atoms can be considered as *known* from the training set. However, although the chemical kernel $\kappa_{\mathrm{GaGa}} = 1$ assigns the maximal possible kernel value to the Ga-Ga comparison and thus the maximum weight in the sum in the above equation, its contribution to the total energy of the GaN compound might still play a minor role among possibly a large number of terms that are based on the similarities between the atomic environments in GaN and the ones in the training set. In contrast, in case of the CTP the energy contribution of Ga-Ga interactions is one of three terms (Eq. 4.40).

A further significant property of the implementation in the ASOAP is the fact that the chemical similarity $\kappa_{\alpha\beta}$ does only compare the neighbours of two atomic environments but not the central atoms themselves. This choice might be a consequence of the concept in the non-alchemical smooth overlap of positions that compares only neighbourhoods with each other. As a result, the ASOAP assigns two atoms of different species types the same atomic energy $\epsilon_i$ if their neighbourhoods are the same. To which extent this choice is a limitation in modeling the total energy $\sum_i^{N_{\mathrm{atoms}}} \epsilon_i$ in a crystal could be investigated in the future.

The crystal graph convolutional neural networks (CGCNN, Sec. 4.6) [42] as used in

this work follows the concept of Ref. [84, 86] initilizing the chemical representation of atom types by random vectors. The chemical information is learned based on the training set using a neural network of multiple layers. While in Ref. [86] it was shown that the random initialization of the atomic features did not (significantly) affect the prediction error of formation energies compared to features based on physical properties of atoms, its limitation in the prediction across chemical composition space is obvious: If neither compounds that consist of Ga nor ones with N are contained in the training set, the prediction of the properties of a GaN compound is random.

# 5 Crystal-structure prediction on sparse data sets: Towards reliable machine-learning models with chemical-transferable potentials

## 5.1 Introduction

First-principles based crystal-structure prediction is the identification of the (meta-)stable crystal structures from (an approximation of) the Born-Oppenheimer potential-energy surface (PES). The potential energy of a collection of atoms is given by the ground-state solution of the electronic Schrödinger equation. The PES describes the potential energy in dependence of the structural arrangement of the atoms and may display a large amount of local minima, the mechanically stable structures. The ultimate goal is to find the global minimum, i.e. the lowest-energy structure or ground-state phase. This task is demanding because the global minimum can only be identified with certainty if all local minima are identified which is typically unfeasible. While not ensuring to find the ground-state phase, structure search algorithms [16–21] have shown to predict new structures that have later been experimentally confirmed [23]. Nevertheless, these algorithms are still limited by the computational cost of the underlying *ab initio* method that determines the PES and the development of surrogate machine-learning (ML) potentials is considered a promising route to accelerate crystal-structure prediction.

The task of predicting the atomization-, formation-, or cohesive energy from the geometry of a material using artificial intelligence has attracted attention during the last decade with a considerable amount of introduced models with accuracies as low as a few meV/atom [42, 72, 73, 84, 88]. However, only some studies [28–31, 39] considered a validation of the models with respect to requirements needed to predict a (meta-)stable crystal structure with a structure-search algorithm that is based on sampling the PES and minimizing the potential energy. Let us first consider one requirement: In a $T = 0$ K approach, the crystal structure has to appear as a minimum (or at least as a saddle point) on the PES. In fact, while we may allow a tolerance for the accuracy on predicting energies of structures, missing a minimum in the right region of the structure space might lead to complete failure of identifying a phase in a crystal-structure search. Fulfilling the mentioned aspect with a ML model is not trivial. Moreover, the validation if it is

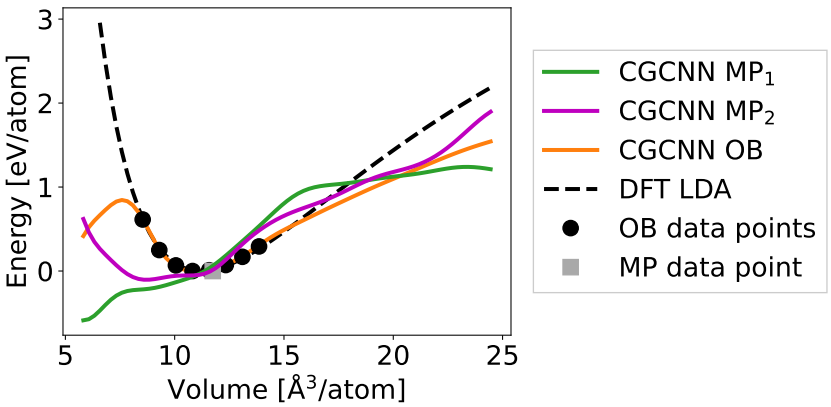



**Figure 5.1:** The energy as a function of the volume is shown for GaN in the ZB phase predicted by different ML models. The presented curves are given by predictions of a crystal graph convolutional neural network (CGCNN) fitted to two different subsets of the materials project (MP) database and to the octet binaries (OB) data set. The data sets $MP_1$ and $MP_2$ consist of 46 744 and 46 747 data points, respectively, differing by three data points: $MP_1$ contains only one GaN compound (in the ZB phase), $MP_2$ consist of further three GaN compounds in other phases. The OB and MP data point(s) represent the respective training data of GaN in the ZB phase. The OB data points and the DFT curve (black dashed line) are obtained from calculations within the LDA while the MP energies are based on the PBE functional. The respective energies are given in relative values to the energy of the reference DFT level at the ZB equilibrium. Note that the LDA- and PBE-equilibrium volume differ by 0.6 Å$^3$/atom.

fulfilled is not necessarily obvious from the analysis of those kind of errors that have typically been used to develop and benchmark ML models for the prediction of the energy from the geometry of a structure: (averaged) errors on predicted energies of a set of given structures [42, 72, 73, 78–81, 84, 88–90]. Crucially, if the analysis stays limited to the evaluation of errors on such selected data points, the quality of the predicted PES outside the data regions or between the given data points (e.g. DFT minima) stays unknown and the ability to perform a successful structure search based on sampling the PES uncertain.

Let us consider an example with a crystal graph convolutional neural network (CGCNN) [42, 86], a state-of-the-art ML model for predicting energies of materials. The model is based on a multi-layer neural network that maps the atomic species types and bond distances in the crystal onto its energy (see theory in Sec. 4.6). We have fitted a CGCNN to the formation energies of the Materials Project (MP) database[1] [9]. The total data that we considered consists of 46 744 relaxed structures (local minima on the PES) covering 87 elements, 7 lattice systems, 216 space groups, and primitive unit cells ranging from 1 to 200 atoms. Splitting the data into 60% training, 20% validation, and 20% test set,

[1]For the model hyperparameters, we used the ones of a pretrained example model published together with the respecitve code [91] by T. Xie and J. Grossman who introduced the CGCNN in Ref. [42, 86]. Note that the MP database is continually updating. To allow a comparison to the model performances of Ref. [42, 86], we used the same subset [91] of the database as in Ref. [42], i.e. 46 744 structures, in a first step.

the CGCNN achieved a mean absolute error of 35 meV/atom for predicting formation energies on the test set (comparible with 39 meV/atom of the original work [42]). Such level of accuracy can be considered as excellent for predicting energies of structures from a large and heterogeneous materials-database. It would be of great value if the model could be used for crystal-structure prediction using a structure search algorithm. However, the found low mean absolute prediction error does not reveal if the model is capable of identifying the structures as local minima on the predicted PES. We analyzed predicted energy-volume curves of 78 octet binary compounds for four cubic phases, i.e. zinc-blende (ZB), rock-salt (RS), CsCl, and NaTl prototype. We observed that 21 out of 312 curves did not exhibit a minimum in the inspected range of volumes. Let us focus on one example: GaN in the ZB phase. Fig. 5.1 shows that no minimum appears over a wide volume range of ZB structures in the predicted PES, see the curve "CGCNN $MP_1$". Accordingly, in a structure-search, the ZB phase of GaN would not be found, at least not in a volume similar to the one of the reference ZB crystal of GaN (grey marker, "MP data point"). Still, the model predicts the formation energy of the reference GaN-ZB structure with an error of only 61 meV/atom. In other words: Despite an accurate prediction of the formation energy of a selected data point, the predicted PES around the data point can be unphysical. Note that the GaN-ZB structure (at the ZB equilibrium) of the MP database was included in the training set. Moreover, no other GaN structure was present in the data but 89 other binary compounds in the ZB phase.

One origin of the failure in predicting a physical energy-volume shape might be the fact that the training data does not reveal information about the shape of the PES both in regions that are *near* the relaxed ZB structure (e.g. the same phase) and ones that are *far* from it (e.g. other phases). By adding three further GaN structures, in the RS and wurtzite prototype structure and a hexagonal layered structure, to the training set[2], a minimum appears in the ZB constrained energy-volume curve in Fig. 5.1 ("CGCNN $MP_2$"). However, the minimum is found at a volume of 8.7 $Å^3$/atom, while at a volume of 10.6 $Å^3$/atom the curve yields a saddle point. In comparison, the reference equilibrium volume (grey marker) is given by 11.7 $Å^3$/atom. We have trained a further CGCNN model replacing the Materials Project database by a training data set of 78 octet binaries (OB) each in eight different polymorphs with a total number of 4 840 data points [63], including GaN in the ZB phase, with data points close around the equilibria. We found that training a new CGCNN on the OB data set gives an improved prediction of the ZB energy-volume dependence of GaN close around the reference minimum. Still, the PES at small volumes ($V < 8$ $Å^3$/atom) completely misses the repulsive behaviour ($\frac{\partial E}{\partial V} < 0$), but rather predicts a completely unphysical minimum at 4.8 $Å^3$/atom, which lies 0.074 eV/atom lower than the correctly predicted minimum at 10.8 $Å^3$/atom. The additional unphysical minimum is a typical characteristics of PESs predicted by ML models currently in the literature. We will call this characteristics in this work an *artefact* outside of the training region of a model.

[2]The three structures are also taken from the Materials Project database but were not included in the first data subset, i.e the one of Ref. [42].

If some preknowledge about the compound exists, one might define beforehand the volume range in which GaN forms its crystals, and remove the artefact from the crystal-structure search. In fact, we will present a scheme in this work to estimate the *realistic* ranges of a compound. The range presented in Fig. 5.1 is the result of our estimator. However, the presence of unphysical minima with low energies on the predicted PES hints at a further potential risk of applying ML models to a crystal-structure search, a problem that we have not discussed yet. The fact that minima on the DFT PES appear also as minima on the ML PES is not a sufficient condition for the correct prediction of the ground-state phase. For instance, if an artefact minimum is (significantly) lower than all other minima on the predicted PES, the identification of the DFT ground-state phase as the most-stable phase on the ML PES has failed. Basically, the tendency of ML models to predict artefacts with possibly low energies and the resulting failure in predicting the DFT ground-state phase as the most stable structure is a central observation of this work. This tendency as well as the demonstrated failure of identifying the ZB phase of GaN highlights the need for examinations of ML models that go beyond the typical energy errors on predefined sets of structures to reveal the reliability of the ML approaches for an application in crystal-structure search[3]. This chapter puts a focus on understanding when ML models (those ones that map the geometry to the potential energy) fail in a structure-search scenario. In particular, it considers the challenging task of constructing robust models from sparse training data sets. This task is complementary to improving ML methodologies by using error estimates and dynamic data-construction or active learning techniques [28, 39, 93].

Recent publications by the group of G. Csanyi have critically examined Gaussian-Approximation-Potential-based ML models [32] for sampling the PES in order to predict (meta-)stable crystal structures [27, 29–31, 39]. The potentials were fitted to large data sets that, in contrast to the CGCNN applications above, included only a single species type, i.e. C, Si, B , P, or Ti. The models are based on a hierarchical combination of $n$-body potentials (explicitly in the additive form as introduced in Def. 4.3). Most of the physics is described by a two-body, a smaller amount by a three-body, and the rest by a many-body potential[4]. A model dominated by an appropriate two-body term might predict the repulsion at small volumes, which was missed by the CGCNN models trained by us, and reduce the risk of artefacts, at least at small volumes. The inclusion of a two-body potential (characterized by a *strong* repulsion) into the model is, in some sense, a step towards defining

[3] At this point, we would like to note that we do not expect that the missing minimum or repulsive behavior on the predicted PESs of the CGCNN models is specific to the CGCNN but rather to ML models, in general, if they are trained on the data sets introduced above. In principle, as our tests indicate, the shape of the PES is strongly influenced by the choice of the data set and not only dependent on the type of the ML model. However, the appearance of multiple minima within a single cubic phase is a typical feature of the CGCNN PESs (at least at the chosen hyperparameters, see for example Fig. F.6 in Appendix) that we have observed in the cubic phases of the octet binaries indicating that the PES also outside of the cubic-symmetry-constrained regions is unphysical. More significantly, the CGCNN is unphysically rough in the PES and, therefore, not applicable to a gradient (forces) based crystal-structure search (more discussion on that will follow in this chapter).

[4] An exception is the work in Ref. [29] which used only a two- and a many-body term.

physically motivated constraints. However, due to the fact that current $n$-body potentials (as in Def. 4.3) depend only parametrically on the species types in a compositions, a large amount of new quantum-mechanical calculations is required for new compositions that include species tuples that were not, or only sparsely, contained in the training set.

In this chapter, we introduce a novel scheme to transfer explicit $n$-body potentials across the chemical composition space using a neural-network based approach. We call these models *chemical-transferable potentials* (CTP). We investigate them for two- and three-body terms on the OB data, which consist of 4 840 data points with 78 compounds in eight different polymorphs. Including two further state-of-the-art ML methods, the alchemical smooth overlap of atomic positions (ASOAP) [43, 67] and CGCNN, that are capable of modeling across both structure and chemical composition space, we explore to which extent the PES of a composition not seen in the training set can be reproduced. This task has, to the best of our knowledge, not been investigated yet. We highlight the risk of model artefacts and analyze the reliability of the models specifically for a structure search, going beyond single-point energy predictions: we perform symmetry-constrained crystal-structure searches (at least for the cubic phases) and an unbiased global structure search. In particular, we investigate to which extent models that are built on sparse training data are able to predict the PES of a composition with an accuracy that allows to identify the ground-state phase in a crystal-structure search. In contrast to the typical demand of ML potentials to have an accuracy of a few meV (on dense data), we aim at coarse predictions that, however, allow to identify the most stable structure. Our goal is a ML-based framework, that predicts a set of *promising* phases. The set consists of the predicted lowest-energy phase plus phases whose energy difference to the lowest phase is below a tolerance of 0.1 eV/atom. In the second step, all promising phases are tested with DFT calculations. Such a framework could greatly reduce the number of DFT calculations in a crystal-structure search.

In Sec. 5.2, we present a short summary of the theoretical framework behind the CTP. Before critically and extensively analyzing the chemical-transfer approach in Sec. 5.4, we validate the structural parts of the CTP, i.e. two- and three-body potentials, on a few selected key materials science problems that depend on compounds containing only a single tuple of atomic species, e.g. a single chemical formula, in Sec. 5.3. We, then, show how the chemical-transfer approach stabilizes the potentials that are specific to a single chemical formula (Sec. 5.4.5). Sec. 5.4.6 presents tests that evaluate to which extent specific regions of the PES of a composition that was not seen in the training set can be predicted reliably for a (constrained) crystal-structure search, for any of the 78 octet binary compounds, i.e. in a cross validation where every compound is once left out from the training set. In this test, also the CGCNN and ASOAP are included. In Sec. 5.4.7, we demonstrate and analyze a global crystal-structure search performing a random-structure search [21, 92] for GaN with CTP, CGCNN, and ASOAP models trained on all octet binary compounds of the OB data set but GaN. As as comparison, we

perform the random-structure search also with DFT. Furthermore, we introduce a novel idea on extracting an interpretable materials descriptor to detect for which materials the ML models failed in predicting the ground-state phase (Sec. 5.4.8). This idea presents a potential scheme to estimate the domain of applicability of the ML models in the chemical composition space, on the one hand. On the other hand, it opens up the route to understand which conceptual feature the ML models miss such that they can be improved.

## 5.2 Theoretical framework

This section presents a short summary of the theoretical framework behind the CTP. More details can be found in Chapter 4.
We consider interatomic potentials with the two fundamental properties that the potential energy $E$ is written in a sum of atomic contributions $\epsilon_i$ and that only interactions between atoms whose distance $r_{ij}$ is below a certain cutoff $r_{\mathrm{cut}}$ are taken into account:

$$E = \sum_{i}^{N_{\mathrm{atoms}}} \epsilon_i(\{\boldsymbol{r}_{ij}\}_{r_{ij}<r_{\mathrm{cut}}}). \tag{5.1}$$

We decompose the interaction orders within the so-called *many-body expansion* that expresses the local energies $\epsilon_i$ formally through

$$\epsilon_i = \epsilon_{1\mathrm{b}} + \sum_j \epsilon_{2\mathrm{b}}(q_{ij}) + \sum_{j,k} \epsilon_{3\mathrm{b}}(\boldsymbol{q}_{ijk}), \tag{5.2}$$

if truncated at third order, and use $q_{ij} = r_{ij}$ and $\boldsymbol{q}_{ijk} = (r_{ij}, r_{ik}, r_{jk})$ as a two- and three-body descriptor, respectively. The specification of appropriate functional forms for the $n$-body contributions $\epsilon_{n\mathrm{b}}$ is a crucial step in the potential design. One choice is given by a basis set expansion with linear regression coefficients $\alpha_\mu$ and basis functions like polynomials or, as used in this work, Gaussians [32]:

$$\epsilon_{n\mathrm{b}}(\boldsymbol{q}) = \sum_{\mu}^{N_{\mathrm{basis}}} \alpha_\mu \exp(-\frac{(\boldsymbol{q}-\boldsymbol{r}_\mu)^2}{2\sigma_\mu^2}). \tag{5.3}$$

The Gaussian centers $\boldsymbol{r}_\mu$ and widths $\sigma_\mu$ are hyperparameters. They should be chosen such that the Gaussians overlap sufficiently and *fill* the descriptor space. However, their number is a factor of computational expense and we will, furthermore, show in Sec. 5.5 that their choice is not trivial. In a multi-species system, we model every species-tuple interaction with a different function $\epsilon_{n\mathrm{b}}(\boldsymbol{q})$ (more details in Sec. 4.3.1) where the terms are distinguished only by the different regression coefficients $\alpha_\mu$. Then, the potentials $\epsilon_{n\mathrm{b}}(\boldsymbol{q}, \{\boldsymbol{Z}\})$ depend explicitly on the structural environment and parametrically on the

species tuples. Here, we represent the species tuples by the vector of the corresponding atomic numbers $\boldsymbol{Z}$, i.e. $\boldsymbol{Z} = (Z_i, Z_j)$ or $\boldsymbol{Z} = (Z_i, Z_j, Z_k)$ depending on the order of the potential, and the parametrical dependence is highlighted by the brackets {}. The potentials become chemically transferable if the atomic numbers enter the potentials explicitly as variables realized by transforming the coefficients to functions of the atomic numbers:

$$\epsilon_{n\mathrm{b}}(\boldsymbol{q}, \boldsymbol{Z}) = \sum_{\mu}^{N_{\mathrm{basis}}} \alpha_{\mu}(\boldsymbol{Z}) \exp(-\frac{(\boldsymbol{q} - \boldsymbol{r}_{\mu})^2}{2\sigma_{\mu}^2}). \tag{5.4}$$

We introduce a chemical descriptor $\boldsymbol{d}(\boldsymbol{Z})$ of the species types based on atomic information such as atomic number, group and row in the periodic table, ionization potential, electron affinity and orbital based radii, and map the descriptor via a neural network with one hidden layer onto the coefficients:

$$\alpha_{\mu}(\boldsymbol{Z}) = \sum_{h}^{N_{\mathrm{neurons}}} W_{\mu,h}^{(2)} \phi \left[ \sum_{j}^{N_d} W_{h,j}^{(1)} d_j(\boldsymbol{Z}) + b_j^{(1)} \right] + b_{\mu}^{(2)} \tag{5.5}$$

Note that symmetries with respect to exchanging atoms need to be incorporated into the descriptor (see. Eq. 4.15 - 4.17). Furthermore, in case of the two-body potential we constrain the signs of the coefficients as described in Sec. 4.4.3. The weights of the neural network are trained using backpropagation and stochastic gradient descent (see Sec. 4.4.2). We provide our implementation of the CTP on GitHub [94].

## 5.3 Validation of the potential form for fixed compositions

The local-energy approximation (Eq. 5.1) bears a limit on the accuracy of the potential [32], the three-body descriptor in Eq. 5.2 does not ensure a unique description of the structural environment [75], and the set of basis functions in Eq. 5.3 do not span the complete function space. Therefore, the choice of constraints put on the potentials need to be validated for the systems to be investigated. Before applying the chemical-transfer approach, we have tested the described potential forms above for three applications: a) the prediction of two key properties, e.g. band gap and formation energies, relevant for optoelectronic applications considering $(\mathrm{Al}_x\mathrm{Ga}_y\mathrm{In}_z)_2\mathrm{O}_3$ compounds, b) the description of silicon in nine phases including an analysis if a phase not seen in the training set can be predicted correctly, and c) the prediction of the transition temperature of $\mathrm{ZrO}_2$ between two phases, a quantitiy that can involve millions of calculations.
Note that the total number of regression coefficients and therefore the complexity of the model increases with the number of species types in the system (see Sec. 4.3.1)[5]. We

[5]Nevertheless, it is not clear to which extent also the complexity of the PES to be described increases with the number of species. An investigation if, for example, a three-body potential can describe most

| Ranking | Model type | Band gap energy | | Formation energy | |
|---|---|---|---|---|---|
| | | RMSLE | MAE [meV] | RMSLE | MAE [meV/cation] |
| 1st | $n$-gram+KRR | 0.077 | 114 | 0.021 | 15 |
| 2nd | c/BOP+LGBM | 0.081 | 93 | 0.022 | 15 |
| 3rd | SOAP+NN | 0.081 | 98 | 0.021 | 13 |
| | **This work** | **0.079** | **104** | **0.021** | **13** |

**Table 5.1:** Prediction errors of the three winning models in the NOMAD 2018 Kaggle competition and the potential of this work. Shown are the root mean square log error (RMSLE) and mean absolute error.

follow the route of varying hyperparameters (cutoff, and number, widths and centers of the Gaussians) with the data to be described targeting the most simple expression the investigated system allows. The selected hyperparameters are listed in App. D. Note that when modeling the thermodynamics $ZrO_2$ the implemention of the QUIP package [69] was used while the rest of the work was performed with our own implementation[6].

### 5.3.1 The NOMAD 2018 Kaggle competition

The NOMAD 2018 Kaggle competition [72] introduced a data-analytics challenge aiming to find the best ML method that can predict the band gap and formation energies of $(Al_xGa_yIn_z)_2O_3$ compounds. Among the 883 submitted models, the winner was based on a crystal-graph representation called $n$-grams not used in materials science before. The data set consist of 2 400 training and 600 test points with structures from six different crystal space groups and unit cells ranging from 10 to 80 atoms.
We found that a potential with a basis set of Gaussians all centered at zero ($\boldsymbol{r}_\mu = 0$) in both the two-body and three-body part achieved an accuracy similar to the ones of the top models. In fact, in the challenge, it would have ranked second with a root-mean-square-log error (RMSLE) of 0.079 on the band gap and 0.021 on the formation energies, right behind the $n$-gram model with 0.077 and 0.021 (Tab. 5.1). Note that while the mean absolute error (MAE) is reported in Tab. 5.1, only the average of the RMSLE determined the ranking metric[7]. Furthermore, note that extending our two- and three-body potential to the prediction of quantities beyond the potential energy (e.g. band gap energy) is technically straightforward as long as to every structural input a scalar output is assigned.

of the solid phases of a ternary while being insufficient for an elemental is to our best knowledge missing in the literature.

[6] Note that after the symmetrization, QUIP implements a normalization of the three-body terms, which modifies Eq. 5.2.

[7] The purpose of using the RMSLE as a metric was to enable the comparison between two quantities with different ranges of values.

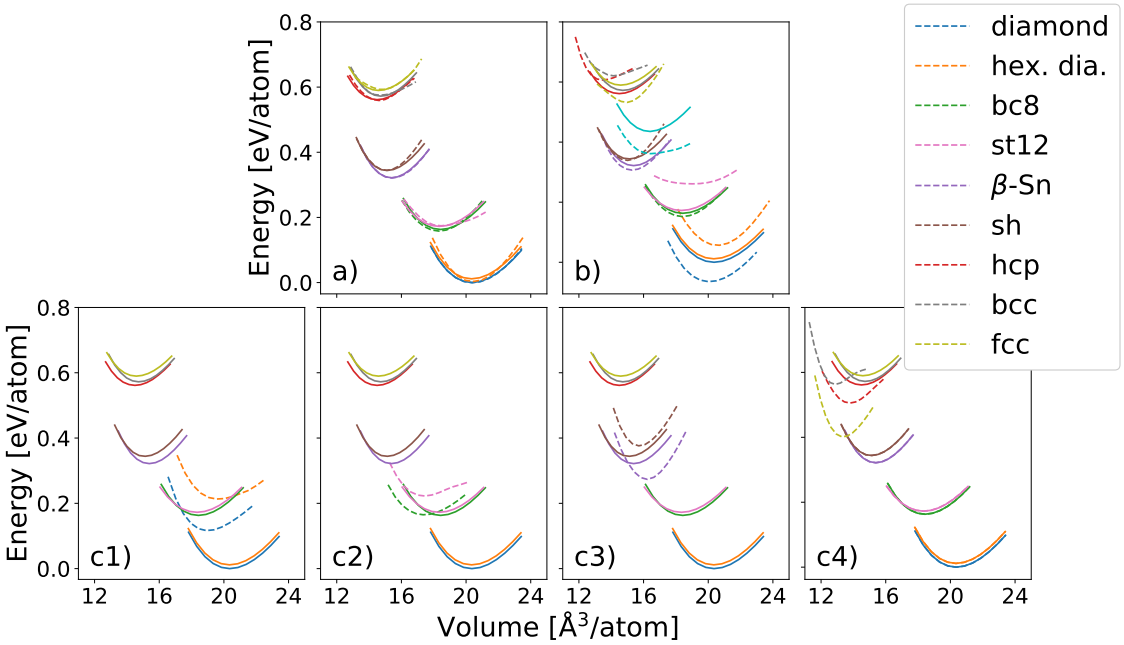



**Figure 5.2:** Energy-volume curves for nine phases of silicon. The solid lines represent the reference DFT energies calculated in Ref. [29] by A. P. Bartok *et al.* using the PW91 exchange-correlation functional [102]. The dashed lines represent the predictions of our ML potential. The curves are obtained by varying the volume and performing at each volume a constant-volume and fixed-crystal-symmetry relaxation. a) shows the predictions of a potential fitted to data of all nine phases. In b) each phase was once left out from the training set and predicted from a potential trained on the remaining phases. Furthermore, the prediction of a second hcp phase (hcp') was added using a potential trained on the initial nine phases. In c1) - c4) the nine phases are divided, based on volume and energy similarity, into four groups: (diamond, hex. dia.), (bc8, st12), ($\beta$-Sn, sh), and (hcp, bcc, fcc). Each group is once left out into a test set and predicted by a potential trained on the remaining phases. The predicted values for the cohesive energy, volume, bulk modulus, derivative of the bulk modulus, and the diamond-to-$\beta$-Sn transition pressure are listed in Tab. E.2 - E.4.

### 5.3.2 Phases of silicon

The history of analytical interatomic potentials is probably most shaped by the description of silicon [64, 95–100]. One basic benchmark for a newly developed potential has often been the performance of predicting the different bulk phases of silicon. We have fitted a ML potential to the energies, forces, and virials of 1 451 structures from nine different bulk phases including low-energy and high-pressure phases. The data was taken from Ref. [29]. The structure types cover the diamond, hexagonal diamond (hex. dia.), $\beta$-Sn, simple hexagonal (sh), bc8, st12, hexagonal close packed (hcp), body-centered cubic (bcc), and face-centered cubic (fcc) phase. For each phase, an energy-volume curve is calculated by varying the volume and performing at each volume a constant-volume and fixed-crystal-symmetry relaxation.

Our potential was able to predict the equilibrium energies and volumes of the nine phases with a MAE of 2 meV/atom and 0.1 Å$^3$/atom, respectively (Fig. 5.2 a)). In contrast,

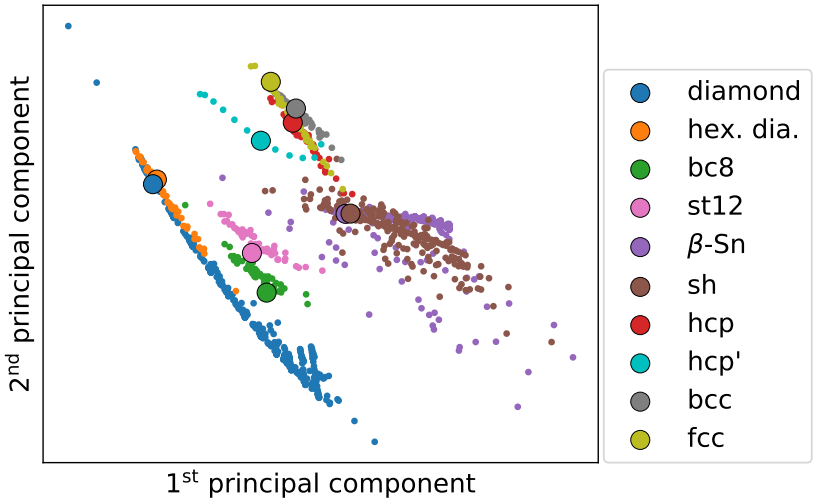



**Figure 5.3:** The two leading principal components of the structural descriptors used to fit 1 451 silicon structures in nine phases. Further twelve structures in the second hcp phase are added. The big edged markers represent the equilibrium structures. For more information about the principal component analysis, including a figure with the three leading principal components, see App. E.1.1.

many other non-ML models, e.g. well-known empirical analytical potentials including ReaxFF [100, 101], the Stillinger-Weber potential [64], or the modified embedded atom method [98, 99], were not able to predict the two equilibrium properties of some phases at that accuracy, when performing the same test, see Ref. [29].

To analyze to which extent the accuracy of the ML potential in predicting the energy-volume curve of a phase is influenced if that phase is excluded from the training set, we have performed a leave-one-phase-out cross validation: each phase is once left out into a test set and its energy-volume curve is predicted by a potential fitted to the remaining phases. In addition, also the prediction of a second phase in the hcp crystal symmetry not in the initial data set was included into the test. The second hcp phase (hcp') is distinguished from the first one by its smaller lattice parameter ratio $c/a$. The potentials achieved a MAE of 45 meV/atom and 0.4 Å$^3$/atom for equilibirium energies and volumes, see Fig. 5.2 b). In contrast to the reference ML potential in Ref. [29], a two-body- and SOAP-based Gaussian approximation potential (GAP), which did not include information about the hcp' phase either, our potential was not able to predict the equilibrium energy and volume of the hcp' phase at the same accuracy, i.e. the errors on predicting the equilibrium energy are 1 meV/atom and 74 meV/atom for the reference and our model, respectively. Note that the GAP was fitted to a larger database that contains also amorphous bulks, liquids, as well as point, line, and plane defects. We found that if the GAP (using the same potential parameters) is retrained on the same data set as used in this work, i.e. nine bulk phases, the error for predicting the equilibrium energy of the hcp' phase is 121 meV/atom.

The leave-one-phase-out cross validation demonstrates that not seen phases could be

predicted with relatively high accuracy, e.g. the maximum error is 89 meV/atom (for st12) and the energetical ranking was only within the three highest phases hcp, bcc, and fcc wrongly reproduced when one of them was left out of the training set. The prediction of a left out phase at that accuracy was possible because its structural environments are similar to the ones in the training data. A structural descriptor map reveals that each of the nine phases in the initial data set (without the hcp' phase) is similar to at least one other phase in the data set, see Fig. 5.3. For instance, phases that are similar in the energy-volume plots (Fig. 5.2) are also clustered together in the descriptor space (Fig. 5.3). For more information about the distribution of the phases in the structural descriptor space or a figure with the three leading principal components, see App. E.1.1.
In order to investigate if a phase can also be predicted from phases that are less similar to it, we have grouped the nine phases according to their similarity and performed a leave-one-group-out cross validation. The groups are given by: (diamond, hexagonal diamond), (bc8, st12), ($\beta$-Sn, sh), and (hcp, bcc, fcc). Due to its relatively isolated location in Fig. 5.3, we consider the hcp' phase as a separate group which was already predicted in the test before. The leave-one-group-out cross validation yields a MAE of 78 meV/atom and 1.1 Å$^3$/atom for equilibrium energies and volumes, see Fig. 5.2 c1) - c4). While the errors have approximately doubled when comparing to the leave-one-phase-out test, the energetical ranking is still roughly kept for the groups. For instance, in a ground state structure search the diamond phase would have been identified correctly or phases and higher energy (hcp, bcc, fcc) would have been identified as less stable ones.

However, the increase of the errors when removing phases out of the training set that are similar to the ones in the test set (i.e. going from test b) to c)) highlights the requirement for a reliable technique to estimate similarities of data points. A reliable error-estimator is a key component of an active-learning framework that explores the descriptor space to adapt the domain of applicability of the model to regions in the structure space relevant for the targeted application. The development of such a technique is, nevertheless, a challenge: diamond and hexagonal diamond are predicted to have a much higher volume and energy difference than DFT reveals (97 meV/atom vs. 12 meV/atom for the energy difference) although being located approximately on the same spot in the structure map. For instance, the structural environments of the two phases differ only by a few neighbour atoms (Fig. E.2). More information about the comparison between diamond and hexagonal diamond can be found in App. E.1.2.
In App. E.1.3, we have performed the tests described in Fig. 5.2 for the GAP of Ref. [29] and tabulated the errors including the ones of our potential, i.e. we have retrained the GAP on the data sets of our tests using the reference GAP parameters. Despite comparable accuracy, e.g. 78 meV/atom for our potential vs 85 meV/atom for GAP in predicting equilibrium energies in test c), the predictions exhibit qualitative differences: the reference potential is able to better recognize similar structures as similar, e.g. the MAE for the energy differences within the groups in test c) is 12 meV/atom for GAP and 78 meV/atom

for the potential of this work[8]. Nevertheless, in test b) GAP mispredicts the volume of diamond by 1.8 Å$^3$/atom although containing information of the similar hexagonal diamond phase.
The reasons behind the deviations of the predictions for similar phases is not clear. In general, (extensive) studies that investigate to which extent ML potentials can predict phases not seen in the training set are currently missing. A work that evaluates different ML methods on different compounds/data sets and yields insights into the influence of the mathematical terms on limitations in the prediction performances is urgently required to develop reliable ML potentials for the exploration of the structure space.
Note that in our tests, we have focused on target quantities that are related to crystal-structure prediction at $T = 0$ K. The reliability of the models for predicting other materials properties like thermodynamic quantities within these leave-some-materials out tests needs to be investigated in future studies as well.

### 5.3.3 Thermodynamics of $ZrO_2$

The computational cost of methods based on DFT is a limiting factor for the prediction of quantities that are based on the evaluation of thermodynamic averages. In this section, we overcome this limitation by building a ML potential from a small number of DFT calculations performed with a hybrid exchange-correlation functional. In particular, we demonstrate how the monoclinic-tetragonal transition temperature of zirconia can be obtained from thermodynamic integration by performing millions of energy and force evaluations with the ML potential. Moreover, we show that the ML potential is able to reproduce the ferroelastic switching in zirconia recently predicted with DFT [103].

To determine the Helmholtz free energy $F_{\mathrm{ML}}$ with our ML potential at a certain temperature and phase, we evaluate the free-energy difference to a reference system of which the free energy can be calculated (analytically), i.e. the harmonic potential $F_{\mathrm{H}}$, using thermodynamic integration [105]:

$$F_{\mathrm{ML}} - F_{\mathrm{H}} = \int_0^1 < U_{\mathrm{ML}} - U_{\mathrm{H}} >_\lambda \, d\lambda. \tag{5.6}$$

The ensemble average $< \cdot >_\lambda$ is built on configurations sampled using the hybrid potential $U_\lambda = \lambda U_{\mathrm{ML}} + (1-\lambda) U_{\mathrm{H}}$ at a specified $\lambda$. The change of the free energy of a system with the temperature is obtained from an integration along the inverse temperature using [105]

$$\frac{\partial(\beta F_{\mathrm{ML}})}{\partial \beta} =< \mathcal{H}_{\mathrm{ML}} >, \tag{5.7}$$

[8] In case of the group with three phases the energy difference is built with respect to the lowest energy phase determined by DFT.

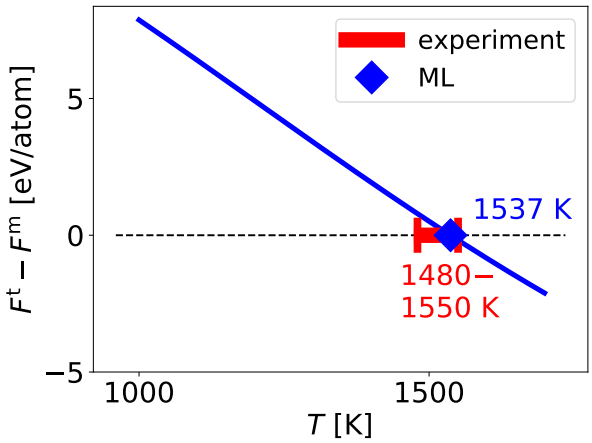
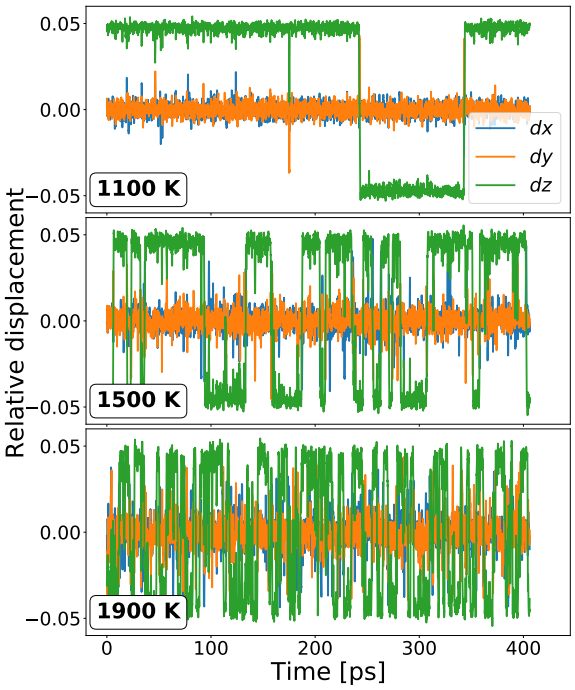



**Figure 5.4:** Predicted free energy difference $F^{\mathrm{t}} - F^{\mathrm{m}}$ between the tetragonal and the monoclinic phase in dependence of the temperature $T$ is shown. The blue marker represents the predicted transition temperature.

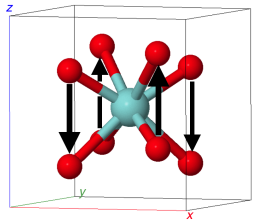
where $\mathcal{H}_{\mathrm{ML}}$ represents the ML Hamiltonian and $\beta = \frac{1}{k_{\mathrm{B}}T}$ is given by the Boltzmann constant $k_{\mathrm{B}}$ and the temperature $T$. The intersection between the free energies of the tetragonal and monoclinic phase, $F^t_{\mathrm{ML}}(T) - F^m_{\mathrm{ML}}(T) = 0$, yields the transitions temperature.

The data set covers 24 000 structures in three phases: monoclinic, tetragonal, and cubic (see Fig. E.4 in appendix). The configurations are outcomes of molecular-dynamics simulations in a 96-atoms supercell at different temperatures (Fig. E.4) using the PBEsol functional, light settings, default tier with additional $f$ functions on the oxygen, NVT molecular dynamics with stochastic velocity rescaling [108] and 2-fs time steps, as implemented in the FHI-aims code [109]. The trajectories in the cubic and tetragonal phase were calculated in Ref. [103, 110] and downloaded from the NOMAD Repository. The simulations in the monoclinic phase were performed in this work.
The data set for training the ML potentials consists of 350 data points where 200 were randomly selected from the molecular-dynamics trajectories and 150 data points were taken from the cubic-tetragonal PES (Fig. 5.5c). The cubic-tetragonal PES was sampled both once with fixed unit cell and once with optimized lattice vectors, as in Ref. [103]. The 350 training data points were recalculated with the HSE06 functional. We have fitted two ML potentials, one to the PBEsol energies and forces, the other one to HSE06 values. As we used the implementation of Gaussian Approximation Potentials (GAP) [32] in the QUIP package [69], we will term the two potentials GAP(PBEsol) and GAP(HSE06). GAP(PBEsol) was only used for validation on the remaining 23 800 configurations from the molecular-dynamics trajectories that were not included in the training set. All molecular-dynamics simulations performed with GAP(HSE06) were carried out using 96-atoms supercells, 2-fs time steps, and the Langevin thermostat as implemented in the QUIP package.

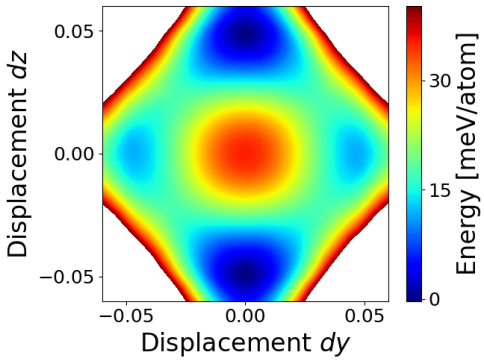
The GAP(PBEsol) predicted the 23 800 PBEsol energies with a mean absolute error




**(a)** Time dependence of the switching.




**(b)** Direction of oxygen displacements.




**(c)** Cubic-tetragonal potential-energy surface.

**Figure 5.5:** Simulation of the ferroelastic switching of $ZrO_2$ in the tetragonal phase. Molecular-dynamics simulations at constant volume and temperature were performed for a 96-atoms supercell in the tetragonal phase deformed along the $z$-axis (a). Accordingly the dominance of oxygen displacements in $z$-direction (b) is observed originated in the unsymmetric cubic-tetragonal potential-energy surface (c). The displacements are given in relative coordinates of the (orthogonal) lattice vectors and averaged over the crystal.

(MAE) of 2 meV/atom. However, the accuracy depends linearly on the simulation temperature (disorder of the structure). The MAEs on structures from simulations run at 300 K, 1 500 K and 2 400 K are 1 meV/atom, 3 meV/atom, and 5 meV/atom for energies and 0.11 eV/Å, 0.27 eV/Å, and 0.37 eV/Å for forces (Fig. E.5b in appendix). The monoclinic-tetragonal energy difference (for relaxed structures) of 26 meV/atoms for GAP(HSE06) is in good agreement with the DFT one of 25 meV/atom.
Using GAP(HSE06), thermodynamic integrations (Eq. 5.6) at 1 000 K were performed to calculate the free energies for the monoclinic and tetragonal phase. By integrating along the inverse temperature (Eq. 5.7), the free energy difference in dependence of the temperature was predicted, as shown in Fig. 5.4. For both phases, the volume was kept fixed at the $T = 0$ K equilibrium. The transition temperature is found at 1 537 K which is in good agreement with the experimental values at 1 480 K - 1 550 K [111, 112]. More details about the presented scheme to predict the monoclinic-tetragonal transition temperature are in App. E.2.1.
GAP(HSE06) was, furthermore, able to reproduce the spontaneous realignment of oxygen atoms in the tetragonal phase (Fig. 5.5), which was predicted in Ref. [103] using DFT. In a molecular-dynamics simulation over 400 ps with fixed tetragonal lattice, switching frequencies of 5 $ns^{-1}$, 55 $ns^{-1}$, and 155 $ns^{-1}$ for 1 100 K, 1 500 K, and 1 900 K were observed. The avarage time (standard deviation) the oxygens stay in a valley is 172 ps (71 ps), 18 ps (16 ps), and 7 ps (5 ps), respectively. Accordingly, at high temperature the outcome of the time avarage over the displacements are oxygen atoms centered between the domains the oxygen atoms align, ($dx = dy = dz = 0$), which becomes after a unit cell optimization the cubic phase measured stable above 2 650 K [113]. At this point we note that we performed 3 000 000 calculations with the ML potential and the speed up with respect to calculations at hybrid-DFT level is given by a factor of 40 000 for the 96-atoms unit cell.

More intensive tests and the extension of the approach to doped zirconia could help to gain insights into the mechanisms that drive the stabilization and toughening in the tetragonal phase, important for the development of advanced thermal barrier coatings.

## 5.4 Prediction across chemical composition space

While the development of purley structural descriptors and models has become a major topic in data-driven materials research, only some studies involve a combination with chemical representations [42, 43, 78, 84]. In 2016, an approach to extend the smooth overlap of atomic positions (SOAP) descriptor by an alchemical kernel (ASOAP) was introduced [43]. The SOAP is maybe the most widley applied many-body descriptor for ML potentials [27, 29, 30, 39, 72, 73, 76, 81, 88, 114]. A conceptually different approach designed to describe several materials properties is given by the crystal graph convolutional neural network (CGCNN) [42], a deep-learning based method demonstrated to predict

the formation energies of large and heterogeneous data sets with an impressive accuracy. Including the ASOAP and the CGCNN into the tests, we investigate to which extent the prediction of the potential-energy surface (PES) of a compound not seen in the training set is possible. To our knowledge, this is a tasks that has not been investigated yet. Moreover, studies examining those kind of models that can predict across both structure and chemical composition space have hitherto only considered errors on predicted energies of a set of provided or known structures. Accordingly, the quality of the predicted PES outside of the considered data regions or between the given data points (e.g. DFT minima) remains unknown. Therefore, the ability to identify stable structures via sampling the PES of the models is uncertain, as highlighted in the introduction of this chapter (Sec. 5.1). We go a significant step beyond the examinations of other studies that considered models capable of predicting across both structure and chemical composition space: we analyze the shapes of the predicted PES (at least for cubic phases) and their potentially negative influence on the success of identifying the ground-state phase of a compound correctly. Moreover, we demonstrate a crystal-structure search for GaN performed with models (CTP, ASOAP, CGCNN) trained on 77 octet binary compounds, GaN not included. GaN is not special, but rather a showcase example to analyze the performance of the ML models.

### 5.4.1 Data set

The prediction across the chemical composition space is investigated on the OB data set. The data set was created within the studies of Ref. [63]. It consists of 4 840 structures and corresponding DFT cohesive energies with 78 octet binary compounds in eight different crystal-structure prototypes: zinc-blende (ZB), rock-salt (RS), CsCl, NaTl, NiAs, CoSn, NbP, CrB. The data is charactarized by eight-point energy-volume curves, as shown in the central panel of Fig. 5.6. An exception is the CrB phase where only one data point per compound is given. For 56 compounds, there are two energy-volume curves present in the CoSn phase, each with exchanged occupation of the sites in the crystal by the atomic species types (AB and BA). In all our tests we will consider CoSn(AB) and CoSn(BA) as two different phases. The data was calculated using the local-spin-density approximation and downloaded from the NOMAD Repository [6]. Further details about the data set, including the reason for the varying number of data points of the phases, can be found in App. C.

In the case of the cubic phases, we have used Birch-Murnaghan fits to extrapolate the enery-volume curves to a wider volume range[9]. The distribution of energy differences is shown in Fig. 5.6, left. The relative energies are built for each compound

[9]The lattice-parameter interval was increased from 5% to 15% of the equilibrium parameter. The convergence tests in Ref. [63] show that the change of the equilibrium energy obtained from the Birch-Murnaghan fits when varying the interval between 1% and 15% is below 1 meV/atom. The reason behind expanding the range is to let the ML models learn better the repulsion and attraction over a wider interval of distances in the crystal as demonstrated in Fig. 5.1 for the CGCNN.

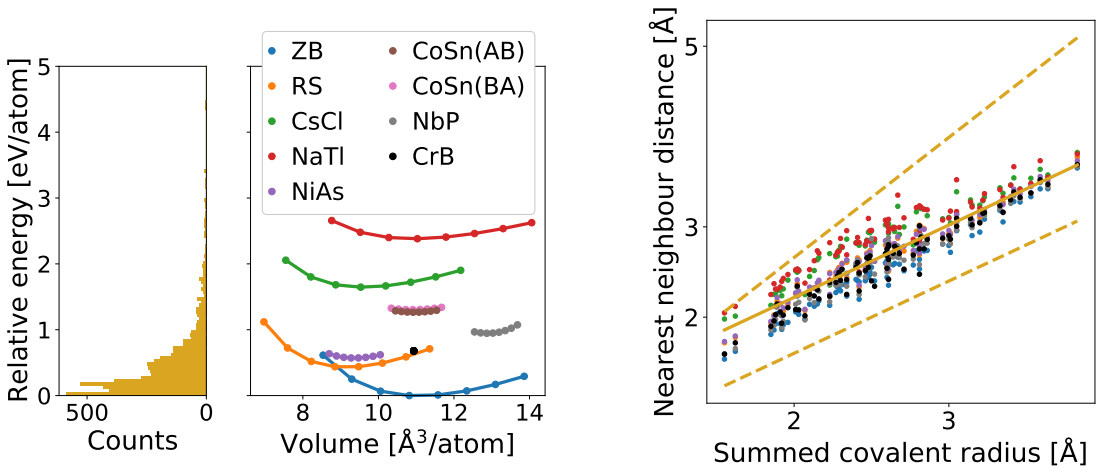



**Figure 5.6:** Distributions in the OB data. The distribution of energy differences is shown left. The energy differences are built with respect to the lowest energy in each compound. In the center, energy-volume curves are shown for GaN. On the right, nearest neighbour distances of the crystals are plotted against the summed covalent radii of the compounds. Only equilibria of the phases are shown. The summed covalent radius is given by $r_{\text{cov}} = (r_{\text{cov}}(A) + r_{\text{cov}}(B))$. The covalent radii are based on the statistical analysis [115] of experimental bond lengths in the Cambridge Structural Database [116]. The solid line represents a linear fit through the data points. The dashed lines define the borders of the search ranges used in our studies for the cubic equilibria, see text in Sec. 5.4.4. A figure with all training data points can be found in Fig. C.1.

with respect to the data point with the lowest energy present for the compound.

### 5.4.2 Choosing the model parameters

The values of all ML-model parameters used in this chapter are tabulated in App. D. For the CTP, we set $\epsilon_{1b}$ in the many-body expansion in Eq. 5.2 to zero. Given that the target property to be learned in this work is the cohesive energy (DFT total energy of a structure minus the DFT energy of the gas-phase atomic constituents), setting $\epsilon_{1b}$ to zero is equivalent to learning the total energy of a structure and setting a different $\epsilon_{1b}$ for every compound where $\epsilon_{1b}$ is determined by the summed energies of the corresponding gas-phase atomic constituents. For both the two-body and three-body part of the CTP, we use a neural network with one hidden layer and 500 neurons (see validation in App. F.1). The weights of the neural networks are trained by using stochastic gradient descent and the backpropagation algorithm, which is a common choice for neural networks.

In Sec. 5.4.6, we include also the crystal graph convolutional neural networks (CGCNN, see theory in Sec. 4.6) and alchemical smooth overlap of positions (ASOAP, see theory in Sec. 4.5.1 and 4.5.2) into the numerical tests. We use a CGCNN code [91] provided by the authors of the studies [42,86] that introduced the CGCNN and an ASOAP code [117]

provided by the authours of Ref. [43, 67] that introduced and investigated the ASOAP. For the CGCNN, we use the hyperparameters of an example pretrained model published together with the respective code. The example model was trained by the authors of Ref. [42, 86] on formation energies of the Materials Project database.
While the CGCNN learns the atomic representations during the training process, in case of the CTP and ASOAP we use atomic properties such as the atomic number, group and row in the periodic table, ionization potential, electron affinity and radii of the valence $s$ and $p$ orbitals where the radial probability density is maximal. With this choice of atomic properties, we improved the ASOAP based on the Pauling electronegativity and van-der-Waals radius as used in previous works for the ASOAP [43, 67, 114][10]. The superior performance of our ASOAP is demonstrated in a leave-one-compound-out cross validation on a smaller data set where only four crystal-structure types were considered (see test results in App. F.2). Moreover, in contrast to the typical cutoff of around 3 Å used in the works that investigated the ASOAP, we used one of 5 Å. The value is a compromise between having a cutoff large enough to include at least the first neighbours shell into the cutoff for all structures in the octet-binaries data set[11] and keeping the cutoff low to maintain the computational efficiency high. Note that we use the averaged kernel (Eq. 4.35) for the comparison of structures.

The weights in the CTP and CGCNN are optimized using a validation set for early stopping. In case of the ASOAP based model, the regression coefficients are determined by kernel ridge regression and the hyperparameters, i.e. the regularization parameter of the ridge model and the Gaussian width in the chemical kernel $\kappa_{\alpha\beta}$ in Eq. 4.33, via a square-grid search using a validation set.
The validation set is given by 10% of the (compound, structure type) tuples in the training set, meaning that if a (compound, structure type) tuple is left out into the validation set all data points in that phase are left out for this compound. When performing the leave-one-compound-out cross validation in Sec. 5.4.6, for all three methods the same validation sets are used.

### 5.4.3 Scaling of the structural CTP parameters

Our experience is that setting the structural parameters (the ones used to describe the PES of a fixed compound) as compound dependent improves the CTP. In practice, we keep the Gaussian widths fixed and scale only the pairwise distances in the two- and three body descriptors by the nearest neighbour distance predicted for the compounds, besides scaling the cutoff radius. This is equivalent to using larger cutoffs and Gaussian widths for compounds that tend to exhibit larger bond lengths and vice versa. The (average) nearest

[10] But only the studies [67, 114] considered the prediction of energies.

[11] A further criterion is that the neighbours within the first shell are not in the smooth decay region of the cutoff function. Such a region is, for example, given by $r_c - r_w < r < r_c$ in Eq. 4.12, where $r_c$ is the cutoff and $r_w$ defines the width of the region.

neighbour distance can be well described by a linear relationship to the summed covalent radii $r_{\text{cov}} = r_{\text{cov}}(A) + r_{\text{cov}}(B)$ of an $AB$ compound[12], as demonstrated by the linear fit in the right panel of Fig. 5.6. The scaling factor $r_{\text{nn}}/r_{\text{ref}}$ for the descriptors of a compound is a ratio of the predicted nearest neighbour distance $r_{\text{nn}}$ and a reference distance $r_{\text{ref}}$ that we set to the average of the nearest neighbour distances over all structures in the training set. A different choice than the covalent radii for estimating the nearest neighbour distance could be given by equilibrium dimer distances or a quantity that depends on orbital based radii, see Fig. C.2.
Using the scaling procedure, the prediction of the PES is, in some sense, decoupled into a prestep that estimates the bond distance in a compound and the actual prediction of the (scaled) PES.

### 5.4.4 Evaluation of the prediction performance

An ultimate goal in data-driven materials design is to predict the ground-state structure of a compound not present in databases. A meaningful framework would be based on predicting a set of *promising* (lowest-energy) structures that are further inspected with DFT. Thus, in our tests we evaluate the ground-state-prediction performance on left out compounds from the training set by validating if the ground-state phase as determined by the reference DFT method is inside a predicted set of *promising* phases. Such a set consists of the predicted lowest-energy phase plus phases whose energy difference to the lowest phase is below a tolerance of 0.1 eV/atom. For the energy assigned to a phase, we consider the energy at the equilibrium of the phase, i.e. obtained from a relaxation. Accordingly, to evaluate the prediction performance of the ML models, we will focus on the accuracy of the predicted energies at the equilibrium structures obtained from relaxations with the ML models. Moreover, we will report errors rather on energy differences than on absolute energies, i.e. the predicted energy differences between the phases are compared to the energy differences as given by DFT. Note that the exact parameters of the equilibrium structures (atomic coordinates and lattice parameters) of DFT and the ones of the models do not necessarily coincide as the equilibria are obtained from relaxations with the respective method. The energy difference is built with respect to the ground-state phase of the corresponding compound where also for the model predictions the ground-state phase determined by DFT is considered. For example, even if the model predicts a different phase than ZB as the ground state for GaN, the energy difference for any GaN data point is built with respect to the predicted equilibrium of the ZB phase, because ZB is the ground-state phase according to DFT. Note that we use the label "ground-state" relative to the considered set of eight phases in the octet-binaries data set, i.e. the label is assigned to the energetically lowest phase inside the set of eight phases. For instance, GaN is more stable in the wurtzite than in the ZB phase according to the reference DFT calculations, but wurtzite is not included in our set of phases. The tests

[12] The covalent radii are taken from Ref. [115]. They are based on the statistical analysis of experimental bond lengths in the Cambridge Structural Database [116].

in the next two sections, Sec. 5.4.5 and Sec. 5.4.6, will focus on this set of phases. In Sec. 5.4.7, we will go beyond the limited set of phases and perform an unbiased global crystal-structure search for the example of GaN.
As the equilibria of the ML potentials can be located outside of the volume ranges considered in the DFT data we perform a lowest-energy search with the models for a wide compound-dependent range of volumes, at least in case of the cubic phases. Such an energy-volume curve is shown in Fig. 5.1. The search is carried out on a grid with 0.2 Å$^3$/atom volume steps. The volume interval is determined by the nearest neighbour distance $0.8r_{\mathrm{cov}} < d_{\mathrm{nn}} < 1.3r_{\mathrm{cov}}$ in the crystal where $r_{\mathrm{cov}} = r_{\mathrm{cov}}(A) + r_{\mathrm{cov}}(B)$ is the the summed covalent radius of an $AB$ compound. The borders for the cubic search ranges are represented by the dashed lines in the right panel of Fig. 5.6. Note that the figure shows the distribution only of equilibirum structures. For the same plot, however, with all training data points, see Fig. C.1. The choice of the cubic search interval is based on the linear correlation of the average covalent radius with the nearest neighbour distance in a compound (Fig. 5.6, right). For instance, the minimum and maximum ratio $d_{\mathrm{nn}}/r_{\mathrm{cov}}$ calculated in the training data set is 0.8 and 1.4. Similarly, in a recent work, the minimum interatomic distance for genarating structures in a random-structure search for silicon was chosen to be 1.7 Å [29], which corresponds to $0.77r_{\mathrm{cov}}$.
For non-cubic phases, an (extensive) search for the phase equilibrium is not performed. For the CrB phase, the single existing DFT structure of a compound is used. For all remaining phases, the energies of all eight DFT structures inside a phase are predicted with the ML model and the energetically lowest is used for each phase.

### 5.4.5 Stabilization of the machine-learning potentials by connecting the chemical space

The OB data set contains only limited information about the interactions between the atoms, e.g. not more than 65 structures are given per compound, the maximum number of atoms in the unit cell is six, and the crystals are highly symmetric (in contrast to the distorted structures of a molecular-dynamics trajectory if more atoms are present in the unit cell). We will show that for a ML model trained on such a sparsely covered structure space without the chemical-transfer (CT) approach, i.e. for every compound an independent potential is fitted, accurately predicting the (equilibrium) energy of a phase not seen in the training set is a challenging task. In contrast, we observed accurate results for the CT approach (Eq. 5.4 and Eq. 5.5) that enabled fitting one single ML model to all compounds at the same time by "connecting" the chemical composition space. For each of the two approaches, with and without CT learning, we have designed a cross-validation test that allows for a comparison between them.

We say "a phase is left out" if all structures that belong to that phase (crystal-structure

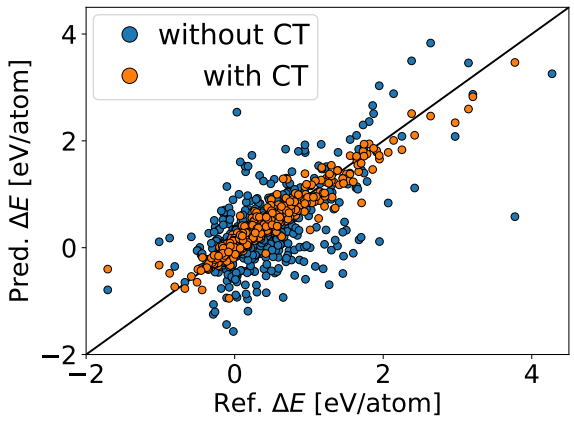



**Figure 5.7:** Predicted vs reference energy differences from cross-validation tests that evaluate the prediction performance of the (equilibrium) energies of phases left out from the training sets. The tests were performed for the chemical-transfer (CT) and non-CT approach.

prototype) are left out of the training set.
In case of the non-CT approach, a separate leave-one-phase-out cross validation for every compound was implemented by leaving one phase out, training the ML model on the structures of the remaining phases (between 48 and 64 data points), searching the equilibrium structure (as described in Sec. 5.4.4) and predicting the corresponding cohesive energy of the left out phase. This procedure is iterated over every phase to be left out and every compound.
For the CT approach, one test with nine cross-validation steps was performed. In each step, for every compound, one phase is left out, where the phase is selected for every compound randomly and may differ among the compounds. This means that a phase that was left out for one compound is present in the training set for many other compounds. The model is trained on the data of all compounds and their phases that were not left out, at the same time. The equilibria of all left out (compound, phase) tuples are searched and the corresponding cohesive energies predicted. Furthermore, we ensure that for every compound every phase was at least once left out in the nine cross-validation steps.
As described in Sec. 5.4.4, we evaluate the errors on the predicted energy differences, i.e. the difference between the predicted energy at the equilibrium (or relaxed) geometry of the left-out phase and the predicted energy at the equilibrium (or relaxed) geometry of the ground-state phase, where the ground-state-phase label is given by the reference DFT method. If the left-out phase is the ground-state phase, we report the energy difference with respect to the second lowest-energy phase[13], again given by DFT. In both cross-validation schemes, the errors are averaged over all compounds and left out phases.

The predicted energy differences for the two methods are compared in Fig. 5.7. In case of

[13]This means that the energy difference, then, becomes negative as opposed to a difference that is built with respect to the ground-state phase.

the non-CT approach, the energy differences of the left out phases were predicted with a mean absolute error (MAE) of 0.35 eV/atom and root mean square error (RMSE) of 0.52 eV/atom. The MAE and RMSE of the CT approach are 0.10 eV/atom and 0.16 eV/atom, respectively. Similar to the outcomes of the multi-task learning approach in Chapter 3 that highlight how a unified model for all tasks (data subsets) improve the prediction accuracy, the results of the CTP demonstrate how the prediction of a left out phase of a compound can benefit from the knowledge of this phase in other compounds.
The MAE and RMSE for the predicted cohesive energies are similar in their magnitudes compared to the ones of the energy differences. In case of the CT approach, the MAE and RMSE on the cohesive energies are given by 0.09 eV/atom and 0.15 eV/atom, in case of the non-CT one by 0.35 eV/atom and 0.52 eV/atom. Note that the relative errors with respect to the distribution of the reference data is higher for the energy differences, i.e. the standard deviation of the DFT energy differences is given by 0.59 eV/atom while the one of the DFT cohesive energies is given by 0.97 eV/atom (compare Fig. F.2 in appendix to Fig. 5.7).

### 5.4.6 Predicting the potential-energy surface of a new compound

We have performed a leave-one-compound-out cross validation for the 78 compounds of the OB data set. At each cross-validation step, the models were trained on the data of 77 compounds and the PES of the left-out compound was predicted for the eight phases as described in Sec. 5.4.4. For all three ML methods, the same training and validation sets were used. A central result of this work is that the PES of a compound can indeed be predicted from other compounds (at least in our data set) with an accuracy that allows to identify the ground-state structure with the constrained structure search described in Sec. 5.4.4. The key quantity with which we evaluate the models in the cross validation is the ground-state-prediction success rate, i.e. the fraction of compounds that had their ground-state phases predicted correctly within a tolerance of 0.1 eV/atom (see description in Sec. 5.4.4).
The ground-state-phase prediction test yielded a success rate of 74/78 for the CTP, 67/78 for the ASOAP, and 68/78 for the CGCNN, see left column in Tab. 5.2. As a comparison, a probabilistic model based on statistical averages of the data (as described in App. F.4) achieves a success rate of 44/78[14]. With a tolerance of 0.11 eV/atom instead of 0.10 eV/atom for determining the set of possible ground states, the CTP would have identified the ground states of two further compounds correctly (total 76/78). In fact, the ground state of one of the two compounds was missed by only 1 meV/atom which highlights the sensitivity of the success rate to the chosen tolerance at the magnitude of

[14]Note that in contrast to a purely random prediction, the probabilistic model uses some information of the data set. As opposed to the investigated ML models, the probabilistic model just predicts the likelihood of a set of promising ground states without the usage of energy-structure or energy-compound relationships.

| | a) LOCOCV | | | b) Ga, N | | | c) Sr, O | | |
|---|---|---|---|---|---|---|---|---|---|
| | $r$ | $\Delta E$ | $V$ | $r$ | $\Delta E$ | $V$ | $r$ | $\Delta E$ | $V$ |
| CTP | 74/78 | 0.25 | 2.0 | 6/7 | 0.64 | 2.2 | 10/10 | 0.25 | 2.2 |
| ASOAP | 67/78 | 0.42 | 2.5 | 6/7 | 1.14 | 3.6 | 7/10 | 0.57 | 3.7 |
| CGCNN | 68/78 | 0.44 | 6.7 | - | - | - | - | - | - |

**Table 5.2:** Prediction results of a) a leave-one-compound-out cross validation (LOCOCV), b) a test in which all compounds that include either Ga, N or both are left out to be predicted, and c) one in in which all compounds with Sr or O are left out. The success rate $r$ shows how many compounds had their ground states predicted correctly within a tolerance of 0.1 eV/atom. $\Delta E$ and $V$ represent the root-mean-square errors for the predicted energy differences and volumes, respectively. $\Delta E$ is given in [eV/atom] and $V$ in [Å$^3$/atom].

0.1 eV/atom[15]. While we consider the specific success rates of the ML methods, despite the gap between CTP and both other methods, as relatively similar we highlight a crucial difference between the predicted PESs: In contrast to the ASOAP and CGCNN, the CTP predicted the cubic surfaces to have one single minimum with higher probability, i.e. an energy-volume curve mainly characterized by a repulsion dominating at small volumes and an attraction at higher volumes (compare to the CTP curve in Fig. 5.1). Only in 4% of the (288 cubic) cases a second minimum was found. In contrast, the ASOAP predicted 23% and the CGCNN 61% of the surfaces with two or even more minima (in the case of CGCNN even up to five), see the right panel of Fig. 5.8. The figures F.4, F.5, and F.6 in the appendix show the qualitative difference of the predicted PESs. The tendency towards predicting PESs with multiple minima already for simple cubic symmetries suggests that overall the predicted PES is unphysical[16]. The fact that unphysical multiple minima in the PES can hinder an (accurate) structure search is also visible in some cases of the CGCNN predictions: if one limits the search to volumes in the surroundings of the DFT minima, the ground states of further five compounds are predicted correctly by the CGCNN and an additional one by the ASOAP. For instance, the predicted minima in the surroundings of the DFT minima are typically accurate in terms of predicted cohesive energy, however, an extended search reveals other minima (artefacts) that are possibly lower in energy. Without the pre-knowledge of the DFT reference, the structure search by CGCNN and ASOAP leads to completely unphysical predictions in some cases.

Next, we will discuss the prediction errors on relative energies. In case of the cubic phases, we will consider errors also on volumes. Recall that the errors are based on predicted properties of geometries that were (at least in case of the cubic phases) identified as the equilibrium structures in the structure search (as described in Sec. 5.4.4). Therefore the DFT and predicted equilibrium geometries can deviate significantly. The left panel of

[15] The sensitivity of the success rate to the tolerance depends on the value of the tolerance, e.g. at a low tolerance small changes of the tolerance affect the success rate more strongly than at a high tolerance.

[16] We will show in Sec. 5.4.7 that the ASOAP fails in a random-structure search for GaN, while the CTP was able to identify the most stable structures of GaN correctly. There are no results for the CGCNN in this test because the model is not even smooth in the PES.

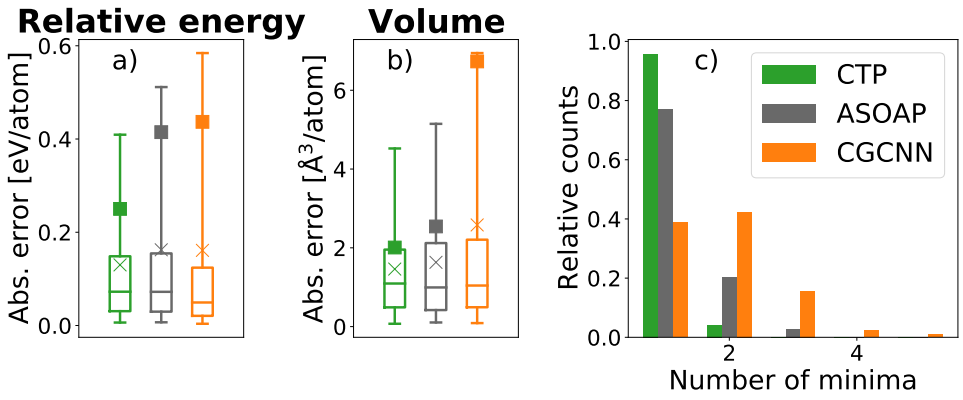



**Figure 5.8:** Prediction results of a leave-one-compound-out cross validation. a) and b) show box plots for the absolute errors of the predicted energy differences and volumes. The errors on the volumes are evaluated only for the cubic phases. The box plots mark the 25th and 75th percentiles (extrema of the rectangle), the 5th and 95th percentiles (extrema of the "whiskers"), and the median (horizontal line inside the rectangle). Shown are also the mean absolute error (MAE, cross) and the root mean square error (RMSE, solid square). c) shows the number of predicted minima inside the inspected volume ranges for the cubic phases.

Fig. 5.8 shows that the three methods achieved comparable mean absolute errors in predicting relative energies, e.g. 0.16 eV/atom, 0.16 eV/atom and 0.13 eV/atom for CGCNN, ASOAP and CTP, respectively. While the CGCNN predicted most of the relative energies with a lower error than the ASOAP and CTP, i.e. the median and the 75th percentile are lower (e.g. 0.12 eV/atom vs 0.15 eV/atom and 0.15 eV/atom, respectively for the 75th percentile), the root mean square error (RMSE) lies higher, e.g. 0.44 eV/atom vs 0.42 eV/atom and 0.25 eV/atom. In case of the CGCNN and ASOAP a large contribution to the higher RMSE originates in the fact that artefacts were identified as equilibria of the cubic phases. If again only the ones close to the DFT volume ranges are considered the RMSEs decrease to 0.34 eV/atom and 0.36 eV/atom for the CGCNN and ASOAP, respectively. The fact that the minima were identified wrongly emerges, in case of the the CGCNN significantly, also in the higher RMSE on predicted equilibrium volumes for the cubic phases (central panel of Fig. 5.8), e.g. 6.7 Å/atom for CGCNN, 2.5 Å/atom for ASOAP, and 2.0 Å/atom for CTP. Note that a dependence of the absolute errors on the percentiles is shown in Fig. F.3 in the appendix. We will analyze possible origins of the model-prediction errors in Sec. 5.4.8.

It is expected that the accuracy of predicting the PES of a new compound (not seen in the training data) depends on the materials distribution in the training data, e.g. if there are compounds in the training data that are similar to the new compound to be predicted. We would like to find out to which extent specifically the fact that atomic constituents of the new compound are also present in compounds of the training data supports the accurate prediction of the new PES. Let us consider the example in which GaN (with

all its structures) is left out from the training set to be predicted. When considering the CTP, the atomic energy contribution to the total energy of a GaN structure from a Ga atom surrounded by Ga atoms, or a N atom surrounded by N neighbours, is already modeled in compounds from the training set, such as in GaP and GaAs or BN and AlP (see Sec. 4.7). In the following, we test the ability of the models to predict a PES if such information is not available. For this, we leave all compounds that contain either Ga, N, or both atom types out from the training set and predict their PESs. We repeat the same test for compounds that contain either Sr, O, or both atom types. Note that this test is performed only for the CTP and ASOAP, because the CGCNN used in this work is based on randomly initialized atomic representations, i.e. there is no physical connection between those CGCNN-input vectors that represent the atomic species types. Accordingly if neither Ga nor N is present in the training data, the choice of their representation vectors will not have an influence on the training procedure but on the prediction of the GaN PES. In other words, different (random) choices of the Ga and N representation vectors will lead to different random predictions of the GaN PES. An analysis about how specific chemical information is incorporated into the models and a discussion to which extent a left-out compound is considered as extrapolated is presented in Sec. 4.7.
We found that the ground states of six of seven compounds that include Ga or N could be identified correctly by both the CTP and ASOAP, see central column in Tab. 5.2. For compounds including Sr or O the success rates are 10/10 and 7/10 for the CTP and ASOAP, respectively. Interestingly, the RMSE for the predicted relative energies does not differ significantly from the one of the leave-one-compound-out cross validation for the same compounds (see Tab. F.2). For compounds that include Ga or N atoms the error has not changed with respect to the leave-one-compound-out cross validation in case of the CTP, i.e. it stayed 0.64 eV/atom, while for the ASOAP it has increased only slightly from 1.10 eV/atom to 1.14 eV/atom. For compounds containing Sr or O atoms, the values changed from 0.23 eV/atom to 0.25 eV/atom and 0.50 eV/atom to 0.57 eV/atom for the CTP and ASOAP, respectively. When comparing the ground-state-prediction success rates of the tests where information about Ga and N or Sr and O was missing in the training set to the success rates of the leave-one-compound-out cross validation, again for the same compounds, the success rates have changed only for the CTP, i.e. from 9/10 to 10/10 for compounds that include Sr or O, where the ground state of the compound BeO is, now, predicted correctly. Note that the ground state of BeO was predicted also by the ASOAP wrongly in both the leave-one-compound-out cross validation and the test with excluded compounds that consist of Sr or O atoms. The improvement of the CTP in predicting the ground state of BeO correctly is surprising because not only the size of the training data set has decreased but specifically compounds that include Be or O atoms were excluded. In contrast, the prediction of AlN energies suffered from the missing information about N atoms, i.e the phase CsCl was identified to be most stable with 0.25 eV/atom lower than the true ground-state phase ZB. In the leave-one-compound-out cross validation the ZB phase was predicted correctly and more stable than the CsCl phase by 1.52 eV/atom, which is in good agreement with the DFT energy difference of 1.67

eV/atom. However, the energy-volume behaviour in the CsCl phase exhibits two minima (see Fig. F.4) and, moreover, the minimum that deviates stronger from the DFT volume was predicted energetically lower for both the leave-one-compound-out cross validation and the test with missing Ga and N atoms in the training set. Accordingly, we consider this second CsCl equilibrium in both cases as wrong despite the low error on the predicted CsCl-ZB energy difference in the leave-one-compound out cross validation. The fact that the second minimum lead to a mispredicted ground state in the test of this paragraph is, similar to the multiple minima that influenced the success rates of the CGCNN in the leave-one-compound cross validation, a further indication of the fact that *artefacts* can increase the risk of an unsuccessful ground-state search.
Compared to the leave-one-compound cross validation, the tests of leaving out compounds that contain Ga or N (Sr or O) atoms from the training set and predicting their PESs showed a significant loss of performance only in case of reproducing the ground-state phase of AlN correctly. Thus the hypothesis that predicting the PES of a compound can benefit from the presence of its atomic constituents in compounds of the training set could not be clearly proven.

### 5.4.7 A global structure search for GaN

The tests in Sec. 5.4.6 evaluated the ability of the ML potentials to predict the PES of a compound not seen in the training set. However, the investigation of the predicted PES was limited to a set of selected phases. Accordingly, the label "ground-state phase" assigned to the lowest-energy phase is relative to this set. For instance, while we have labeled ZB as the ground state of GaN, the reference DFT method predicts the wurtzite (WZ) structure, which is not in our set, to be the ground state, predicted to be 9 meV/atom more stable than ZB.
To demonstrate a completely unbiased global search for the ground-state phase of a compound, we have performed a random-structure search (RSS) [21, 92] for GaN, using a CTP and ASOAP trained on the data of all octet binary compounds except GaN, i.e. a test that again evaluates the ability of a ML model to predict the PES of a new compound. In addition, we performed a RSS also with DFT. The RSS was carried out by generating 300 random structures (the same for all three methods) and relaxing them (both atomic positions and unit cells) with the considered methods. We considered only structures with eight atoms in the unit cell, constrained to have a minimum pairwise distance of 1.54 Å ($0.8r_{\mathrm{cov}}$, as motivated in Sec. 5.4.4). For more details, e.g. on the generation rules of the structures, see App. F.6. The upper panel of Fig. 5.9 shows the crystal structures, which were identified in the RSS, in a energy-volume scatter plot. Note that we do not report results for the CGCNN in this test because none of its 300 relaxations resulted in a minimum[17], given that the PES predicted by the CGCNN is

[17] More precisely, within 1 500 relaxation steps, the forces did not converge below a tolerance of 0.01 $\mathrm{Å}^2$/atom.

unphysically rough. This behavior is mainly attributed to a construction choice of the CGCNN, where strictly only 12 neighbours are considered in the representation of the model, therefore causing uncontrolled oscillatory behavior in the forces when different neighbours enter the first shell in a relaxation step (see explanation in App. F.7). However, we will present results for the CGCNN in an adapted RSS in a second test, see below.

A key result of the RSS is that the ASOAP failed in predicting the PES with an accuracy that would be needed to identify the most stable crystal structures of GaN with the RSS. We consider the predicted PES as significantly wrong. Neither WZ nor ZB were identified. All found structures are energetically lower than ZB, with a noticeable average energy difference of -1.57 eV/atom. A separate relaxation of the WZ structure resulted in a structure that was still in the WZ crystal symmetry and 1.39 eV/atom lower than ZB, however, higher than most of the structures found in the RSS and 0.76 eV/atom higher than the energetically lowest structure. Moreover, in 87 structures the minimum distance between two atoms is unphysically small, below 0.5 Å. The maximum space group is 15 (monoclinic crystal-system), see Fig. F.7 in the appendix for the crystal-system distribution. This observation is in accordance with the general expectation that a potential based solely on a many-body descriptor such as SOAP carries higher a risk of predicting an unphysical PES, especially when the training data set is limited.
As mentioned above, the predicted PES by the CGCNN is unphysically rough. We would like to investigate if the CGCNN still predicts ZB or WZ as the global minimum on its PES. Due to the roughness of the PES, the forces in the crystals did not converge during the relaxations in the RSS approach (at least not within 1 500 relaxation steps, see Fig. F.10 in appendix). As an alternative, we have performed the structure optimization using a random walk on the PES, where a step is only accepted if it leads to a lower energy. That is, at each step of the algorithm, a configuration with randomly slightly changed coordinates of the atomic positions and lattice vectors is suggested and only accepted if its energy is lower than the one of the previous configuration (the one without changed coordinates). This is equivalent to perform a Metropolis Monte Carlo random-walk at $T = 0$ K. As a consequence, the strong energy fluctuations observed in the forces-based optimization technique (Fig. F.10) are avoided and a (fast) convergence becomes possible. In order to validate the random-walk based optimization, we have performed it also for the CTP and ASOAP. The results for all three machine-learning models are shown in Fig. F.9 in the appendix. The energy-volume distributions of the CTP and ASOAP in Fig. F.9 are comparable to the ones in the upper panel of Fig. 5.9, which validates the random-walk based optimization[18]. Analyzing the results of the CGCNN, shows that 25% of the optimized structures are more stable than ZB (and also than WZ), 12% are more stable by at least 0.1 eV/atom. Note that neither ZB nor WZ were identified by CGCNN

[18]Note, that two optimized structures obtained from the two optimization techniques, the forces-based and random-walk-based one, do not necessarily coincide, if they had the same initial structure. However, in case of the CTP, we find that also the space-group distributions of the optimized structures are comparable for both optimization techniques.

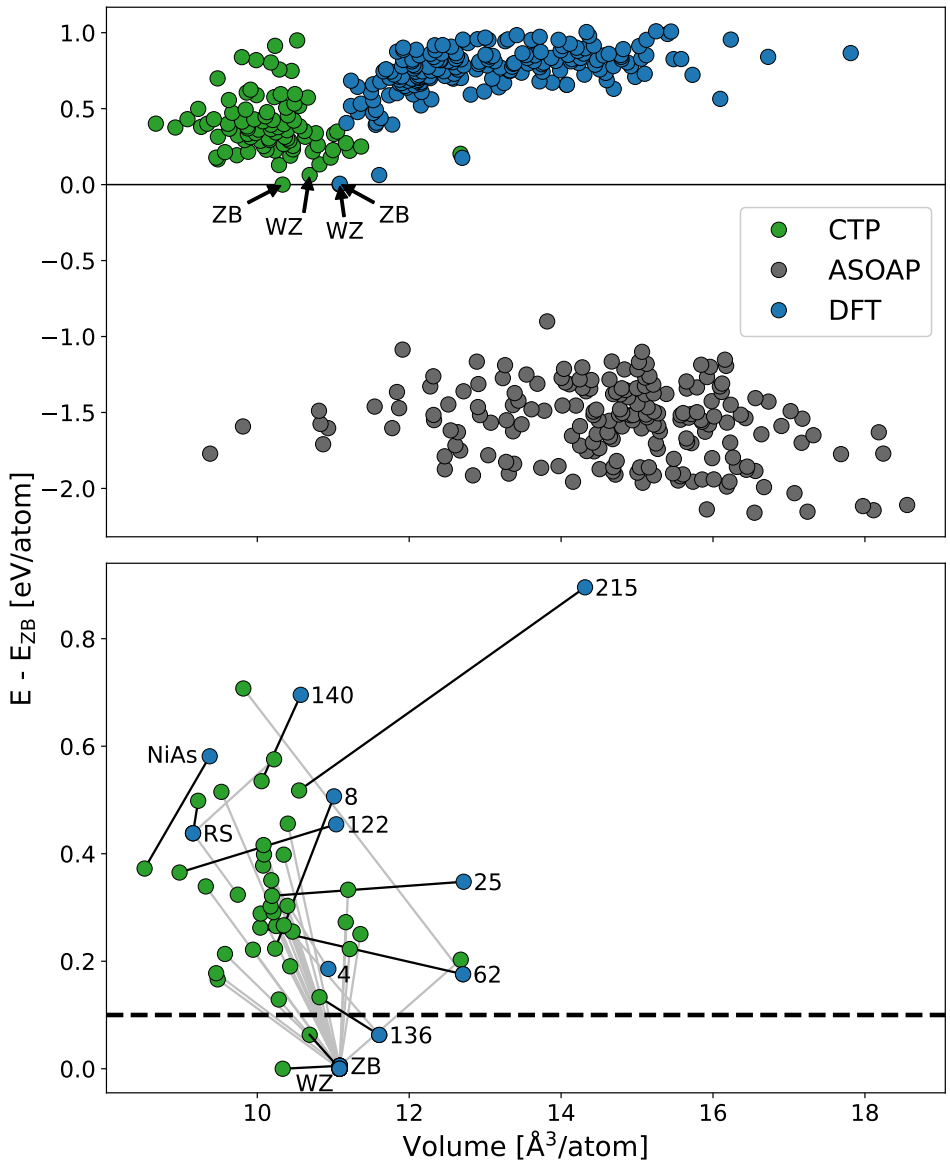



**Figure 5.9:** GaN structures from random-structure searches (RSS) in energy-volume scatter plots. The upper panel shows the result of the RSS with 300 initial random structures performed for the chemical-transferable potentials (CTP), alchemical smooth overlap of atomic positions (ASOAP), and DFT. The CTP and ASOAP were trained on structures of 77 octet binary compounds, GaN not included. The RSS performed for the CTP resulted in 58 unique structures and 40 of them are mechanically stable (no imaginary phonon frequencies were found). In the lower panel, the 40 unique and stable "CTP structures" (green markers) were further relaxed with DFT, resulting in structures represented by the blue markers. The lines between the markers show which CTP structure resulted in which further-DFT-relaxed structure. The black lines highlight that a DFT relaxation did not significantly change a CTP structure (see text). The structures obtained from the DFT relaxations are labeled either by their space group or by the name of their structure prototype. The black dashed line represents the relative-energy base of 0.1 eV/atom.

in the structure search. In fact, all found structures have a space group of 1. Moreover, most of them are unphysical: three have a minimum pairwise distance of below 0.5 Å and 170 of below 1.5 Å (1.5 Å is still unphysically small for GaN).
From now, we will again consider the results of the forces-based RSS only.

The CTP identified both ZB and WZ as the most stable structures, just as predicted by DFT, although predicting WZ higher than ZB by 63 meV/atom. We remind that DFT predicts WZ lower by 9 meV. The next higher structure predicted by CTP has a difference of 129 meV/atom to ZB. This means that in a search where only the lowest energy structure and the ones with maximally 100 meV/atom above were considered, only WZ and ZB would have been chosen for a DFT check.
An interesting result is that the CTP tended to predict structures with relatively high crystal symmetry, while most structures identified by DFT have space group of 1, compare crystal-system distributions in Fig. F.7. ZB was identified 104 times by CTP but only 11 times by DFT. WZ was identified 32 times by CTP vs two times by DFT. Moreover, 95% of the "DFT structures" have at least one atom whose next-neighbour shell contains only the same atom types as itself, while none of the CTP structures show this property. A possible reason for the higher symmetries of the CTP structures and the fact that the structures do not have atoms whose next-neighbour shell contains only the same atom types as itself might be that the two characteristics apply also to the octet-binaries data set on which the CTP was trained.

In Sec. 5.4.6, we have shown that the PESs predicted by the ML models were often characterized by artefacts (minima that are not ones on the DFT PES) although relatively rare in case of the CTP. In order to analyze if the predicted phases by the CTP in the RSS are also minima (or saddle points) on the DFT PES, we have relaxed them further with DFT. The RSS performed for the CTP resulted in 58 unique structures[19] and 40 of them are mechanically stable, as we did not find imaginary phonon frequencies for them. The lower panel of Fig. 5.9 compares in an energy-volume scatter plot the 40 unique and stable structures predicted by the CTP in the RSS with their DFT-relaxed forms. The lines in the plot connecting the the green (CTP structures) blue marker (DFT-relaxed structures) highlight which CTP structure resulted in which DFT-relaxed one. A line is colored black if the structure (crystal symmetry) did not change (significantly[20]). The 29 grey lines represent the cases where a CTP structure is not stable on the DFT PES.
Out of 29 CTP structures that are not stable in DFT, 24 relaxed relaxed into ZB or WZ. What the 29 CTP structures have in common is that they tend to have low crystal symmetry. For example 23 of the 29 structures are triclinic or monoclinic (space group of

[19] The decision if two structures are the same is based on both a) if the space groups are the same and b) if volumes and energies are same (similar).

[20] In one case, a CTP structure has space group 58 and the corresponding DFT relaxed one has 136. However, the structures are similar, i.e., the CTP structure (space group 58) is just a stretched (or with fewer symmetries) version of the DFT-relaxed one (space group 136). This similarity can be appreciated by setting a loose tolerance in spglib, where the space group 58 structure is actually recognized as space group 136.

16 or below), and 28 of 29 are of space group 62 or below. In contrast, out of the eleven CTP structures that are stable (minimum or saddle point) also in DFT, only one has a space group of 16 or below and four have one of 62 or below. Thus, the test indicates that for the CTP, predictions of structures with higher crystal-symmetry are reliable[21]. We remind that the lowest space group in the training database of the CTP is 62, given by the CrB prototype.
The fact that the CTP predicted each of the eleven structures, whose form did not change significantly in a relaxation by DFT, with a smaller volume than DFT, indicates that the CTP underestimated the volumes of GaN phases collectively. The mean absolute volume difference between DFT and CTP for the eleven structures is 1.4 $Å^3$/atom. We assume that the collective underestimation of the volumes (at this extent) did not have an effect on the task of identifying a phase in the RSS.

Four (ZB, WZ, RS, 136) of the eleven DFT-relaxed structures, which do not differ significantly from their corresponding CTP structures, are known phases of GaN, where only two (ZB and RS) of the four are contained in the training set of the CTP. The prediction of WZ is not surprising, due to its similarity to ZB[22]. However, the structure with the space group 136 was just recently predicted for GaN using DFT [118]. The DFT energy-difference of the strucutre with space group 136 to WZ is 0.063 eV/atom. Note that, before the work in Ref. [118], besides ZB and WZ, only two further phases had been reported for GaN: RS [119, 120] and a hexagonal layered type [121, 122][23].

The tests of this section have shown that the CTP trained on the OB data set where information about GaN was excluded was able to predict, in a RSS, the most stable crystal structures of GaN. In a crystal-structure search, where only the lowest-energy structure and the ones with maximally 0.1 eV/atom above were considered, the CTP would identify WZ and ZB. In a test with a higher tolerance of 0.15 eV/atom, also the just recently predicted phase with a space group of 136 would have been identified, a phase not contained in the training set of the CTP.
Evaluating the benefit of using the CTP also for the prediction of higher-energy phases is not trivial, as the results of DFT and CTP differ from each other. For instance, the identified structures in the RSS of DFT mainly deviate from the ones of CTP and, furthermore, the knowledge of which DFT structures, not found with CTP, are *promising*

[21] Note that the CTP structures are mechanically stable while the DFT-relaxed ones might only be saddle points.

[22] The ZB and wurtzite structure based on two atomic species types become diamond and hexagonal diamond, respectively, if both atomic species types are replaces by a single type, resulting in an an elemental solid. The similarity between diamond and hexagonal diamond are discussed for silicon in App. E.1.2.

[23] Also in the AFLOW database, there are only ZB, WZ, RS and a hexagonal layered type available for GaN. In the Materials Project database the same are given plus a further triclinic system. However, while the hexagonal layered phases of AFLOW and the Materials Project are the same, they differ from the one that has been discussed in Ref. [118, 121, 122].

metastable states requires further extensive tests. Still, we observed that the CTP was able to identify all structures of space group above 75 that were also identified with DFT, except one[24]. Therefore, the results demonstrate the applicability of the CTP (trained on the considered data set) to find high-energy phases that have high crystal symmetries.

### 5.4.8 Error analysis in the chemical composition space

The reliable estimate of the error expected for the prediction of a ML model is a critical goal in data-driven materials science. One important component of an error-estimator is the partitioning of the input space of a model, e.g. the structural and chemical composition space, into *known* and *unknown* regions depending on the population of the training data points in the input space. The prediction of the energy of a material in an unknown region or at the border of an known region would be considered as an extrapolation of the model and the prediction accuracy estimated to be low. A further crucial component independent of the population of the training data in the materials space is the estimation if it is harder to accurately describe the PES of a material due to the nature of the material. For instance, in the leave-one-compound-out cross validation in Sec. 5.4.6 the compounds AgF and CuF had their ground-states mispredicted by all three ML approaches, the CTP, ASOAP and CGCNN. Moreover, we find that the likelihood that the ground state of a material is mispredicted by at least one of the three models correlates with a materials descriptor that depends on the sizes (radii) of the atomic constituents (see below).
Beyond discussing the correlation between this materials descriptor and the failed predictions of the models, this section demonstrate how this descriptor is identified using the SISSO algorithm [41]. Such a descriptor does not only provide a potential estimator for the domain of applicability of the ML model in the chemical composition space. Due to its interpretable form, it also opens the route to analyze and understand which conceptual feature a ML model misses such that it can be improved. A similar idea to find such a descriptor was just recently presented [123], however using subgroup discovery.

17 of the 78 octet binary compounds had their ground states mispredicted by at least one of the three models, CTP, ASOAP, and CGCNN. The goal is to find a descriptor with SISSO for classification that separates the 17 compounds from the remaining ones based on the atomic properties that were used as the input of the CTP and ASOAP models. However, we expect that a highly accurate classifier, e.g. a descriptor that perfectly separates the two compound classes, would overfit to a sort of noise: If the ground state of a compound is considered as mispredicted depends on a tolerance (described in Sec. 5.4.4). As discussed in Sec. 5.4.6, the set of compounds which had their ground states mispredicted is sensitive on small changes of the tolerance. Therefore, we searched for a descriptor that is rather *simple* and allows for misclassifications.

[24]The structure has space group 187 and is similar to wurtzite, see Fig. F.8 in the appendix.

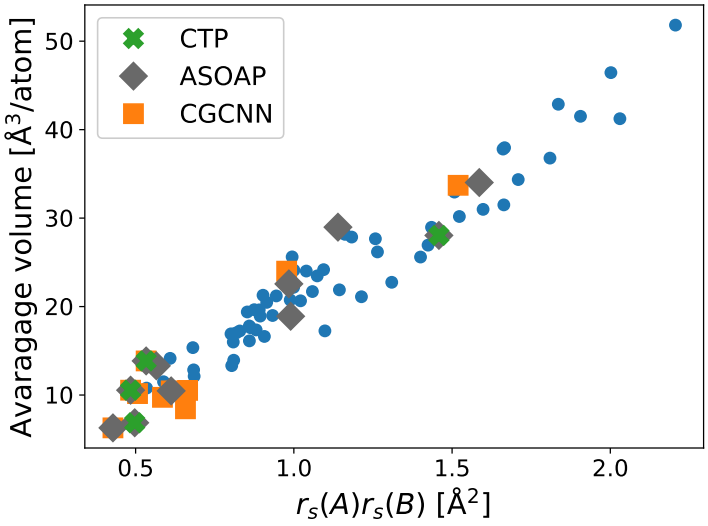



**Figure 5.10:** Dependence of the ground-state prediction success on the location of a compound in the chemical-compound representation $r_s(A)r_s(B)$. The figure highlights compounds that had their ground states mispredicted by three ML methods in a leave-one-compound-out cross validation (Sec. 5.4.6). The three methods are: chemical-transferable potentials (CTP), alchemical smooth overlap of atomic positions (ASOAP), and crystal graph convolutional neural networks (CGCNN). The blue markers represent the materials whose ground states were predicted correctly by all three methods. The descriptor $r_s(A)r_s(B)$ was identified in a classification task with the sure independence screening and sparsifying operator (SISSO). The aim was to separate most of the 17 compounds which had their ground-states mispredicted by at least one of the three ML methods from the remaining compounds. The descriptor $r_s(A)r_s(B)$ is linearly correlated with the average volume of a compound ($y$-axis). The average is built over the volumes of the different phases of the compounds, each at the equilibrium of its crystal symmetry.

We identified a one-dimensional descriptor: $r_s(A)$ $r_s(B)$. The atomic properties $r_s(A)$ and $r_s(B)$ denote the radii, where the radial probability density of the valence $s$ orbitals are maximal for the elements $A$ and $B$ of the $AB$ compounds. Note that $A$ labels the element with the smaller electronegativity. Fig. 5.10 shows how the compounds are distributed in the descriptor space $r_s(A)$ $r_s(B)$ ($x$-axis). We find that $r_s(A)r_s(B)$ is linearly correlated with the volume in which a compound tends to crystallize, as shown in the upper panel of Fig. 5.10.

Ten compounds that had their ground states mispredicted by at least one of the three ML methods are located at a low value of $r_s(A)r_s(B)$, i.e. below 0.7 Å$^2$. This means that the PESs of compounds, whose (stable) crystal structures are characterized by small volumes, are harder to be predicted accurately. The ten compounds are CdO, AgF, ZnO, CuF, AlN, MgO, BN, BP, BeO, and LiF.
The origin of the erroneous description of *small-volume compounds* is, however, not clear. Due to the fact that the CGCNN learns the atomic representations in the training process, the choice of atomic descriptors in the case of the CTP and ASOAP might be excluded as a (leading) origin. A possible factor of statistical nature that might influence the prediction performance is the higher variance of the target (energies) of many compounds with smaller volumes. For instance, the left panel of Fig. F.11 demonstrates that the variance is more likely to be high for small $r_s(A)r_s(B)$. The variance was calculated for each compound over the relative energies of the different phases with respect to the ground state phase of the compound. More precisely, the variance assigned to a compound $AB$ is given by $\mathrm{Var}_{AB} = \sum_i \frac{(\Delta E_i - \overline{\Delta E})^2}{n-1}$ where $n$ represents the number of phases of the compound $AB$, $i$ runs over all phases but the ground-state phase, $\Delta E_i$ denotes the energy difference of a phase to the ground-state phase, and $\overline{\Delta E}$ is the average of the energy differences. The three highest variances are given by BN, AlN and BeO with a value of approximately 0.87 eV$^2$/atom$^2$. At the same time, the errors on the predicted energies are more likely to increase when $r_s(A)r_s(B)$ decreases (Fig. F.11, center). However, while, in fact, the errors tend to increase with the variance of energies, relatively high errors (a root mean square error above 0.2 eV/atom) are also found for variances of below 0.11, including some compounds that are in the region $r_s(A)r_s(B) < 0.7$ Å$^2$, e.g. LiF, CuF, AgF, NaF, BAs, and BP. Therefore, the variance of the energies alone does not completely explain the more challenging description of small-volume compounds. Out of the scope of this work, a deeper investigation of the correlation between materials descriptors and prediction errors might help to design reliable error-estimators for the exploration of the materials space.

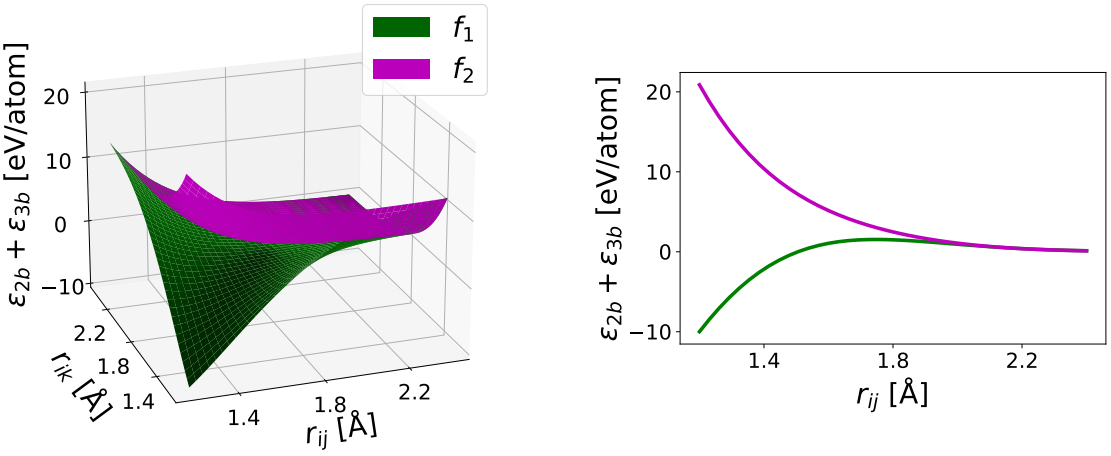



**Figure 5.11:** Potential energy of an atom surrounded by two neighbours (i.e. an isolated trimer) in dependence of the distances $r_{ij}$ and $r_{ik}$ to the neighbours for two different models $f1$ and $f2$ fitted to the same Si structures (Sec. 5.3.2). The distance $r_{jk}$ between the two neighbours is fixed at 2.2 Å. The potential energy of the trimer depends on the sum of a two- and three- body contribution, $\epsilon_{2b}$ and $\epsilon_{3b}$, and is shown for two different models $f_1$ and $f_2$ which differ by the centers $\boldsymbol{r}_\mu$ of the Gaussian basis functions $\exp(-\frac{[(r_{ij},r_{ik},r_{jk})-\boldsymbol{r}_\mu]^2}{2\sigma_\mu^2})$ in the three body term, see Eq. 5.3. In both cases, the centers $\boldsymbol{r}_\mu$ are distributed on a uniform grid in the input space $(r_{ij}, r_{ik}, r_{jk})$, where in $f_1$ the "smallest" grid point is given by $(1.0, 1.0, 1.0)$ and in $f_2$ by $(1.7, 1.7, 1.7)$. Furthermore, the Gaussian widths are $\sigma_\mu = 1.4$ and $\sigma_\mu = 1.3$ (for all $\mu$), respectively. The two different analytical forms of $f_1$ and $f_2$ fitted to the same data set show significantly different potential energies in the shown domain. The left panel shows the dependence of the potential on both $r_{ij}$ and $r_{ik}$ in a three-dimensional plot. The right panel presents the potential along the diagonal $r_{ij} = r_{ik}$, in a two-dimensional plot.

## 5.5 Technical challenges

### 5.5.1 Sensitivity of the model reliablity to small changes in the potential-basis functions

The analysis of the leave-one-compound-out cross validation in Sec. 5.4.6 performed for the CTP, ASOAP and CGCNN shows that the predictions of different models can correlate in specific ranges of the input space but deviate significantly in other regions. Similarly, even when considering one method, different choices of (hyper-)parameters may cause a similar effect. This is expected because of the fundamental requirement that ML algorithms shall select a model from a *diverse* function space. The optimization of all hyperparameters of a model, including the choice of the basis functions, is a high-dimensional problem, and some are typically chosen by the scientist based on intuition or experience [124]. Even if an optimization of the hyperparameters is performed, the optimality of the results is valid only for the training range.

For instance, let us consider the choice of the basis functions for fitting the solid phases of silicon (Sec. 5.3.2). We have built two models $f_1$ and $f_2$ which are based on a two-

and three-body potential and only distinguished by the basis functions of the three-body terms. In both cases, Gaussian basis functions $\exp(-\frac{[(r_{ij},r_{ik},r_{jk})-\boldsymbol{r}_\mu]^2}{2\sigma_\mu^2})$ were used (see Eq. 5.3) with the centers $\boldsymbol{r}_\mu$ distributed on a uniform grid in the input space $(r_{ij}, r_{ik}, r_{jk})$, where in $f_1$ the "smallest" (smallest $\ell_2$ norm) grid point is given by $(1.0, 1.0, 1.0)$ and in $f_2$ by $(1.7, 1.7, 1.7)$. Furthermore, the Gaussian widths are $\sigma_\mu = 1.4$ and $\sigma_\mu = 1.3$ (for all $\mu$), respectively. Note that the nearest neighbour distance in the diamond phase of silicon is 2.36 Å. Clearly, the difference of the prediction accuracies of the two models $f1$ and $f2$ on the considered data set of nine silicon phases is negligible: 44 meV/atom vs 45 meV/atom in a leave-one-phase-out cross validation (the test is described in Sec. 5.3.2). However, outside of the training range they differ significantly, where one of them predicts a distinctly unphysical PES. Fig. 5.11 shows the predicted energy of an isolated trimer with small neighbour distances, where the energy consists of both, the two- and three-body contribution. Recall that the energy of a structure is modeled as a sum of triplet energies. The figure highlights that a relaxation of the isolated trimer with $f_1$ might decrease the neighbour distances to unphysically small values if the starting geometry is given by rather small neighbour distances. For instance, while the two-body potentials of both models exhibit a repulsion ($\frac{\partial \epsilon_{2b}}{\partial r} < 0$ at small distances $r$), the three-body potential of $f_1$ exhibits a strong attraction behaviour dominating the trimer energy in the considered region. We found that when relaxing with $f_1$ a random bulk structure with eight silicon atoms in the unit cell and rather small pairwise distances where, for example, one atom is surrounded by two neighbours with distances between 1.7 Å and 1.8 Å, four atoms collapsed into each other, i.e. they had the same positions. Moreover, the energy of the relaxed structure is 3 500 eV/atom below the one of diamond[25]. In contrast, when using $f_2$, a relaxation of the same starting geometry lead to a structure located inside the region of the reference DFT distribution shown in Ref. [29] with a volume of 23.3 Å$^3$/atom and an energy difference of 0.293 eV/atom to the diamond phase.

There is no trimer in the training data set that has two neighbours at such small distances ($< 1.8$ Å) and an appropriate similarity model could detect that the considered structure is out of the training range. Still, the fact that only relatively small adjustments of the three-body basis functions lead to a highly unphysical behaviour of the model underlines the challenge of designing reliable ML models. A possible way out of the problem, could be a stronger regularization of the three-body term. With a penalized three-body term, the contribution of the three-body potential to the total energy could be kept low such that the repulsive behaviour of the two-body potential could dominate at small atomic distances. Alternatively, a constraint on the coefficients that prevents the thee-body potential forming an attractive behaviour at small distances, similar to the one in the two-body part of the CTP (Sec. 4.4.3), could be investigated in the future.

[25] Note that a relaxation of the isolated trimer is only partially representing the relaxation of a bulk that contains the same three-atoms cluster (triplet). For example, in Fig. 5.11 we assigned a two-body contribution to the triplet energy but the two-body contributions in the crystal are independent of the three-body terms and not distributed to triplets. Furthermore, only the collection of all triplet contributions in the crystal fully describes the behaviour of a bulk in a relaxation.

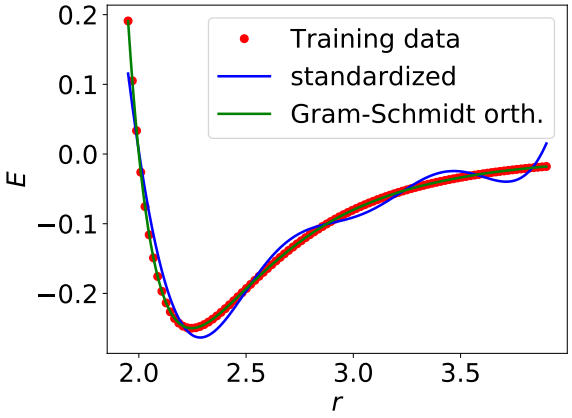



**Figure 5.12:** Influence of the data-processing type on the learning of a linear model with gradient descent. The modes are fitted to a toy Lennard-Jones potential $E = \epsilon[(\sigma/r)^{12} - (\sigma/r)^6]$ with dimensionless parameters $\epsilon = 1$ and $\sigma = 2$, represented by 100 training data points, i.e. distance-energy tuples $\{(r_i, E_i)\}$. The linear models $f = \sum_\mu \alpha_\mu b_\mu$ consist of 14 Gaussian basis functions $b_\mu = \exp(-\frac{[r-r_\mu]^2}{2\sigma_\mu^2})$, with centers $r_\mu$ from a uniform grid on the training range and $\sigma_\mu = 0.5$ for all $\mu$. Two different models were trained with gradient descent, one with standardized, the other one with Gram-Schmidt-orthogonalized data. In the former case, 30 000 epochs were used and in the latter 500 epochs.

### 5.5.2 Limitations on the fitting accuracy through gradient descent

The consequence of the linearity of our two- and three body potentials in their regression coefficients $\alpha_i$ results in the fact that the modeled energy $E$ of a crystal is also linear in these regression coefficients. When modeling the energies of structures that depend on a fixed set of atomic species types, e.g. a single chemical formula (no chemical-transfer learning), the final linear system is written $\boldsymbol{E} = \boldsymbol{B}\boldsymbol{\alpha}$, where the matrix $\boldsymbol{B}$ represents the sums of the basis functions over all atomic environments in the structures (see derivation in Sec. 4.3.2). The standard optimization method to determine the coefficients $\boldsymbol{\alpha}$ is (regularized) linear regression. When extending the potentials to CTP by introducing an explicit dependence of the coefficients $\boldsymbol{\alpha}$ on the species types that is modeled with a neural network, the regression parameters to be optimized become the weights of the neural network. Given that the energy $E$ is, then, not linear anymore in the regression parameters, linear regression is not applicable anymore. The typical optimization method for a neural network is a gradient-descent based one, as used in this work. The CTP construction does not change the shape of the equations $\boldsymbol{E} = \boldsymbol{B}\boldsymbol{\alpha}$. In principle, they are still written for every chemical formula separately where the only difference to the non-chemical-transfer-learning approach is that the coefficients $\boldsymbol{\alpha}$ are determined by a neural network and gradient descent. However, in practice, the solution for the individual coefficients $\boldsymbol{\alpha}$ of the CTP for every chemical compound obtained by optimizing the neural network with gradient descent is not necessarily optimal, while the solutions of linear regression applied to $\boldsymbol{E} = \boldsymbol{B}\boldsymbol{\alpha}$ for every chemical formula separately (non-chemical-transfer

learning) are[26]. One reason lies in the fact that converging the solution of gradient-descent in its iterative procedure towards the exact solution can have numerical barriers.
For instance, assume we want to fit only a linear and *simple* function depending on one variable with gradient descent (non-chemical-transfer approach), i.e. a toy Lennard-Jones potential $E(r) = \epsilon[(\sigma/r)^{12} - (\sigma/r)^6]$ with dimensionless parameters $\epsilon = 1$ and $\sigma = 2$. Using as a model our two-body ML potential $f = \sum_\mu \alpha_\mu b_\mu$ with Gaussian basis functions $b_\mu = \exp(-\frac{[r-r_\mu]^2}{2\sigma_\mu^2})$ and a set of 100 samples $\{(r_i, E_i)\}$ from the Lennard-Jones potential to which we want to fit the ML potential, we again obtain the linear system $\boldsymbol{E} = \boldsymbol{B}\boldsymbol{\alpha}$ that we have to solve for $\boldsymbol{\alpha}$. For different learning rates and random distributions for initializing $\boldsymbol{\alpha}$, we were not able to obtain a good fit to the Lennard-Jones potential with gradient descent after 30 000 epochs, which is a relatively high number considering the simplicity of the problem (linear, only 100 data points, only 14 coefficients), see the blue curve in Fig. 5.12. Note that we have standardized the columns of the matrix $\boldsymbol{B}$ to have zero mean and variance one[27]. In contrast, we would be able to describe the curve accurately if we determined $\boldsymbol{\alpha}$ with a least-squares regression. The reason for the poor performance of the gradient descent is given by the fact that the Gaussian basis functions are correlated. More precisely, given that we used basis functions with Gaussian centers $r_\mu$ from a uniform grid on the training range, $1.95 \leq r_\mu \leq 3.9$, and $\sigma_\mu = 0.5$ for all $\mu$, two neighbouring basis functions (similar $r_\mu$) are strongly correlated, e.g. the two corresponding columns of $\boldsymbol{B}$ have a Pearson correlation coefficient of 0.99. As a result, when coming closer to the least-squares solution during evolvement of the epochs, the gradients and accordingly the changes of the coefficients $\boldsymbol{\alpha}$ become smaller and smaller. While, theoretically, gradient descent leads to the optimal solution due to the convexity of the least-squares problem, in practice the needed epoch to reach the optimal solution might be computational unfavourably high or the absolute value of the gradients might come below a numerical tolerance. In the following we discuss three different approaches that improve the accuracy of a linear model learnt by gradient descent.
If we decorrelate the columns of $\boldsymbol{B}$ by orthogonalizing them via the Gram-Schmidt process instead of standardizing them, already after a few hundred epochs we were able to obtain an accurate model, see the green curve in Fig. 5.12. Note that the Gram-Schmidt orthogonalization of the column vectors does not change the column space of $\boldsymbol{B}$ (or the function space from which the linear regression *selects* a model)[28]. Nevertheless, while the Gram-Schmidt orthogonalization has provided a way to solve the considered problem (efficiently) with gradient descent, its integration into the CTP approach is not straightforward, however, could be investigated in a future work.
We found that orthogonalizing the basis functions (not the columns of $\boldsymbol{B}$) numerically also improves the gradient-descent result. In practice, we perform the Gram-Schmidt

[26] If the standard squared-error loss is used, the least-squares solution (non-regularized linear regression) is proven to minimize the loss.

[27] Furthermore, the model includes an intercept. This means that the model consists of a further regression coefficient $\alpha_0$, such that $f = \sum_\mu \alpha_\mu b_\mu + \alpha_0$.

[28] This means that for any coefficient vector $\boldsymbol{\alpha}$ assigned to a model of not orthogonalized columns of $\boldsymbol{B}$, there is a $\boldsymbol{\alpha}^*$ for a model based on $\boldsymbol{B}^*$ with orthogonalized columns, that yield the same model.

process on an interval $[a, b]$, where the scalar product between two basis functions is given by $< b_i(r), b_j(r) >= \int_a^b b_i(r) b_j(r) dr$ . In contrast to the orthogonalization of the columns of $\boldsymbol{B}$, the incorporation of the orthogonal basis functions into the CTP implementation is straightforward. An alternative way to the numerically orthogonalized basis functions is to use orthogonal basis functions of whom the analytical forms are tabulated (or recursively obtained), e.g. orthogonal polynomials.
In general, the choice of the basis functions can be a key factor for obtaining an accurate fit when using gradient descent. While also polynomial basis functions gave a poor result when the data was only standardized, we found that with a Gaussian basis set, where all Gaussians are centered at zero ($r_\mu = 0$) and distinguished by their widths $\sigma_\mu$, we were able to fit the Lennard-Jones potential with the gradient-descent solution accurately without any orthogonalization. In fact, the advantage of using zero-centered Gaussians as basis functions when fitting a Lennard-Jones potential with gradient descent was one reason for implementing them into the CTP of this work, given also that we expect that two-body potentials, typically, adopt after fitting a shape similar to a Lennard-Jones potential (i.e. repulsion at small and attraction at larger distances)[29].

Still, despite the effort to improve the models when being fitted with gradient descent, we may obtain a significant deviation from the optimal solution given by least-squares regression. Considering a more realistic example, i.e. 1 541 silicon structures and the same two- and three-body basis functions as used in Sec. 5.3.2, the mean absolute errors on the fitted energies are 5.5 meV/atom and 0.9 meV/atom for standardized and orthogonalized columns of $\boldsymbol{B}$, respectively, when using gradient descent, versus 0.9 meV/atom in case of the least-squares regression. Although the deviation between the solution of gradient descent with standardized data and the other two methods might be small relative to our goal of predicting the ground state of a compound with an accuracy of 100 meV/atom, it is not clear to which extent this deviation might increase for larger and more complex data sets when using the CTP approach.

## 5.6 Conclusions and outlook

In this work, we have considered machine-learning (ML) potentials. Such models have been used to approximate the potential-energy surface (PES) for a set of a few atomic species types. By making ML potentials (specifically $n$-body potentials) explicitly species-type dependent, we have investigated to which extent ML potentials can be generalized towards the prediction across chemical composition space. We have termed these models *chemical-transferable potentials* (CTP). In particular, we have evaluated the applicability

[29] When fitting Si (Sec. 5.3.2) or $ZrO_2$ (Sec. 5.3.3) structures with a two- and three body potential (using linear regression), we observed that the two-body parts of the potentials adopted this kind of shape, independent of the choice of basis functions (except that they allow for an accurate fit).

of the CTP to predict the ground-state and metastable crystal structures of materials.

First we validated the ML potentials without the chemical-transfer approach.

- In Sec. 5.3.1, models for the prediction of formation energies and band gaps of $(Al_xGa_yIn_z)_2O_3$ compounds were built using a data set from the NOMAD 2018 Kaggle competition [72]. The models achieved a prediction accuracy which would have ranked the models second place in the competition.
- We have considered nine solid phases of silicon and discussed to which extent the equilibrium energy of a phase could be accurately predicted if the phase was not in the training set (Sec. 5.3.2). We demonstrated that the predicted energetical ranking of the phases, some of which were not in the training set, was accurately reproduced.
- We have shown that a ML potential fitted to $ZrO_2$ compounds in Sec. 5.3.3 predicted a sufficiently reliable cubic-tetragonal and monoclinic potential-energy surface (PES) such that molecular-dynamics based properties or phenomena below 2 000 K like the ferroelastic switching and monoclinic-tetragonal transition temperature could be well reproduced.

The chemical-transfer approach was analyzed on a data set of 78 binary compounds in eight different phases. This data is sparse and limited in the structure space. Building reliable models from sparse data sets is a challenge with ML potentials that have been introduced so far. The goal of this work was to construct models with the focus on a coarser but robust prediction of the PES, particularly when only sparse training data is available.

- In a cross-validation test in Sec. 5.4.5, we have shown that the single unified model determined with the CTP for all compounds stabilized the predictions of the separate potentials determined for every compound independently.
- In order to evaluate the ability of the CTP to accurately predict the PES of a new compound, we have performed a leave-one-compound-out cross validation in Sec. 5.4.6. Performing (at least for the cubic phases) symmetry-constrained relaxations to search for phase equilibria, the CTP were able to correctly identify the energetically lowest phase (within the set of eight phases) for 74 out of 78 left-out compounds. We considered a phase of a compound as correctly identified if the true ground-state phase (as determined by DFT) was inside a set of phases that were predicted to be promising, i.e. the predicted lowest phase plus phases whose energy difference to the lowest phase is below a tolerance of 0.1 eV/atom. If this tolerance was increased to 0.11 eV/atom, the ground-state-prediction success rate of the CTP would be 76/78. For a comparison, we have included two state-of-the-art models into the tests, i.e. the alchemical smooth overlap of atomic positions (ASOAP) and crystal graph convolutional neural networks (CGCNN). Their respective ground-state-prediction success rates are 67/78 and 68/78.

- We analyzed the shapes of the energy-volume curves of the cubic phases that were predicted by the CTP, ASOAP, and CGCNN in the leave-one-compound-out cross-validation in Sec. 5.4.6. We found that the ASOAP and CGCNN predicted many spurious minima for the cubic phases, indicating that the predicted PESs of the two models might be overall unphysical.
- We have performed a random-structure search (a global search) for GaN with the CTP, ASOAP, and CGCNN in Sec. 5.4.7. The three ML models were trained on all binary compounds but GaN. In case of the ASOAP and CGCNN, many of the found structures were energetically lower than the true ground-state phase wurtzite and many structures exhibited unphysically small atomic pairwise distances. Moreover, neither wurtzite and zinc-blende were identified in the random-structure search nor any other phase from the training set. Furthermore, in case of the CGCNN, the forces-based relaxations in the inital random-structure-search method could not be performed due to the unphysical roughness of the CGCNN. Therefore, we introduced a random-walk-based method for the relaxations to still obtain insights into the CGCNN PES. In contrast to the ASOAP and CGCNN, the CTP was able to identify the three most stable phases zinc-blende, wurtzite, and one with a space group 136. The latter phase was not considered in the training set. Moreover, it was not known for GaN before a recent prediction with DFT [118].
- By using the SISSO algorithm, we have demonstrated in Sec. 5.4.8 how descriptors can be identified that predict which materials are harder to be described by the three considered ML methods, CTP, CGCNN, ASOAP.

Our motivation behind using $n$-body potentials (explicitly in the simple additive form as introduced in Def. 4.3) in the CTP was the intention to build models based on lower terms in the $n$-body expansion to better control and inspect the flexibility of the models. For instance, the presence of an explicit two-body potential allowed us to further implement physical constraints that promote the appearance of a repulsion in the two-body terms which decreases the risk that a relaxation leads to structures with unphysically small atomic pairwise distances as in the case of the ASOAP and CGCNN in the random-structure search for GaN. Incorporating (more) physics (or experience) into the mathematical formulas of the models, that is not specific to a general system, such that the human intervention is kept limited for an autonomous application to large materials databases, is a challenge and highly needed (see tests and discussion in Sec. 5.5.1).
The literature dealing with studies that demonstrate successful ML applications in materials science is expanding fast. However, the knowledge of why a ML model was successful in a shown application and what risks it might carry in a modified application stays generally limited. In this work, we analyzed the requirement to better understand the limitations of the models in order to develop more reliable models and our first step was given by demonstrating that the typical quantity used to develop and benchmark ML potentials, i.e. (averaged) errors on predicted energies of a set of given structures, is not necessarily sufficient to evaluate the success and reliability of a potential in a crystal-structure search.

One possible step towards gaining more insights into the ability of ML potentials for the inter- and extrapolation of the structure space could be realized by a future work that evaluates different ML methods on different compounds/data sets and yields insights into the influence of the mathematical terms on limitations on reproducing the PES accurately enough for the task of crystal-structure prediction.
Studies that have considered databases like the Materials Project, AFLOW, and Open Quantum Materials Database [8] for training ML models have typically used only relaxed structures. In the introduction (Sec. 5.1) of this chapter, we have shown (only for the CGCNN) the risk that a data set consisting only of relaxed structures might lead to models that have not learned that a structure is (mechanically) stable. If the CTP is trained on databases as the ones mentioned above, one needs to use relaxed and unrelaxed structures. If only relaxation paths are sufficient or also molecular-dynamics trajectories need to be taken into account must be investigated in the future. Note that, for example, the AFLOW database provides also relaxation paths and the NOMAD Repository is, in general, not limited to relaxations, e.g. data also from molecular-dynamics simulations are available.
We have analyzed the CTP for two- and three-body potentials. Future applications might require the inclusion of higher-body terms. One possible example, whose implementation into the CTP framework is straightforward, is the (many-body) "embedded-atom-method like" descriptor [125]. A further limitation of the CTP is their dependence on local-environment descriptors. An explicit treatment of long-range effects is missing. A generalization of the CTP towards the description of electrostatic effects beyond the cutoff radius might be realized by learning the partial charges of materials [85,136,137].

# 6 Conclusions

In this thesis, we have introduced two novel artificial-intelligence-based approaches for the prediction of the ground-state and metastable crystal structures of materials from quantum-mechanical materials data.

The first approach is a multi-task-learning extension of a symbolic-regression- and compressed-sensing-based scheme that identifies low-dimensional, meaningful, and interpretable materials descriptors from a space of billions of candidate descriptors. We demonstrated how only the multi-task extension, which determines a unified model describing multiple crystal-structures with a single descriptor, enabled the prediction of a well-defined structural stability and, therefore, the design of a low-dimensional crystal-structure map. Furthermore, we highlighted how multi-task learning stabilized the models on incomplete data sets considering the specific example of the stability of octet binary compounds among five different phases. As opposed to single-task learning, our multi-task learning approach was able to determine accurate predictive models also with high levels of incompleteness (e.g. when 50% or more of the information was randomly missing). Moreover, we discussed possible steps and related challenges towards a more global description of the structure space.

In the second part of the thesis, we introduced the chemical-transferable potentials (CTP). We demonstrated how the CTP transfer a certain class of machine-learning (ML) potentials across chemical composition space by using a neural-networks-based scheme.
First we validated the class of ML potentials itself on different relevant materials-science problems, i.e. excluding the chemical transfer approach and focusing on the prediction across structure space for fixed compositions (a few set of atomic species types). We have shown that our models would have ranked second in a Kaggle competition for predicting the formation energies and band gaps of $(Al_xGa_yIn_z)_2O_3$ compounds [72]. Moreover, we have presented that a ML potential fitted to $ZrO_2$ structures predicted a sufficiently reliable cubic-tetragonal and monoclinic potential-energy surface (PES) such that molecular-dynamics-based properties (phenomena) below 2 000 K like the ferroelastic switching and monoclinic-tetragonal transition temperature could be well reproduced. In a further example, we demonstrated that the predicted energetical ranking of nine silicon phases, some of which were not in the training set, was accurately reproduced.
After validating the structural parts of the CTP, we considered a sparse data set of octet binary compounds in eight different phases to analyze the prediction across chemical composition space. We found that the chemical-transfer approach that identified a unified model for all binary compounds stabilized the ML potentials that were specific to a single

compound. The observation that the description of both structure and composition space by a unified model leads to more reliable predictions than in case of models that focus only on one of the two spaces is found in the application of both introduced approaches of the thesis (multi-task learning and CTP). This is a central result of our work.
Next, we have performed extensive tests to evaluate to which extent the PES of a compound not seen in the training set can be reproduced with an accuracy that allows to identify the most stable structure(s) of that compound. One example is a cross-validation test, where the structural stability of a test compound among eight phases was predicted by a model trained on all but the test compound. For this purpose, at least the cubic structures were obtained from a structure search with the ML model to determine the equilibrium within a given cubic crystal symmetry. The CTP were able to correctly predict the ground-state phases of 74 out of 78 test compounds. In contrast, two state-of-the-art models for predicting potential energies from geometries, the ASOAP and CGCNN, achieved ground-state-prediction success rates of 68/78 and 69/78, respectively. A crucial distinction between the CTP and the two other methods was observed in the shapes of the predicted cubic PESs: the ASOAP and CGCNN are more likely to predict an unphysical energy-volume relationship than the CTP.
The tendency of ASOAP and CGCNN towards predicting PESs with spurious minima already for simple cubic symmetries suggested that overall the predicted PESs are unphysical. By performing a random-structure search only for GaN using models that were trained on all compounds but GaN, we were able to demonstrate that the predicted PESs of CGCNN and ASOAP were indeed overall unphysical: many of the found structures in the random-structure search were energetically lower than the true ground-state phase wurtzite and many structures exhibited unphysically small atomic pairwise distances. Furthermore, neither the two most stable structures wurtzite and zinc-blende were identified in the random-structure search nor any other phase from the training set. Moreover, the relaxations of the random structures with the CGCNN needed to be performed with a random-walk based approach as the unphysically roughness of the predicted CGCNN PES did not allow to perform a forces based relaxation. In contrast to the ASOAP and CGCNN, the CTP was able to identify the three most stable phases zinc-blende, wurtzite, and one with a space group 136. The latter phase was not considered in the training set. Moreover, it was not known for GaN before a recent prediction with DFT [118].
We have, furthermore, shown using the SISSO method how descriptors can be identified that predict which materials are harder to be described by the three considered ML methods, CTP, CGCNN, ASOAP. The introduced scheme to find such descriptors presents a route to identify the applicability boundaries of a model in the materials space and understand what further development step is needed for improving the model.
In summary, our work has demonstrated how ML potentials that were able to predict only across structure space could be extended towards the prediction across chemical composition space. The results are promising: the PES of a new compound can indeed be predicted with an accuracy that is needed to identify the most stable structures. However, the considered data set is limited and our analysis provides

only a proof-of-concept. Aiming at a wide-scale exploration of the materials space, our approach needs to be analyzed and possibly further developed on a larger data set including more complex systems, in the future. In particular, such analysis must involve a critical examination of the issue to which extent the ML based approach accelerates crystal-structure prediction and materials discovery compared to using DFT only.

# Bibliography


[1] F. J. DiSalvo, *Challenges and opportunities in solid-state chemistry.* Pure Appl. Chem. **72**, 1799-1807 (2000).

[2] P. Villars and S. Iwata, *PAULING FILE verifies / reveals 12 principles in materials science supporting four cornerstones given by Nature.* Chem. Met. Alloys **6**, 81-108 (2013).

[3] T. W. Eagar, *Bringing new materials to market.* Technol. Rev. **98**, 43 (1995).

[4] C. Wadia, https://mgi.gov/sites/default/files/documents/wadia_mgi_talk.pdf, Accessed 8 January 2021.

[5] G. Hautier, A. Jain, and S. P. Ong, *From the computer to the laboratory: materials discovery and design using first-principles calculations.* J. Mater. Sci. **47**, 7317 (2012).

[6] NOMAD (Novel Materials Discovery) repository, http://repository.nomad-coe.eu.

[7] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. B. Nardelli, N. Mingo, and O. Levy, *AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations.* Comput. Mater. Sci. **58**, 227-235 (2012).

[8] 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).

[9] 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).

[10] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, *The high-throughput highway to computational materials design.* Nat. Mater. **12**, 191-201 (2013).

[11] 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).

[12] G. Cao, R. Ouyang, L. M. Ghiringhelli, M. Scheffler, H. Liu, C. Carbogno, and Z. Zhang, *Artificial intelligence for high-throughput discovery of topological insulators: The example of alloyed tetradymites.* Phys. Rev. Mater. **4**, 034204 (2020).

[13] J. E. Saal, A. O. Oliynyk, and B. Meredig, *Machine Learning in Materials Discovery: Confirmed Predictions and Their Underlying Approaches.* Ann. Rev. of Mater. Res. **50**, 49-69 (2020).

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

[15] M. T. Yin and M. L. Cohen, *Theory of static structural properties, crystal stability, and phase transformations: Application to Si and Ge.* Phys. Rev. B **26**, 5668 (1982).

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

[17] D. C. Lonie and E. Zurek, *XtalOpt version r7: An open-source evolutionary algorithm for crystal structure prediction.* Comput. Phys. Commun. **182**, 372-387 (2011).

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

[19] D. J. Wales and H. A. Scheraga, *Global Optimization of Clusters, Crystals, and Biomolecules.* Science **285**, 1369 (1999).

[20] 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 (2004).

[21] C. J. Pickard and R. J. Needs, *High-Pressure Phases of Silane.* Phys. Rev. Lett. **97**, 045504 (2006).

[22] A. R. Oganov, C. J. Pickard, Q. Zhu, and R. J. Needs, *Structure prediction drives materials discovery.* Nat. Rev. Mater. **4**, 331-348 (2019).

[23] Y. Wang and Y. Ma, *Perspective: Crystal structure prediction at high pressures.* J. Chem. Phys. **140**, 040901 (2014).

[24] A. R. Oganov, *Crystal structure prediction: reflections on present status and challenges.* Faraday Discuss. **211**, 643-660 (2018).

[25] V. L. Deringer, M. A. Caro, and G. Csanyi, *Machine Learning Interatomic Potentials as Emerging Tools for Materials Science.* Science. Adv. Mater. **31**, 1902765 (2019).

[26] M. Born and R. Oppenheimer, *Zur Quantentheorie der Molekeln.* Ann. Phys. **389**, 457-484 (1927).

[27] V. L. Deringer and G. Csanyi, *Machine learning based interatomic potential for amorphous carbon.* Phys. Rev. B **95**, 094203 (2017).

[28] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, *Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning.* Phys. Rev. B **99**, 064114 (2019).

[29] A. P. Bartok, J. Kermode, N. Bernstein, and G. Csanyi, *Machine Learning a General-Purpose Interatomic Potential for Silicon.* Phys. Rev. X **8**, 041048 (2018).

[30] V. L. Deringer, C. J. Pickard, and G. Csanyi, *Data-Driven Learning of Total and Local Energies in Elemental Boron.* Phys. Rev. Lett. **120**, 156001 (2018).

[31] V. L. Deringer, D. M. Proserpio, G. Csanyi, and C. J. Pickard, *Data-driven learning and prediction ofinorganic crystal structures.* Faraday Discuss. **211**, 45 (2018).

[32] A. P. Bartok, M. C. Payne, R. Kondor, and G. Csanyi, *Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons.* Phys. Rev. Lett. **104**, 136403 (2010).

[33] R. Ouyang, Y. Xie, and D. Jiang, *Global minimization of gold clusters by combining neural network potentials and the basin-hopping method.* Nanoscale **7**, 14817 (2015).

[34] S. Jindal, S. Chiriki, and S. S. Bulusu, *Spherical harmonics based descriptor for neural network potentials: Structure and dynamics of $Au_{147}$ nanocluster.* J. Chem. Phys. **146**, 204301 (2017).

[35] Q. Tong, L. Xue, J. Lv, Y. Wang, and Y. Ma, , *Accelerating CALYPSO structure prediction by data-driven learning of a potential energy surface.* Faraday Discuss. **211**, 31 (2018).

[36] E. L. Kolsbjerg, A. A. Peterson, and B. Hammer, *Neural-network-enhanced evolutionary algorithm applied to supported metal nanoparticles.* Phys. Rev. B **97**, 19542 (2018).

[37] S. Faraji, S. A. Ghasemi, B. Parsaeifard, and S. Goedecker, *Surface reconstructions and premelting of the (100) $CaF_2$ surface..* Phys. Chem. Chem. Phys. **21**, 16270-16281 (2019).

[38] H. A. Eivari, S. A. Ghasemi, H. Tahmasbi, S. Rostami, S. Faraji, R. Rasoulkhani, S. Goedecker, and M. Amsler, *Two-Dimensional Hexagonal Sheet of $TiO_2$.* Chem. Mater. **29**, 8594 (2017).

[39] N. Bernstein, G. Csanyi, and V. L. Deringer, *De novo exploration and self-guided learning of potential-energy surfaces.* npj Comput. Mater. **5**, 99 (2019).

[40] 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).

[41] 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).

[42] 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).

[43] S. De, A. P. Bartok, G. Csanyi, and M. Ceriotti, *Comparing molecules and solids across structural and alchemical space.* Phys. Chem. Chem. Phys. **18**, 13754-13769 (2016).

[44] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K. Müller, *Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies.* J. Chem. Theory Comput. **9**, 3404-3419 (2013).

[45] B. Schölkopf, R. Herbrich, and A. J. Smola, *A Generalized Representer Theorem.* Proceedings of the 14th Annual Conferenceon Computational Learning Theory, volume **2111** of Lecture Notes in Artificial Intelligence, 416-426 (2001).

[46] R. Rojas, *Neural Networks: A Systematic Introduction.* Springer-Verlag, Berlin, New York (1996).

[47] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning.* MIT Press, Cambridge (2016).

[48] R. Tibshirani, *Regression Shrinkage and Selection via the Lasso.* J. R. Stat. Soc. B **58**, 267 (1996).

[49] 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).

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

[51] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research (2009).

[52] 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.* Sci. Adv. **5**, eaav0693 (2019).

[53] C. J. Bartel, S. L. Millican, A. M. Deml, J. R. Rumptz, W. Tumas, A. W. Weimer, S. Lany, V. Stevanovic, C. B. Musgrave, and A. M. Holder, *Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry.* Nat. Commun. **9**, 4168 (2018).

[54] R. Caruana, *Multitask Learning.* Mach. Learn. **28**, 41 (1997).

[55] G. Obozinski, B. Taskar, and M. Jordan, *Multi-task feature selection,.* Tech. Rep., Department of Statistics, University of California Berkeley (2006).

[56] A. Argyriou, T. Evgeniou, and M. Pontil, *Convex multi-task feature learning.* Mach. Learn. **73**, 243 (2008).

[57] X. Yin and X. Liu, *Multi-Task Convolutional Neural Network for Pose-Invariant Face Recognition.* IEEE Trans. Image Process. **27**, 964 (2018).

[58] P. Gong, J. Ye, and C. Zhang, *Multi-Stage Multi-Task Feature Learning.* J. Mach. Learn. Res. **14**, 2979 (2013).

[59] B. Huang, D. Ke, H. Zheng, B. Xu, Y. Xu, and K. Su, *Multi-task Learning Deep Neural Networks For Speech Feature Denoising.* Proc. INTERSPEECH, (Dresden, Germany), 2464-2468 (2015).

[60] K. Thung and C. Wee, *A brief review on multi-task learning.* Multimed. Tools Appl. **77**, 29705 (2018).

[61] Y. Zhang and Q. Yang, *An overview of multi-task learning.* Natl. Sci. Rev. **5**, 30 (2018).

[62] R. Ouyang, E. Ahmetcik, C. Carbogno, M. Scheffler, and L. M. Ghiringhelli, *Simultaneous learning of several materials properties from incomplete databases with multitask SISSO.* J. Phys. Mater. **2**, 024002 (2019).

[63] E. Ahmetcik, *Machine Learning of the Stability of Octet Binaries.* Master Thesis, Fritz-Haber-Institut der Max-Planck-Gesellschaft, https://th.fhi-berlin.mpg.de/site/uploads/Publications/Masterthesis_AhmetcikEmre.pdf (2016).

[64] F. H. Stillinger and T. A. Weber, *Computer simulation of local order in condensed phases of silicon.* Phys. Rev. B **31**, 5262 (1985).

[65] J. Tersoff, *New empirical approach for the structure and energy of covalent systems.* Phys. Rev. B **37**, 6991 (1988).

[66] D. W. Brenner, *The Art and Science of an Analytic Potential.* Phys. Status Solidi B **217**, 23 (2000).

[67] M. J. Willatt, F. Musil, and M. Ceriotti, *Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements.* Phys. Chem. Chem. Phys. **20**, 29661-29668 (2018).

[68] A. P. Bartok and G. Csanyi, *Gaussian Approximation Potentials: A Brief Tutorial Introduction.* Int. J. Quantum Chem. **115**, 1051-1057 (2015).

[69] https://github.com/libAtoms/QUIP.

[70] A. P. Bartok (private communication).

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

[72] C. Sutton, L. M. Ghiringhelli1, T. Yamamoto, Y. Lysogorskiy, L. Blumenthal, T. Hammerschmidt, J. R. Golebiowski, X. Liu, A. Ziletti, and M. Scheffler, *Crowd-sourcing materials-science challenges with the NOMAD 2018 Kaggle competition.* npj Comput. Mater. **5**, 111 (2019).

[73] M. F. Langer, A. Goeßmann, and M. Rupp, *Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning.* arXiv:2003.12081 (2020).

[74] M. J. Willatt, F. Musil, and M. Ceriotti, *Atom-density representations for machine learning.* J. Chem. Phys. **150**, 154110 (2019).

[75] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartok, C. Ortner, G. Csanyi, and M. Ceriotti, *On the Completeness of Atomic Structure Representations.* arXiv:2001.11696 (2020).

[76] A. P. Bartok, R. Kondor, and G. Csanyi, *On representing chemical environments.* Phys. Rev. B **87**, 184115 (2013).

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

[78] F. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld, *Alchemical and structural distribution based representation for universal quantum machine learning.* J. Chem. Phys. **148**, 241717 (2018).

[79] M. Rupp, A. Tkatchenko, K. Müller, and O. A. von Lilienfeld, *Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning.* Phys. Rev. Lett. **108**, 058301 (2012).

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

[81] H. Huo and M. Rupp, *Unified Representation for Machine Learning of Molecules and Crystals.* arXiv:1704.06439 (2017).

[82] A. Shapeev, *Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials.* Multiscale Model. Simul. **14**, 1153 (2016).

[83] T. Mueller, A. Hernandez, and C. Wang, *Machine learning for interatomic potential models.* J. Chem. Phys. **152**, 050902 (2020).

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

[85] 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).

[86] T. Xie and J. C. Grossman, *Hierarchical visualization of materials space with graph convolutional neural networks.* J. Chem. Phys. **149**, 174111 (2018).

[87] F. H. Stillinger, *Exponential multiplicity of inherent structure.* Phys. Rev. E **59**, 48 (1999).

[88] C. Nyshadham, M. Rupp, B. Bekker, A. V. Shapeev, T. Mueller, C. W. Rosenbrock, G. Csanyi, D. W. Wingate, and G. L. W. Hart, *Machine-learned multi-system surrogate models for materials prediction.* npj Comput. Mater. **5**, 51 (2019).

[89] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K. Müller, *Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies.* J. Chem. Theory Comput. **9**, 3404-3419 (2013).

[90] 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 (2015).

[91] https://github.com/txie-93/cgcnn.git, HEAD:b98c1cd109da971f599dd492ac7a5dad2c29e886.

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

[93] Volker L. Deringer, Albert P. Bartok, Noam Bernstein, David M. Wilkins, Michele Ceriotti, and Gabor Csanyi, *Gaussian Process Regression for Materials and Molecules.* Chem. Rev. **121**, 10073-10141 (2021)

[94] https://github.com/ahmetcik/Chemical-Transferable-Potentials.

[95] J. Tersoff, *New empirical model for the structural properties of silicon.* Phys. Rev. Lett. **56**, 632 (1986).

[96] M. I. Baskes, *Application of the Embedded-Atom Method to Covalent Materials: A Semiempirical Potential for Silicon.* Phys. Rev. Lett. **59**, 23 (1987).

[97] H. Balamane, T. Halicioglu, and W. A. Tiller, *Comparative study of silicon empirical interatomic potentials.* Phys. Rev. B **46**, 2250 (1992).

[98] M. I. Baskes, *Modified embedded-atom potentials for cubic materials and impurities.* Phys. Rev. B **46**, 2727 (1992).

[99] T. J. Lenosky, B. Sadigh, E. Alonso, V. V. Bulatov, T. D. d. l. Rubia, J. Kim, A. F. Voter, and J. D. Kress, *Highly optimized empirical potential model of silicon.* Model. Simul. Mater. Sci. Eng. **8**, 825 (2000).

[100] A. C. T. van Duin, A. Strachan, S. Stewman, Q. Zhang, X. Xu, and W. A. G. III, *$ReaxFF_{SiO}$ Reactive Force Field for Silicon and Silicon Oxide Systems.* J. Phys. Chem. A **107**, 3803-3811 (2003).

[101] M. J. Buehler, A. C. T. van Duin, and W. A. G. III, *Multiparadigm Modeling of Dynamical Crack Propagation in Silicon Using a Reactive Force Field.* Phys. Rev. Lett. **96**, 095505 (2006).

[102] 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 (1992).

[103] C. Carbogno, C. G. Levi, C. G. V. d. Walle, and M. Scheffler, *Ferroelastic switching of doped zirconia: Modeling and understanding from first principles.* Phys. Rev. B **90**, 144109 (2014).

[104] A. Evans, D. Clarke, and C. Levi, *The influence of oxides on the performance of advanced gas turbines.* J. Eur. Ceram. Soc. **28**, 1405 (2008).

[105] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications.* Academic press (2001).

[106] B. Cheng, E. A. Engel, J. Behler, C. Dellago, and M. Ceriotti, *Ab initio thermodynamics of liquid and solid water.* Proc. Natl. Acad. Sci. U. S. A. **116**, 1110-1115 (2019).

[107] D. A. Kofke, *Direct evaluation of phase coexistence by molecular simulation via integration along the saturation line.* J. Chem. Phys. **98**, 4149 (1993).

[108] G. Bussi, D. Donadio, and M. Parrinello, *Canonical sampling through velocity rescaling.* J. Chem. Phys. **126**, 014101 (2007).

[109] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, *Ab initio molecular simulations with numeric atom-centered orbitals.* Comput. Phys. Commun. **180**, 2175 (2009).

[110] C. Carbogno, R. Ramprasad, and M. Scheffler, *Ab Initio Green-Kubo Approach for the Thermal Conductivity of Solids.* Phys. Rev. Lett. **118**, 175901 (2017).

[111] C. Ricca, A. Ringuede, M. Cassir, C. Adamo, and F. Labat, *A comprehensive DFT investigation of bulk and low-index surfaces of $ZrO_2$ polymorphs.* J. Comput. Chem. **36**, 9-21 (2015).

[112] F. Frey, H. Boysen, and T. Vogt, *Neutron powder investigation of the monoclinic to tetragonal phase transformation in undoped zirconia.* Acta Crystallogr. B **46**, 724-730 (1990).

[113] C. J. Howard, R. J. Hill, and B. E. Reichert, *Structures of the $ZrO_2$ Polymorphs at Room Temperature by High-Resolution Neutron Powder Diffraction.* Acta Crystallogr. B **44**, 116 (1988).

[114] A. P. Bartok, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csanyi, and M. Ceriotti, *Machine learning unifies the modeling of materials and molecules.* Sci. Adv. **3**, e1701816 (2017).

[115] B. Cordero, V. Gomez, A. E. Platero-Prats, M. Reves, J. Echeverria, E. Cremades, F. Barragan, and S. Alvarez, *Covalent radii revisited.* Dalton Trans., 2832-2838 (2008).

[116] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, *The Cambridge Structural Database.* Acta Cryst. B **72**, 171-179 (2016).

[117] https://github.com/cosmo-epfl/glosim2.git, HEAD:f59d6c046340bf4cac71f189f4b6146d05d2b33c.

[118] A. Sun, S. Gao, and G. Gu, *Stability and electronic properties of GaN phases with inversion symmetry to inherently inhibit polarization.* Phys. Rev. Mater. **3**, 104604 (2019).

[119] A. Munoz and K. Kunc, *High-pressure phase of gallium nitride.* Phys. Rev. B **44**, 10372 (1991).

[120] H. Xia, Q. Xia, and A. L. Ruoff, *High-pressure structure of gallium nitride: Wurtzite-to-rocksalt phase transition.* Phys. Rev. B **47**, 12925 (1993).

[121] K. Sarasamak, A. J. Kulkarni, M. Zhou, and S. Limpijumnong, *Stability of wurtzite, unbuckled wurtzite, and rocksalt phases of SiC, GaN, InN, ZnO, and CdSe under loading of different triaxialities.* Phys. Rev. B **77**, 024104 (2008).

[122] C. L. Freeman, F. Claeyssens, N. L. Allan, and J. H. Harding, *Graphitic Nanofilms as Precursors to Wurtzite Films: Theory.* Phys. Rev. Lett. **96**, 066102 (2006).

[123] C. Sutton, M. Boley, L. M. Ghiringhelli, M. Rupp, J. Vreeken, and M. Scheffler, *Identifying domains of applicability of machine learning models for materials science.* Nat. Commun. **11**, 4428 (2020).

[124] P. Rowe, V. L. Deringer, P. Gasparotto, G. Csaanyi, and A. Michaelides, *An accurate and transferable machine learning potential for carbon.* J. Chem. Phys. **153**, 034702 (2020).

[125] C. Zeni, *Gaussian process regression for nonparametric force fields.* PhD Thesis, King's College London, https://kclpure.kcl.ac.uk/portal/files/131638733/2020_Zeni_Claudio_1556527_ethesis.pdf (2020).

[126] http://web.archive.org/web/20110902084216/http://cst-www.nrl.navy.mil/lattice/index.html..

[127] http://www.libatoms.org/Home/DataRepository.

[128] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods.* Cambridge University Press, Cambridge (2010).

[129] P. Hohenberg and W. Kohn, *Inhomogeneous Electron Gas.* Phys. Rev. **136**, B864 (1964).

[130] W. Kohn and L. J. Sham, *Self-Consistent Equations Including Exchange and Correlation Effects.* Phys. Rev. **140**, A1133 (1965).

[131] P. A. M. Dirac, *Note on Exchange Phenomena in the Thomas Atom.* Math. Proc. Cambridge Philos. Soc. **26**, 376 (1930).

[132] D. M. Ceperley and B. J. Alder, *Ground State of the Electron Gas by a Stochastic Method.* Phy. Rev. Lett. **45**, 566 (1980).

[133] J. P. Perdew, K. Burke, and M. Ernzerhof, *Generalized Gradient Approximation Made Simple.* Phys. Rev. Lett. **77**, 3865 (1996).

[134] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, *Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces.* Phys. Rev. Lett. **100**, 136406 (2008).

[135] J. Heyd and G. E. Scuseria, *Hybrid functionals based on a screened Coulomb potential.* J. Chem. Phys. **118**, 8207 (2003).

[136] T. Wai Ko, J. A. Finkler, S. Goedecker, and J. Behler, *A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer.* Nat. Commun. **12**, 398 (2021).

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

# A Derivation of the MT-SISSO correlation measure

MT-SISSO determines models in an iterative process. At each SIS step, it evaluates which single descriptors are closest to the current residuals $\boldsymbol{R}^q$. We determine the closest descriptor $j^*$ by:

$$j^* = \arg\min_j \min_{\boldsymbol{C}_j \in \mathbb{R}^Q} \sum_{q=1}^{Q} \frac{1}{N_q} \left\| \boldsymbol{R}^q - \boldsymbol{D}_j^q C_j^q \right\|_2^2 . \tag{A.1}$$

Let us define the least-squares solution with a vector of lower case letters:

$$(c_j^1, c_j^2, ..., c_j^Q) = \arg\min_{\boldsymbol{C}_j \in \mathbb{R}^Q} \sum_{q=1}^{Q} \frac{1}{N_q} \left\| \boldsymbol{R}^q - \boldsymbol{D}_j^q C_j^q \right\|_2^2 . \tag{A.2}$$

Then, for all $j \neq j^*$:

$$\sum_{q=1}^{Q} \frac{1}{N_q} \left\| \boldsymbol{R}^q - \boldsymbol{D}_{j^*}^q c_{j^*}^q \right\|_2^2 \leq \sum_{q=1}^{Q} \frac{1}{N_q} \left\| \boldsymbol{R}^q - \boldsymbol{D}_j^q c_j^q \right\|_2^2 . \tag{A.3}$$

$$\Leftrightarrow \sum_{q=1}^{Q} \frac{1}{N_q} \left\langle \boldsymbol{R}^q - \boldsymbol{D}_{j^*}^q c_{j^*}^q, \boldsymbol{R}^q - \boldsymbol{D}_{j^*}^q c_{j^*}^q \right\rangle \leq \sum_{q=1}^{Q} \frac{1}{N_q} \left\langle \boldsymbol{R}^q - \boldsymbol{D}_j^q c_j^q, \boldsymbol{R}^q - \boldsymbol{D}_j^q c_j^q \right\rangle . \tag{A.4}$$

In the least-squares regression, $\boldsymbol{R}^q$ is orthogonally projected onto the space spanned by the vector $\boldsymbol{D}_j^q$. Thus, $\langle \boldsymbol{R}^q - \boldsymbol{D}_j^q c_j^q, \boldsymbol{D}_j^q c_j^q \rangle = 0$. We obtain:

$$\Leftrightarrow \sum_{q=1}^{Q} \frac{1}{N_q} \left\langle \boldsymbol{R}^q - \boldsymbol{D}_{j^*}^q c_{j^*}^q, \boldsymbol{R}^q \right\rangle \leq \sum_{q=1}^{Q} \frac{1}{N_q} \left\langle \boldsymbol{R}^q - \boldsymbol{D}_j^q c_j^q, \boldsymbol{R}^q \right\rangle \tag{A.5}$$

$$\Leftrightarrow \sum_{q=1}^{Q} \frac{1}{N_q} \left( \langle \boldsymbol{R}^q, \boldsymbol{R}^q \rangle - \left\langle \boldsymbol{D}_{j^*}^q c_{j^*}^q, \boldsymbol{R}^q \right\rangle \right) \leq \sum_{q=1}^{Q} \frac{1}{N_q} \left( \langle \boldsymbol{R}^q, \boldsymbol{R}^q \rangle - \left\langle \boldsymbol{D}_j^q c_j^q, \boldsymbol{R}^q \right\rangle \right) \tag{A.6}$$

$$\Leftrightarrow - \sum_{q=1}^{Q} \frac{1}{N_q} \left\langle \boldsymbol{D}_{j^*}^q c_{j^*}^q, \boldsymbol{R}^q \right\rangle \leq - \sum_{q=1}^{Q} \frac{1}{N_q} \left\langle \boldsymbol{D}_j^q c_j^q, \boldsymbol{R}^q \right\rangle \tag{A.7}$$

$$\Leftrightarrow \sum_{q=1}^{Q} \frac{1}{N_q} \left\langle \boldsymbol{D}_{j^*}^q c_{j^*}^q, \boldsymbol{R}^q \right\rangle \geq \sum_{q=1}^{Q} \frac{1}{N_q} \left\langle \boldsymbol{D}_j^q c_j^q, \boldsymbol{R}^q \right\rangle \tag{A.8}$$

$$\Leftrightarrow \sum_{q=1}^{Q} \frac{1}{N_q} \left\langle \boldsymbol{D}_{j^*}^q, \boldsymbol{R}^q \right\rangle c_{j^*}^q \geq \sum_{q=1}^{Q} \frac{1}{N_q} \left\langle \boldsymbol{D}_j^q, \boldsymbol{R}^q \right\rangle c_j^q . \tag{A.9}$$

The least-squares solution is given by $c_j = \frac{1}{\|\boldsymbol{D}_j^q\|_2^2}\langle \boldsymbol{D}_j^q, \boldsymbol{R}^q \rangle$. Hence:

$$\Leftrightarrow \sum_{q=1}^{Q} \frac{1}{N_q} \frac{1}{\|\boldsymbol{D}_{j^*}^q\|_2^2} \left\langle \boldsymbol{D}_{j^*}^q, \boldsymbol{R}^q \right\rangle^2 \geq \sum_{q=1}^{Q} \frac{1}{N_q} \frac{1}{\|\boldsymbol{D}_j^q\|_2^2} \left\langle \boldsymbol{D}_j^q, \boldsymbol{R}^q \right\rangle^2 . \tag{A.10}$$

All $\boldsymbol{D}_j^q$ have same lengths $\left\|\boldsymbol{D}_j^q\right\|_2$ if the vectors are standardized. Thus:

$$\Leftrightarrow \sum_{q=1}^{Q} \frac{1}{N_q} \left\langle \boldsymbol{D}_{j^*}^q, \boldsymbol{R}^q \right\rangle^2 \geq \sum_{q=1}^{Q} \frac{1}{N_q} \left\langle \boldsymbol{D}_j^q, \boldsymbol{R}^q \right\rangle^2 \tag{A.11}$$

$$\Leftrightarrow \sqrt{\sum_{q=1}^{Q} \left\langle \boldsymbol{D}_{j^*}^q, \boldsymbol{R}^q \right\rangle^2 / N_q} \geq \sqrt{\sum_{q=1}^{Q} \left\langle \boldsymbol{D}_j^q, \boldsymbol{R}^q \right\rangle^2 / N_q} \tag{A.12}$$

$$\Leftrightarrow \theta_{j^*} \geq \theta_j . \tag{A.13}$$

Thus, at each SIS step, we search for the descriptors with the largest linear correlation scores $\theta_j$.

# B Determining well-defined phase diagrams with multi-task SISSO

MT-SISSO determines the same descriptors for all tasks (relative stabilities). As a consequence, the crystal-structure stability among five phases can be uniquely described by four independent relative stabilities. For example, for any three structures $\alpha$, $\beta$, $\gamma$, the difference of the predicted energies $E(\alpha) - E(\gamma)$ is by construction equal to $(E(\alpha) - E(\beta)) - (E(\gamma) - E(\beta))$. This can be shown by considering the least-squares solution

$$\boldsymbol{c}^{\alpha,\gamma} = (\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})^{-1}\boldsymbol{D}^{\mathrm{T}}(\boldsymbol{E}^{\alpha} - \boldsymbol{E}^{\gamma}), \tag{B.1}$$

which determines during the $\ell_0$ step of the MT-SISSO algorithm the model of the task corresponding to the target $E(\alpha) - E(\gamma)$. Here, $\boldsymbol{E}^{\mu}$ denotes the vector of training energies of structure $\mu$ and $\boldsymbol{D}$ is the descriptor matrix made of the $\Omega$-dimensional descriptor that was identified by MT-SISSO. The prediction of any (new) data point represented by $\boldsymbol{d}$ (e.g. a row of $\boldsymbol{D}$) yields:

$$\begin{aligned}
(E(\alpha) - E(\beta)) - (E(\gamma) - E(\beta)) &= \boldsymbol{d}\boldsymbol{c}^{\alpha,\beta} - \boldsymbol{d}\boldsymbol{c}^{\gamma,\beta} && \text{(B.2)}\\
&= \boldsymbol{d}\left[\boldsymbol{c}^{\alpha,\beta} - \boldsymbol{c}^{\gamma,\beta}\right] && \text{(B.3)}\\
&= \boldsymbol{d}(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})^{-1}\boldsymbol{D}^{\mathrm{T}}\left[(\boldsymbol{E}^{\alpha} - \boldsymbol{E}^{\beta}) - (\boldsymbol{E}^{\gamma} - \boldsymbol{E}^{\beta})\right] && \text{(B.4)}\\
&= \boldsymbol{d}(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})^{-1}\boldsymbol{D}^{\mathrm{T}}\left[\boldsymbol{E}^{\alpha} - \boldsymbol{E}^{\gamma}\right] && \text{(B.5)}\\
&= \boldsymbol{d}\boldsymbol{c}^{\alpha,\gamma} && \text{(B.6)}\\
&= E(\alpha) - E(\gamma). && \text{(B.7)}
\end{aligned}$$

The important condition to derive the above relationship is that for both (all) tasks the same descriptors are used, which is the case in MT-SISSO. Note that we have assumed that for both tasks (target properties) $(E(\alpha) - E(\beta))$ and $(E(\gamma) - E(\beta))$ the same materials were included in the training data set. However, in an incomplete database, this might not be the case. In such a case, in a pre-step, missing target-property values need to be filled by the predictions of models that were fitted to the incomplete data. For instance, refitting the model to the filled data set will not change the least-squares solution.

# C Octet binaries data set

The OB data set consists of 4 840 structures and DFT cohesive energies with 78 octet binary compounds in eight different crystal structure prototypes: zinc-blende (ZB), rock-salt (RS), CsCl, NaTl, NiAs, CoSn, NbP, CrB. Details about the space groups and number of atoms in the unit cell are listed in in Tab. C.1. Information about the atomic and lattice-vector coordinates can be found in [126]. The 78 compounds are given by:
LiF, LiCl, LiBr, LiI, NaF, NaCl, NaBr, NaI, KF, KCl, KBr, KI, RbF, RbCl, RbBr, RbI,CsF, CsCl, CsBr, CsI, AgF, AgCl, AgBr, AgI, CuF, CuCl, CuBr, CuI, BeO, BeS, BeSe,BeTe, MgO, MgS, MgSe, MgTe, CaO, CaS, CaSe, CaTe, SrO, SrS, SrSe, SrTe, BaO, BaS,BaSe, BaTe, ZnO, ZnS, ZnSe, ZnTe, CdO, CdS, CdSe, CdTe, BN, BP, BAs, BSb, AlN,AlP, AlAs, AlSb, GaN, GaP, GaAs, GaSb, InN, InP, InAs, InSb, SnGe, SnSi, SnC, GeSi, GeC, SiC.
Note that in Chapter 3, also the four elemental solids C, Si, Ge, Sn are included in the data set and considered as octet binaries.

Every (compound, phase) tuple is characterized by eight-point energy-volume curves. An exception is the CrB phase, where only one data point per compound is given, because the reference work [63] used a different optimization technique for the CrB phase than for the other phases. The equilibria of the seven phases, whose crystal symmetries depend on one or two degrees of freedom, were determined by Birch-Murnaghan fits. Using the Birch-Murnaghan fit also for the CrB phase was, however, not possible because its crystal symmetry depends on five degrees of freedom. Instead, a symmetry-constrained relaxation was performed and we used only the equilibrium structures.
For 56 compounds, there are two energy-volume curves present in the CoSn phase, each with exchanged occupation of the sites in the crystal by the atomic species types (AB and BA). The data set does not contain both CoSn(AB) and CoSn(BA) for all 78 compounds because the reference work [63] aimed to find only the energetically lower structure of the two types. For compounds who showed already in estimations, performed with light numerical DFT settings, a large energy gap between the two types, only the lower one was calculated with more accurate settings.

Fig. C.2 shows the correlation of the nearest neighbour distance inside a structure with the dimer equilibrium distance and a descriptor that depends on orbital based radii. Only structures at the equilibria of the corresponding phases are shown.

| Prototype | Space group | Crystal system | $N_{\text{atoms}}$ in unit cell |
|---|---|---|---|
| ZB | 216 | cubic | 2 |
| RS | 225 | cubic | 2 |
| CsCl | 221 | cubic | 2 |
| NaTl | 227 | cubic | 4 |
| NiAs | 194 | hexagonal | 4 |
| CoSn | 191 | hexagonal | 6 |
| NbP | 141 | tetragonal | 4 |
| CrB | 63 | orthorhombic | 4 |

**Table C.1:** Details about the crystal-structure types present in the octet binaries data set.

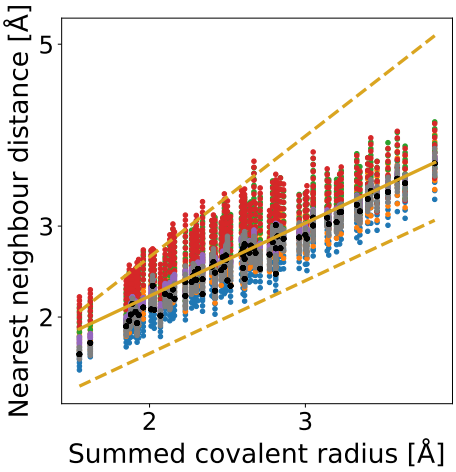



**Figure C.1:** Same figure as the right panel of Fig. 5.6, however, with all training data points instead of only equilibrium structures.

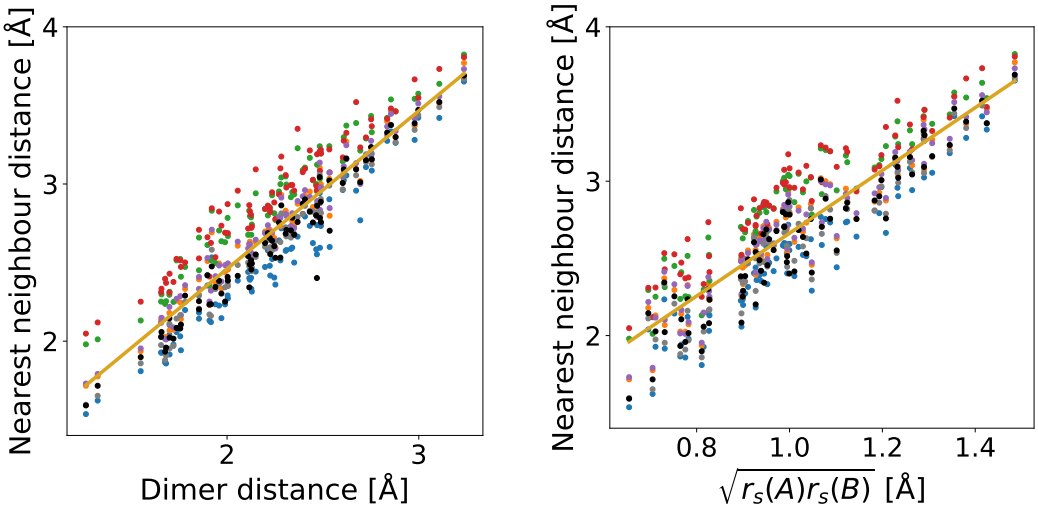



**Figure C.2:** Correlation of the nearest neighbour distances with the equilibrium dimer distances (left) and $\sqrt{r_s(A)r_s(B)}$ (right) for the octet binaries data set. The dimers are given by the same species tuples A-B that determine the $AB$ compounds. $r_s$ represents the radius where the radial probability density of the valence $s$ orbitals is maximal. More details about the figure can be found in Fig. 5.6.

# D Lists of model hyperparameters

| | |
|---|---|
| cutoff radius | 6 Å |
| cutoff-transition width | 2 Å |
| widths of zero-centered two-body Gaussians | 0.7, 1.0, 1.3, 1.5, 2.0, 3.0, 4.0 |
| widths of zero-centered three-body Gaussians | 0.6, 0.8, 1.0, 1.3, 1.5, 1.7, 2.0, 2.5, 3.0, 4.0, 5.0 |

**Table D.1:** Hyperparameters of the potentials applied to the NOMAD 2018 Kaggle competition (Sec. 5.3.1).

| | |
|---|---|
| cutoff radius | 5 Å |
| cutoff-transition width | 2 Å |
| widths of zero-centered two-body Gaussians | 0.5, 0.7, 1.0, 1.3, 1.5, 2.0, 3.0, 4.0 |
| number of three-body Gaussians | 432 |
| widths of three-body Gaussians | 1.3 |

**Table D.2:** Hyperparameters of the potentials for the description of silicon in its different phases (Sec. 5.3.2).

| | |
|---|---|
| cutoff radius for the 2b | 5 Å |
| cutoff radius for the 3b | 4 Å |
| number of sparse points 2b | 12 |
| number of sparse points 3b | 300 |
| weight $\delta_2 b$ on 2b potential | 2 |
| weight $\delta_2 b$ on 2b potential | 0.6 |
| sparse method | "uniform" |
| regularization parameter $\sigma_{\text{energy}}$ | 0.0014 |
| regularization parameter $\sigma_{\text{force}}$ | 0.1 |

**Table D.3:** Hyperparameters of the potentials for the description of the thermodynamics of zirconia (Sec. 5.3.3), as implemented in the QUIP package [69].

| | |
|---|---|
| cutoff radius | 6 Å |
| cutoff-transition width | 6 Å |
| widths of zero-centered two-body Gaussians | 0.5, 0.7, 1.0, 1.5, 2.0, 3.0 |
| two-body Gaussians with negative $\alpha_\mu$ | 0.5, 0.7 |
| two-body Gaussians with positive $\alpha_\mu$ | 1.0, 1.5, 2.0, 3.0 |
| number of three-body Gaussians | 128 |
| widths of three-body Gaussians | 1.5 |
| number of hidden layers in two-body neural network | 1 |
| number of hidden layers in three-body neural network | 1 |
| number of neurons in two-body neural network | 500 |
| number of neurons in three-body neural network | 500 |
| activation function in two-body neural network | ReLU |
| activation function in three-body neural network | ReLU |
| batch size | 32 |
| number of epochs | 400 |

**Table D.4:** Hyperparameters of the chemical-transferable potentials as used in Sec. 5.4.6.

| | |
|---|---|
| cutoff radius | 5 Å |
| cutoff-transition width | 1 Å |
| $n_{\max}$ | 9 |
| $l_{\max}$ | 9 |
| Gaussian width | 0.5 |
| $\zeta$ | 2 |

**Table D.5:** Hyperparameters of the alchemical smooth overlap of atomic positions (ASOAP) as used in Sec. 5.4.6

| | |
|---|---|
| max. number of neighbours | 12 |
| cutoff radius | 8 Å |
| minimum radius | 0 Å |
| radius steps for the Gaussian centers | 0.2 Å |
| length of atomic feature vector | 64 |
| length of hidden-layer feature vector | 64 |
| number of convolutional layers | 4 |
| batch size | 256 |
| number of epochs | 5000 |

**Table D.6:** Hyperparameters of the crystal graph convolutional neural networks (CGCNN) as used in Sec. 5.4.6.

# E Machine-learning potentials for fixed compositions

## E.1 Phases of silicon

### E.1.1 Structural descriptor map

The regression coefficients $\boldsymbol{\alpha}$ of the ML potentials used in this work are optimized within a linear regression problem, as described in Sec. 4.3.2. The linear system to fit the energies is written:

$$\boldsymbol{E} = \boldsymbol{B}\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{B}_{2\mathrm{b}} & \boldsymbol{B}_{3\mathrm{b}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_{2\mathrm{b}} \\ \boldsymbol{\alpha}_{3\mathrm{b}} \end{pmatrix} \tag{E.1}$$

The components of the matrix $\boldsymbol{B}$ contain structural information of the crystals represented by sums of basis functions over the the atomic environments in the crystal. In other words, a row of $\boldsymbol{B}$ is a vectorial descriptor representing a structure. Descriptors based on basis functions that are summed over the atomic environments were used to represent crystal structures and predict energies with kernel ridge regression in Ref. [81]. Given the fact that Eq. E.1 determines a linear model, a principal-component analysis is well suited to decompose the summed structural information, i.e. into contributions with the highest variances. The two leading principal components of the structural descriptors used to fit the energies of 1 451 silicon structures in nine phases are shown in 5.3. The three leading principal components are shown in Fig. E.1. The first two components together describe 98.8% of the variance of the data, the first four components each describe 73.2%, 25.6%, 1.0%, and 0.1%.

The structure map in Fig. 5.3 is only partially visualizing the distribution of the structures in the descriptor space. In fact, the two-dimensional descriptor is not able to separate the phases into non-overlapping domains, and two structures of different phases might not be distinguishable. We find, however, that when using the full high dimensional descriptor $\boldsymbol{B}$ instead of its first principal components, the phases are separable into non-overlapping domains. Performing a classification of the data into their phase labels with a linear support vector machine already yields an almost perfect pairwise separation of the phases. Only twelve data points were misclassified and belong to the $\beta$-Sn and sh phase. We checked that with a Gaussian kernel based support vector machine the two phases can be separated perfectly into non-overlapping domains

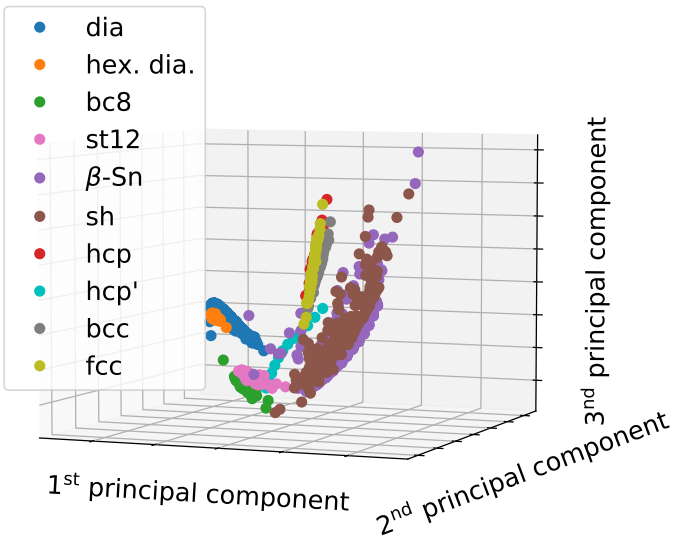



**Figure E.1:** The three leading principal components of the structural descriptors used to fit the energies of 1 451 silicon structures in nine phases. Further twelve structures in the second hcp phase are added.

which are each fully connected. The connectivity is determined by performing a walk along a line (in the descriptor space) between every pair of structures of the same phase and ensuring that the predicted class does not change during the walk. Note that one cannot conclude from the perfect classification alone that two regions are non-overlapping because with a sufficiently small Gaussian width, a Gaussian kernel based support vector machine is, in general, able to classify any data set perfectly, if two points of different classes do not have the same descriptor. The connectivity-test is needed to clarify if the model is not alternating the predicted class between two points of the same class.

### E.1.2 Comparison of the diamond and hexagonal diamond structure

Fig. E.2 shows the neighbours of an atom in the diamond and hexagonal diamond structure, both at the equilibrium. More precisely, the neighbours within the cutoff of the ML potential are shown. Note that every atom within a structure has the same structural environment when considering diamond or hexagonal diamond. When comparing diamond to hexagonal diamond, the relaxed structures of the two phases have (almost) the same 22 neighbours (brown atoms in Fig. E.2), i.e. the positions differ from each other on average by 0.01 Å. Further six neighbours in the case of diamond (blue atoms) and four in hexagonal diamond (orange atoms) yield the major difference of the both structures

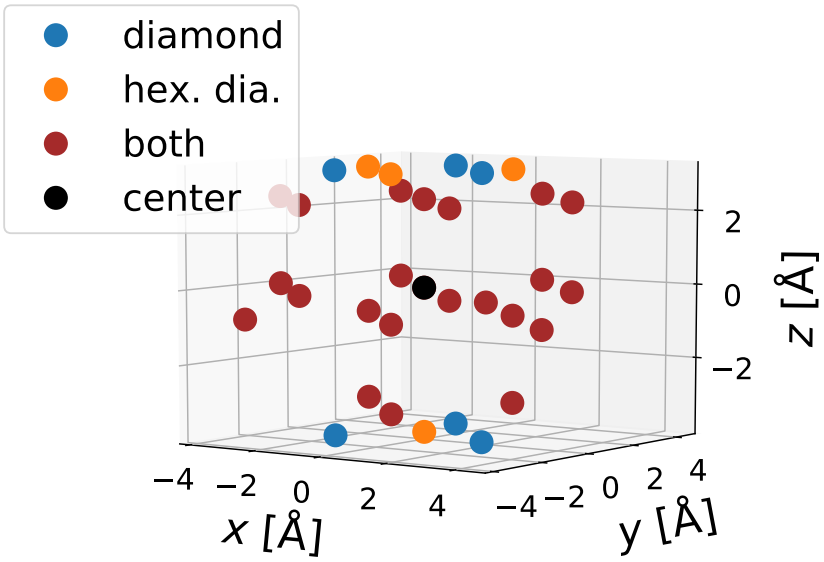



**Figure E.2:** Positions of neighbours around a central atom in the diamond and hexagonal diamond phase. The central atom is located at the origin.

and the apparent volume and energy difference predicted by our ML potential.
Let us consider a model trained on all phases but diamond (Fig. 5.2 b)). The volume difference between hexagonal diamond and diamond is predicted to be 0.27 Å$^3$/atom and the energy difference 71 meV/atom. In comparison, the reference DFT values are 0.01 Å$^3$/atom and 11 meV/atom. Removing the energy and forces contributions of the ML model that are based on the neighbours that distinguish the two structures significantly (the blue and orange markers in Fig. E.2), the differences become 0.01 Å$^3$/atom and 3 meV/atom for rerelaxed structures.
Similarly, we observed deviations of the predicted quantities of diamond and hexagonal dimanond for the GAP of Ref. [29] that we retrained, in this work, only on bulk phases without diamond, see Fig. E.3 b). While the energy difference was predicted with 36 meV/atom more accurately than in case of our model, the volume difference is given by 1.73 Å$^3$/atom. If the GAP is trained on all (solid) phases but both diamond and hexagonal diamond, the predictions become more accurate (Fig. E.3 c)), i.e. 0.4 Å$^3$/atom and 20 meV/atom. It is, however, not clear why the additional exclusion of the hexagonal diamond phase from the training data improves the predictions of the GAP.

### E.1.3 Performance of the reference machine-learning potential

In order to compare the performances of the ML potential of this work and the GAP of Ref. [29], we performed the tests demonstrated in Fig. 5.2 for the GAP, using the same training data of the solid phases. The results are shown in Fig. E.3. The errors of both potentials, our ML potential and retrained GAP, are tabulated in Tab. E.1.

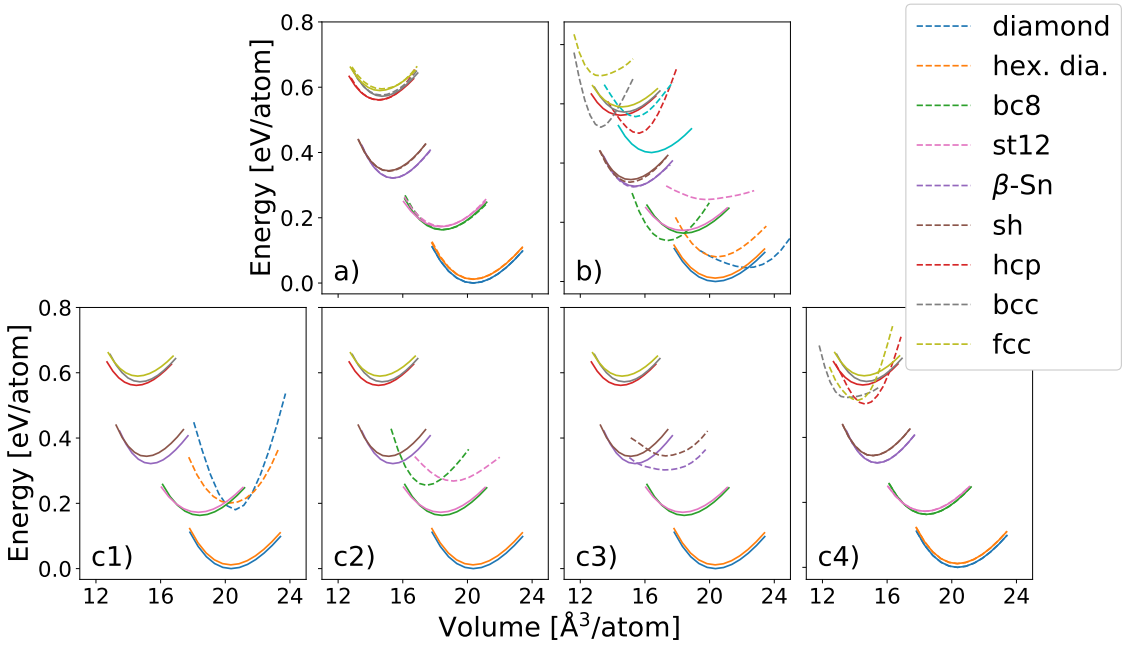



**Figure E.3:** Predictions of the reference Gaussian approximation potential [29] as a comparison to the results of our potential (Fig. 5.2). The figure shows energy-volume curves for nine phases of silicon. The solid lines represent the reference DFT energies, the dashed lines the predictions of the ML potential. The curves are obtained by varying the volume and performing at each volume a constant-volume and fixed-crystal-symmetry relaxation. a) shows the predictions of a potential fitted to a data of all nine phases. In b) each phase was once left out from the training set and predicted from a potential trained on the remaining phases. Furthermore, the prediction of a second hcp phase (hcp') was added using a potential trained on the initial nine phases. In c1) - c4) the nine phases are divided, based on volume and energy similarity, into four groups: (diamond, hex. dia.), (bc8, st12), ($\beta$-Sn, sh), and (hcp, bcc, fcc). Each group is once left out into a test set and predicted by a potential trained on the remaining phases.

The performance of the two potentials is comparable. While the GAP was able to yield a slightly more accurate model when all phases were included in the training set (3 meV/atom vs 1 meV/atom for MAE of predicted energies in test a)), our potential yields a lower average prediction error on phases not seen in the training set, i.e. 45 meV/atom vs 60 meV/atom for energies in test b) and 78 meV/atom vs 85 meV/atom for energies in test c). However, the reference potential is able to better recognize similar structures as similar, e.g. the MAE for the energy differences within the groups in test c) is 12 meV/atom for the reference potential and 78 meV/atom for the one of this work. Note that in case of the group with the three phases the energy difference is built with respect to the lowest energy phase as determined by DFT.

| | Energy [meV/atom] | | | | | | Volume [Å$^3$/atom] | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | a) | | b) | | c) | | a) | | b) | | c) | |
| | Our | Ref. | Our | Ref. | Our | Ref. | Our | Ref. | Our | Ref. | Our | Ref |
| diamond | 0 | 0 | 64 | 47 | 116 | 180 | 0.0 | 0.0 | 0.4 | 1.8 | 1.4 | 0.3 |
| hex. dia. | 8 | 0 | 45 | 71 | 202 | 189 | 0.1 | 0.0 | 0.3 | 0.1 | 0.8 | 0.1 |
| bc8 | 5 | 0 | 11 | 25 | 2 | 92 | 0.1 | 0.0 | 0.0 | 1.1 | 1.1 | 1.0 |
| st12 | 2 | 1 | 89 | 104 | 50 | 96 | 0.1 | 0.1 | 0.7 | 1.4 | 0.9 | 0.8 |
| $\beta$-Sn | 1 | 0 | 15 | 2 | 48 | 19 | 0.0 | 0.0 | 0.1 | 0.1 | 0.8 | 1.8 |
| sh | 0 | 1 | 5 | 9 | 32 | 1 | 0.1 | 0.0 | 0.1 | 0.1 | 0.6 | 2.1 |
| hcp | 2 | 0 | 45 | 61 | 56 | 59 | 0.1 | 0.1 | 1.0 | 1.1 | 0.7 | 0.2 |
| hcp' | - | - | 74 | 121 | - | - | - | - | 0.0 | 1.0 | - | - |
| bcc | 3 | 3 | 48 | 51 | 9 | 49 | 0.0 | 0.0 | 0.6 | 1.4 | 1.8 | 1.2 |
| fcc | 3 | 5 | 57 | 105 | 188 | 75 | 0.4 | 0.1 | 0.4 | 1.3 | 1.3 | 0.4 |
| MAE | 3 | 1 | 45 | 60 | 78 | 85 | 0.1 | 0.0 | 0.4 | 0.9 | 1.1 | 0.9 |

**Table E.1:** Prediction errors of our ML potential and the Gaussian approximation potential trained in this work for the tests described in Fig. 5.2.

### E.1.4 Tables of predicted values

The predicted cohesive energy, volume, bulk modulus, and derivative of the bulk modulus within the tests described in Fig. 5.2 are listed in E.2 and E.3. The bulk modulus and its derivative were obtained from a fit to the Birch-Murnaghan equation of states. Using the Maxwell construction, we have calculated the diamond-to-$\beta$-Sn transition pressure, see Tab. E.4.

| | Cohesive energy [eV/atom] | | | | Volume [Å$^3$/atom] | | | |
|---|---|---|---|---|---|---|---|---|
| | DFT | Our ML potential | | | DFT | Our ML potential | | |
| | | a) | b) | c) | | a) | b) | c) |
| diamond | 4.633 | 4.633 | 4.697 | 4.517 | 20.4 | 20.4 | 20.0 | 18.9 |
| hex. dia. | 4.622 | 4.630 | 4.576 | 4.420 | 20.3 | 20.5 | 20.7 | 19.5 |
| bc8 | 4.470 | 4.475 | 4.481 | 4.468 | 18.4 | 18.3 | 18.4 | 17.3 |
| st12 | 4.461 | 4.459 | 4.372 | 4.411 | 18.3 | 18.4 | 19.0 | 17.5 |
| $\beta$-Sn | 4.311 | 4.312 | 4.326 | 4.360 | 15.4 | 15.4 | 15.3 | 16.2 |
| sh | 4.289 | 4.289 | 4.294 | 4.257 | 15.1 | 15.1 | 15.0 | 15.7 |
| hcp | 4.072 | 4.074 | 4.026 | 4.127 | 14.5 | 14.6 | 13.5 | 13.8 |
| hcp' | 4.198 | - | 4.272 | - | 16.4 | - | 16.4 | - |
| bcc | 4.061 | 4.058 | 4.012 | 4.070 | 14.7 | 14.7 | 14.1 | 12.9 |
| fcc | 4.043 | 4.040 | 4.101 | 4.231 | 14.6 | 15.0 | 15.0 | 13.3 |

**Table E.2:** Predicted equilibrium cohesive energies and volumes of our ML potential for the tests described in Fig. 5.2.

| | $B$ [Mbar] | | | | $B$' | | | |
|---|---|---|---|---|---|---|---|---|
| | DFT | Our ML potential | | | DFT | Our ML potential | | |
| | | a) | b) | c) | | a) | b) | c) |
| diamond | 0.88 | 0.89 | 0.98 | 0.95 | 4.3 | 3.4 | 5.4 | 9.6 |
| bc8 | 0.83 | 0.89 | 0.94 | 0.73 | 4.3 | 2.9 | 3.1 | 6.5 |
| bcc | 0.89 | 0.67 | 0.59 | 1.14 | 4.3 | 9.2 | 10.7 | 14.1 |
| fcc | 0.77 | 0.68 | 1.01 | 1.66 | 4.5 | 2.2 | 2.1 | 7.4 |

**Table E.3:** Predicted bulk modulus $B$ and its derivative $B$' (with respect to pressure) of our ML potential for cubic phases within the tests described in Fig. 5.2.

| DFT | a) | b) | | c) | |
|---|---|---|---|---|---|
| | | left out $\beta$-Sn | left out diamond | left out $\beta$-Sn | left out diamond |
| 112 | 112 | 106 | 140 | 123 | 92 |

**Table E.4:** Predicted diamond-to-$\beta$-Sn transition pressure (in kbar) of our ML potential for the tests described in Fig. 5.2.

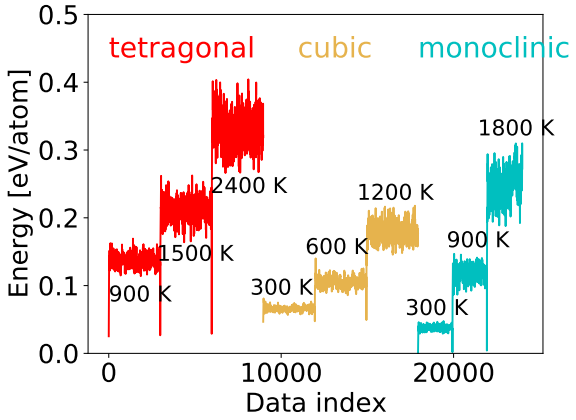
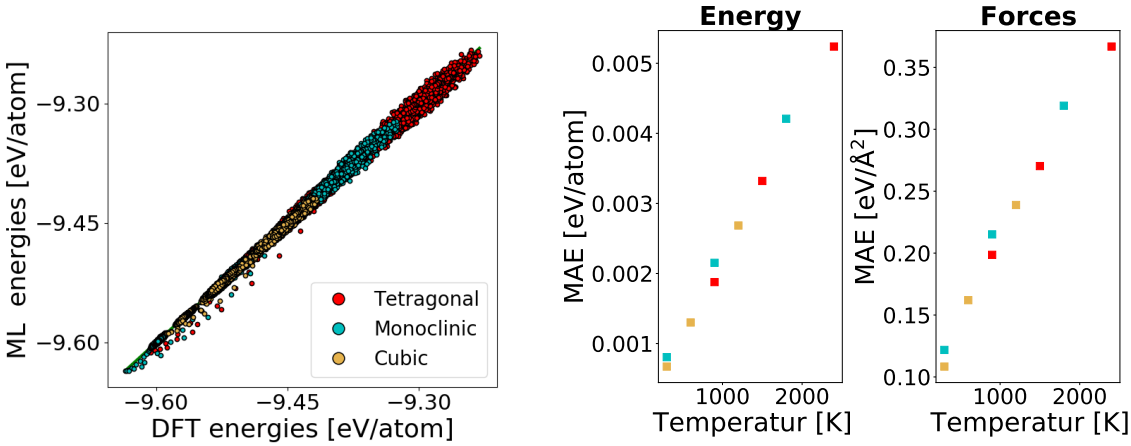



**Figure E.4:** DFT-PBEsol energies of configurations from molecular-dynamics simulations for $ZrO_2$ in the tetragonal, cubic and monoclinic phase, each for three different temperatures using 96 atoms in the unit cell.




**(a)** ML-versus-DFT scatter plot for energies. **(b)** Dependence of the mean absolute error (MAE) on temperature and crystal structure.

**Figure E.5:** Prediction performance of GAP(PBEsol) evaluated for energies and forces on a test set of 23 800 data points.

## E.2 Thermodynamics of $ZrO_2$

A data set of 24 000 $ZrO_2$ configurations from molecular-dynamics simulations performed using DFT-PBEsol was considered, see Fig. E.4. A training data set of 350 configurations was constructed where 200 were randomly selected from the molecular-dynamics trajectories and 150 data points were taken from the cubic-tetragonal PES (Fig. 5.5c). We fitted a two-body and three-body Gaussian Approximation Potential [32] to the energies and forces of this training data. The prediction errors of the potential, termed GAP(PBEsol), on the remaining 23 800 configurations of the molecular-dynamics trajectories are shown in Fig. E.5.

| | $\Delta E_{\mathrm{DFT}}$ [meV/atom] | $\Delta E_{\mathrm{ML}}$ [meV/atom] |
|---|---|---|
| PBEsol | 23 | 25 |
| HSE06 | 25 | 26 |

**Table E.5:** Tetragonal-monoclinic energy differences for relaxed structures using DFT and the ML potential.

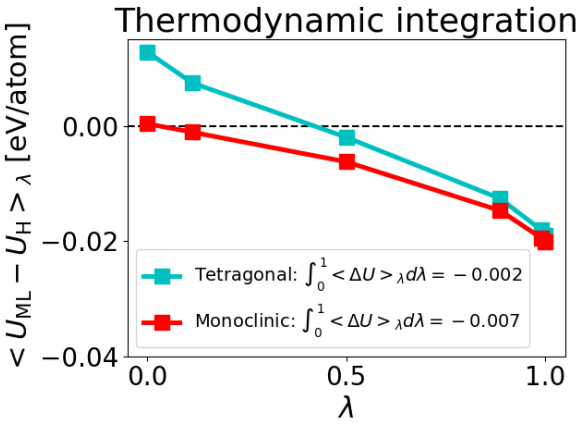



**Figure E.6:** Thermodynamic integration performed for the monoclinic and tetragonal phase at 1 000 K. The reference potential is given by the harmonic potential. A 96-atoms unit cell was used.

### E.2.1 Monoclinic-tetragonal phase transition

Crystal-structure relaxations were performed in the tetragonal and monoclinic phase for both DFT and the ML potential. The tetragonal-monoclinic energy differences $\Delta E = E_{\mathrm{t}} - E_{\mathrm{m}}$ of the relaxed structures predicted by the ML potentials are in good agreement with the DFT ones (Tab. E.5). In order to calculate the monoclinic-tetragonal transition temperature at hybrid level, first the free energies at 1 000 K were calculated via thermodynamic integrations (Eq. 5.6) for both the monoclinic and tetragonal phase using the GAP(HSE06) and reference harmonic potentials based on the relaxed structures. The integration grid is given by five $\lambda$-points, i.e. 0 and 1 plus three sample points according to the Gauss-Legendre quadrature. The results give anharmonic contributions of -0.002 eV/atom for the tetragonal and -0.007 eV/atom for the monoclinic case (Fig. E.6). By integrating the calculated free energies along the (inverse) temperature (Eq. 5.7) the transition temperature was determined.

# F Prediction across chemical composition space

## F.1 Choosing parameters of the chemical-transferable potentials

The parameters can be separated into two categories, the parameters of the potential (for a fixed composition) and the parameters of the neural network for the chemical-transfer learning. For the tests in Sec. 5.4.5 and 5.4.6, the potential parameters were chosen to keep the potentials through a smaller number of basis functions simple (compare CTP parameters in Tab. D.4 to the ones of the silicon potential in Tab. D.2).
We have chosen a neural network with one hidden layer that consists of 500 neurons with ReLU activation functions, see left panel of Fig. F.1. While deeper architectures could be investigated, we consider the performance of the neural network with one hidden layer in Sec. 5.4.5 and 5.4.6 as sufficient. Furthermore, the hidden layer finds a nonlinear representation of the input layer (atomic descriptors) to interpolate the chemical similarity of only 78 compounds, a relatively small number of data points compared to standard deep-learning applications.
The right panel of Fig. F.1 shows a convergence of the validation error at about the 400th epoch. In contrast, the training errors keeps decreasing at an epoch of 2 000. In the test in Sec. 5.4.6, we have run 400 epochs in every cross-validation step, but used the weights determined at an epoch that optimizes the validation error within the 400 epochs for the final model. While one might take the *optimal* weights from a longer run of, for example, 10 000 epochs with the aim to further decrease the validation error (only) slightly, its influence on overfitting the model to the training and validation set is not clear. For instance, the distribution of information in the validation set does not represent the one in the test set. More precisely, the validation set consists of randomly selected (compound, phase) tuples, where no compound is left completely out of the training set, while the test set in Sec. 5.4.6 is a compound left out completely with all its phases. The reason of not designing a validation set of compounds that are completely left out of the training set is the small sample size of compounds in the octet binaries data set and considering the demanding task of predicting the potential-energy surface of a left out compound.

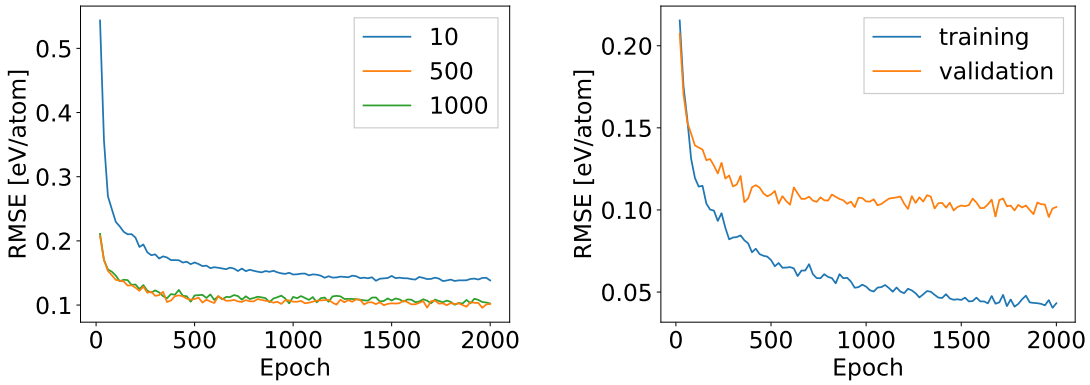



**Figure F.1:** Root mean square errors (RMSE) on energies in dependence of neural network parameters of the chemical-transferable potentials (CTP). The errors are evaluated on a validation set that consist of 10% of randomly selected (compound, phase) tuples of the octet binaries data set. Every 20th epoch the minimum error of the past 20 epochs is plotted only. The same neural-network architecture is used for both the two- and three-body part of the potential. A neural network with one hidden layer and ReLU activation functions is implemented. The left panel shows the dependence of the error on the number of nodes in the hidden layer. The right panel compares training and validation error for a neural network with 500 nodes. The validation error is converged at around the 400th epoch.

## F.2 Choosing parameters of the alchemical smooth overlap of positions

We have varied selected hyperparameters of the alchemical smooth overlap of positions (ASOAP) in a leave-one-compound-out cross validation. The leave-one-compound-out cross validation is described in Sec. 5.4.6, the evaluation of the performance in Sec. 5.4.4. However, in contrast to the investigation in Sec. 5.4.6, the dependence on the hyperparameters is tested on a smaller data set of only four instead of eight crystal-structure types, i.e. ZB, RS, CsCl, and NiAs. Besides the dependence of the model on the cutoff radius and the exponent $\zeta$ in Eq. 4.25, we considered two sets of atomic descriptors:

$$D_1 = \{Z, G, R, IP, EA, r_s, r_p\}, \tag{F.1}$$

$$D_2 = \{r_{\mathrm{vdW}}, \chi\}. \tag{F.2}$$

The sets consist of the following atomic properties: the atomic number $Z$, group $G$ and row $R$ in the periodic table, ionization potential $IP$, electron affinity $EA$, radii $r_s$ and $r_p$ of the valence $s$ and $p$ orbitals where the radial probability density is maximal, van-der-Waals radius $r_{\mathrm{vdW}}$, and Pauling electronegativity $\chi$. For the alchemical similarity $\kappa_{\alpha\beta}$ in Eq. 4.33 we used a Gaussian kernel. In case of $D_2$ the Gaussian kernel depends on two Gaussians widths, one for each atomic property, following Ref. [67]. Accordingly, if considering also the regularization parameter of the kernel ridge model, the hyperparameter optimization

| Atomic descriptors | $r_c$ [Å] | $\zeta$ | $r$ | RMSE($\Delta E$) [eV/atom] | $R_{\text{min, 1}}$ |
|---|---|---|---|---|---|
| $D_1$ | 6 | 2 | 67/78 | 1.26 | 0.69 |
| $D_1$ | 4 | 2 | 73/78 | 0.24 | 0.75 |
| $D_1$ | 4 | 1 | 70/78 | 0.21 | 0.73 |
| $D_2$ | 4 | 2 | 61/78 | 0.51 | 0.45 |
| $D_2$ | 4 | 1 | 60/78 | 0.40 | 0.60 |

**Table F.1:** Prediction results of the alchemical smooth overlap of positions (ASOAP) in a leave-one-compound-out cross validation in dependence of hyperparameters. The test differs from the leave-one-out-compound cross validation in 5.4.6 by the smaller data set of only four instead of eight crystal structure types, i.e. ZB, RS, CsCl, and NiAs. The set of atomic descriptors $D1$ and $D2$ are defined in the text (App. F.2), $r_c$ represents the cutoff radius, and $\zeta$ is the exponent in Eq. 4.25. The success rate $r$ shows how many compounds had their ground states predicted correctly within a tolerance of 0.1 eV/atom. RMSE($\Delta E$) represent the root-mean-square errors for the predicted energy differences. $R_{\text{min, 1}}$ gives the fraction of cubic surfaces that were predicted to have only one minimum.

is performed on a cubic grid. In contrast, in case of $D_1$ we implemented only a single Gaussian width for all atomic properties, to keep the computational expense of the hyperparameter-grid search low, i.e. the Gaussian width and the regularization parameter of the kernel ridge model were optimized on a square grid. The results are tabulated in Tab. F.1.

We found a significant better performance of the ASOAP when using $D_1$ instead of $D_2$. Furthermore, a model based on a cutoff radius of 4 Å is more accurate than one of 6 Å. Varying $\zeta$ between 1 and 2 does not influence the model performance significantly when using the set of atomic descriptors $D_1$.

## F.3 Stabilization of the machine-learning potentials by connecting the chemical space

In order to demonstrate the benefit of the CT approach to using no CT learning, we have performed for each of the two methods a different cross-validation test that allows for the comparison of the two methods, see Sec. 5.4.5. Here, we only show the results of the test using errors on cohesive energies (Fig. F.2), as opposed to Sec. 5.4.5 where we used errors on energy differences (Fig. 5.7).

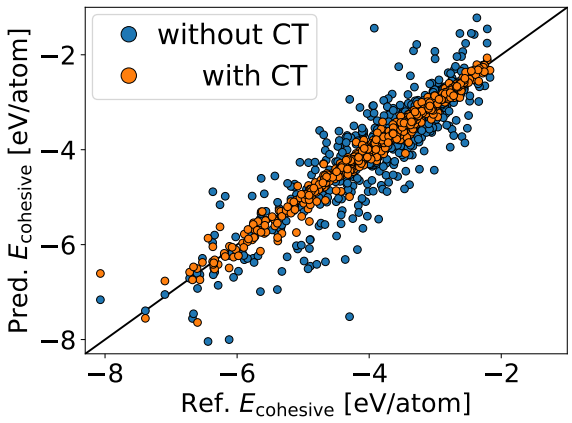



**Figure F.2:** Predicted vs reference cohesive energies from cross-validation tests that evaluate the prediction performance of the (equilibrium) energies of phases left out from the training sets. The tests were performed for the chemical-transfer (CT) and non-CT approach. For more information see Sec. 5.4.5.

## F.4 Probabilistic model for the leave-one-compound out cross validation

As described in Sec. 5.4.4, we consider the prediction of a ground state as successful if the reference (DFT) ground-state phase is inside the set of predicted ones where the set consists of the predicted lowest energy phase and the ones which lie maximally 0.1 eV/atom above. The success rates for identifying the ground states of the compounds in the cross-validation test determined for the three ML methods in Sec. 5.4.6 provide information for a comparison of the performance between the three methods. However, the three rates alone do not reveal to which extent the methods were capable of solving a demanding problem. For instance, a model that would predict the same energy for every data point would achieved a success rate of 78/78 because the predictions would count any phase into the set of possible ground states. Using the reference value of $1/8 = 9/72$ (one of eight phases) as a baseline to benchmark the ML models is one meaningful choice. Still the ML methods predict in some cases more than one phase and a better reference might be a value based on the distribution of the data set, as described in the following.
We define a model that assigns to every combination of phases a probability to be the set of possible ground-state phases, where the probability is given by the frequency of a set with competing ground-state phases (as given by DFT energies) within a tolerance of 0.1 eV/atom, count over the 78 compounds in the reference data. For example, the tuple (RS, ZB) defines a set of competing ground-state phases for a compound if one of the two phases is the ground state, the energy difference between them is not higher than 0.1 eV/atom, and the energy difference of all other phases to the ground state is higher than 0.1 eV/atom. We find that (RS, ZB) is the set of competing ground-state phases for 5/78 compounds. Then, for any compound, the probabilistic model predicts that (RS, ZB) is the

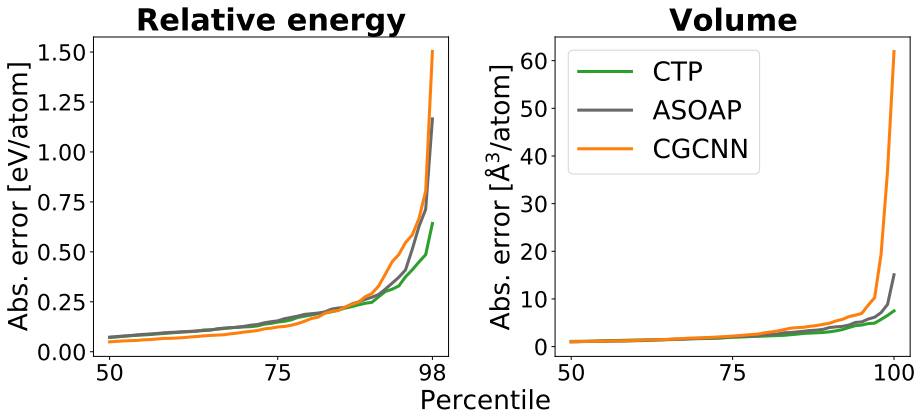



**Figure F.3:** Absolute errors of a leave-one-compound-out cross validation in dependence of the percentile.

set of promising ground-states with a probability of 5/78. The success rate in the leave-one-compound-out cross validation corresponds to the outcome of predicting at each step the set of promising ground-state phases and repeating the cross-validation until the appearance rate of the phase combinations converges against the respective probability. A prediction is counted as correct if the DFT ground state is inside the set of promising candidates. The resulting success rate is approximately 44/78. Note that only 13 of 28 possible phase tuples appear in the reference data. Accordingly the probability for 15 combinations is zero.

## F.5 Leaving out Ga-or-N or Sr-or-O based compounds versus leave-one-compound-out cross validation

In Sec. 5.4.6, Ga-or-N or Sr-or-O based compounds were left out to be tested. Tab. F.2 compares the result of this test to results of the same compounds in the leave-one-compound-out cross validation, also performed in Sec. 5.4.6.

## F.6 Random-structure search for GaN

Let $\boldsymbol{A}_{\text{ref}}$ be the cell matrix (lattice vectors as rows) of the eight-atoms cubic supercell of ZB-GaN relaxed by DFT. The lengths of the lattices vectors are given by 4.5 Å. 300 random structures are generated based on $\boldsymbol{A}_{\text{ref}}$. Each random structure is constructed in the following way:
First a strain matrix $\boldsymbol{F} = \boldsymbol{I} + \boldsymbol{R}$ is built, by adding to the identity matrix $\boldsymbol{I}$ a random matrix $\boldsymbol{R}$ with values from a uniform distribution between -0.2 and 0.2, and, furthermore, adding the constraint that the mean of the diagonal elements of $\boldsymbol{F}$

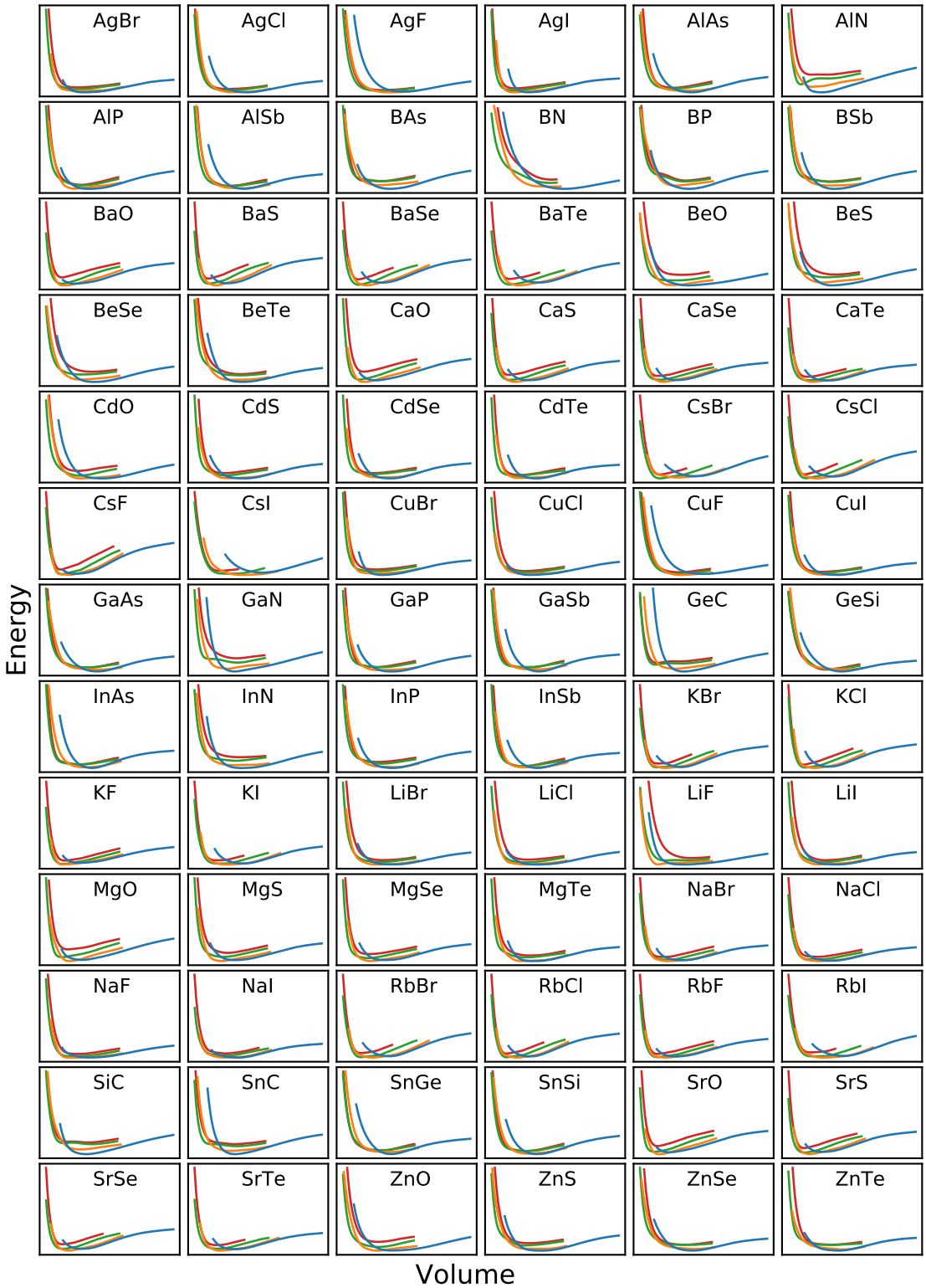



**Figure F.4:** Energy-volume curves predicted by the chemical-transferable potentials (CTP). 78 binary compounds in four cubic phases are considered. The phases are: ZB (blue), RS (orange), CsCl (green), NaTl (red). The volume intervals are given by $0.8r_{\mathrm{cov}} < d_{\mathrm{nn}} < 1.3r_{\mathrm{cov}}$. Furthermore, the curves are constrained by a maximum energy value of 10 eV/atom.

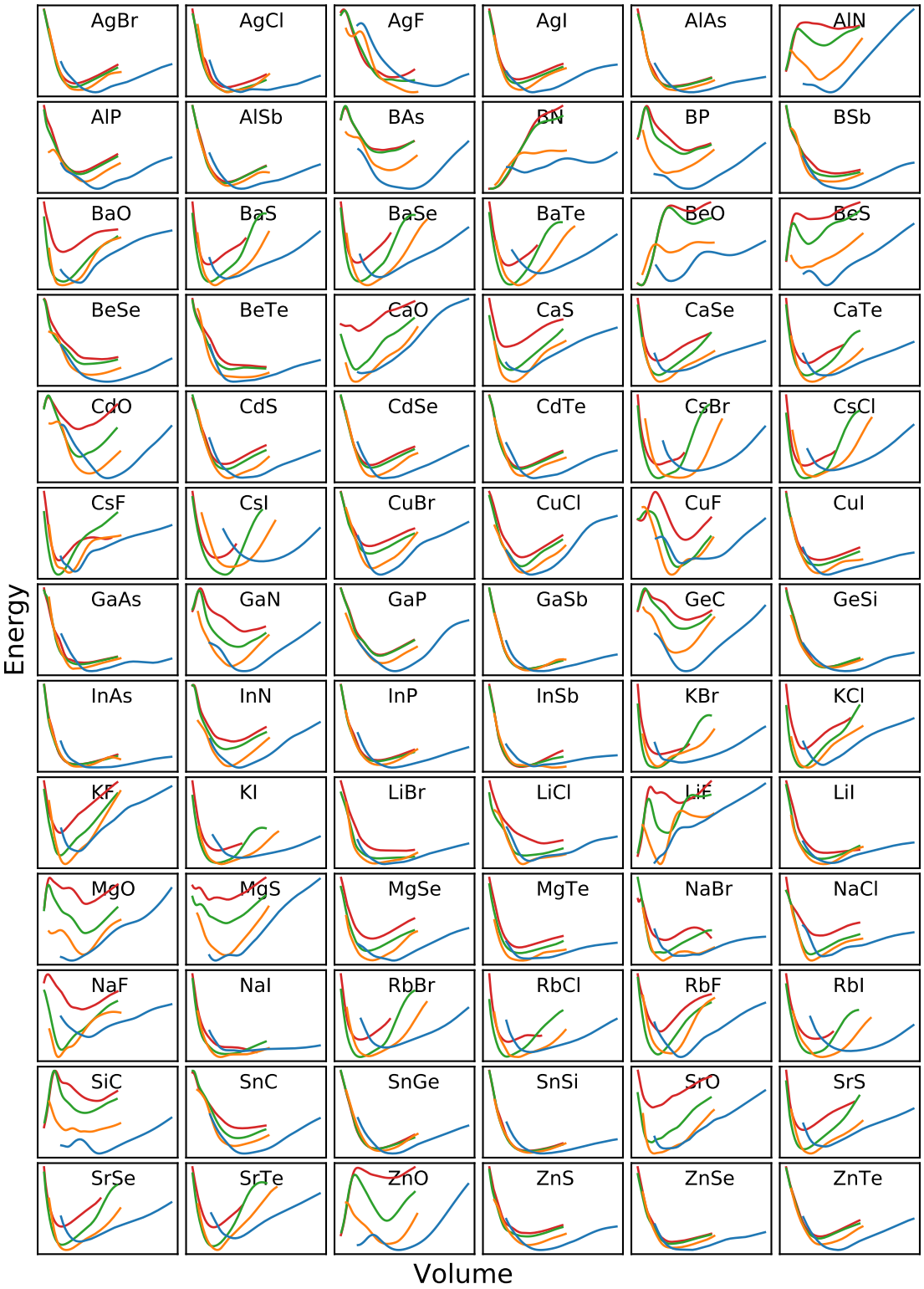



**Figure F.5:** Energy-volume curves predicted by the alchemical smooth overlap of atomic positions (ASOAP). 78 binary compounds in four cubic phases are considered. The phases are: ZB (blue), RS (orange), CsCl (green), NaTl (red). The volume intervals are given by $0.8r_{\text{cov}} < d_{\text{nn}} < 1.3r_{\text{cov}}$. Furthermore, the curves are constrained by a maximum energy value of 10 eV/atom.

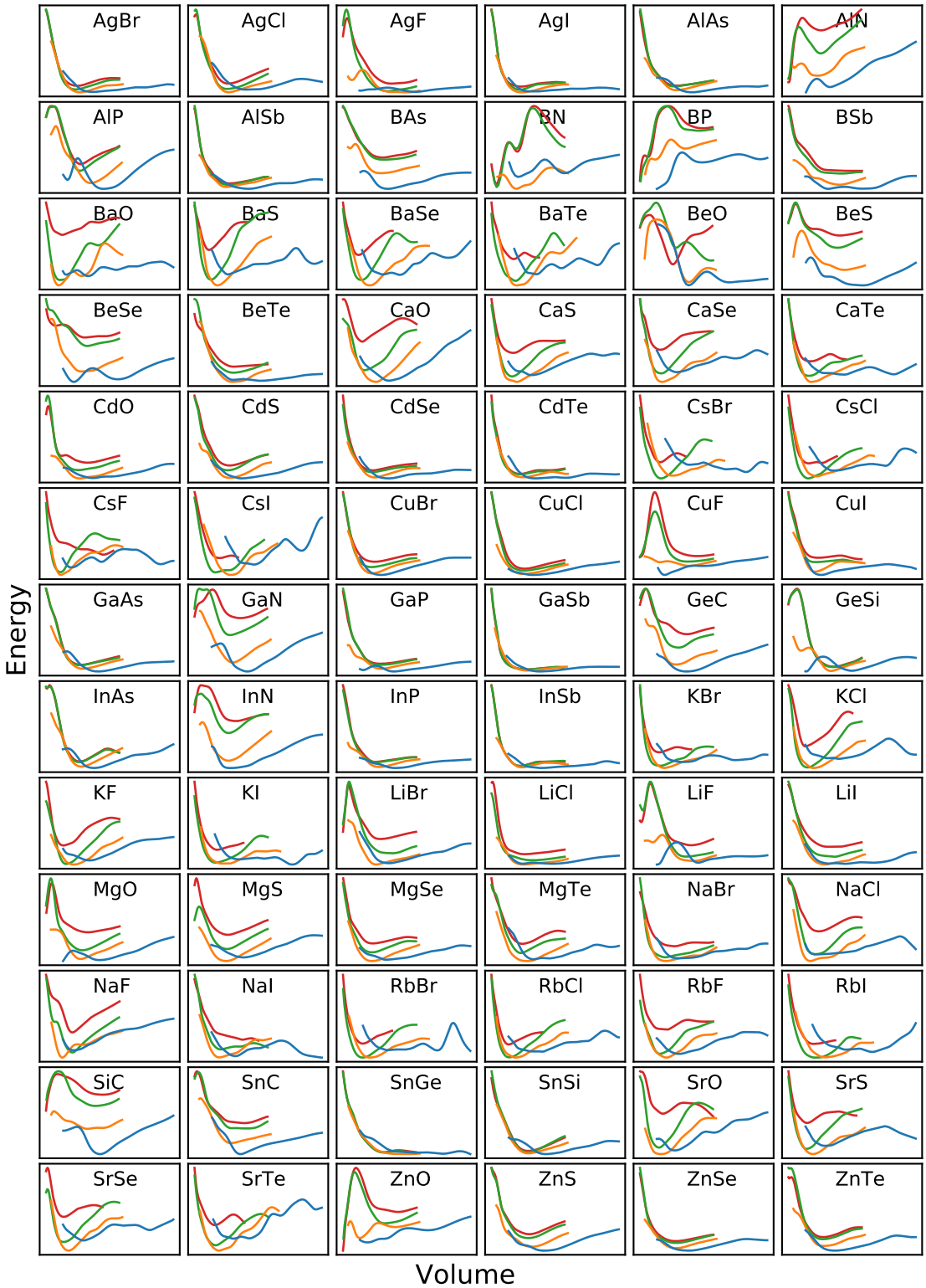



**Figure F.6:** Energy-volume curves predicted by the crystal graph convolutional neural networks (CGCNN). 78 binary compounds in four cubic phases are considered. The phases are: ZB (blue), RS (orange), CsCl (green), NaTl (red). The volume intervals are given by $0.8r_{\mathrm{cov}} < d_{\mathrm{nn}} < 1.3r_{\mathrm{cov}}$. Furthermore, the curves are constrained by a maximum energy value of 10 eV/atom.

| | Ga, N | | | Sr, O | | |
|---|---|---|---|---|---|---|
| | $r$ | $\Delta E$ | $V$ | $r$ | $\Delta E$ | $V$ |
| CTP | 6/7 (7/7) | 0.64 (0.64) | 2.2 (2.1) | 10/10 (9/10) | 0.25 (0.23) | 2.2 (1.1) |
| ASOAP | 6/7 (6/7) | 1.14 (1.10) | 3.6 (2.5) | 7/10 (7/10) | 0.57 (0.50) | 3.7 (2.6) |

**Table F.2:** Prediction results of the chemical-transferable potentials (CTP) and alchemical smooth overlap of positions (ASOAP) in a test in which all compounds that include either Ga, N or both are left out to be predicted and one in in which all compounds with Sr or O are left out. The success rate $r$ shows how many compounds had their ground states predicted correctly within a tolerance of 0.1 eV/atom. $\Delta E$ and $V$ represent the root-mean-square errors for the predicted energy differences and volumes, respectively. The values in the brackets correspond to the results for the same compounds in a leave-one-compound-out cross validation (Sec. 5.4.6).

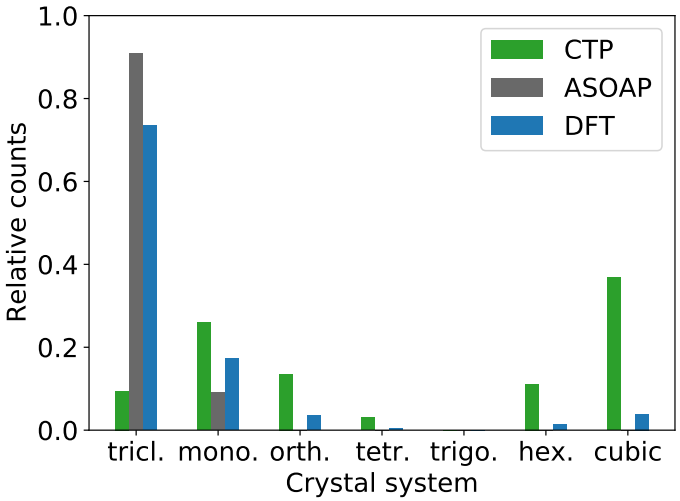



**Figure F.7:** Crystal-system distribution of the identified structures in the random-structure searches of Sec. 5.4.7 performed for the CTP, ASOAP and DFT. The tick-labels on the $x$-axis are the abbreviations for (from left to right): triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal, and cubic.

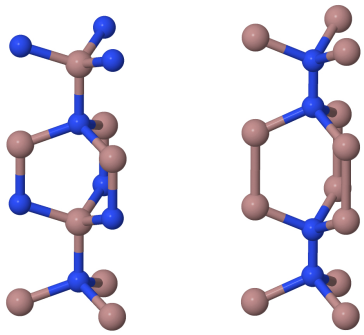


**Figure F.8:** "Chains" of two hexagonal crystal structures identified in a random-structure search (RSS) for GaN using DFT. Left: the wurtzite (WZ, space group 186) structure is represented. Right: a similar structure with space group 187. In principle, the right structure can be obtained from WZ by only exchanging some atoms (plus further adjustments of the exact lattice lengths and one atomic-positions related degree of freedom). While WZ is characterized by an alternating Ga-N order along the vertical axis, in case of the right structure there is either a Ga-Ga or N-N order. Furthermore, the smallest bond in the right structure is given by the N-N bonding (blue). The next neighbour of a N atom is given only by a N atom, i.e. the same atom type. The property, that the next-neighbour shell of an atom is only given by atoms of its atom type, was not found in any of the structures identified by the CTP in the RSS in Sec. 5.4.7. The DFT energy difference of the right structure to WZ is 0.556 eV/atom.

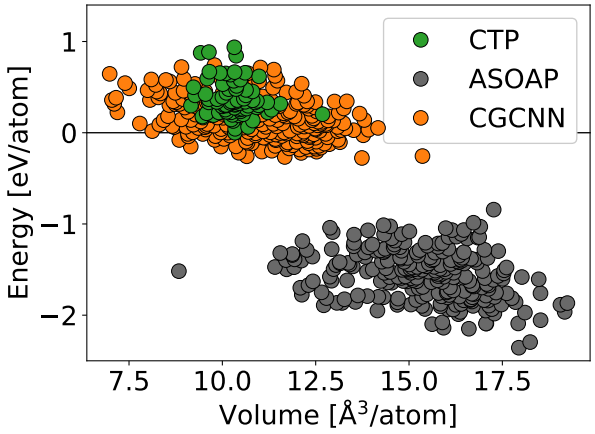



**Figure F.9:** GaN structures from random-structure searches (RSS) in an energy-volume scatter plot. The result of the RSS with 300 initial random structures performed for the CTP, ASOAP, and CGCNN is shown. In contrast to the searches in Fig. 5.9, which were performed via relaxations using gradients (forces), here the PES was optimized using a random walk on the PES. That is, at each step of the algorithm, a configuration with randomly changed coordinates of the atomic positions and lattice vectors is suggested and only accepted if its energy is lower than the previous configuration. Note that the energies of some structures might have not fully converged. However, the energy-volume distributions of the CTP and ASOAP are comparable with the ones in Fig. 5.9.

should equal one. Moreover, $\boldsymbol{F}$ is symmetrized such that $F_{ij} = F_{ji}$. The cell of the random structure is, then, defined by the matrix multiplication $\boldsymbol{A} = \boldsymbol{A}_{\text{ref}}\boldsymbol{F}$. The fractional coordinates of eight atomic positions are given by random numbers obtained from a uniform distribution between zero and one. The only constraint on the atomic positions is that the number of Ga and N atoms should be equal and the distance between two atoms should not be smaller than 1.54 Å ($0.8r_{\text{cov}}$, see Sec. 5.4.4).

To evaluate the similarity of the generated random cell shapes $\boldsymbol{A}$ to a cubic one, we introduce a two-dimensional descriptor, one element for the lattice vector lengths $a$, $b$, $c$ and one for the angles $\alpha$, $\beta$, $\gamma$ between the lattice vectors. The descriptor is then given by: $(\max_3(a,b,c) - \min_3(a,b,c), \max_3(\alpha,\,\beta,\,\gamma) - \min_3(\alpha,\,\beta,\,\gamma)$, where $\max_3$ ($\min_3$) returns the maximum (minimum) value among its three variables. In case of a cubic cell, the descriptor yields $(0,0)$. Note that for a more precise similarity to the cubic cell also the difference of the angles to 90° need to be considered. With the current descriptor, (0,0) is assigned, for example, to cells that are characterized by $a = b = c$ and $\alpha = \beta = \gamma$ while only the additional criterion $\alpha = 90°$ gives the cubic cell.
The elementwise average, standard deviation and maximum value for the descriptors of the 300 structures[1] are (0.9 Å, 15°), (0.4 Å, 8°), and (1.7 Å, 40°), respectively. While the structures were generated as (slightly) distorted cubic cells, we consider many of the cell shapes as not "close" to cubic, mainly due to larger angle differences.

The results of the RSS for the CTP, ASOAP and DFT are presented in Sec. 5.4.7. Here, we show only the distribution of the crystal systems for the structures identified in the RSS (after relaxation), see Fig. F.7.

## F.7 Unphysical roughness of the crystal graph convolutional neural networks in the potential-energy surface

The predicted PES of the CGCNN is not smooth. The main reason for this is the fact that the CGCNN considers for every atom only pairwise distances to the twelve nearest neighbours, see theory in Sec. 4.6.
Consider the example of a GaN compound. If a small change in atomic positions or lattice vectors results in the change of an atomic environment such that a new Ga atom enters the neighbourhood defined by the twelve nearest neighbours and replaces a N atom on neighbour index $i$, then the species representation assigned to neighbour index $i$ will change. The "sudden" replacement of the representation vector in the model will lead to a (possibly) significant jump on the PES, see example in Fig.

[1] The average, standard deviation and maximum value are taken along each of the two independent coordinates.

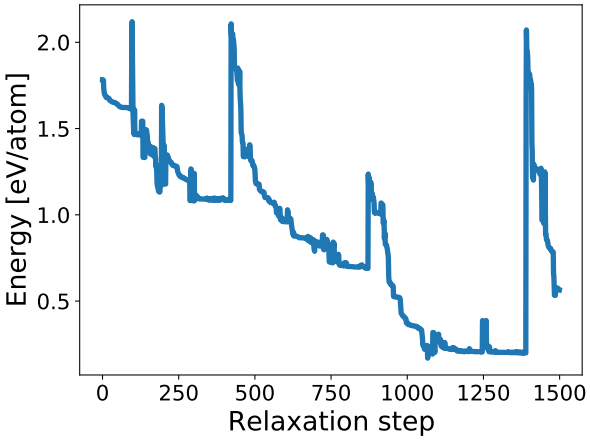



**Figure F.10:** Energy evolution along a relaxation path for a CGCNN and for an initial random GaN structure. The CGCNN was trained on the structures of 77 octet binaries, GaN not included. The energy is given relative to the relaxed ZB structure as predicted by the CGCNN.

F.10. Note that we have evaluated the forces and stresses for the relaxation numerically.

## F.8 Error analysis

In Sec. 5.4.6 a leave-one-compound-out cross validation was performed for the CTP, ASAOP, and CGCNN. Fig. F.11 shows the relationships between the prediction errors on the energies within this test for the three models, the variance of the reference energies, and a materials descriptor that was determined in Sec. 5.4.8 to identify materials that are harder to be described by the models. Fig. F.12 presents the dependence of the errors and reference energies on the crystal-structure types.

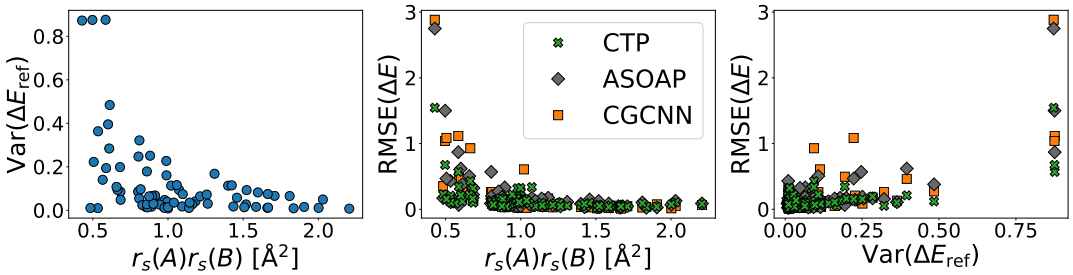



**Figure F.11:** Relationship between the materials descriptor $r_s(A)r_s(B)$, the variance $\mathrm{Var}(\Delta E_{\mathrm{ref}})$ of the reference relative energies, and prediction root mean square errors $\mathrm{RMSE}(\Delta E)$ on the relative energies. Representing 78 compounds, 78 markers are shown in the left panel, and $3 \cdot 78$ in the central and right one. $r_s(A)r_s(B)$ is the first component of the two-dimensional descriptor in Fig. 5.10. Both the variance $\mathrm{Var}(\Delta E_{\mathrm{ref}})$ and $\mathrm{RMSE}(\Delta E)$ are built for each compound separately over the energy differences of the phases with respect to the ground-state phase of the compound. The $\mathrm{RMSE}(\Delta E)$ are results of a leave-one-compound-out cross validation (Sec. 5.4.6) for the chemical-transferable potentials (CTP), the alchemical smooth overlap of atomic positions (ASOAP), and the crystal graph convolutional neural network (CGCNN). Energies, thus also the RMSE, are given in eV/atom.

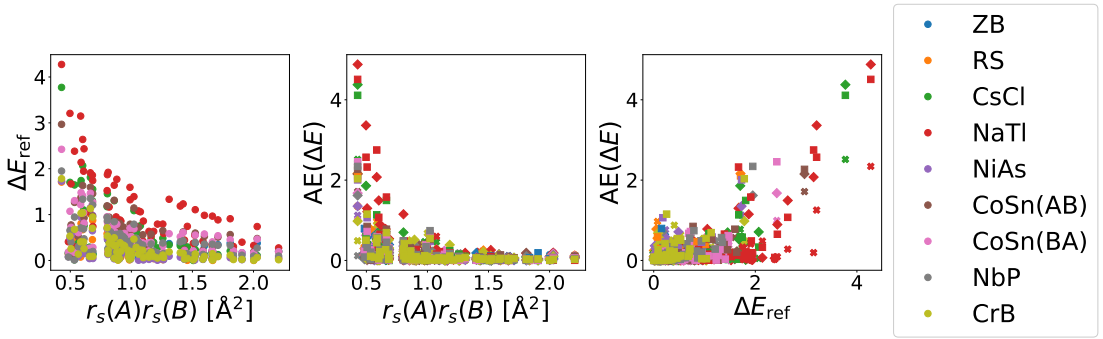



**Figure F.12:** Relationship between the materials descriptor $r_s(A)r_s(B)$, the reference relative energies $\Delta E_{\mathrm{ref}}$ , and the absolute prediction error $\mathrm{AE}(\Delta E)$ on the relative energies in dependence of the phases. The $\mathrm{AE}(\Delta E)$ are results of a leave-one-compound-out cross validation (Sec. 5.4.6) for the chemical-transferable potentials (CTP), the alchemical smooth overlap of atomic positions (ASOAP), and the crystal graph convolutional neural network (CGCNN). Energies, thus also the AE, are given in eV/atom. Note that the cross, diamond and square symbols in the central and right panel represent the errors of the CTP, ASOAP, and CGCNN, respectively.

# G Density-functional theory and approximations

The physical mechanisms that drive crystal-structure stability of materials are learned from calculations based on density-functional theory (DFT) and approximations to it. While most of the training data of our machine-learning models are extracted from the NOMAD Laboratory, in some parts of the thesis we have performed DFT calculations. For example, we have run molecular-dynamics simulations for monoclinic $ZrO_2$ using DFT and the PBEsol functional (Sec. 5.3.3). Furthermore, in order to fit a machine-learning potential to hybrid-DFT-level data, energies and forces of a set of configurations was calculated using the HSE06 functional. In Sec. 5.4.7, we have performed a random-structure search using DFT within the local-density approximation.
This chapter briefly summarizes the concepts of density-functional theory and approximations to it.

## G.1 The many-body problem

In order to understand the physics of solids and molecular systems, a set of interacting atoms consisting of $N_{\text{el}}$ electrons and $N_{\text{nuc}}$ nuclei, we consider the time-independent Schrödinger equation:

$$\hat{H}\Psi = E\Psi. \tag{G.1}$$

If relativistic effects are neglected we may write the Hamiltonian [128]

$$\hat{H} = \hat{T}_{\text{el}} + \hat{T}_{\text{nuc}} + \hat{V}_{\text{el-el}} + \hat{V}_{\text{el-nuc}} + \hat{V}_{\text{nuc-nuc}}, \tag{G.2}$$

where $\hat{T}_{\text{el}}$ and $\hat{T}_{\text{nuc}}$ represent the kinetic energy operators of the electrons and nuclei and $\hat{V}_{\text{el-el}}$, $\hat{V}_{\text{el-nuc}}$, and $\hat{V}_{\text{nuc-nuc}}$ the electron-electron, electron-nuclear, and nuclear-nuclear interaction operators, respectively. Using atomic units, the operators are given by

$$\hat{T}_{\text{el}} = -\frac{1}{2}\sum_{i}^{N_{\text{el}}} \Delta_i, \tag{G.3}$$

$$\hat{T}_{\text{nuc}} = -\frac{1}{2}\sum_{I}^{N_{\text{nuc}}} \frac{\Delta_I}{M_I}, \tag{G.4}$$

$$\hat{V}_{\text{el-el}} = \frac{1}{2}\sum_{i}^{N_{\text{el}}}\sum_{j\neq i}^{N_{\text{el}}}\frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|}, \tag{G.5}$$

$$\hat{V}_{\text{el-nuc}} = -\frac{1}{2}\sum_{i}^{N_{\text{el}}}\sum_{I}^{N_{\text{nuc}}}\frac{Z_I}{|\boldsymbol{r}_i - \boldsymbol{R}_I|}, \tag{G.6}$$

$$\hat{V}_{\text{nuc-nuc}} = \frac{1}{2}\sum_{I}^{N_{\text{nuc}}}\sum_{J\neq I}^{N_{\text{nuc}}}\frac{Z_I Z_J}{|\boldsymbol{R}_I - \boldsymbol{R}_J|}, \tag{G.7}$$

where $\boldsymbol{r}_i$ and $\boldsymbol{R}_I$ are the electronic and nuclear coordinates, respectively, and $\Delta$ represents the Laplacian.
An exact analytical solution of the eigenvalue problem G.1 exists only for a few cases (e.g. H, $He^+$, or $H_2^+$). A first step towards approximating the solution to Eq. G.1 is realized by the Born-Oppenheimer approximation: due to the mass ratio $\frac{m_i}{M_I} < 10^{-3}$ it is assumed the electrons adapt instantaneously to the movement of the nuclei, which allows us to factorize the wave function into an electronic and a nuclear part

$$\Psi = \Psi_{\text{el}}(\boldsymbol{r};\{\boldsymbol{R}\})\Psi_{\text{nuc}}(\boldsymbol{R}), \tag{G.8}$$

where $\boldsymbol{r}$ and $\boldsymbol{R}$ denote the set of positions of the $N_{\text{el}}$ electrons and $N_{\text{nuc}}$ nuclei, respectively, and $\Psi_{\text{el}}$ depends only parametrically on $\boldsymbol{R}$ as highlighted by the brackets {}. As a result, the Schrödinger equation G.1 can be decoupled into an electronic equation

$$[\hat{T}_{\text{el}} + \hat{V}_{\text{el-el}} + \hat{V}_{\text{el-nuc}}]\Psi_{\text{el}} = E_{\text{el}}\Psi_{\text{el}} \tag{G.9}$$

and into a nuclear one:

$$\left[-\frac{1}{2}\sum_{i}^{N_{\text{nuc}}}\frac{\Delta_I}{M_I} + \hat{V}_{\text{nuc-nuc}} + E_{\text{el}}(\{\boldsymbol{R}\})\right]\Psi_{\text{nuc}} = E_{\text{nuc}}\Psi_{\text{nuc}}. \tag{G.10}$$

In this work we will consider only the ground-state solution of the electronic Schrödinger equation G.9. Together with the nuclei-nuclei interaction and for different sets of nuclear positions it provides the Born-Oppenheimer potential-energy surface (PES)

$$E_{\text{PES}} = E_{\text{el}} + V_{\text{nuc-nuc}} \tag{G.11}$$

on which the nuclei move.

## G.2 The Hohenberg-Kohn theorems

Despite the simplification of the Schrödinger equation through the Born-Oppenheimer approximation, determining the ground-state solution of the electronic Schrödinger equation G.9 stays complicated as the function $\Psi_{\text{el}}$ depends on $3N_{\text{el}}$ coordinates. In the formalism of density-functional theory (DFT), the energy of the electrons

is written as a functional of the electron density $n(\boldsymbol{r})$ and the ground-state energy is obtained from the minimum of the functional via variational principle. As a result, the complexity of the problem is reduced, since the electron density

$$n(\boldsymbol{r}) = N_{\mathrm{el}} \int ... \int |\Psi_{\mathrm{el}}(\boldsymbol{r}, \boldsymbol{r}_2, ..., \boldsymbol{r}_{N_{\mathrm{el}}})|^2 \, \mathrm{d}\boldsymbol{r}_{2...N_{\mathrm{el}}} \tag{G.12}$$

depends on only 3 coorindates. The theory is build on the two Hohenberg-Kohn theorems [129]:

**First Hohenberg-Kohn theorem:** *For any system of interacting particles in an external potential, the external potential is determined up to a constant by the ground-state electron density.*

**Second Hohenberg-Kohn theorem:** *A universal functional $E[n]$ of the electron density exists. The minimum of the functional via variation of $n(\boldsymbol{r})$ is the ground-state energy at the ground state electron density.*

The energy functional is defined as

$$E[n] = F[n] + \int \nu_{\mathrm{ext}}(\boldsymbol{r}) n(\boldsymbol{r}) \mathrm{d}\boldsymbol{r} \tag{G.13}$$

with the external potential $\nu_{\mathrm{ext}}(\boldsymbol{r})$ and the universal functional

$$F[n] = \langle g | \hat{T}_{\mathrm{el}} + \hat{V}_{\text{el-el}} | g \rangle = T_{\mathrm{el}}[n] + E_{\text{el-el}}[n]. \tag{G.14}$$

Here, $\langle \cdot | \cdot \rangle$ denotes the usual bra-ket notation and $|g\rangle$ the ground state.
While the formalism that results from the Hohenberg-Kohn theorems is exact, it is not helpful in solving the many-electron problem because it leaves the universal functional unknown. For instance, no analytical expression has been found for $F[n]$, so far. This problem can be overcome with the Kohn-Sham formalism, as discussed in the next section.

## G.3 Kohn-Sham equations

The idea behind the approach of Kohn and Sham [130] is that we map the problem of the system of interacting electrons onto an auxiliary system of non-interacting electrons. Kohn and Sham suggested to write the energy $E[n]$ of the interacting system as

$$E[n] = T_{\mathrm{s}}[n] + E_{\mathrm{H}}[n] + \int \nu_{\mathrm{ext}}(\boldsymbol{r}) n(\boldsymbol{r}) \mathrm{d}\boldsymbol{r} + E_{\mathrm{xc}}[n] \tag{G.15}$$

with the kinetic energy $T_s[n]$ of the non-interacting electron system, the Hartree energy

$$E_{\mathrm{H}}[n] = \frac{1}{2}\int\int \frac{n(\boldsymbol{r})n(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}, \mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{r'}, \tag{G.16}$$

and the exchange-correlation functional

$$E_{\mathrm{xc}}[n] = F[n] - T_{\mathrm{s}}[n] - E_{\mathrm{H}}[n]. \tag{G.17}$$

The variational problem for Eq. G.15 is given by

$$\frac{\delta T_{\mathrm{s}}[n]}{\delta n(\boldsymbol{r})} + \frac{\delta E_{\mathrm{H}}[n]}{\delta n(\boldsymbol{r})} + \nu_{\mathrm{ext}}(\boldsymbol{r}) + \frac{\delta E_{\mathrm{xc}}[n]}{\delta n(\boldsymbol{r})} - \frac{\delta}{\delta n(\boldsymbol{r})}(\mu\int n(\boldsymbol{r'})\mathrm{d}\boldsymbol{r'}) = 0 \tag{G.18}$$

$$\frac{\delta T_{\mathrm{s}}[n]}{\delta n(\boldsymbol{r})} + \underbrace{\nu_{\mathrm{H}}[n](\boldsymbol{r}) + \nu_{\mathrm{ext}}[n](\boldsymbol{r}) + \nu_{\mathrm{xc}}[n](\boldsymbol{r})}_{\nu_{\mathrm{KS}}(\boldsymbol{r})} - \mu = 0 \tag{G.19}$$

whereby the Lagrange multiplier $\mu$ imposes the constraint that the number of particles

$$N_{\mathrm{el}} = \int n(\boldsymbol{r'})\mathrm{d}\boldsymbol{r'} \tag{G.20}$$

shall be fixed.
Eq. G.19 has the form of an equation that would have been obtained for a system of non-interacting particles moving in an external potential $\nu_{\mathrm{KS}}$. This non-interacting system can be expressed by the electron density

$$n(\boldsymbol{r}) = \sum_{i}^{N_{\mathrm{el}}} \langle\phi_i|\phi_i\rangle\,. \tag{G.21}$$

and the Schrödinger equation

$$\left[-\frac{1}{2}\Delta + v_{\mathrm{KS}}(\boldsymbol{r})\right]\phi_i(\boldsymbol{r}) = \epsilon_i\phi_i(\boldsymbol{r}). \tag{G.22}$$

Eq. G.21 and Eq. G.22 are called the *Kohn-Sham equations*. They are solved in a self-consistent manner. The kinetic energy of the non-interacting particles is determined by

$$T_s[n] = \sum_{i}^{N_{\mathrm{el}}} \epsilon_i - \int \nu_{\mathrm{KS}}(\boldsymbol{r})n(\boldsymbol{r})\mathrm{d}\boldsymbol{r}. \tag{G.23}$$

## G.4 Approximations to the exchange-correlation functional

The only unknown part in the energy functional G.15 is the exchange-correlation functional. There exist different approximations to it with different levels of accuracy and computational cost. In the following, we will only introduce the *local-density approximation*, *generalized gradient approximation*, and hybrid functionals.

### G.4.1 Local-density approximation

The simplest approximation is the local-density approximation (LDA), in which the electron density is locally approximated by a homogeneous electron gas. The exchange-correlation functional is written

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[n] = \int n(\boldsymbol{r}) \epsilon_{\mathrm{xc}}^{\mathrm{LDA}}(n(\boldsymbol{r})) \mathrm{d}\boldsymbol{r}. \tag{G.24}$$

At each point in the space, $\epsilon_{\mathrm{xc}}^{\mathrm{LDA}}$ yields the respective value of homogeneous electron gas. Moreover $\epsilon_{\mathrm{xc}}^{\mathrm{LDA}}$ is decomposed linearly into exchange and correlation term. The expression for the exchange part is known exactly [131]. Accurate paraemtrizations for the correlation part were introduced on the basis of quantum Monte Carlo simulations [132].

### G.4.2 Generalized gradient approximation

The generalized gradient approximation introduces an explicit dependence on the gradient of the density in the exchange-correlation functional

$$E_{\mathrm{xc}}^{\mathrm{GGA}}[n] = \int n(\boldsymbol{r}) \epsilon_{\mathrm{xc}}^{\mathrm{LDA}}(n(\boldsymbol{r})) K_{\mathrm{xc}}(n(\boldsymbol{r}), \nabla n(\boldsymbol{r})) \mathrm{d}\boldsymbol{r}, \tag{G.25}$$

where $K_{\mathrm{xc}}(n(\boldsymbol{r}), \nabla n(\boldsymbol{r}))$ is a factor modifying $\epsilon_{\mathrm{xc}}^{\mathrm{LDA}}(n)$ in dependence of $n(\boldsymbol{r})$ and $\nabla n(\boldsymbol{r}))$. Examples are the PBE [133] or PBEsol [134] functional.

### G.4.3 Hybrid functionals

Further improvements are obtained by using hybrid functionals, which reduce the so-called self-interaction error. The exchange functional is given by

$$E_{\mathrm{x}}^{\mathrm{hyb}} = \alpha E_{\mathrm{x}}^{\mathrm{HF}} + (1-\alpha) E_{\mathrm{x}}^{\mathrm{DFA}}, \tag{G.26}$$

with the Hartree-Fock (HF) exchange

$$E_{\mathrm{x}}^{\mathrm{HF}} = -\frac{1}{2} \sum_{i,j}^{N_{\mathrm{el}}} \int \int \frac{\phi_i^*(\boldsymbol{r}) \phi_j(\boldsymbol{r}) \phi_j^*(\boldsymbol{r'}) \phi_i(\boldsymbol{r'})}{|\boldsymbol{r} - \boldsymbol{r'}|}, \mathrm{d}\boldsymbol{r} \mathrm{d}\boldsymbol{r'} \tag{G.27}$$

and the parameter $\alpha \in [0, 1]$, determining the weights of the HF and a density-functional-approximation (DFA) contribution. A popular hybrid functional is HES06 [135], in which additionally the Coulomb interaction in Eq. eq:hartee-hse is screened.