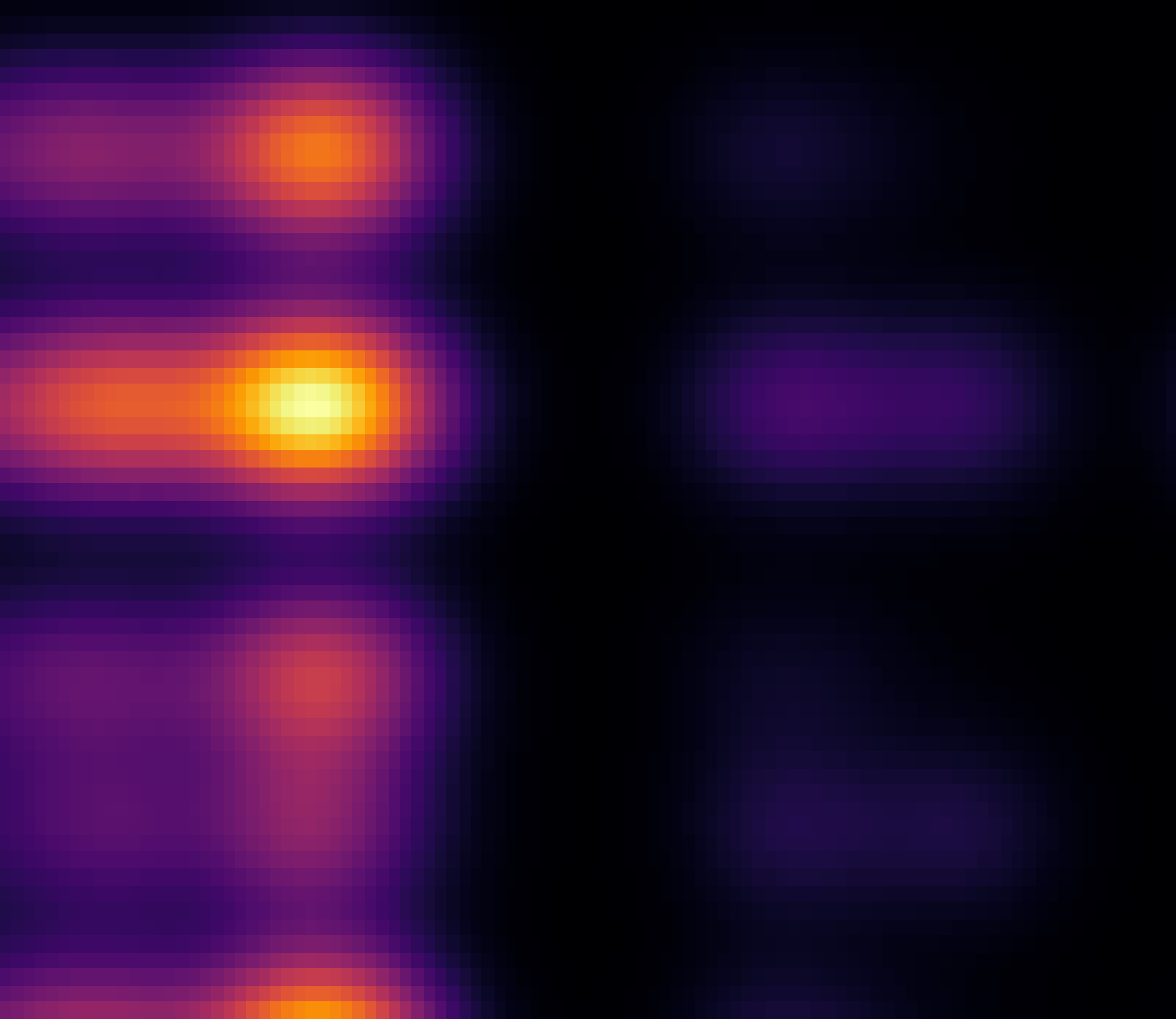
DTU Physics
Department of Physics

Ph.D. Thesis

Machine Learning Quantum Mechanics for Materials Science

Nikolaj Rørbæk Knøsgaard, November 2022



DTU Physics
Department of Physics

Ph.D. Thesis

Machine Learning Quantum Mechanics for Materials Science

Nikolaj Rørbæk Knøsgaard

Kongens Lyngby, November 2022

Machine Learning Quantum Mechanics for Materials Science

Ph.D. Thesis
Nikolaj Rørbæk Knøsgaard
November 30, 2022

Cover: Radially decomposed projected density of states (RAD-PDOS) fingerprint for MoS_2 (see Publication [I]).

Technical University of Denmark
DTU Physics
Department of Physics

Fysikvej 311
2800 Kongens Lyngby, Denmark
Phone: +45 4525 3344
info@fysik.dtu.dk

www.fysik.dtu.dk

Preface

This thesis is submitted in candidacy for a Ph.D. degree in Physics from the Technical University of Denmark. The Ph.D. project was carried out at the Center for Atomic-scale Materials Design (CAMD) at the DTU Department of Physics in the period between November 2019 and November 2022. The project was supervised by main supervisor Professor Kristian Sommer Thygesen and co-supervisor Professor Karsten Wedel Jacobsen, both from DTU Physics.

Part of the Ph.D. project was conducted during an external stay at Aalto University in Helsinki, Finland. The external stay was hosted by Professor Patrick Rinke from November 2021 to January 2022.

Nikolaj Knøsgaard

Nikolaj Rørbæk Knøsgaard

Kongens Lyngby, November 30, 2022

Abstract

Computational materials science plays an important role in the discovery and design of new materials. With accurate and efficient methods for computing the properties of materials it is possible to search through large compositional spaces with the purpose of screening for materials with specific properties. In recent years, the use of machine learning methods in materials science has become increasingly useful. This is a result of the vast amounts of materials data generated using first-principles methods such as density functional theory (DFT), but also the development of new machine learning methods for materials science has had an impact.

Finding good representations or fingerprints of materials as inputs to machine learning models is essential. This thesis presents novel fingerprint methods utilizing additional information from the electronic density and wavefunction obtainable from standard DFT calculations besides the atomic structure. More specifically, the energy decomposed operator matrix elements (ENDOME) fingerprint is constructed using matrix elements of quantum mechanical operators, e.g. the position and momentum operators. Additionally, the radially decomposed projected density of states (RAD-PDOS) fingerprint is developed using projections of DFT wavefunctions onto atoms and angular orbitals. The presented methods differs from other fingerprints by encoding individual quantum states.

The ENDOME and RAD-PDOS fingerprints of individual states are then applied in a machine learning model. The model predicts the difference in state eigenenergies between a low-fidelity DFT calculation and a high-fidelity G_0W_0 calculation for 2D materials. The model predicts the G_0W_0 correction energies for individual states with a mean absolute error (MAE) of 0.11 eV. This converts to a MAE of 0.15 eV on the G_0W_0 band gap by using the model to compute full G_0W_0 band structures.

Additionally, the RAD-PDOS fingerprint is used to evaluate the dynamical stability of 2D materials. This is done by training a binary machine learning classification model predicting the stability. The model achieves an excellent receiver operating characteristic with an area under the curve of 0.90, and the model can thus be used to screen materials for dynamic stability without performing expensive phonon calculations.

The dynamical stability is further investigated by developing approximative methods for calculating electron-phonon coupling matrix elements. The methods are based on replacing the DFT effective potential with a potential set up from atomic potentials. With this approximation, the matrix elements are quantitatively similar to the true DFT matrix elements. The approach is further improved by using machine learning to reconstruct the DFT potentials from the atomic potentials, which reduces the error by a factor of ≈ 2.

Resumé

Databaseret materialevidenskab spiller en vigtig rolle i opdagelsen og design af nye materialer. Med nøjagtige og effektive metoder til beregning af materialers egenskaber er det muligt at gennemsøge store kompositionsrum med det formål at screene for materialer med specifikke egenskaber. I de senere år er brugen af maskinlæringsmetoder i materialevidenskab blevet mere og mere nyttig. Dette er et resultat af de enorme mængder af materialedata, der er genereret ved hjælp af *ab-initio*-metoder såsom tæthedsfunktionalteori (DFT), mens også udviklingen af nye maskinlæringsmetoder indenfor materialevidenskab har haft en indflydelse.

Det er vigtigt at finde gode repræsentationer eller fingeraftryk af materialer som input til maskinlæringsmodeller. Denne afhandling præsenterer nye fingeraftryksmetoder, der anvender yderligere information fra den elektroniske tæthed og bølgefunktion, der kan opnås ved standard DFT-beregninger, udover atomstrukturen. Mere specifikt er fingeraftrykket for energi-dekomponerede operatormatrixelementer (ENDOME) konstrueret ved hjælp af matrixelementer af kvantemekaniske operatorer, f.eks. positions- og momentum-operatorer. Derudover er det radialt dekomponerede projekterede tilstandstæthed (RAD-PDOS) fingeraftryk udviklet ved hjælp af projektioner af DFT-bølgefunktioner på atomer og angulære orbitaler. De præsenterede metoder adskiller sig fra andre fingeraftryk ved at beskrive individuelle kvantetilstande.

ENDOME- og RAD-PDOS-fingeraftrykkene for individuelle tilstande anvendes derefter i en maskinlæringsmodel. Modellen forudsiger forskellen i tilstandsegenenergier mellem en mindre nøjagtig DFT-beregning og en mere nøjagtig G_0W_0-beregning for 2D-materialer. Modellen forudsiger G_0W_0 korrektionsenergierne for individuelle tilstande med en gennemsnitlig absolut fejl (MAE) på 0.11 eV. Dette konverteres til en MAE på 0.15 eV på G_0W_0-båndgabet ved at bruge modellen til at beregne hele G_0W_0-båndstrukturer.

Derudover bruges RAD-PDOS-fingeraftrykket til at evaluere den dynamiske stabilitet af 2D-materialer. Dette gøres ved at træne en binær maskinlæringsklassifikationsmodel, der forudsiger stabiliteten. Modellen opnår en udmærket ROC med et areal under kurven på 0.90, og modellen kan således bruges til at screene materialer for dynamisk stabilitet uden at udføre dyre fononberegninger.

Den dynamiske stabilitet undersøges yderligere ved at udvikle approksimative metoder til beregning af elektron-fonon-kobling-smatrixelementer. Metoderne er baseret på at erstatte det effektive potentiale fra DFT med et potentiale sat op fra atomare potentialer. Med denne tilnærmelse svarer matrixelementerne kvantitativt til de sande DFT-matrixelementer. Fremgangsmåden forbedres yderligere ved at bruge maskinlæring til at rekonstruere DFT-potentialerne ud fra de atomare potentialer, hvilket reducerer fejlen med en faktor på ≈ 2.

Acknowledgements

Doing a Ph.D. for the past three years has been a challenge but also a great experience for which I owe a huge thanks to many people.

First of all, I want to thank my supervisor Kristian for giving me the opportunity to return to DTU to do a Ph.D. and for the great guidance throughout the project. I have definitely learned a lot from working with you.

Furthermore, I want to thank Patrick Rinke for hosting my external stay at Aalto University and to all the people in the CEST group for making me feel very welcome in your group. I really enjoyed my time in Finland, both personally and professionally.

I also want to thank all the people in the CAMD group, past and present, for creating a really good working place: Mark, Fabian, Mads, Matthew, Stefano, Hadeel, Thorbjørn, Simone, Morten, Estefania, Asbjørn, Daniele, Anders, Alireza, Sajid, Joachim, Peder, Cuau, Sahar, Casper, Julian, Sami, Urko, Jiban, Ask, Fredrik, Mikael, Tara, Martin, Karsten, Jacob, Thomas. Additionally, I want to thank Mark, Peder and Mads for helping me proofreading this thesis.

For providing and sustaining the excellent software and hardware, and for answering all the technical questions, I want to thank Jens Jørgen and Ole. Also, a big thank you to Bettina and Lone for helping with all the administrative tasks of a Ph.D. project.

To my family, thank you for all the support, not just for the past three years but through all of my life.

Finally, to Sofie. Thank you so much! When we both started our Ph.D. three years ago, I don't think we knew what we were going into. There has definitely been ups and downs in our projects, but it has really been invaluable to have someone at home who truly understood the challenges. Now that I'm done with the hard part, I'm looking forward to support you in the next months so we can cross the finish line together.

List of Publications

Publication [I] is the main outcome of this Ph.D. project, where the novel fingerprints of individual electronic states using operator matrix elements and projected density of states are introduced, and applied to train a machine learning model predicting the G_0W_0 band structures of 2D materials.

Publication [II] tests the electronic structure fingerprints in a context of classifying dynamically unstable 2D materials in terms of unstable phonon modes.

Finally, Publication [III] introduces new materials and properties of the computational 2D materials database (C2DB). This paper is the result of a collaboration of the entire CAMD section, where my contribution is a benchmark of the ability to machine learn various properties using different fingerprints.

[I] **N. R. Knøsgaard** and K. S. Thygesen,
"Representing individual electronic states for machine learning GW band structures of 2D materials,"
Nat Commun **13**, 468 (2022).

Printed in copy from page 72.

[II] S. Manti, M. K. Svendsen, **N. R. Knøsgaard**, P. M. Lyngby and K. S. Thygesen,
"Predicting and machine learning structural instabilities in 2D materials,"
arXiv:2201.08091v1 **[cond-mat.mtrl-sci]**.
(Accepted by *npj Computational Materials*)

Printed in copy from page 83.

[III] M. Gjerding, A. Taghizadeh, A. Rasmussen, S. Ali, F. Bertoldo, T. Deilmann, **N. R. Knøsgaard**, M. Kruse, A. H. Larsen, S. Manti, T. G. Pedersen, U. Petralanda, T. Skovhus, M. K. Svendsen, J. J. Mortensen, T. Olsen, and K. S. Thygesen,
"Recent Progress of the Computational 2D Materials Database (C2DB),"
2D Mater. 2021, **8**, 044002.

Printed in copy from page 96.

Contents

CHAPTER 1
Introduction

The global society is in constant need of new materials. New materials to facilitate the increasing energy consumption and to enable the transition from a fossil-fuel based energy consumption to renewable energy sources. We seek materials to support new technologies and to optimize existing technologies. New materials to replace the current materials relying on scarce resources or materials with negative health effects.

As the title reveals, the work of the Ph.D. project presented in this thesis intersects the subjects of **machine learning**, **quantum mechanics** and **materials science**. These three subjects are individual complex research fields, and therefore the purpose of this introduction is to define the playground of this Ph.D. project and point in the direction of relevant literature.

Materials science is the interdisciplinary study of materials design and discovery. This covers both the experimental work of synthesising and testing materials, and the theoretical aspects of understanding and explaining the underlying phenomena and calculating the desired properties of the materials. The work in this thesis focuses on the computational part of materials science, where the generation of databases of material properties is essential [1–3].

Quantum mechanics is the fundamental theory describing the physical properties at the scale of atoms and elementary particles such as electrons [4–6]. The characteristic length scale of the materials investigated in this thesis is of atomic scale, and therefore quantum mechanics is a necessity for accurately describing and calculating the properties of the materials.

Machine learning is the field of mathematical algorithms learning from data, and this is the main tool for the work presented in this thesis. Machine learning in the context of materials science often relates to predicting material properties based on the atomic structure of the materials [7–11]. This project also focuses on machine learning material properties, but with the aim to incorporate information from quantum mechanics into the machine learning models. Using machine learning methods is a pathway to accelerate the materials discovery cycle, since the machine learning models are orders of magnitudes faster at predicting properties compared to quantum mechanical modelling methods such as density functional theory (DFT).

The starting point of any project doing machine learning for materials science is the data. While experimental data of materials provides the true baseline, it is cumbersome to achieve suitable amounts of data needed for machine learning. Also, experimental data does not eliminate uncertainties, and therefore the use of alternative data sources such as computational material databases is highly relevant [12]. A frequently used first-principles method for calculating properties of atomic-scale materials is DFT

[13, 14]. DFT is in principle an exact method for determining the ground state electronic structure of many-body systems such as molecules and solid state materials [15]. DFT is used in many high-throughput projects, where large amounts of materials are screened and stored in databases [16–18]. This thesis focuses mainly on the specific class of materials of two-dimensional (2D) materials. 2D materials are atomically thin solid state materials of which graphene (single layer of graphite) is a classical example. 2D materials have a variety of different properties, and they open for the opportunity for stacking different 2D layers. In this way, the properties of the materials can be engineered for specific purposes [19, 20].

The natural representation of a material is the atomic structure, i.e. lists of the chemical elements in the material and their corresponding positions. The atomic structure is often used as the basic level of information when encoding materials for machine learning, but the encoding can be enriched using information from atomic properties [21–25].

The general method of using machine learning to predict material properties is applicable across a wide range of different properties. Common properties predicted with machine learning include total energies and formation energies. These are widely available properties which can be predicted with high accuracy compared to the DFT data [26–28]. While formation energies are relevant for screening materials for thermodynamic stability, the electronic band gap is a property with relevance for specific applications. The band gap is an important property when designing photovoltaic cells, lasers, transistors and other semi-conductor applications. Typically, the band gap is a challenging property to map directly from atomic structures, and a lot of effort is made for developing machine learning methods specifically for band gaps [29–31].

This thesis presents the development of machine learning methods focusing on utilizing information from quantum mechanical simulations of materials. Besides the atomic structure, information from the electronic density and wavefunction is coded into the representations used as input to machine learning models. The novel representation methods are then used for predicting material properties such as electronic band gaps and band structures, elastic properties and dynamic stability of materials, in particular 2D materials.

1.1 Thesis overview

Following this introduction, **Chapter 2** introduces the general machine learning concepts, including an overview of the machine learning algorithms and methods used in this project. Also, an introduction to representing materials and atomic structures is given.

Chapter 3 provides the basic electronic structure theories and methods used for data generation and construction of specific electronic fingerprints of materials. This includes an introduction to density functional theory (DFT) and a brief overview of the *GW* approximation used in many-body perturbation theory. Finally, a short introduction to electron-phonon interactions and its reference to the dynamical stability of materials is presented.

Chapter 4 introduces the computational 2D-materials database (C2DB), which is the primary data set of this project. Additionally, a small exercise of combining unsupervised and supervised machine learning methods based on data from C2DB is presented.

Chapter 5 presents the main results of this thesis which are related to a novel set of fingerprints encoding electronic structure information for individual electronic states. These fingerprints are then applied in a machine learning model to predict G_0W_0 eigenenergies and band gaps. Additionally, the electronic fingerprints are applied to predict the dynamical stability of 2D materials using a classification model. Finally, the electronic fingerprint is benchmarked against structural fingerprints on mulitple properties of 2D materials. As a brief excursion from 2D materials to molecules, a small study applying similar methods for predicting G_0W_0 energies of molecules is presented.

Chapter 6 continues on the work in the previous chapter related to the dynamical stability of materials. This is done by introducing a method that approximates the electron-phonon coupling matrix elements using machine learning.

Chapter 7 makes an overall conclusion of the thesis and gives an outlook of the open questions raised by this project.

1.1.1 Reading guide

This thesis aims to present the output of the Ph.D. project in a results and applications oriented manner, but also in an easy-to-read way enabling readers with various levels of expertise in the different fields to gain something from the reading experience. This means that only the most relevant theory is included in order to build the foundation of understanding the results.

If the reader is familiar with machine learning in general, most of Chapter 2 can be skipped. For some introduction to machine learning in a context of materials science, it can be advantageous to read Section 2.5 regarding representations of atomic structures for machine learning.

For readers with some knowledge of electronic structure methods, it can be advantageous to skip Chapter 3.

For readers with experience with machine learning for materials science, it is suggested to start from chapter 4 for an introduction to the specific data used in this project.

CHAPTER 2
Theory: Machine learning in materials science

In this chapter, the basic concepts of machine learning is explained to support the understanding of the results and applications in later chapters. Additionally, machine learning in a context of materials science is introduced by giving an overview of some of the most frequently used structural fingerprints.

In the context of this project, machine learning is interpreted as the study of algorithms that learn from experience in form of data in an automated way. Within machine learning several fundamental terms exist and sometimes multiple terms are used interchangeably for similar concepts, with this thesis being no exception. Therefore, a thought example of a machine learning application is used to explain some of the terminology.

Consider an algorithm that predicts the price of a house. Some *data* is needed to build such algorithm using machine learning. This data should consist of *observations* of houses and their corresponding prices, which is the *target variable* of the algorithm, i.e. the variable being predicted by the algorithm. Each observation needs to represented by a set of *features*, e.g. construction year, size in square meters, number of floors, postal code, distance to different points of interest etc. These features can be both continuous, integers or categorical. The set of features may also be called the *fingerprint* or *representation* of the houses. The key element of this algorithm is the *model* or mathematical *function*, that takes the features as *input* variables and returns the predicted price as the *output* variable. This model could for instance be a simple linear model, which needs to be *trained* or *fitted* to the data. The training involves determining the optimal set of model *parameters*, i.e. the coefficients and intercept of the linear model. Depending on the choice of model, some *hyperparameters* are introduced which cannot be determined by fitting to the data. This could for example be the order of a polynomial function. These hyperparameters can be determined by comparing an *objective* or *evaluation metric* such as the average prediction error for multiple models with different hyperparameters. Finally, we have a machine learned model for predicting house prices.

In the following sections, some important concepts of machine learning will be introduced with perspectives to applications in material science.

2.1 Supervised learning

This project involves two different tasks of machine learning, namely supervised and unsupervised learning. In supervised learning, all observations in the data set are associated with a target variable, and the purpose is then to give predictions of the target variable based on features of the observations.

Supervised learning tasks are distinguished by the nature of the target variable, resulting in the important terms of classification and regression.

2.1.1 Classification and regression

In classification, the target variable is discrete, i.e. for each observation the output is the predicted class that the observation belongs to. The classification models can be binary or multi-class.

In a context of material science, classification models could be used to predict using e.g. the atomic structure as input if a specific material is

- Metallic
- Dynamically and thermodynamically stable
- Magnetic

For regression models, a continuous response is predicted for each observation. Similar to the classification examples, regression models can give predictions of the values of continuous material properties such as:

- Total energy
- Heat of formation
- Electronic band gap
- Stiffness

2.1.2 Learning algorithms

There is a wide variety of available algorithms for supervised learning differing in model complexity, computational costs, number of hyperparameters, abilities to estimate prediction uncertainties etc. There is the famous ”No free lunch” theorem [32] stating that any algorithm showing an elevated performance over one class of problems is offset by the performance over another class. In the context of machine learning algorithms this means that there is no universal algorithm that always performs better than others. Therefore, for each specific problem a suitable algorithm must be chosen. In the following sections, some of algorithms relevant for the work of this thesis is presented.

Common for all supervised learning tasks is the aim to find a model to predict the target variable y from the observations $\mathbf{x}$:

$$y = f(\mathbf{x}, \mathbf{w}) + \varepsilon \tag{2.1}$$

where $\mathbf{w}$ is the parameters of the model f and ε is the noise of the predictions.

2.1.2.1 Linear models

The perhaps simplest family of machine learning models are the linear models:

$$f(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + ... + w_M x_M = \mathbf{w}^T \mathbf{x} \tag{2.2}$$

where the prediction is simply a linear combination of the input features. This can be expanded to a linear combination of some basis functions (e.g. polynomials $\{1, x, x^2, ..., x^M\}$), which are feature transformations of $\mathbf{x}$

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\Phi} \tag{2.3}$$

There are several ways to find the optimal set of parameters $\mathbf{w}^*$. One way is to define an objective function of the predictions, e.g. the sum of squares $L(\mathbf{y}, \hat{\mathbf{y}}) = \sum_i (y_i - \hat{y}_i)^2$, and then minimize the objective function. The solution $\mathbf{w}^*$ is:

$$\mathbf{w}^* = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{y} \tag{2.4}$$

This is referred to as the ordinary least squares (OLS) solution. OLS might lead to significant overfitting, which is often due to very large parameters chosen by OLS, and therefore it is useful to introduce some regularisation of the parameters $\mathbf{w}$. This can be done by adding a term to the objective function that tends to decrease the size of the parameters. This is typically done by adding a L_1 or L_2 norm term to the objective function with regularisation strengths α and λ:

$$L_1(\mathbf{y}, \hat{\mathbf{y}}) = \sum_i (y_i - \hat{y}_i)^2 + \alpha \sum_j |w_j| \tag{2.5}$$

$$L_2(\mathbf{y}, \hat{\mathbf{y}}) = \sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_j w_j^2 \tag{2.6}$$

The first method is often referred to as LASSO (least absolute shrinkage and selection operator) regression while the second is called ridge regression. Both tend to give smaller parameters than OLS, while LASSO has the additional advantage that it forces some of the parameters to be strictly zero. Therefore LASSO can be used as a feature selection method.

2.1.2.2 Artificial neural networks

Artificial neural networks (ANN) were invented as a mathematical model of the information processing in the neurons of the human brain [33]. In ANN, a neuron is the basic processing unit, and the neurons are organized in connected layers. Figure 2.1

shows a simple schematic of a feed-forward neural network (FFNN) with one input layer with a neuron per input feature, two hidden layers and one output layer. The FFNN maps from M input features to D output variables, but the hidden layers can have more neurons.

Information is passed through the network using the forward pass algorithm. In the input layer, the activities for the neurons are simply the value of the input features $\mathbf{x}$. The j-th neuron in the first hidden layer has activity

$$a_j^{(1)} = \mathbf{x}^T \mathbf{w}_j^{(1)} \tag{2.7}$$

where $\mathbf{w}_j^{(1)}$ holds the weights of all input neurons. This is then passed through an activation function $h(x)$ giving

$$z_j^{(1)} = h(a_j^{(1)}), \qquad \mathbf{z}^{(1)} = \left[z_1^{(1)} z_2^{(1)} \dots z_H^{(1)} \right] \tag{2.8}$$

Neuron k in the next layer then has activation $a_k^{(2)} = (\mathbf{z}^{(1)})^T \mathbf{w}_k^{(2)}$. In general

$$\mathbf{z}^{(l)} = h^{(l)} \left((\mathbf{z}^{(l-1)})^T \mathbf{W}^{(l)}) \right) \tag{2.9}$$

where $\mathbf{W}^{(l)}$ is the weight matrix connecting all neurons between layer $l-1$ and l. The activation functions $h(x)$ play an important role in the ANN since these are responsible for introducing the non-linearity to the models. There are endless options for the activation functions, but two commonly used functions are the hyperbolic tangent $h(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, which maps any number x to the interval $[-1, 1]$, and the rectified linear unit $h(x) = \mathrm{ReLU}(x) = 0$ if $x < 0$ else x.

Training a neural network involves finding the optimal set of weights $\mathbf{W}^*$. This is typically done by defining an objective function for the output values and then using gradient descent backpropagation. In backpropagation, the partial derivatives (gradients) of the objective function with respect to the parameters are propagated backwards through the network from the output layer through the hidden layers to the input layer, and the weights are then updated using gradient descent. The details of backpropagation are elegant but out of scope for this introduction.

So far, only the most simple architecture of an ANN has been considered, i.e. the feed-forward neural network, but there is a lot of flexibility in the design of neural network architectures. One specific design, which has also been used briefly in this project, is the convolutional neural network (CNN). CNNs are often used for image processing. The key difference from a FFNN is that weights are shared across neurons in a layer, typically by sliding a filter of weights across the feature space. This can extract information that is invariant to the position in feature space, which is often useful in image processing, where e.g. the specific location of an object in an image is not relevant, but only the presence of that object is relevant.

2.1.2.3 Decision trees and ensembles

Another frequently used family of algorithms is those based on decision trees such as classification and regression trees (CART). In a CART classes or values are assigned

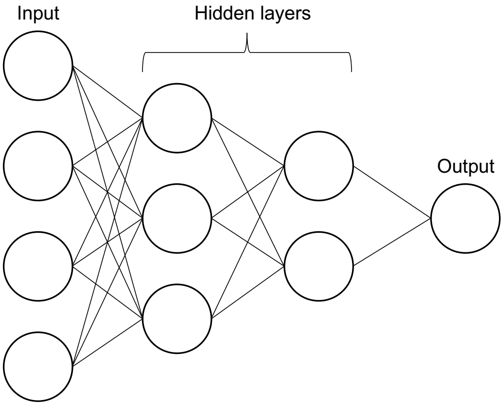


Figure 2.1: Schematic of a feed-forward neural network. Information from the input neurons are passed through the neurons in the two hidden layers before reaching the output layer.

to the observations by hierarchically splitting the observations based on the feature values. Figure 2.2 shows an example of a CART where a value is assigned based on splits of three features. The interpretation of a decision tree is straightforward, at least for smaller trees, since it is based directly on splits of features, and in some cases this can be used to extract knowledge from the model.

Even though a single decision tree may be used as a machine learning model, it is often used in an ensemble method. In decision tree ensembles, multiple trees are

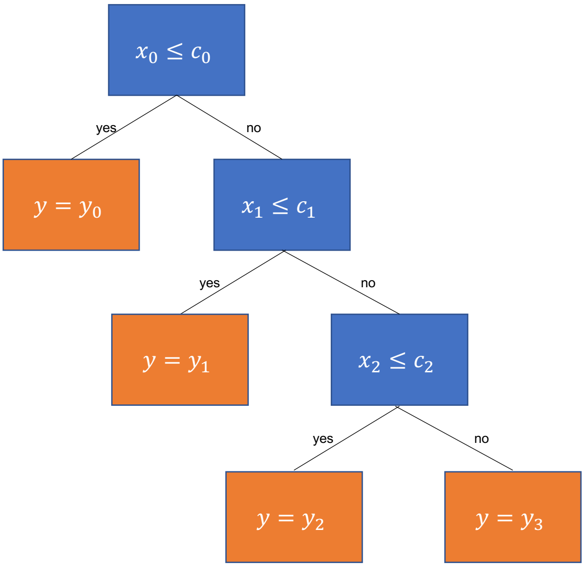


Figure 2.2: Schematic of a simple classification and regression tree (CART). The tree output y depends on the values of features x_0, x_1 and x_2.

trained and the final result is then typically the majority vote (classification) or the average prediction (regression) of the trees in the ensemble. With ensemble methods some of the interpretability vanishes due the larger number of trees and splits.

Several decision tree ensemble methods exist, such as bagging trees and random forests [34, 35], but for this project the focus is on gradient boosted trees and more specifically the software package *XGBoost* [36, 37]. The core model in XGBoost is the decision tree ensemble, i.e. the predicted value $\hat{y}_i$ for the observation $\boldsymbol{x}_i$ is a sum over trees

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i), f_k \in \mathcal{F} \tag{2.10}$$

where K is the number of trees in the ensemble, f_k is a single decision tree and $\mathcal{F}$ is the functional space of all CARTs. As for all machine learning algorithms, the model is trained by defining an objective function and optimizing it. The objective function for XGBoost is a combination of training loss and regularization:

$$\text{obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \omega(f_k) \tag{2.11}$$

where $l(y_i, \hat{y}_i)$ is the loss function, which is typically mean squared error (MSE) for regression and logistic loss for classification, and $\omega(f_k)$ is a measure of the complexity of the tree f_k. The trees are trained using additive learning, which means that a single tree is added at a time with the previous trees being fixed:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(\boldsymbol{x}_i) = \hat{y}_i^{(t-1)} + f_t(\boldsymbol{x}_i) \tag{2.12}$$

and for each step the tree that optimizes the objective function is added:

$$\begin{aligned} \text{obj}^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{K} \omega(f_k) && (2.13) \\ &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\boldsymbol{x}_i)) + \omega(f_t) + c && (2.14) \end{aligned}$$

where the complexity of the previous trees is simply a constant. In the case of a general loss function, the Taylor expansion of the loss function up to second order is taken:

$$\text{obj}^{(t)} = \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(\boldsymbol{x}_i) + \frac{1}{2} h_i f_t(\boldsymbol{x}_i)^2 \right] + \omega(f_t) + c \tag{2.15}$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ and $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$. Since $l(y_i, \hat{y}_i^{(t-1)})$ is a constant, the objective becomes (after removing all constants):

$$\text{obj}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(\boldsymbol{x}_i) + \frac{1}{2} h_i f_t(\boldsymbol{x}_i)^2 \right] + \omega(f_t) \tag{2.16}$$

So far, the complexity has been only vaguely introduced, since no universal measure of the complexity of a decision tree exists. In XGBoost it is defined as:

$$\omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2.17}$$

where T is the number of leaves in the tree, and w_j is the value assigned to leaf j. This resembles the L_1 and L_2 regularization of a linear model.

By using the defined complexity and changing the summation to sum over leafs, the objective is rewritten as

$$\begin{aligned} \text{obj}^{(t)} &= \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2 \right] + \gamma T & (2.18) \\ &= \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T & (2.19) \end{aligned}$$

where I_j is the set of observations in the j-th leaf, and this objective has the optimal set of leaf scores w_j given by

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{2.20}$$

and the objective value

$$\text{obj}^* = \frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{2.21}$$

This is used to determine how good a specific tree is, and in principle all possible trees should be examined. In practice, this is intractable and therefore a tree is build one branch at a time. By looking at one leaf, it is to be split if the gain of splitting the leaf is larger than some threshold value. The gain is given by:

$$\text{Gain} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \tag{2.22}$$

where the subscripts refer to the left and right parts of the split. This way the tree is build one step at a time until no more gain is achievable and the tree is pruned.

The reason for choosing XGBoost for many of the machine learning models in this thesis is a combination of its ability to handle large amounts of data and that the importances of features can be interpreted. Compared to neural networks which also handles large amounts of data, XGBoost is easier to tune in terms of hyperparameters.

2.1.2.4 Gaussian process regression

As an alternative to the previously described machine learning algorithms, Gaussian process regression (GPR) is also used throughout this project. GPR has the main advantage compared to most ML algorithms that it is probabilistic, i.e. uncertainties are provided along with the predicted value. Another advantage is that it is generally possible to fit a reasonable GPR model with a relatively small amount of data, while e.g. a neural net typically requires much more data. On the other hand, a GPR is not well suited for large amounts of data since both the training and prediction involves inverting a matrix of size $N \times N$ with N being the number of observations.

A Gaussian process is a collection of random variables of which any finite set has a joint Gaussian distribution [38]. A Gaussian process is completely determined by a mean function $m(x)$ and a covariance function or kernel function $k(\mathbf{x}, \mathbf{x}')$, i.e. a function $f(x)$ is expressed as a Gaussian process

$$f(x) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{2.23}$$

If $\mathbf{x}$ is a set of training points with target values $\mathbf{y}$, then predictions for a set of test points $\mathbf{x}_*$ are calculated as

$$\mathbf{y}(\mathbf{x}_*) = \mathbf{m}(\mathbf{x}_*) + \mathbf{k}(\mathbf{x}_*, \mathbf{x})^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} (\mathbf{y} - \mathbf{m}(\mathbf{x})) \tag{2.24}$$

with variances

$$\boldsymbol{\sigma}^2(\mathbf{x}_*) = \mathbf{k}(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}(\mathbf{x}_*, \mathbf{x})^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}(\mathbf{x}_*, \mathbf{x}) \tag{2.25}$$

Here, $\mathbf{K}_{nm} = k(\mathbf{x}_n, \mathbf{x}_m)$ is the kernel matrix for points in the training set, $\mathbf{I}$ is the idenity matrix and σ is the noise parameter. There is a long list of possible kernel functions, but the most frequently used is the radial basis function:

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp(-||\mathbf{x} - \mathbf{x}'||^2 / 2l^2) \tag{2.26}$$

with a prefactor α and length scale l.

GPR are very useful machine learning models since reasonable models can typically be trained with much smaller amounts of data compared to e.g. neural networks and decision trees. Additionally, GPR provides estimates of the prediction uncertainties, which is often as valuable as the actual prediction. In this project, GPRs are mainly used in Bayesian optimization methods, e.g. for optimizing hyperparameters.

2.2 Unsupervised learning

In unsupervised learning, the observations are not associated with a target variable and therefore prediction models in terms of classification or regression are not an option.

In this project, the two main applications of unsupervised learning are clustering and dimensionality reduction.

2.2.1 Clustering

In clustering, labels are assigned to the observations rather than given from the data. This is typically done based on a distance metric, e.g. the euclidean distance, in the feature space of the observations. A common obstacle in many clustering algorithms is that the number of clusters is a hyperparameter of the algorithm, and since no target variable is taken into consideration in unsupervised learning, the optimal number of clusters is typically unknown.

2.2.1.1 *k*-means clustering

One of the frequently used clustering algorithm is the k-means algorithm which labels the observations based on the euclidean distance to the closest cluster centroid [39–41]. The algorithm takes the number of clusters as a hyperparameter and then optimizes the cluster means by aiming to minimize the momentum, i.e. the within-cluster sum of squares :

$$\text{argmin}_S \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} ||\mathbf{x} - \mu_i||^2 \tag{2.27}$$

This problem is computationally difficult, so typically the *naive k-means* method is applied. This consists of an assignment step where the observations are assigned to the cluster with the nearest mean, and then the new cluster means are computed. These two steps are the iterated until some convergence criteria is met. An example of applying k-means clustering on the structures of 2D materials is presented in Section 4.2.

2.2.2 Dimensionality reduction

Another aspect of unsupervised learning is dimensionality reduction, which is very useful both for visualisation purposes, e.g. being able to show high-dimensional feature vectors in a low-dimensional space, but also as a data processing step in a supervised learning approach. Reducing the number of features prior to training a supervised model can both decrease the computational costs and additionally increase the prediction accuracy. For some of the material fingerprints presented later in this chapter and the following chapters, the number of features can be quite high and also higher than the number of observations, which makes it impossible to fit e.g. a linear model using least squares.

2.2.2.1 PCA

In principal component analysis (PCA), the high-dimensional feature vectors are projected onto subspaces using linear transformations. These subspaces are refered to as the principal components and they are chosen such that they are orthogonal and that the variance of the projected data is maximised.

Suppose there are N observations of dimension M, $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N \in \mathbb{R}^M$. In PCA, the aim is to find a new n-dimensional representation $\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_N \in \mathbb{R}^n$ where $n < M$.

Here $\mathbf{b}_i$ is the projection of $\mathbf{x}_i$ onto a subspace V:

$$\mathbf{b}_i^T = \tilde{\mathbf{x}}_i^T \mathbf{V}, \qquad \tilde{\mathbf{x}}_i = \mathbf{x}_i - \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \tag{2.28}$$

The projection matrix $\mathbf{V} = [\mathbf{v}_1 \mathbf{v}_2 \cdots \mathbf{v}_n]$ is chosen such that the variance

$$W = \sum_{i=1}^{N} (b_i - \bar{b})^2 \tag{2.29}$$

is maximised. In practice this is done using singular value decomposition where any $N \times M$ matrix $\mathbf{X}$ can be decomposed as $\mathbf{X} = \mathbf{U \Sigma V}^T$ where $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\mathbf{\Sigma}$ is a diagonal matrix with elements $\sigma_1, \sigma_2, ..., \sigma_M$. The n first principal components are then chosen from $\mathbf{V}$.

2.2.2.2 t-SNE

An alternative to the linear PCA method is t-distributed stochastic neighbor embedding (t-SNE) [42]. Here, the similarity between high-dimensional observations $\mathbf{x}_i$ and $\mathbf{x}_j$ is modelled as the conditional probability that $\mathbf{x}_i$ chooses $\mathbf{x}_j$ as its neighbor using a Gaussian probability density:

$$p_{j|i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma_i)}{\sum_{k\neq i}\exp(-||\mathbf{x}_i - \mathbf{x}_k||^2/2\sigma_i)} \tag{2.30}$$

The observation $\mathbf{x}_i$ is mapped to $\mathbf{y}_i$ in the low-dimensional space, and in this space the similarity between points are modelled using the Student t-distribution with one degree of freedom

$$q_{ij} = \frac{(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_k \sum_{l\neq k}(1 + ||\mathbf{y}_k - \mathbf{y}_l||^2)^{-1}} \tag{2.31}$$

By defining the joint probability $p_{ij} = \frac{p_{j|i}+p_{i|j}}{2N}$ for the N observations, the aim of the t-SNE algorithm is to make the two joint probabilities similar and this is done by minimizing the Kullback-Leibler divergence:

$$KL(P||Q) = \sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2.32}$$

which is minimized using gradient descent. It should be noted that t-SNE can only be used as a visualization method and not as a dimensionality reduction step in a machine learning process.

2.2.2.3 Auto-encoder

A simple feed-forward neural network can also be used in an unsupervised learning setting using an network architecture called an auto-encoder. Figure 2.3 shows such architecture which consists of an encoder that compresses the input into a latent space, and then a decoder that tries to reconstruct the input. The network is trained using back-propagation using a loss function that measures the difference between the input and output (typically just L2-loss $L = \sum_i ||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2$). The latent space can be used for visualisation where the dimension of the latent space is typically chosen to be $2-3$, but the decoder can also be used as a data pre-processing step before training a supervised regression or classification model.

2.3 Model selection and evaluation

The goal of machine learning is not to find the model that fits the data most accurately, both to find the model that generalises the best to new data not seen by the model in the training phase. Therefore the data set is split in a train and a test set, and the model is trained only on a subset of the data. The prediction accuracy on the test set is then a measure of the models ability to generalise to new data.

Usually a lot of different models are compared, corresponding to different choices of algorithms or different hyperparameter values, in order to find the optimal model for the specific problem. In this case, the test set should not be used to select the best model, because then information from the test set will leak into the model selection routine. Therefore, it is useful to further split the train set to create a validation set. This can for example be done using a leave-out method such as k-fold cross validation (CV), which is sketched in Figure 2.4. In k-fold CV, the train set is split in k subsets. The training routine is then repeated k times, where for each iteration or fold, the model is trained on $k-1$ subsets and validated on the k-th set. This method also

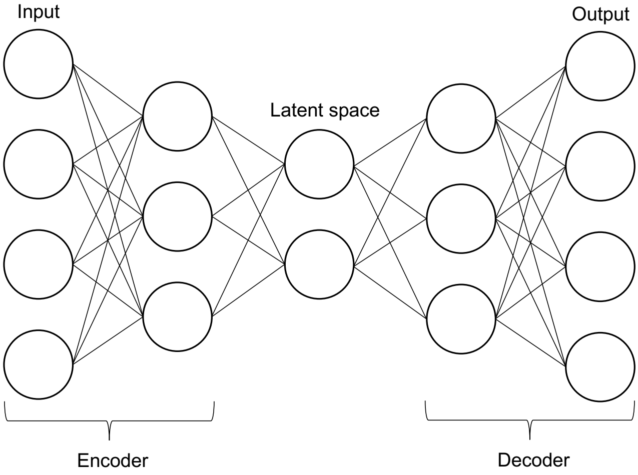


Figure 2.3: Schematic of a neural net auto-encoder. The encoder compresses the input information to the low-dimensional latent space, while the decoder aims to reconstruct the original input.

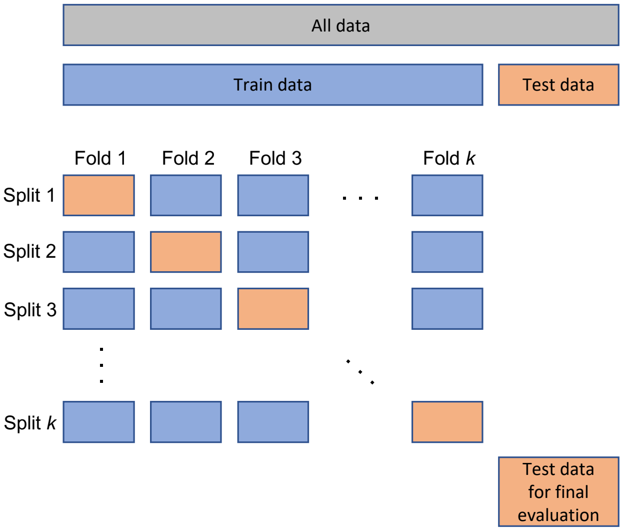


Figure 2.4: Schematic of a train-test split strategy for training machine learning models. First, the data set is split in train and test sets. The train set is then further split using k-fold cross validation.

has the advantage that statistics can be calculated based on the k prediction accuracy scores.

In this project, the standard approach is to split the data set in a 80% train and 20% test set, and then perform 5- or 10-fold cross validation on the train set.

2.3.1 Prediction error metrics

A key element in machine learning is choosing a metric to evaluate the prediction errors. The choice of such metric will depend on the learning task (classification or regression) but also more specifically on the nature of the problem being solved, e.g. how sensitive the metric should be to outliers in a regression setting or how to deal with imbalanced classes in a classification setting.

2.3.1.1 Regression metrics

For regression, mainly two metrics are used in this project. The mean absolute error (MAE) for a prediction $\hat{y}_i$ and the corresponding true value y_i:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i}^{N}|y_i - \hat{y}_i| \tag{2.33}$$

and the root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2} \tag{2.34}$$

The main difference between the two is that RMSE is more sensitive to outliers.

2.3.1.2 Classification metrics

The most simple metric for a classification problem is the accuracy which is simply the fraction of correct labels:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{correct}} + N_{\text{wrong}}} \tag{2.35}$$

For a binary classification problem with a positive and negative class, there are four different scenarios for a prediction. A true positive/negative (TP/TN) is a correctly labeled positive/negative sample, a false positive (FP) is a negative sample incorrectly labeled as positive while a false negative (FN) is a positive sample labeled as negative. From these definitions several metrics can be derived:

$$\text{Sensitivity or true positive rate (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} = 1 - \text{FNR} \tag{2.36}$$

$$\text{Specificity or true negative rate (TNR)} = \frac{\text{TN}}{\text{TN} + \text{FP}} = 1 - \text{FPR} \tag{2.37}$$

$$\text{Fall-out or false positive rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{2.38}$$

$$\text{Miss-rate or false negative rate (FNR)} = \frac{\text{FN}}{\text{FN} + \text{TP}} \tag{2.39}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.40}$$

Additionaly, the receiver operating characteristic (ROC) curve is a plot showing the true positive rate (TPR) vs. the false positive rate (FPR) for various probability thresholds of the underlying model [43]. From this, the area under the curve (AUC) is a very useful metric for classification models. A perfect classifier will have a ROC-AUC of 1 while random guessing yields a ROC-AUC of 0.5. This metric is especially useful when dealing with imbalanced classes.

2.3.2 Hyperparameter optimization

Finding the optimal set of hyperparameters is an essential yet difficult part of machine learning. Typically, the space of hyperparameters can be quite large, and depending on the choice of learning algorithm it can be infeasible to explore the full space. For projects in this thesis, mainly two different approaches to hyperparameter optimization has been used:

- **Grid search.** Define a one-dimensional grid for each parameter and iterate through all combinations of parameters calculating the objective function (e.g. the validation set accuracy of a ML model).
- **Bayesian optimization.** Train a Gaussian Process Regression model to predict the objective as function of the hyperparameter values and then use this surrogate model to perform an optimization to select new set of parameters based on a acquisition function. Repeat until some convergence criteria is met.

If the hyperparameter space is sufficiently small and the training routine is computationally cheap it is preferred to do a full grid search, but in other cases it can be necessary to settle for a method, e.g. Bayesian optimisation, that does not necessarily explore the entire space but samples the space in a smart way depending on the choice of acquisition function.

2.4 Interpretation of features

When doing machine learning projects, the overall aim is often to train a model as accurate as possible with the data available. But as an additional outcome, information about the importance of features in the model is desirable. This is both to build trust in the model by being able to verify some of the choices made by the machine learning algorithm, but also to possibly gain some knowledge from the model, e.g. how the model output depends on the value of a specific input.

For a linear model, the importance of features can be directly evaluated by looking at the coefficients in the model, but for more complex and non-linear models, this task is more cumbersome. In this section, methods for evaluating feature importance and dependence by first discussing the specific case of decision tree models and then for a general model using SHAP analysis.

2.4.1 Feature importance for decision trees

As mentioned in the introduction of decision tree methods in Section 2.1.2.3, decision tree models are in principle easy to interpret, but for ensembles of trees it is more complicated. It is, however, still possible to evaluate the relative importance of features by looking at statistics of which features are used in the model. In XGBoost, three different feature importance methods are implemented [36]:

Weight. The feature importance is calculated as the number of times a feature is used to perform a split.

Gain. The average gain as calculated in (2.22) is used as measure of feature importance.

Coverage. The average coverage is defined as the average number of samples affected by splits performed by a specific feature, i.e. this method typically favors features used near the root of the trees.

These three methods naturally give different results, which is to be considered when concluding on feature importance.

2.4.2 SHAP analysis for explainable ML

The SHAP method aims to explain why a model makes a certain prediction [44]. For a single prediction $f(x)$ based on a single point x, SHAP uses an additive explanation model g to the real model f that is a linear combination of binary variables:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{2.41}$$

where $z' \in \{0,1\}^M$ is a simplified binary input that maps to the original input through some mapping function. M is the number of simplified inputs. This additive explanation model assigns an effect ϕ_i to each feature. The chosen method for determining the set of parameters ϕ yields the differences between various explanation methods. The SHAP method inherits from classic Shapley regression values where

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)] \tag{2.42}$$

Here, $|z'|$ is the number of non-zero elements in z', $z' \subseteq x'$ is all z' where the elements are a subset of non-zero elements in x' and $z' \setminus i$ means setting $z'_i = 0$. The exact computation of these SHAP values is complicated and beyond the scope of this project, but the interpretation of the SHAP values is very clear. The SHAP value of a feature for a specific prediction approximates the effect on the prediction output of having that feature in the model compared to a model without the feature.

2.5 Representation of atomic structures

So far, the application of machine learning in a context of materials science has only been discussed on an abstract level. Using machine learning methods for materials requires first of all that the data (typically just in terms of the atomic structure, e.g. a list of atomic numbers and their position in some unit cell) is represented as a feature vector serving as input to the machine learning algorithms. In this section, examples of such representations/fingerprints of atomic structures will be introduced along with an outline of the general fingerprint requirements, that should be considered when developing new fingerprint methods.

Many different fingerprints methods have been developed [11, 45–48], also methods which uses other information sources than just the atomic structures (e.g. atom specific properties), but the focus here is on fingerprints encoding just the atomic structure. This section does not aim to provide a full overview of atomic structure fingerprints, but merely an introduction to a few selected fingerprints. For this project, the Python package DScribe has been especially useful, since it contains implementations of some of the most common structural fingerprints [49].

2.5.1 Fingerprint requirements

In principle, it should be possible to map any material property from the atomic structure, i.e. the list of atoms and their positions should be enough information for a machine learning model to predict any property given that a reasonable dataset exists. This means that if we just have enough data and a suitable complex machine learning model, then it should be possible to learn any property directly from the atomic structure. The reason that the atomic structure is not used directly as a fingerprint is that it does not fulfill some basic requirements for fingerprints. These requirements represent the lowest level of domain knowledge that are coded into the fingerprint methods, which makes it easier for the machine model to learn the mapping from structure to property.

A material fingerprint should exhibit the following requirements [11, 22, 47]:

1. **Invariance**. The fingerprint should be invariant to transformations that preserve the property being predicted, e.g. rotations, translations, permutations of atom indexing. For periodic systems, a supercell (repetition of the unit cell) should have the same fingerprint as the unit cell.

2. **Uniqueness**. The fingerprint should be able to describe materials with different properties uniquely, i.e. if two materials have different properties they should have different fingerprints.

3. **Descriptive**. If two materials have similar properties they should also be close in fingerprint space.

4. **Continuity**. The fingerprint should be continuous and ideally differentiable with respect to the atomic coordinates.

5. **Computational efficiency**. The computational costs of generating the fingerprint should be significantly cheaper than the reference method, e.g. if the model predicts a property calculated using DFT, then costs of generating the fingerprint and predicting the property using machine learning should be cheaper than the corresponding DFT calculation.

6. **Generality**. The fingerprint should be able to encode any type of material (atoms, molecules, periodic systems of both one, two and three dimensions). Also, the fingerprint should ideally be applicable for learning several different properties.

2.5.2 Coulomb and Ewald sum matrices

A very simple structural fingerprint is the Coulomb matrix where the individual matrix elements are [22]:

$$M_{ij} = \begin{cases} 0.5 Z_i^{2.4} & \text{for } i = j \\ \\ \frac{Z_i Z_j}{R_{ij}} & \text{for } i \neq j \end{cases} \tag{2.43}$$

Here Z_i is the atomic number of atom i and R_{ij} is the euclidean distance between atoms i and j for atoms in the unit cell. Thus, the Coulomb matrix is not very suitable for periodic systems, since no interactions with neighboring cells are included. An extension of the Coulomb matrix for periodic systems is the Ewald sum matrix. The basic idea here is to represent the full Coulomb interaction energy corresponding to all infinite repetitions of the periodic lattice, i.e. the matrix element is a double infinite sum over all atoms:

$$x_{ij} = \frac{1}{N} Z_i Z_j \sum_{k,l,k \neq l} \frac{1}{R_{kl}} \tag{2.44}$$

This type of sum has some convergence issues, and therefore it is split into parts of two rappidly converging sums and one constant:

$$x_{ij} = x_{ij}^{(r)} + x_{ij}^{(m)} + x_{ij}^{0} \tag{2.45}$$

where $x_{ij}^{(r)}$ is the short-term interaction calculated in real space, $x_{ij}^{(m)}$ is the long-term interactions calculated in reciprocal space and x_{ij}^{0} is a constant. The short-term part is given by

$$x_{ij}^{(r)} = Z_i Z_j \sum_{\boldsymbol{L}} \frac{\operatorname{erfc}(a||\boldsymbol{r}_i - \boldsymbol{r}_j + \boldsymbol{L}||_2)}{||\boldsymbol{r}_i - \boldsymbol{r}_j + \boldsymbol{L}||_2} \tag{2.46}$$

where the sum runs over lattice vectors $\boldsymbol{L}$ within a sphere defined by $L_{\max}$. The long-range term is

$$x_{ij}^{(m)} = \frac{Z_i Z_j}{\pi V} \sum_{\boldsymbol{G}} \frac{\exp\left(\frac{-||\boldsymbol{G}||_2^2}{(2a)^2}\right)}{\boldsymbol{G}||_2^2} \cos\left(\boldsymbol{G} \cdot (\boldsymbol{r}_i - \boldsymbol{r}_j)\right) \tag{2.47}$$

where the sum runs over all reciprocal lattice vectors $\boldsymbol{G}$ in a sphere of radius $G_{\max}$ and V is the unit cell volume. The last constant term is

$$x_{ij}^{0} = -\frac{a}{\sqrt{pi}}(Z_i^2 + Z_j^2) - (Z_i + Z_j)^2 \frac{\pi}{2Va^2} \tag{2.48}$$

where the first part is the Ewald self-terms and the second is a background compensating term. The Ewald sum matrix components are defined by the screening parameter a which affects how quickly the sums converge.

A common artefact of the Coulomb and Ewald matrix representation methods is that the size of the matrix depends on the number of atoms in the unit cell, which is generally not suitable for machine learning algorithms. Therefore, the matrix size is defined by the system with the highest number of atoms in the unit cell, and for all other systems the matrix is simply padded by zeros. Also, to make the fingerprints invariant to the change of atom indexing, the matrix rows/columns can be sorted according to e.g. their L_2 norms.

2.5.3 Many-body Tensor Representation (MBTR)

The many-body tensor representation (MBTR) is a collection of broadened distributions of k-body terms described by the general formula [11, 47]:

$$f_k(x, z_1, z_2, ..., z_k) = \sum_{i_1,...,i_k} w_k \mathcal{N}(x|g_k, \sigma) \prod_{j=1}^{k} \delta_{z_j, Z_{i_j}} \tag{2.49}$$

Here w_k is a weighting function reducing the contribution from atoms far away from each other, g_k is a k-body function depending on the atoms $i_1, ..., i_k$. $\mathcal{N}(x|\mu, \sigma)$ is a normal distribution with mean μ and variance σ^2 evaluated at x, and $\delta_{i,j}$ is the Kronecker delta function. Though this is a general k-body formula, the method is typically applied to encode one-body (atomic numbers), two-body (distances, inverse distances) or three-body (dihedral angles) functions.

Since MBTR uses the atomic number Z to encode different types of atoms, the dimension of the fingerprint scales with Z^k. When using MBTR for machine learning model, all relevant Z must be encoded in the fingerprint. This typically results in very high-dimensional fingerprints.

2.5.4 Smooth Overlap of Atomic Positions (SOAP)

The Smooth Overlap of Atomic Positions (SOAP) uses an expansion of an atoms local neighborhood density approximated by Gaussian functions located at atom positions onto orthogonal radial basis functions and spherical harmonics [47, 48]. The SOAP fingerprint uses a partial power spectrum defined as

$$p(\boldsymbol{r})^{Z_1 Z_2}_{nn'l} = \pi \sqrt{\frac{8}{2l+1}} \sum_m c^{Z_1}_{nlm}(\boldsymbol{r})^* c^{Z_2}_{n'lm}(\boldsymbol{r}) \tag{2.50}$$

where the coefficients are

$$c^{Z}_{nlm}(\boldsymbol{r}) = \int_{\mathcal{R}^3} \mathrm{d}V g_n(r) Y_{lm}(\theta, \phi) \rho^Z(\boldsymbol{r}) \tag{2.51}$$

Here $g_n(r)$ are radial basis functions, $Y_{lm}(\theta, \phi)$ are spherical harmonics and $\rho^Z(\boldsymbol{r})$ is the Gaussian smoothed density at $\boldsymbol{r}$ for atoms with atomic number Z. For a global fingerprint, the power spectrum is averaged over the atomic sites in the (periodic) system. This averaging can be done either before summing up the magnetic quantum numbers such that $p^{Z_1 Z_2}_{nn'l} \sim \sum_m (\frac{1}{n} \sum_i c^{i,Z_1}_{n'lm})^* (\frac{1}{n} \sum_i c^{i,Z_2}_{n'lm})$ or by averaging the power spectrum of different sites as $p^{Z_1 Z_2}_{nn'l} \sim \frac{1}{n} \sum_i \sum_m (c^{i,Z_1}_{n'lm})^* (c^{i,Z_2}_{n'lm})$. As for MBTR, SOAP fingerprints are typically also high-dimensional due to use of the atomic number Z to encode atoms.

CHAPTER 3

Theory: Electronic Structure Methods

The second dimension in the theoretical foundation of this project regards electronic structure methods, which have both been used for generating the data used for machine learning but also to develop machine learning fingerprints with some domain knowledge originating from the theories and concepts of quantum mechanics and electronic structure methods.

This chapter starts by introducing the basics of density functional theory (DFT), which is an approach for calculating ground state properties of interacting many-body systems. Next, the *GW* approximation for calculating excited state properties of systems using many-body Green's functions is briefly introduced. Finally, a short introduction to the theory of the electron-phonon coupling is given, which is used later for designing a fingerprint with the specific purpose of learning the structural stability of materials.

3.1 Density functional theory

The basic equation for developing the concepts of density functional theory is the time-independent Schrödinger equation [15]:

$$\hat{H}\Psi_n(\mathbf{r},\mathbf{R}) = \varepsilon_n \Psi_n(\mathbf{r},\mathbf{R}) \tag{3.1}$$

where $\Psi_n(\mathbf{r},\mathbf{R})$ are the eigenstates with eigenenergies ε_n for a system with electronic coordinates of N electrons $\mathbf{r} = \mathbf{r}_1, ..., \mathbf{r}_N$ and nuclear coordinates $\mathbf{R} = \mathbf{R}_1, ..., \mathbf{R}_P$. The Hamiltonian $\hat{H}$ includes the kinetic energies of electrons and nuclei as well as all Coulomb interactions between charged particles (nuclei-nuclei, electron-electron and electron-nuclei):

$$\begin{aligned}
\hat{H} = &-\sum_{I=1}^{P} \frac{\hbar}{2M_I}\nabla_I^2 - \sum_{i=1}^{N} \frac{\hbar}{2m}\nabla_i^2 + \frac{e^2}{2}\sum_{I=1}^{P}\sum_{J\neq I}^{P} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} \\
&+ \frac{e^2}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} - \frac{e^2}{2}\sum_{I=1}^{P}\sum_{i=1}^{N} \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}_i|} \qquad (3.2) \\
&= \hat{T}_n + \hat{T}_e + \hat{V}_{nn} + \hat{V}_{ee} + \hat{V}_{ne} \qquad (3.3)
\end{aligned}$$

The full electron+nuclear wavefunction $\Psi_n(\mathbf{r},\mathbf{R})$ is a function of $3(N+P)$ variables and this is not feasible (nor possible) to solve mainly due to the two-body Coulomb interactions making the Schrödinger equation not seperable and simplifications/approximations are thus needed.

3.1.1 Born-Oppenheimer approximation

The first level of approximations is the Born-Oppenheimer approximation which originates from the fact that the large difference in masses between electrons and nuclei makes the dynamics of the electrons much faster than that of the nuclei. This means that the wavefunction $\Psi(\mathbf{r},\mathbf{R})$ can be written as a product state $\Psi(\mathbf{r},\mathbf{R}) = \Phi(\mathbf{R})\phi(\mathbf{r},\mathbf{R})$ where $\Phi(\mathbf{R})$ is the nuclear wavefunction which only depends on the nuclei coordinates while the electronic wavefunction $\phi(\mathbf{r},\mathbf{R})$ depends on the electronic coordinates $\mathbf{r}$ and parametrically on the on the nuclei coordinates $\mathbf{R}$. The electronic Schrödinger equation then reads

$$\hat{h}_e\phi_n(\mathbf{r},\mathbf{R}) = \varepsilon_n\phi_n(\mathbf{r},\mathbf{R}) \tag{3.4}$$

with the electronic Hamiltonian

$$\begin{aligned}\hat{h}_e &= -\sum_{i=1}^{N}\frac{\hbar}{2m}\nabla_i^2 + \frac{e^2}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|} - \frac{e^2}{2}\sum_{I=1}^{P}\sum_{i=1}^{N}\frac{Z_I}{|\mathbf{R}_I-\mathbf{r}_i|}\\ &= \hat{T}_e + \hat{V}_{ee} + \hat{V}_{ne}\end{aligned} \tag{3.5}$$

In other words, the role of the nuclei is merely to set up a potential for the electrons to interact with and this can also be thought of as an external potential $V_{\text{ext}}(\mathbf{R})$.

This electronic problem is still intractable to solve for most realistic systems since the size of the wavefunctions scales exponentially with the number of electrons. Therefore further approximations are needed, of which DFT is the method of choice for this project. The key element in DFT is the electronic density

$$\rho(\mathbf{r}) = N\int\prod_{i=1}^{N}d\mathbf{r}_i\,|\phi(\mathbf{r}_1,...,\mathbf{r},...,\mathbf{r}_N)|^2 \tag{3.6}$$

3.1.2 Hohenberg-Kohn theorems

The two Hohenberg-Kohn theorems [50] build on top of the Thomas-Fermi theory [51, 52] that the ground state energy can be expressed as a functional of the ground state density. The two theorems read:

1. The external potential is determined directly by the electronic density, and this is unique up to an additive constant. As a corollary to this theorem, the ground state wavefunction is indeed also directly determined from the ground state density.

The ground state energy E_0 is then a functional of the ground state density ρ_0:

$$E_0 = E[\rho_0] = T[\rho_0] + V_{ee}[\rho_0] + \int d\mathbf{r} v_{\text{ext}}(\mathbf{r})\rho_0(\mathbf{r}) \tag{3.7}$$

where the two first terms are the kinetic and potential energy functionals of the density.

2. The energy functional gives the ground state energy only if the true ground state density is given as input:

$$E[\rho_0] \leq E[\rho'] \tag{3.8}$$

for $\rho \neq \rho_0$.

These two theorems are the mathematical foundation of DFT. Using this formulation, the ground state density can be found by minimization due to the variational principle of the energy functional, though this is extremely difficult in practice since the universal form of the functional is unknown.

3.1.3 Kohn-Sham equations

Kohn and Sham introduced a scheme for determining the ground state [53]. The general idea is that since the ground state energy is a functional of the density, all systems with the same density will give the same energy, even some self-constructed non-interacting system as long as it has the same density as the full interacting system. The general energy functional is rewritten as

$$E[\rho] = T_{\text{NI}}[\rho] + V_{\text{ext}}[\rho] + V_{\text{H}}[\rho] + E_{\text{XC}}[\rho] \tag{3.9}$$

where $T_{\text{NI}}[\rho]$ is the kinetic energy of the non-interacting system which is much simpler to calculate as the kinetic energy of the interacting system is formally unknown. $V_{\text{ext}}[\rho]$ is the external potential setup by the nuclei and other external perturbations. $V_{\text{H}}[\rho]$ is the Hartree potential describing the electrostatic repulsion of electrons in the system, and finally $E_{\text{XC}}[\rho]$ is the exchange-correlation functional which is designed to account for all the electron-electron interactions missed in the non-interacting terms. This is a much simpler task to solve, since all the unknown terms are grouped in the exchange-correlation functional which needs to be approximated.

For the non-interacting Kohn-Sham system, the functional is constructed as

$$E_{\text{KS}}[\rho] = T_{\text{NI}}[\rho] + V_{\text{KS}}[\rho] \tag{3.10}$$

with the Kohn-Sham potential energy functional $V_{\text{KS}}[\rho]$ which is written as

$$V_{\text{KS}}[\rho] = \int d\mathbf{r}\rho(\mathbf{r})v_{\text{eff}}(\mathbf{r}) \tag{3.11}$$

where the Kohn-Sham effective potential $v_{\text{eff}}(\mathbf{r})$ is designed as

$$v_{\text{eff}}(\mathbf{r}) = v_{\text{ext}}(\mathbf{r}) + v_{\text{H}}(\mathbf{r}) + v_{\text{XC}}(\mathbf{r}) \tag{3.12}$$

such that the non-interacting system has the same ground state density as the interacting system. The exchange-correlation potential is given as the functional derivative of the exchange-correlation energy functional

$$v_{\mathrm{XC}}(\mathbf{r}) = \frac{\delta E_{\mathrm{XC}}[\rho]}{\delta \rho(\mathbf{r})} \tag{3.13}$$

and the Hartree potential is simply

$$v_{\mathrm{H}}(\mathbf{r}) = \int d\mathbf{r}' \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \tag{3.14}$$

The non-interacting Kohn-Sham states $\phi_i(\mathbf{r})$ are the eigenstates of the one-particle Schrödinger equation

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + v_{\mathrm{eff}}(\mathbf{r})\right] \phi_i(\mathbf{r}) = \varepsilon_i \phi_i(\mathbf{r}) \tag{3.15}$$

from which the density of the system with N electrons can be calculated as

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2 \tag{3.16}$$

This closes the circle of the Kohn-Sham formalism as the effective potential depends on the density which depends on the Kohn-Sham states calculated using the effective potential. This means that the solutions are to be determined self-consistently using the following algorithm:

1. Initial guess of the density $\rho(\mathbf{r})$
2. Calculate the Hartree potential $v_{\mathrm{H}}(\mathbf{r}) = \int d\mathbf{r}' \dfrac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}$ and then the effective potential $v_{\mathrm{eff}}(\mathbf{r}) = v_{\mathrm{ext}}(\mathbf{r}) + v_{\mathrm{H}}(\mathbf{r}) + v_{\mathrm{XC}}(\mathbf{r})$
3. Solve the one-particle Schrödinger equations $\left[-\dfrac{\hbar^2}{2m}\nabla^2 + v_{\mathrm{eff}}(\mathbf{r})\right] \phi_i(\mathbf{r}) = \varepsilon_i \phi_i(\mathbf{r})$
4. Recalculate the electronic density $\rho(\mathbf{r}) = \sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2$
5. Compare the new and old densities. Restart from 2 until converged

3.1.4 Exchange-correlation functionals

Using the Kohn-Sham equations in practice comes down to selecting a suitable approximation to the exchange-correlation functional. The simplest example is the local density approximation (LDA), where the exchange-correlation energy is calculated using:

$$E_{\mathrm{XC}}^{\mathrm{LDA}}[\rho] = \int d\mathbf{r} \rho(\mathbf{r}) \epsilon_{\mathrm{XC}}^{\mathrm{hom}}[\rho(\mathbf{r})] \tag{3.17}$$

which is designed to be exact for the homogeneous electron gas. $\epsilon_{\text{XC}}^{\text{hom}}[\rho(\mathbf{r})]$ is the exchange-correlation energy per volume for the homogeneous electron gas, which can be divided into an exchange and a correlation part. The exchange contribution has an analytic form while the correlation contribution is estimated using Monte-Carlo simulations. LDA works well for systems with slowly varying densities and poorly for systems with localized states.

The next level in XC-functional approximations is to also use information of the gradient of the density, which is done in the generalised gradient approximation (GGA):

$$E_{\text{XC}}^{\text{GGA}}[\rho] = \int d\mathbf{r} F_{\text{XC}}^{\text{GGA}}[\rho(\mathbf{r}), \nabla\rho(\mathbf{r})] \tag{3.18}$$

The most commonly used GGA functional is the PBE functional [54, 55], which is also the one used for most of the data in this project.

Both LDA and GGA tends to underestimate band gaps of systems, due to a derivative discontinuity, i.e. derivatives of certain quantities depend on the number of electrons in a discontinuous way while this is continuous for GGA and LDA. One way to reduce this problem is to use hybrid functionals, where the exchange part is mixed with exchange from Hartree Fock

$$E_{\text{XC}}^{\text{hybrid}}[\rho] = \alpha E_{\text{X}}^{\text{HF}}[\rho] + (1-\alpha)E_{\text{X}}^{\text{GGA}}[\rho] + E_{\text{C}}^{\text{GGA}}[\rho] \tag{3.19}$$

where α is a parameter fitted to some dataset. A specific hybrid function often used for more accurate band gaps is the HSE06 functional [56].

3.2 Many-body perturbation theory: The GW approximation

A general issue with Kohn-Sham DFT is that it tends to underestimate band gaps [57, 58]. A commonly used method to acquire more accurate band gaps is many-body perturbation theory and more specifically the GW approximation [59], which is a method well-suited for explaining the processes behind direct and inverse photoemission, i.e. the emission and adsorption of electrons [60]. The experimental observable here is the photocurrent, which is the probability of emitting an electron with a specific kinetic energy within a time interval. This is closely related to the intrinsic spectral function of the electronic system $A(\boldsymbol{r}, \boldsymbol{r}', \omega)$:

$$A(\boldsymbol{r}, \boldsymbol{r}', \omega) = \frac{1}{\pi}\text{Im}G(\boldsymbol{r}, \boldsymbol{r}', \omega)\,\text{sign}(E_F - \omega) \tag{3.20}$$

i.e. the spectral function is given by the imaginary part of the single-particle Green's function $G(\boldsymbol{r}, \boldsymbol{r}', \omega)$, which is the probability that an electron created or destroyed at $\boldsymbol{r}$ is correlated with the inverse process at $\boldsymbol{r}'$.

In the case of non-interacting electrons the spectral function becomes a series of delta functions $A_{ss'}(\omega) = \langle\psi_s| A(\omega) |\psi_{s'}\rangle = \delta_{ss'}\delta(\omega - \epsilon_s)$. When interactions is turned on the matrix elements $A_{ss'}(\omega)$ may reform into peaks appearing as a delta peak broadened by electron-electron interactions, which is interpreted as the excitation of a *quasiparticle*, which is an important concept in many-body perturbation theory. The broadening of the spectral peak of the quasiparticle depends on the lifetime of the excitation due to electron-electron scattering and the area of the peak is the renormalisation or quasiparticle weight Z. The real space image of a quasiparticle is that of an additional electron or hole interacting with its own polarisation cloud, i.e. the electron/hole is screened by the other electrons in the system.

3.2.1 Hedin's GW approximation

The actual calculation of the single particle Green's function is often done by writing is as a Dyson equation:

$$G(1,2) = G_0(1,2) + \int G_0(1,3)\Sigma(3,4)G(4,2)d(3,4) \tag{3.21}$$

where G_0 is the non-interacting Green's function, Σ is the self-energy and the notation (i) refers to the spacetime point $(\boldsymbol{r}_i, t_i)$. The self-energy is constructed to capture the effects of exchange and correlation which corrects the single particle Hamiltonian. The Green's function G is related to other quantities through Hedin's equations, which is a set of 5 integral-differential equations. In the GW approximation, Hedin's equations are (together with Eq. 3.21):

$$\Gamma(1,2,3) = \delta(1,2)\delta(1,3) \tag{3.22}$$

$$\chi_0(1,2) = -iG(1,2)G(2,1) \tag{3.23}$$

$$W(1,2) = V(1,2) + \int V(1,3)\chi_0(3,4)W(4,2)d(3,4) \tag{3.24}$$

$$\Sigma(1,2) = iG(1,2)W(1^+,2) \tag{3.25}$$

Here Γ is a three-point interaction vertex, χ_0 is the irreducible polarizability and W is the screened Coulomb interaction. The GW approximation is named such because the self-energy Σ is approximated as the product of G and W.

3.2.2 The G_0W_0 approximation

Hedin's GW equations can be solved iteratively, but due to large computational costs, the calculations may be iterated only once, which gives the G_0W_0 approximation. G_0W_0 calculations are typically performed on top of a Kohn-Sham DFT calculation, where the G_0W_0 quasiparticle energy can be found as

$$\epsilon_s^{\mathrm{QP}} = \epsilon_s^{\mathrm{KS}} + Z_s \langle\psi_s^{\mathrm{KS}}|\, \Sigma(\epsilon_s^{\mathrm{KS}}) - V_{\mathrm{xc}}\, |\psi_s^{\mathrm{KS}}\rangle \tag{3.26}$$

with the quasiparticle weight given as

$$Z_s = \left(1 - \frac{\partial}{\partial\omega} \langle\psi_s^{\mathrm{KS}}|\, \Sigma(\omega)\, |\psi_s^{\mathrm{KS}}\rangle\bigg|_{\omega=\epsilon_s^{\mathrm{KS}}}\right)^{-1} \tag{3.27}$$

The DFT exchange-correlation energy V_{xc} is subtracted because the self-energy is designed specifically to account for exchange and correlation.

3.3 Electron-phonon interactions

A very central type of interaction between (quasi)-particles in solids is the electron-phonon interaction, which is the source of many interesting physical phenomena. The lowest order process of electron-phonon interactions is that of an electron being scattered by the creation or annihilation of a single phonon. The Hamiltonian modelling such process is [61]:

$$\hat{H} = \hat{H}_{\text{el}} + \hat{H}_{\text{ph}} + \hat{H}_{\text{el-ph}} \tag{3.28}$$

with the electron Hamiltonian

$$\hat{H}_{\text{el}} = \sum_{n\boldsymbol{k}\sigma} \epsilon_{n\boldsymbol{k}} c^{\dagger}_{n\boldsymbol{k}\sigma} c_{n\boldsymbol{k}\sigma} \tag{3.29}$$

where $c^{\dagger}_{n\boldsymbol{k}\sigma}$ ($c_{n\boldsymbol{k}\sigma}$) are the creation (annihilation) operators of electrons with band index n, momentum $\boldsymbol{k}$, spin σ and eigenenergies $\epsilon_{n\boldsymbol{k}}$. The phonon Hamiltonian is expressed as

$$\hat{H}_{\text{ph}} = \sum_{\boldsymbol{q}j} \omega_{\boldsymbol{q}j} \left(b^{\dagger}_{\boldsymbol{q}j} b_{\boldsymbol{q}j} + \frac{1}{2} \right) \tag{3.30}$$

where $b^{\dagger}_{\boldsymbol{q}j}$ ($b_{\boldsymbol{q}j}$) are creation (annihilation) operators of a phonon with momentum $\boldsymbol{q}$, branch index j and energy $\omega_{\boldsymbol{q}j}$. The final term is the coupling between electrons and phonons:

$$\hat{H}_{\text{el-ph}} = \sum_{nn'\boldsymbol{k}\sigma} \sum_{\boldsymbol{q}j} g^{\boldsymbol{q}j}_{n'\boldsymbol{k}+\boldsymbol{q},n\boldsymbol{k}} c^{\dagger}_{n'\boldsymbol{k}+\boldsymbol{q}\sigma} c_{n\boldsymbol{k}\sigma} \left(b_{\boldsymbol{q}j} + b^{\dagger}_{-\boldsymbol{q}j} \right) \tag{3.31}$$

where $g^{\boldsymbol{q}j}_{n'\boldsymbol{k}+\boldsymbol{q},n\boldsymbol{k}}$ is the electron-phonon matrix element describing the probability of the scattering process. To first order, this matrix element is

$$g^{\boldsymbol{q}j}_{n'\boldsymbol{k}+\boldsymbol{q},n\boldsymbol{k}} = \sum_{s\alpha} A^{\boldsymbol{q}j}_{ai} \langle n'\boldsymbol{k}+\boldsymbol{q}\sigma | \, \delta^{\boldsymbol{q}}_{ai} V \, | n\boldsymbol{k} \rangle \tag{3.32}$$

with $A^{\boldsymbol{q}j}_{ai} = \frac{\eta_{ai}(\boldsymbol{q}j)}{\sqrt{2M_s \omega_{\boldsymbol{q}j}}}$ describing the mass-scaled polarization vector where $\eta_{s\alpha}(\boldsymbol{q}j)$ is the eigenvector of phonon mode $\boldsymbol{q}j$ with a and i indexing the atom and cartesian coordinate, respectively.

3.3.1 Dynamical stability from perturbation theory

In the following, u_{aiN} denotes the displacement of atom a along cartesian axis i in unit cell N. $u_s = u_{aiN}$ is used as collective notation of such displacement. The energy from displacements of all atoms u_s to second order is given by:

$$E(\{u_s\}) = E_0 + \frac{1}{2}\sum_{ss'} u_s F_{ss'} u_{s'}, \tag{3.33}$$

where

$$F_{ss'} = F_{s's} = \frac{\partial^2 E}{\partial u_s \partial u_{s'}} \tag{3.34}$$

is the force constant matrix. The first order contribution is zero since the system is assumed relaxed in the unit cell making the forces on the atoms zero.

The potential induced by displacements u_s is

$$V = \sum_s W_s = \sum_s \frac{\partial V(r - R_s)}{\partial u_s}, \tag{3.35}$$

i.e. W_s is the gradient of the potential with respect to displacement of atom a in direction i in cell N. Considering V as a perturbation, the second order energy is then:

$$\begin{aligned} E^{(2)}(\{u_s\}) &= \sum_{ss'} u_s u_{s'} \sum_{n\mathbf{k}} \sum_{m\mathbf{k}'} f_{n\mathbf{k}}(1 - f_{m\mathbf{k}'}) \frac{\langle n\mathbf{k}| W_s |m\mathbf{k}'\rangle \langle m\mathbf{k}'| W_{s'} |n\mathbf{k}\rangle}{\varepsilon_{n\mathbf{k}} - \varepsilon_{m\mathbf{k}'}} && (3.36) \\ &= \sum_{ss'} u_s u_{s'} M_{ss'} && (3.37) \end{aligned}$$

and the force constant matrix is given as

$$F_{ss'} = M_{ss'} + M_{s's} \tag{3.38}$$

The dynamical matrix is the Fourier transform of the force constant matrix:

$$\begin{aligned} D_{ai,a'i'}(\mathbf{q}) &= \sum_{NN'} \exp(i\mathbf{q}\cdot\mathbf{R}_N) F_{aiN,a'i'N'} \exp(-i\mathbf{q}\cdot\mathbf{R}_{N'}) && (3.39) \\ &= M_{ai,a'i'}(\mathbf{q}) + M_{a'i',ai}(\mathbf{q}) && (3.40) \end{aligned}$$

with

$$\begin{aligned} M_{ai,a'i'}(\mathbf{q}) &= \sum_{n\mathbf{k}} \sum_{m\mathbf{k}'} \frac{f_{n\mathbf{k}}(1 - f_{m\mathbf{k}'})}{\varepsilon_{n\mathbf{k}} - \varepsilon_{m\mathbf{k}'}} \langle n\mathbf{k}| \sum_N \exp(i\mathbf{q}\cdot\mathbf{R}_N) W_{aiN} |m\mathbf{k}'\rangle \times \\ &\quad \langle m\mathbf{k}'| \sum_{N'} \exp(-i\mathbf{q}\cdot\mathbf{R}_{N'}) W_{a'i'N'} |n\mathbf{k}\rangle \\ &= \sum_{n\mathbf{k}} \sum_{m\mathbf{k}'} \frac{f_{n\mathbf{k}}(1 - f_{m\mathbf{k}'})}{\varepsilon_{n\mathbf{k}} - \varepsilon_{m\mathbf{k}'}} \langle n\mathbf{k}| W_{ai}(\mathbf{q}) |m\mathbf{k}'\rangle \langle m\mathbf{k}'| W_{a'i'}(-\mathbf{q}) |n\mathbf{k}\rangle \\ &= \sum_{n\mathbf{k}} \sum_{m\mathbf{k}+\mathbf{q}} \frac{f_{n\mathbf{k}}(1 - f_{m\mathbf{k}+\mathbf{q}})}{\varepsilon_{n\mathbf{k}} - \varepsilon_{m\mathbf{k}+\mathbf{q}}} \langle n\mathbf{k}| W_{ai}(\mathbf{q}) |m\mathbf{k}+\mathbf{q}\rangle \langle m\mathbf{k}+\mathbf{q}| W_{a'i'}(-\mathbf{q}) |n\mathbf{k}\rangle \end{aligned} \tag{3.41}$$

By diagonalising $D_{ai,a'i'}(\mathbf{q})$ the electron-phonon coupling frequencies can be retrieved. The eigenvalues of the dynamical matrix ultimately gives the dynamical stability of the system, with negative eigenvalues indicating instability since the energy can be decreased by translating atoms along a phonon mode.

CHAPTER 4

Data: Computational 2D-materials database

With the theoretical foundation of both machine learning and electronic structure methods summarised in the previous chapters, this chapter focuses on presenting the primary data source for the projects of this thesis, namely the Computational 2D Materials Database (C2DB)[62], of which Publication [III] outlines the progress since the first introduction of the database in 2018 [63].

C2DB is a highly curated database differentiating itself from other material databases by its high number of calculated properties. The data has a high level of consistency since all calculations are performed using the same code, parameters and workflow to ensure transparency, reproducibility and consistency. The database consists of $\approx$ 4000 atomically thin materials, of which a minority of the materials have actually been experimentally synthesised in lab. The majority of the materials are *hypothetical* materials generating using lattice decorations.

The general method for materials generation is sketched in Figure 4.1. By using the subset of experimentally known 2D materials, the lattices of these are decorated with atoms from a chemically reasonable subset of the periodic system. This generates a large space of hypothetical materials which are then passed through the computational workflow.

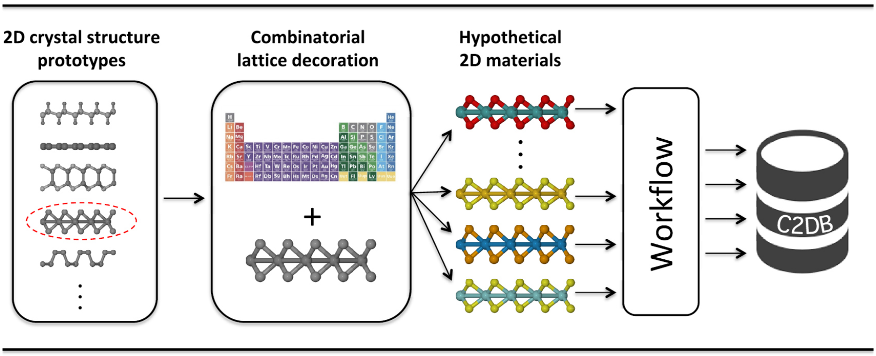

2D crystal structure prototypes
Combinatorial lattice decoration
Hypothetical 2D materials
+
Workflow
C2DB

Figure 4.1: Schematic of the material candidate generation of C2DB (reprinted from [62]). Starting from a set of prototypes from experimentally known 2D materials, lattice decoration is used to generate new hypothetical materials. The generated candidates are then passed through the database workflow.

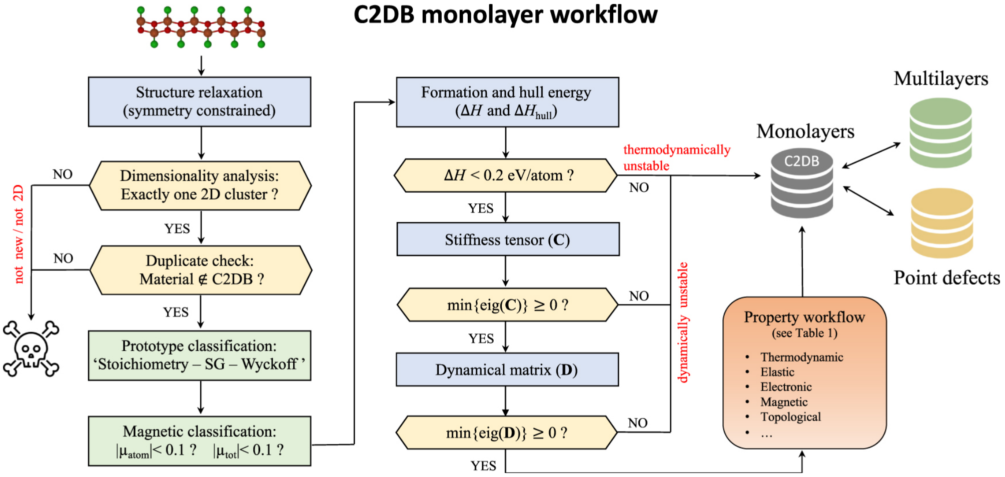

C2DB monolayer workflow
Structure relaxation (symmetry constrained)
Dimensionality analysis: Exactly one 2D cluster ?
NO
YES
Duplicate check: Material ∉ C2DB ?
NO
YES
not new / not 2D
Prototype classification: 'Stoichiometry – SG – Wyckoff'
Magnetic classification: $|\mu_{atom}|$< 0.1 ? $|\mu_{tot}|$< 0.1 ?
Formation and hull energy (ΔH and ΔH_{hull})
ΔH < 0.2 eV/atom ?
thermodynamically unstable
NO
YES
Stiffness tensor (C)
min{eig(C)} ≥ 0 ?
NO
YES
Dynamical matrix (D)
min{eig(D)} ≥ 0 ?
NO
YES
dynamically unstable
Monolayers
C2DB
Multilayers
Point defects
Property workflow
(see Table 1)
Thermodynamic
Elastic
Electronic
Magnetic
Topological
…

Figure 4.2: Schematic of the workflow of C2DB (reprinted from [63]).

4.1 High throughput calculations and workflow

As mentioned, one of the key advantages of the C2DB is the high level of consistency due to all calculations being done using the same code stack and workflow. The code stack consists of GPAW [64, 65] used for DFT calculations, ASE[66] for manipulating atomic structures in Python and ASR[67] for handling workflows. Figure 4.2 shows the workflow for C2DB. For a potential candidate structure, the first step is to relax the structure. This is done in a symmetry constrained way ensuring the symmetries of the prototype structure from the candidate generation method. Next up are two filtering steps which takes out materials that have either disintegrated into a non-2D structure during relaxation or that already exists in the database. After this, the structure is classified based on its stoichiometry, space group and occupied Wyckoff positions, and also its magnetic state. Two central energetic stability properties in terms of the heat of formation and the energy above the convex hull are then calculated, which is used to filter out thermodynamically unstable materials (materials with an energy above the convex hull larger than 0.2 eV/atom). Following the thermodynamical stability check, the dynamical stability of the structure is analysed. Since the structures are relaxed with DFT using symmetry constraints, they may lie on a saddle point on the potential energy surface due to the symmetry constraints or an insufficient number of atoms in the unit cell. The dynamical stability is assessed based on the eigenvalues of the stiffness tensor relating the stress of the material to the applied strain, and the Γ-point Hessian matrix for a 2x2 supercell with no relaxation performed for the supercell. If any of the minimal eigenvalues are negative, the material is labeled as dynamical unstable, since this indicates that the total energy can be reduced by either deforming the unit cell (negative minimum eigenvalue of the stiffness tensor) or displacing some atoms (negative minimum eigenvalue of the Hessian matrix).

The thermodynamically and dynamically stable materials are then passed through the property workflow calculating a wide range of different electronic, magnetic and optical properties. Properties of certain interest for this thesis are electronic band structures and band gaps, projected band structures, effective masses, polarisabilities and phonons. Some of the property methods introduce specific criteria in terms of the number of atoms per unit cell or the electronic band gap, and therefore not all properties are calculated for all materials. Figure 4.3 shows a pairplot of five properties from the C2DB, i.e. the HSE06 band gap, heat of formation, exciton binding energy, in-plane static polarisability calculated in the RPA and averaged over the x and y polarisation directions and the in-plane Voigt modulus.

4.2 Insights from unsupervised clustering of 2D materials

As an initial showcase of machine learning applications using the data from C2DB, an unsupervised clustering based on the atomic structures represented by MBTR fingerprints is performed. The data is split into a 80% train and 20% test set and the k-means clustering algorithm is applied on the train set. The algorithm takes the number of clusters as the primary parameter, and then iteratively assigns labels to the input structures based on the minimum distance to a cluster center and then recomputes the cluster centers until convergence. Since the optimal number of clusters is ill-defined, a range from 1 to 10 clusters is examined. Figure 4.4 shows the number of materials per cluster vs. the number of clusters. For small number of clusters, it tends to create one large cluster holding the majority of materials, but as the number of clusters is increased, the majority cluster is split into separate clusters. In the case of 10 clusters, 8 of the 10 clusters hold more than 200 materials. In order to investigate how the clusters differ, Figure 4.5 shows the mean (μ) vs. standard deviations (σ) of the materials per cluster for six selected properties, which is used to examine differences in the distribution of properties per cluster. The properties are the number of atoms in the unit cell, dynamic stability, heat of formation, energy above the convex hull, PBE band gap and Voigt modulus. The mean axis are scaled to cover the 5-95% percentile range of the property distribution for the entire data set. It is seen that for most of the properties the cluster distributions mainly differ in terms of the standard deviations, though for the heat of formation and Voigt modulus some of the clusters have significantly different means as well.

The generated clusters (with the number of clusters varying from 1 to 10) are then used in a supervised approach to predict the heat of formation and the PBE band gap. A Random Forest regression model is fitted per cluster for the training observations within the cluster, and then predictions are made for the observations in the test set labeled to the same cluster. Finally, the mean absolute error is calculated for the entire test set (with elements from the different cluster regression models). Figure 4.6 and 4.7 shows the results for the cluster-regression method for the heat of formation and the PBE band gap, respectively. The bars in the top panels show the test MAE per cluster with the black dots corresponding to the MAE of the 1-cluster model, i.e. the model trained on all observations. It is seen that for both heat of formation and PBE band gap some of the clusters have MAEs significantly lower than that of the model trained on all data, while other clusters have MAEs similar to or even larger than

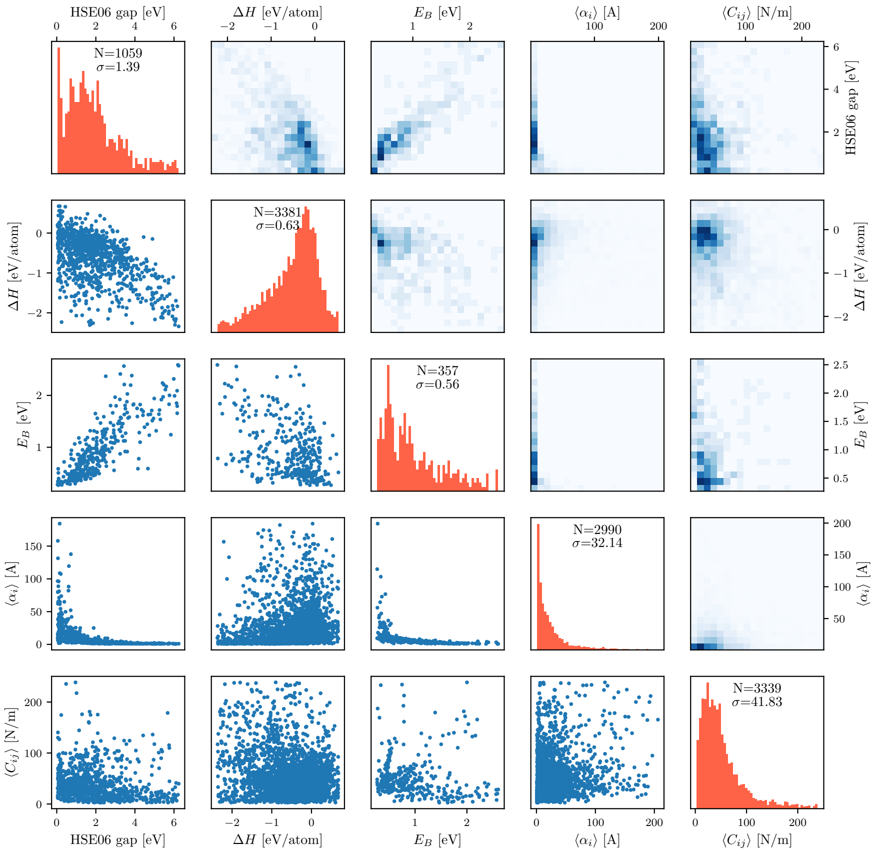


Figure 4.3: Pair plot of five properties from C2DB (HSE06 band gap, heat of formation, exciton binding energy, in-plane static polarisability and Voigt modulus). The diagonal plots show the histograms of the individual properties. Below and above the diagonal are scatter plots and density plots, respectively. Reprinted from Publication [III].

the all-data-model. The bottom panels show the weighted MAEs of the entire test set, which yield that for both properties there are significant improvements in the test MAEs when increasing the number of clusters. This is a quite interesting conclusion, since the general machine learning thesis is that more data gives more accurate models. Obviously, the full amount of data is used to generate the clusters, but even with the reduced amount of data within the clusters, the cluster-specific models are better. This means that training models on data that are more uniform is better than adding more data. Of course, this analysis is limited to only the MBTR fingerprint, k-means clustering and Random Forest regression, so a more thorough analysis should be carried out to determine if this conclusion is merely an artefact of e.g. the chosen fingerprint.

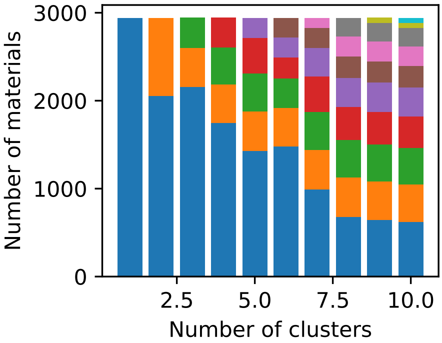


Figure 4.4: The number of materials per cluster vs. the number of clusters found using k-means clustering of materials from C2DB. For low number of clusters, the materials group in one large majority cluster and few minority clusters. As the number of clusters is increased, the majority cluster is split and the cluster sizes become more similar.

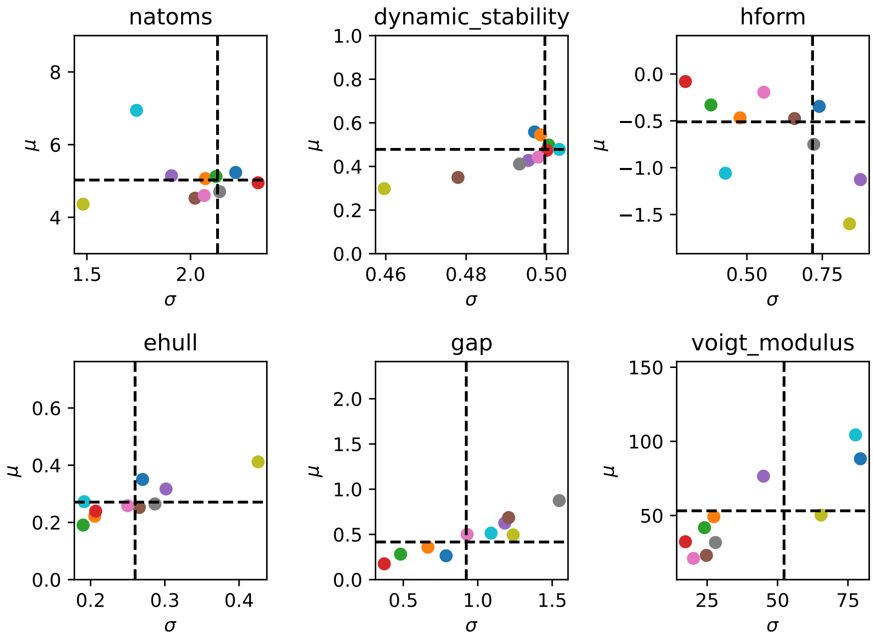


Figure 4.5: Scatter plots of the mean vs. standard deviation of the six properties in the case of 10 clusters.

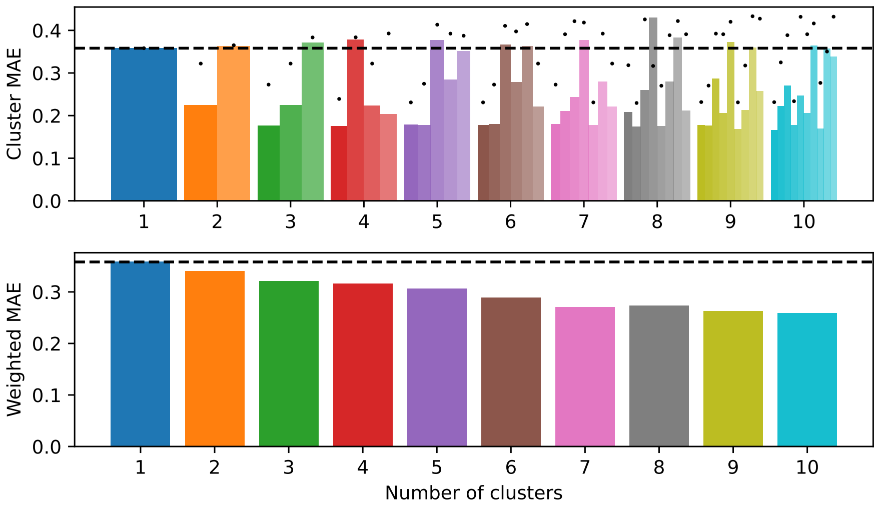


Figure 4.6: Mean absolute errors for the prediction of the heat of formation vs. the number of clusters when fitting a model per cluster. The bars in the top panel shows the MAEs of the single clusters with the black dot showing the corresponding MAE when using the 1-cluster model. The bottom panel shows the MAEs weighted over all clusters.

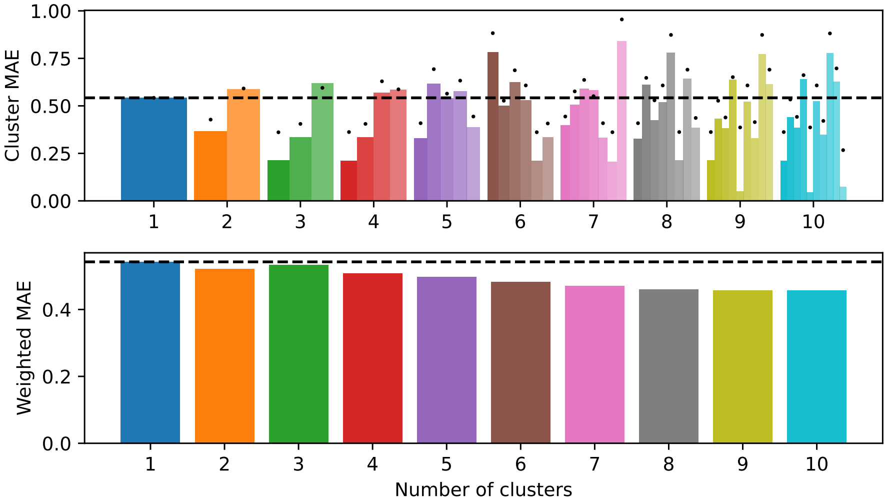


Figure 4.7: Mean absolute errors for the prediction of the PBE band gap vs. the number of clusters when fitting a model per cluster. The bars in the top panel shows the MAEs of the single clusters with the black dot showing the corresponding MAE when using the 1-cluster model. The bottom panel shows the MAEs weighted over all clusters.

CHAPTER 5

Electronic structure fingerprints

Predicting material properties using machine learning requires a method for representing the material in a fingerprint used as input to the machine learning model. This is typically done by using the atomic structure as the main source of information. Several fingerprint methods for encoding the atomic structure already exist (see Section 2.5 for an introduction to structural fingerprints). In most high-throughput studies, the atomic structure is determined based on a DFT relaxation of the structure. This means that the Kohn-Sham wavefunction and electronic structure are already available, which makes it possible to utilize such information directly in the fingerprints.

In this chapter, two novel types of fingerprints (ENDOME and RAD-PDOS) [68] encoding the electronic structure are presented. These fingerprint are some of the main outcomes of this Ph.D. project and they were originally developed for Publication [I]. This is followed by short summaries of the three publications associated with this thesis, each demonstrating different applications of the electronic structure fingerprints. The summaries are intended to briefly showcase the results, while the more detailed discussions are found in the actual papers enclosed in the thesis.

5.1 Global, local and state fingerprints

Fingerprints used in machine learning for materials science are typically encoding either the entire structure of a material (global fingerprint) or a specific point in space with information from its surroundings (local fingerprint). However, when fingerprinting electronic structures, another level of information is relevant i.e. the individual quantum mechanical states of the system. In the following sections, new fingerprints that takes this extra level of information into account are introduced.

5.2 Energy decomposed operator matrix elements (ENDOME)

In quantum mechanics, each measurable physical observable of a system is associated with a hermitian operator acting on the wave function or state. For periodic systems such as 2D materials these states are typically labelled by the band index n and wave vector index k (ideally also a spin index, which is neglected here). Information between states $|nk\rangle$ and $|n'k'\rangle$ in the system can be extracted from operator matrix elements of a given operator $\hat{A}$:

$$A_{nk,n'k'} = |\, \langle nk|\, \hat{A}\, |n'k'\rangle\, |^2 \tag{5.1}$$

The motivation behind the energy decomposed operator matrix elements (ENDOME) fingerprint is that standard DFT wave functions holds a lot of useful information, but the wavefunctions of realistic systems are high-dimensional and unmanageable. Therefore, ENDOME aims to provide some of the information from the wavefunction in a lower dimensional representation by extracting the operator matrix elements from the wavefunctions. The general ENDOME fingerprint for a state (n, k) with eigenenergy ε_{nk} is defined as

$$m^A_{nk}(E) = \sum_{n'k'} A_{nk,n'k'} G\left(E - (\varepsilon_{nk} - \varepsilon_{n'k'}); \delta_E\right) \exp\left(-\alpha_E E\right) \mathrm{sign}(E_F - \varepsilon_{n'k'}) \tag{5.2}$$

The ENDOME fingerprint is thus a function of the energy distance E from the reference state. Here, matrix elements between the reference state and other states are weighted by a Gaussian function $G(x; \delta) = \frac{1}{\delta\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{x^2}{\delta^2}\right)$ centered at $x = 0$ with width δ. Additionally, the contributions are weighted by an exponentially decaying function in the energy E with rate α_E. This decreases the influence from states far away from the reference state. Finally, the sign of the fingerprint component is given by the occupancy of the states $(n'k')$.

The ENDOME fingerprint is represented on e.g. a uniformly spaced grid with N_E points between limits E_{min} and E_{max}. Therefore, the fingerprints introduces the hyperparameters δ_E, α_E, N_E, E_{min} and E_{max} that should be determined depending on the problem.

A global ENDOME fingerprint encoding the entire system instead of a single state can also be constructed as

$$m^A(E_i, E_f) = \sum_{nn'kk'} A_{nk,n'k'} G\left(E_i - (\varepsilon_{nk} - E_F); \delta_E\right) G\left(E_f - (\varepsilon_{n'k'} - E_F); \delta_E\right) \tag{5.3}$$

Here, the two Gaussians are encoding the distance from the Fermi level to states (nk) and $(n'k')$ with the variables E_i and E_f, respectively.

In principal, any quantum mechanical operator can be used in the ENDOME fingerprints. Publication [I] presents a study where the state specific ENDOME fingerprints based on the position operator $\hat{\boldsymbol{r}}$, the nabla operator ∇ and the squared nabla operator ∇^2 are used to predict G_0W_0 eigenenergies.

5.3 Radially decomposed projected density of states (RAD-PDOS)

Another valuable information that can be extracted from the electronic structure is the projections of the wavefunction ψ_{nk} onto atom a and angular orbital ν:

$$\rho_{nk}^{a\nu} = |\left\langle \psi_{nk} | a\nu \right\rangle|^2 \tag{5.4}$$

The radially decomposed projected density of states (RAD-PDOS) fingerprint is constructed using these projections. For a single state (n, k) the RAD-PDOS fingerprint for the coupling between electrons in angular orbitals ν and ν' is defined as:

$$\begin{aligned}\rho_{nk}^{\nu\nu'}(E, R) = \frac{1}{N_e} \sum_{n'k'aa'} & \rho_{nk}^{a\nu} \rho_{n'k'}^{a'\nu'} G\left(R - |R_a - R_{a'}|; \delta_R\right) \exp\left(-\alpha_R R\right) \\ & \times G\left(E - (\varepsilon_{nk} - \varepsilon_{n'k'}); \delta_E\right) \exp\left(-\alpha_E E\right) \operatorname{sign}(E_F - \varepsilon_{n'k'})\end{aligned} \tag{5.5}$$

The RAD-PDOS fingerprint is a function of the radial distance R between atoms a and a' and energy distance E between eigenstates (nk) and $(n'k')$. Similar to ENDOME, the fingerprint contributions are weighted by Gaussians and exponential functions of R and E.

The RAD-PDOS fingerprint introduces additional hyperparameters of δ_R, α_R, N_R, $R_{\min}$ and $R_{\max}$ besides the hyperparameters of ENDOME. The RAD-PDOS fingerprint have different components corresponding to different combinations of atomic orbitals ν and ν'. For the materials in C2DB, only the s, p and d orbitals are relevant giving a total of six RAD-PDOS components (ss, sp, sd, pp, pd and dd).

The RAD-PDOS also comes in a global fingerprint version which uses a radial pair correlation function for atomic orbitals in a given wave function:

$$\rho_{\nu\nu'}^{nk}(R) = \sum_{aa'} \rho_{nk}^{a\nu} \rho_{nk}^{a'\nu'} G\left(R - |R_a - R_{a'}|; \delta_R\right) \exp\left(-\alpha_R R\right) \tag{5.6}$$

The energy dependence is then introduced to achieve the global RAD-PDOS fingerprint:

$$\rho_{\nu\nu'}(E, R) = \sum_{nk} \rho_{\nu\nu'}^{nk}(R) G\left(E - (\varepsilon_{nk} - E_F); \delta_E\right) \exp\left(-\alpha_E E\right) \tag{5.7}$$

Compared to ENDOME, RAD-PDOS has the advantage that it encodes both the electronic structure and the atomic structure.

Figure 5.1 shows examples of the state specific ENDOME and RAD-PDOS fingerprints for the 2D material MoS_2. Panel a) shows the PBE band structure. Panels b) and c) show six different ENDOME fingerprints for the valence band maximum and conduction band minimum at the K-point. The six fingerprints correspond to matrix elements of a unitary DOS operator, the in-plane (xy) and out-of-plane (z) elements of the position operator, similarly for the momentum operator (p_{xy} and p_z) and the square of the momemtum operator (p^2). The panels d)-i) shows the six different RAD-PDOS components.

As an approach to visualise the ENDOME DOS fingerprint of many states in two dimensions, a neural net autoencoder is trained. The autoencoder has a latent space with two neurons, and both the encoder and decoder have a single hidden layer. In Figure 5.2 (left) the observations are plotted in the latent space colorcoded with the G_0W_0 correction energies. Using only the two latent features, the autoencoder clearly separates states with low and high correction energies. Additionally, three random input samples are plotted in the center panels with the corresponding output samples in the right panels. The decoder is able to reproduce the inputs from the two-dimensional latent space with relatively good accuracy.

The ENDOME and RAD-PDOS fingerprints are implemented in Python using GPAW [64, 65] to calculate the operator matrix elements and projections.

5.4 Summary of Publication 1: Representing individual electronic states for machine learning GW band structures of 2D materials

The electronic structure fingerprints ENDOME and RAD-PDOS representing individual states are introduced in this publication, which is printed in copy from page 72. The fingerprints are applied to the problem of predicting the G_0W_0 eigenenergies of individual states using machine learning. The inputs are based on a DFT calculation using the PBE functional, i.e. the ML model is bridging the gap between a computationally cheaper method (PBE) and an expensive method (G_0W_0). Besides the ENDOME and RAD-PDOS fingerprints, the ML model also uses the available properties of the PBE band gap, the occupation of the state, the distance of the state energy to the Fermi level and the in-plane and out-of-plane static polarisabilities as features.

Using a data set of 286 G_0W_0 band structures of non-magnetic two-dimensional materials, which results in a total of 46.000 observations of individual (n, k) states, a machine learning model is trained using the gradient boosting method XGBoost [36, 37] to predict the difference between the G_0W_0 and PBE eigenenergies. The XGBoost algorithm is introduced in Section 2.1.2.3. Figure 5.3 shows an example of PBE and G_0W_0 band structures of MoS_2 in panel a), while panel b) and c) show the total data set distribution of the G_0W_0 correction energies and the absolute correction energies, respectively. The distribution of the G_0W_0 correction energies is seen to consist of two separate distribution corresponding to occupied and unoccupied states.

Figure 5.4 shows low-dimensional visualizations of the ENDOME fingerprint using the in-plane momentum operator p_{xy} and the RAD-PDOS pd fingerprint. These are visualized in two dimensions using the tSNE method, where each observation is colorcoded by the G_0W_0 correction energy. The ENDOME p_{xy} is seen to be able to clearly distinguish between occupied and unoccupied states, and also between states with high and low absolute correction energies. The RAD-PDOS pd shows some of same trends, but it also shows a large blob of observations with mixed correction energies and occupancies. These observations correspond to the materials with no d-electrons resulting

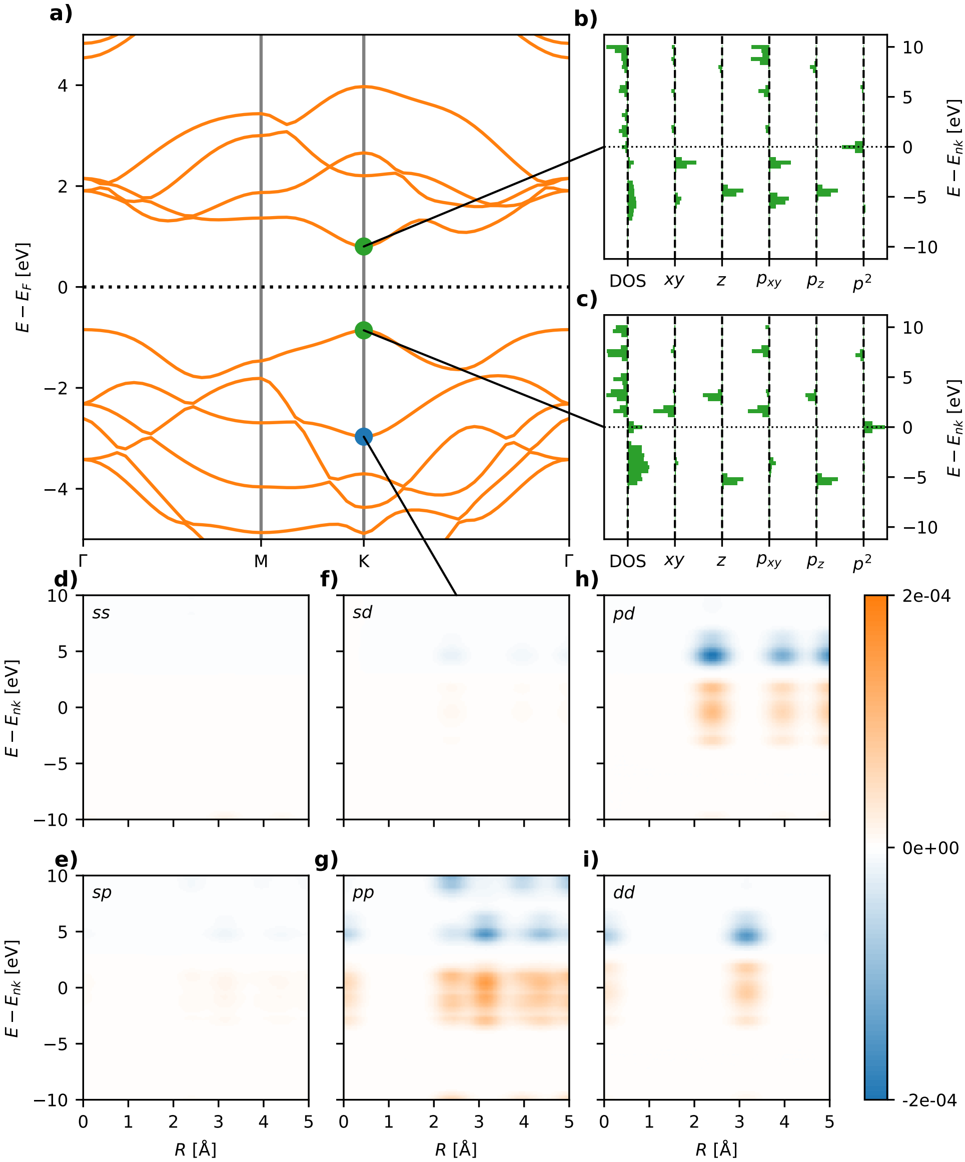


Figure 5.1: Visualization of ENDOME and RAD-PDOS state fingerprints for MoS_2. a) shows the PBE band structure. b) and c) show ENDOME fingerprints of the conduction band minimum and valence band maximum states for the K-point. d)–i) show six RAD-PDOS fingerprints for combinations of s, p, and d orbitals. Reprinted from Publication [I].

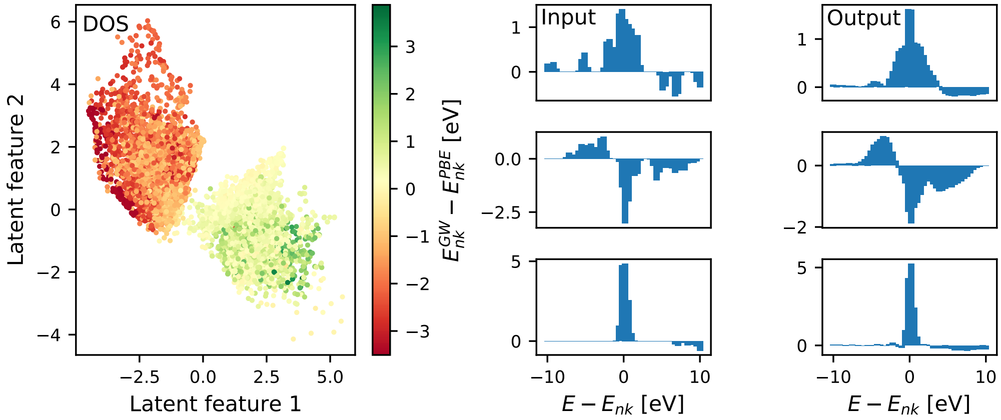


Figure 5.2: Example of using an autoencoder on the ENDOME DOS fingerprint. The left panel shows the latent space observations colorcoded by the G_0W_0 correction energy. The center and right panel show three input samples and the corresponding reconstructed output samples, respectively.

in the RAD-PDOS fingerprint d-components being all zero.

The XGBoost model is fitted to the G_0W_0 correction energies based on the electronic fingerprints. Figure 5.5 shows the predicted vs. true values of the train and test set in panel a), the histograms of the prediction residuals in b) and a learning curve of the ML model performance with respect to the amount of training data in c). Using the full training set, the model yields a MAE of 0.11 eV on the test set.

The ML model can be used to construct full band structures by predicting the G_0W_0

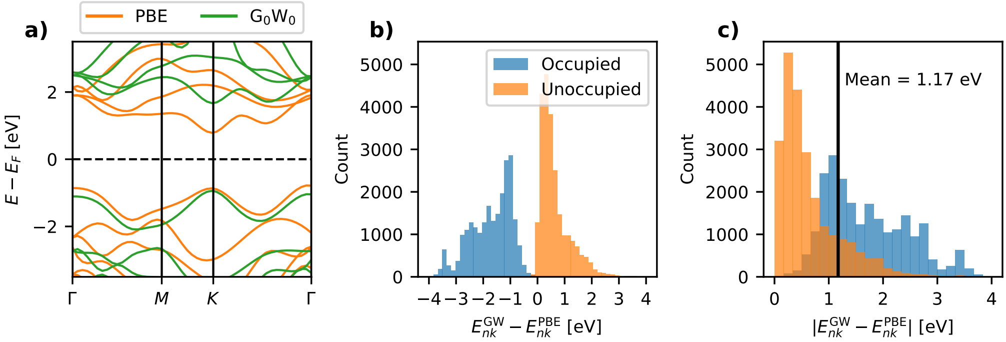


Figure 5.3: a) shows examples of PBE and G_0W_0 of MoS_2. b) shows the distributions of the G_0W_0 correction energies for occupied and unoccupied states. c) shows histograms of the absolute values of the G0W0 corrections with a mean of 1.17 eV. Reprinted from Publication [I].

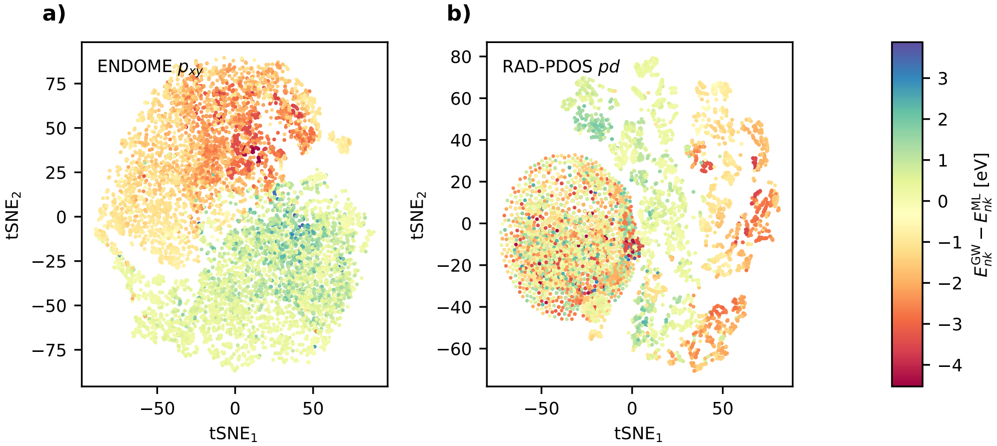


Figure 5.4: tSNE visualizations of ENDOME p_{xy} in a) and RAD-PDOS pd in b) colorcoded with the G_0W_0 correction energies. The large blob of data in b) correspond to the materials without d-electrons where the RAD-PDOS pd fingerprint is all zero. Reprinted from Publication [I].

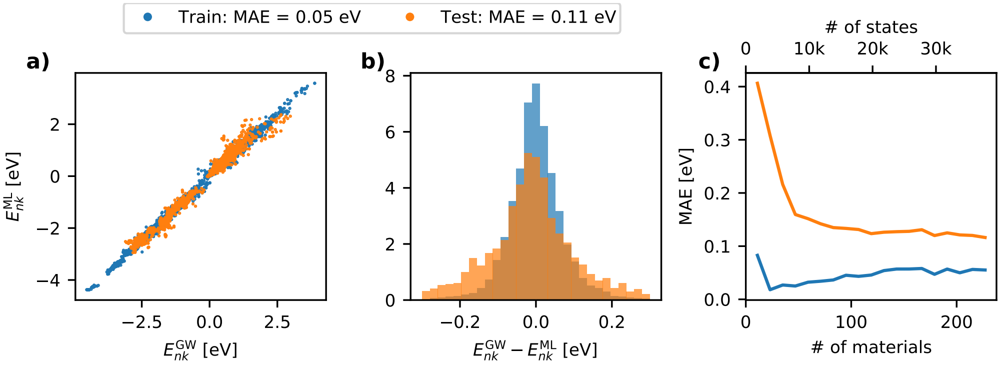


Figure 5.5: Results of the machine learning model predicting G_0W_0 correction energies. a) shows the predicted vs. true values with a test MAE of 0.11 eV. b) shows the prediction residuals. c) shows the learning curve with respect to number of materials/state observations in the train set. Reprinted from Publication [I].

energies along the chosen band path. Figure 5.6 shows examples of the PBE, G_0W_0 and ML band structures for four materials from the test set. The examples show great agreement between the ML model and the target G_0W_0 band structures in most cases, but as seen in d) there can be deviations though the ML model still offers a better estimate than the corresponding PBE band structures.

Using the ML model, the band gaps of the materials can be directly derived from the predictions. Figure 5.7 shows the predicted band gaps vs. true G_0W_0 band gaps for the general ML model, a ML model trained only on valence and conduction bands (VB+CB) and the DFT band gaps using functionals PBE and HSE06. The general ML model yields a test set MAE of 0.18 eV for the prediction of the G_0W_0 band gaps, while the ML model trained only on VB+CB has a MAE of 0.15 eV. This is to be compared to the PBE and HSE06 predictions with MAEs of 1.70 and 0.85 eV, respectively.

With the ability of the ML model to predict the G_0W_0 energies for any k-point grid, the model can be used to estimate effective masses, which is typically challenging with G_0W_0 due to the high computational costs of calculating the energies at a densely sampled k-point grid around the band extrema. The valence and conduction band effective masses have been calculated using the ML model for ≈ 330 materials yielding significant deviations from the corresponding masses calculated using PBE. The mean absolute deviations are $0.31m_0$ and $0.19m_0$ for the VBM and CBM, respectively. There is also a tendency of the ML model to give effective masses with a smaller absolute value than PBE.

In order to assess the importance of the different components of the fingerprints a feature importance analysis is performed using a feature subset hold-out method where the MAE is evaluated when the model is trained with/without the subset. The fingerprint is divided into four major feature groups, i.e. the static polarisabilities, the electronic properties (PBE band gap, occupancy and distance to the Fermi level), the ENDOME components and the RAD-PDOS components. Figure 5.8 shows the results of the feature analysis. In a), the major feature groups are evaluated and in b) these groups are further broken down, i.e. the ENDOME is split to the individual operators and the RAD-PDOS is split to the individual components. The analysis concludes that there is a lot of redundant information in the fingerprint since any feature group can be dropped from the fingerprint without causing a large increase in the MAE. On the other hand, there is also some synergy in the full fingerprint since no individual group of features results in as low an MAE as for the full fingerprint. Looking at the low level feature groups in b) it is seen that dropping the in-plane static polarisability α_{xy} causes the highest increase in MAE suggesting that this feature brings information that is difficult to extract from the rest of the features.

As an additional analysis of the α_{xy} feature, a SHAP analysis is carried out. The SHAP value for a certain feature corresponds to the effect on the prediction value when including this feature as compared to a model without the feature. Figure 5.9 shows the SHAP values of α_{xy} for the prediction of individual state energies in a) and band gaps in b). For materials with a low polarisability, the ML model predicts a lower G_0W_0 correction energy for the occupied states and a higher energy for the empty states. For the high polarisability materials, the model predicts a more positive correction energy for the occupied states while predictions of the empty states are only weakly affected.

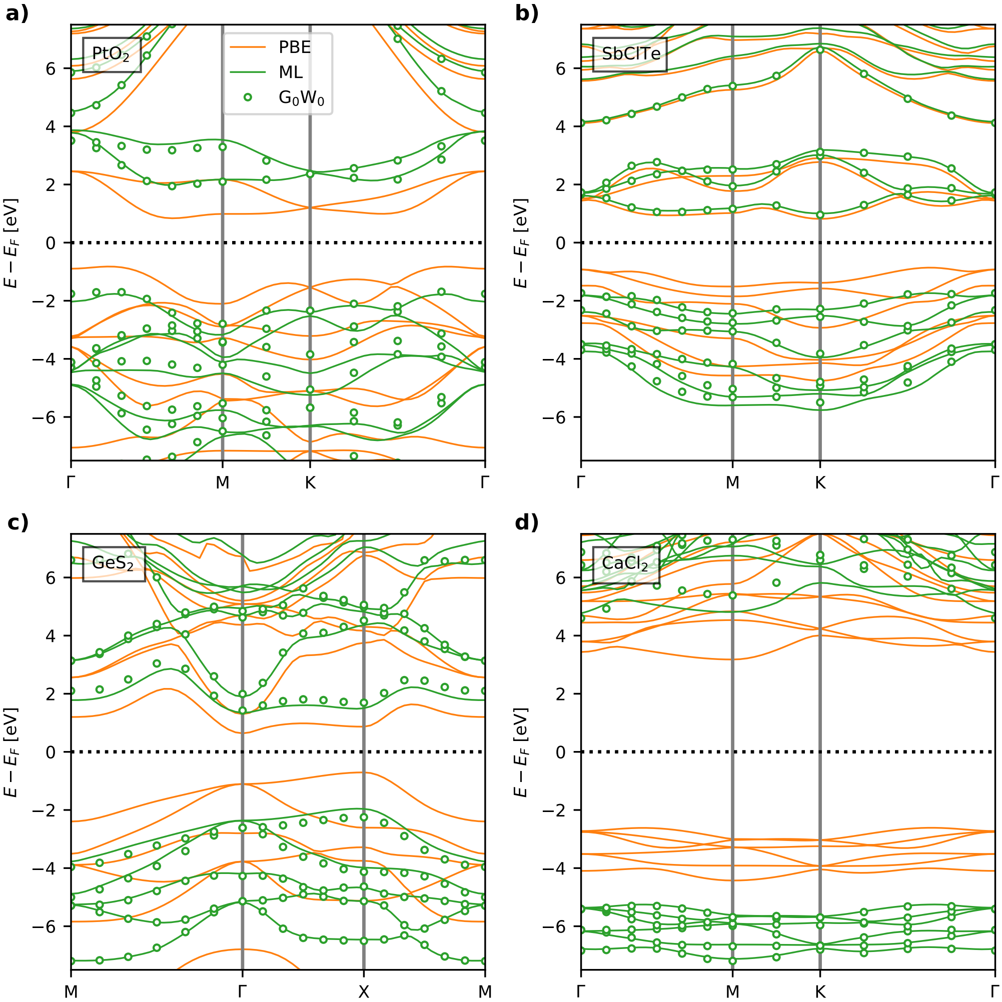


Figure 5.6: Examples of band structures for materials from the test set for a) PtO_2, b) SbClTe, c) GeS_2 and d) $CaCl_2$. The plots show both the PBE, G_0W_0 and ML band structures. Reprinted from Publication [I].

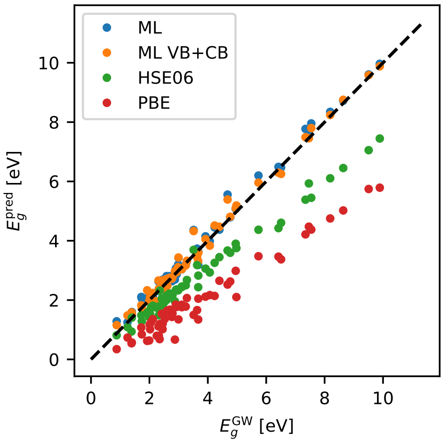


Figure 5.7: Parity plot of predicted band gaps vs. G_0W_0 band gaps. The plot includes PBE and HSE06 band gaps with MAEs of 1.70 eV and 0.85 eV, respectively. The two ML models refer to a model trained on all bands and a model trained on only valence and conduction bands (VB+CB) with MAEs of 0.18 eV and 0.15 eV, respectively. Reprinted from Publication [I].

5.5 Summary of Publication 2: Predicting and machine learning structural instabilities in 2D materials

This publication seeks to address the problem of predicting the dynamical stability of 2D materials without calculating the full phonon band structures [69]. The full paper is appended to the thesis from page 83. Two different approaches to this problem are made; one is the Center and Boundary Phonon (CBP) protocol where phonons are calculated at the center and boundary of the Brillouin zone from which the stability can be well estimated. The CBP protocol then proceeds by displacing the atoms along the potentially unstable phonon mode which results in 49 of 137 investigated dynamically unstable materials being stable after the displacement and relaxation. More information on the CBP protocol is found in the paper.

The other approach to assess dynamical stability is by using machine learning to build a classification model, which is the contribution of this Ph.D. project to the publication. Two models with different sets of fingerprints are trained using a dataset of > 3000 2D materials and their dynamical stability label (stable/unstable) as target variable. Both models are trained using the gradient boosting method XGBoost, which is explained in Section 2.1.2.3. The first model uses the relatively basic electronic properties consisting of the energy above the convex hull, PBE band gap, DOS at the Fermi level, total energy per atom and heat of formation as descriptors in the model. The second model uses the RAD-PDOS fingerprint as input. Figure 5.10 shows the distribution of the features in the first fingerprint for both dynamically stable and unstable materials. None of the features are clearly separating the two classes, but

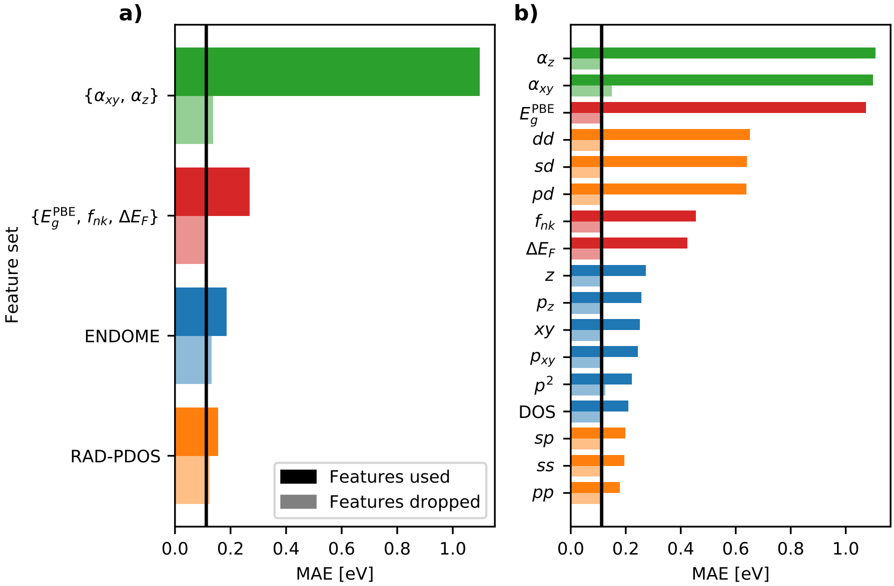


Figure 5.8: Feature importance analysis of the ML model predicting G_0W_0 correction energies. The solid bars refer to models trained with only the given subset of features while the shaded bars refer to models trained without the subset. a) shows high-level feature groups and b) shows low-level feature groups. The green subset holds the in-plane and out-of-plane static polarisabilities, while the red holds the PBE band gap, state occupation and distance to the Fermi level. Reprinted from Publication [I].

both the convex hull energy, the band gap and the DOS at the Fermi level show some correlation with dynamical stability.

The two models are evaluated using the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) as performance metric [43]. Figure 5.11 shows the results, with ROC curves of the two models shown in a) and b), and a feature importance analysis of the RAD-PDOS model in c). Both models have good classification performances with the model using the simple fingerprint yielding an AUC of 0.82 ± 0.01 while the model using RAD-PDOS as fingerprint gives an AUC of 0.90 ± 0.01. Based on a feature importance analysis of the model trained with the RAD-PDOS fingerprint it is found that the most important components of the fingerprint are those corresponding to couplings between two s electrons.

Due to the strong performance of the ML model, it is possible to utilize it as a screening step in a high-throughput workflow where dismissing unstable materials is of importance. The applied ML method opens up for some freedom in choosing a threshold value for dismissing materials depending on how many false classifications one can afford. The curvature of the ROC curve can be used to investigate this. For the ML model trained on the RAD-PDOS fingerprint a true unstable prediction rate of $85 \pm 3\%$ can be achieved if a false unstable prediction rate of 20% can be accepted.

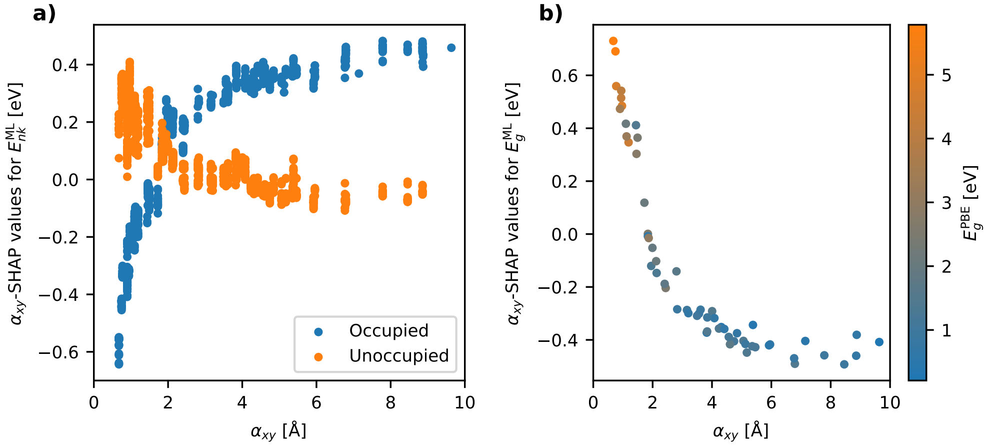


Figure 5.9: SHAP analysis for the in-plane static polarisability α_{xy}. a) shows SHAP values for α_{xy} for the prediction of G_0W_0 correction energies color-coded by occupancy. b) shows SHAP values for α_{xy} for the prediction of G_0W_0 band gaps. For materials with a low polarisability, the ML model predicts a more negative G_0W_0 correction for the occupied states and a more positive correction for the unoccupied states. For materials with a high polarisability, the occupied states are predicted with a more positive correction when using the polarisability as a feature while the unoccupied states are only weakly affected. Reprinted from Publication [I].

Similarly, true unstable rates of $70 \pm 6\%$ and $56 \pm 9\%$ for false prediction rates of 10% and 5%, respectively.

5.6 Summary of Publication 3: Recent progress of the computational 2D materials database (C2DB)

This publication describes the advancements of the Computational 2D-materials Database (C2DB) after its initial publication in 2018 [62]. The publication reports on: (1) General updates in the workflow used for selection, stability classification and assessment of new materials, (2) Developments of the properties already introduced in the first version of the database, (3) New properties, (4) New materials and (5) Examples of applying machine learning methods to predict properties using atomic structures as input.

A more comprehensive introduction to the C2DB workflow is found in chapter 4, while this summary focuses on the machine learning section of the paper, which is based entirely on outputs of this Ph.D. project. The section introduces machine learning as an efficient method to make computationally cheap predictions of material properties based on fingerprints encoding atomic and electronic structure information. A benchmark study is performed investigating how well different fingerprints can be used to

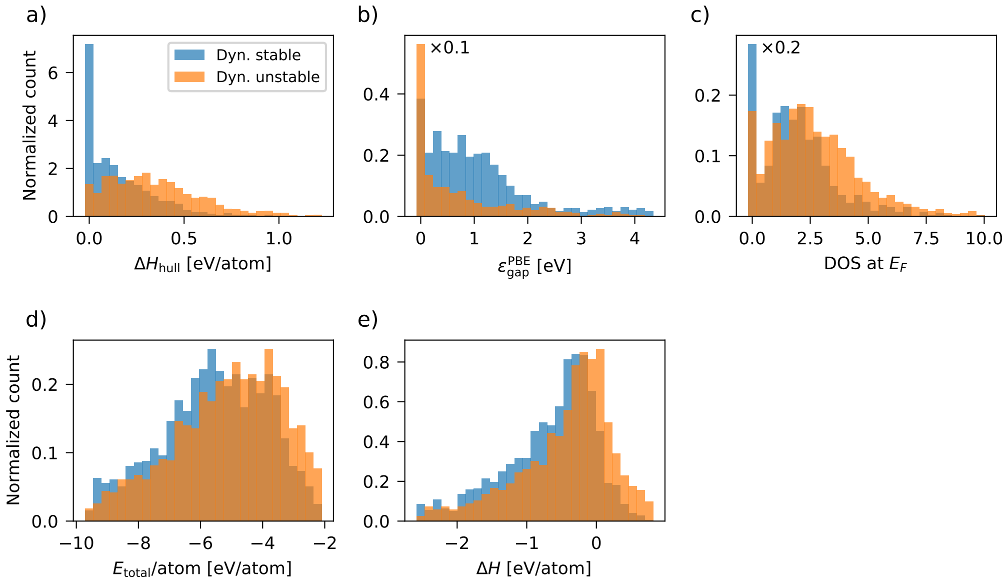


Figure 5.10: Histograms of electronic features for dynamically stable and unstable materials. The properties are a) energy above the convex hull, b) PBE band gap, c) density of states at the Fermi level, d) total energy per atom and e) heat of formation. The energy above convex hull, PBE band gap and DOS at the Fermi level show some correlation with stability. Reprinted from Publication [II].

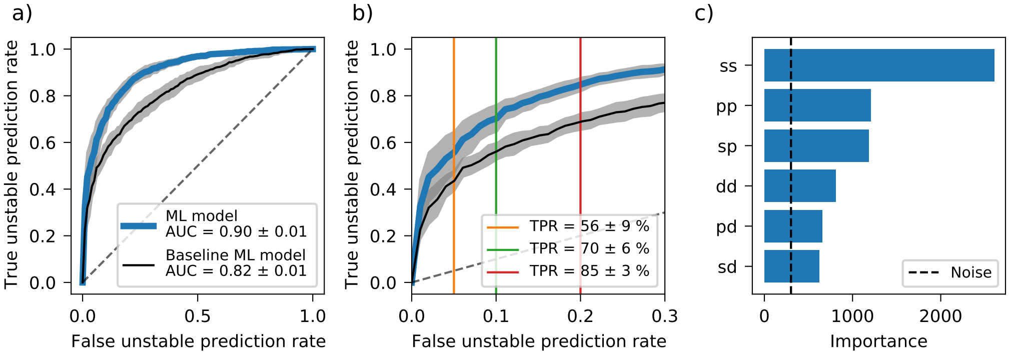


Figure 5.11: Machine learning results for the classification of dynamical unstable materials. a) show the ROC curve for the model trained with the RAD-PDOS fingerprint (AUC $= 0.90 \pm 0.01$) and the baseline model trained on simple electronic features (AUC $= 0.82 \pm 0.01$). b) shows a zoomed version of the ROC curve. c) shows the feature importance of RAD-PDOS components as evaluated by the model. Reprinted from Publication [II].

predict various properties. The properties include the HSE06 band gap, PBE heat of formation, BSE exciton binding energy, in-plane static polarisability calculated with RPA, and Voigt modulus derived from the elastic stiffness tensor. These properties are predicted using three different fingerprints; two structural fingerprints in terms of the Ewald sum matrix and many-body tensor representation (MBTR), and additionally the global RAD-PDOS fingerprint encoding projected density of states is tested. All the properties are predicted using a Gaussian process regression model, which is chosen due to its ability to handle relatively small number of observations.

The conclusion of the ML work in this publication is that across the five properties there is a clear hierarchy between the fingerprints. Fig 5.12 shows the MAEs relative to the sample standard deviations for all fingerprints and properties. The Ewald sum matrix performs the worst while the RAD-PDOS fingerprint outscores the other fingerprints for all the investigated properties.

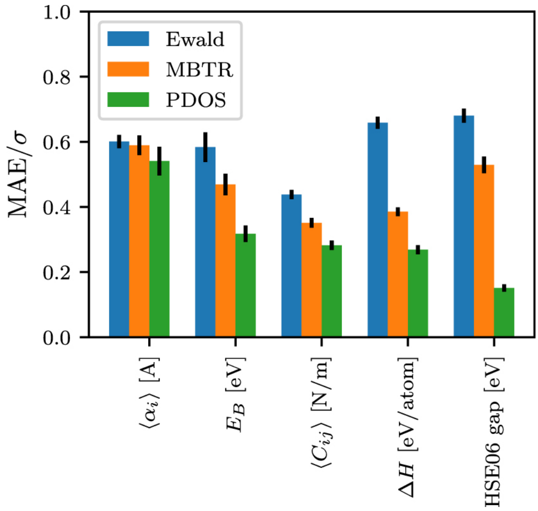


Figure 5.12: Test MAEs normalized to the standard deviation of the property for ML models using Ewald sum matrix, MBTR and RAD-PDOS fingerprints. The properties on the x-axis are the in-plane static polarisability, exciton binding energy, Voigt modulus, heat of formation and HSE06 band gap. Reprinted from Publication [III].

5.7 Excursion: G_0W_0 energies of molecules

So far, the concept of electronic fingerprints has only been showcased in a context of periodic systems, i.e. the 2D materials from C2DB. As an excursion from 2D materials to the world of molecules, this section presents a minor study of predicting G_0W_0 energies of molecules.

Using the OE62 database, which is a spectroscopy benchmark dataset containing 61,489 crystal forming molecules [70]. All molecules have orbital energies reported at PBE level of DFT. Additionally, for 5239 molecules the G_0W_0 quasiparticle energies are calculated. This 5k subset is used for machine learning the G_0W_0 correction to the PBE orbital energies. Besides the energetic properties, the dataset also contains Hirshfeld partial charges for each atom in the molecules [71].

The 5k molecules gives ≈ 300,000 observations of G_0W_0 correction energies. Figure 5.13 visualizes the data. Panels a) and b) show the distribution of PBE and G_0W_0 energies for the occupied and unoccupied states, respectively. Panels c) and d) show the G_0W_0 energies vs. the PBE energies. For all the occupied states, the G_0W_0 energies are shifted down compared to the PBE energy, while the opposite happens for the unoccupied states. Panels e) and f) show the distribution of the G_0W_0 correction energies, which is the target variable of the study.

Unlike the study in Publication [I], it was out of scope to acquire matrix elements or projected density of states for the molecules in this study. Therefore, the fingerprint is based primarily on the PBE energies. The first part of the fingerprint is an ENDOME-like DOS fingerprint, which encodes the state n with PBE energy ε_n as:

$$f_n(E) = \sum_m G(E - (\varepsilon_m - \varepsilon_n); \delta_E)\,\mathrm{sign}(E_{\mathrm{HOMO}} - \varepsilon_m) \tag{5.8}$$

where E_{HOMO} is the energy of the highest occupied molecular orbital. Figure 5.14 shows examples of such DOS fingerprints. Panels a) and b) show the PBE energies for two molecules with three states per molecule highlighted. Panel c) show the DOS fingerprints of these 2×3 states.

The second part of the fingerprint utilizes the Hirshfeld charges available in the dataset. A simple analysis of the distribution of charges for each chemical element is shown in Figure 5.15. It is seen that some chemical elements have a quite distinct distribution. By combining the charges and atomic distances in a fingerprint, the structural information of the molecule is encoded in an enriched way. This fingerprint component is defined as:

$$f(Q, R) = \sum_{ij} G(Q - q_i q_j; \delta_Q) G(R - |\boldsymbol{R}_i - \boldsymbol{R}_j|; \delta_R) \exp(-\alpha_R R) \tag{5.9}$$

where q_i and $\boldsymbol{R}_i$ are the charge and position of atom i. This fingerprint is thus mapped in the QR space of charge products and distances.

A gradient boosting tree ensemble model using XGBoost is trained to predict the G_0W_0 correction energies for the individual states. By evaluating on a test set with 20% of the molecules, the machine learning model yields a test MAE of 0.15 eV. The model is then used to calculate G_0W_0 HOMO-LUMO gaps for all molecules. Figure 5.16 shows the distribution of G_0W_0 gaps in panel a). Panel b) shows the predicted vs. true gaps while panel c) shows the residuals.

This study shows that the electronic fingerprints based on individual state energies is a powerful tool for bridging the gap between a cheap computational method (PBE) and an expensive method (G_0W_0). This is the case for predicting individual state energies but also material properties such as energy gaps.

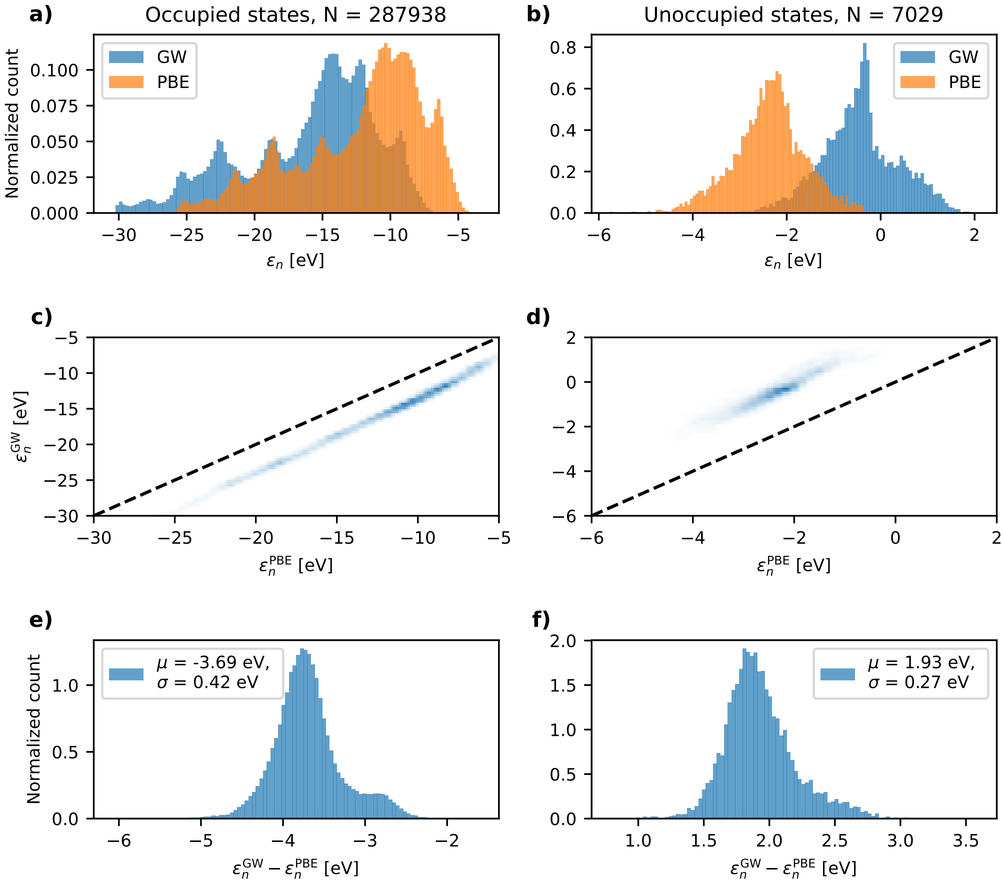


Figure 5.13: Visualization of the 5k molecules with G_0W_0 energies. a) and b) show histograms of PBE and G_0W_0 eigenenergies for occupied and unoccupied states, respectively. c) and d) show G_0W_0 vs PBE energies. e) and f) show the distribution of G_0W_0 correction energies.

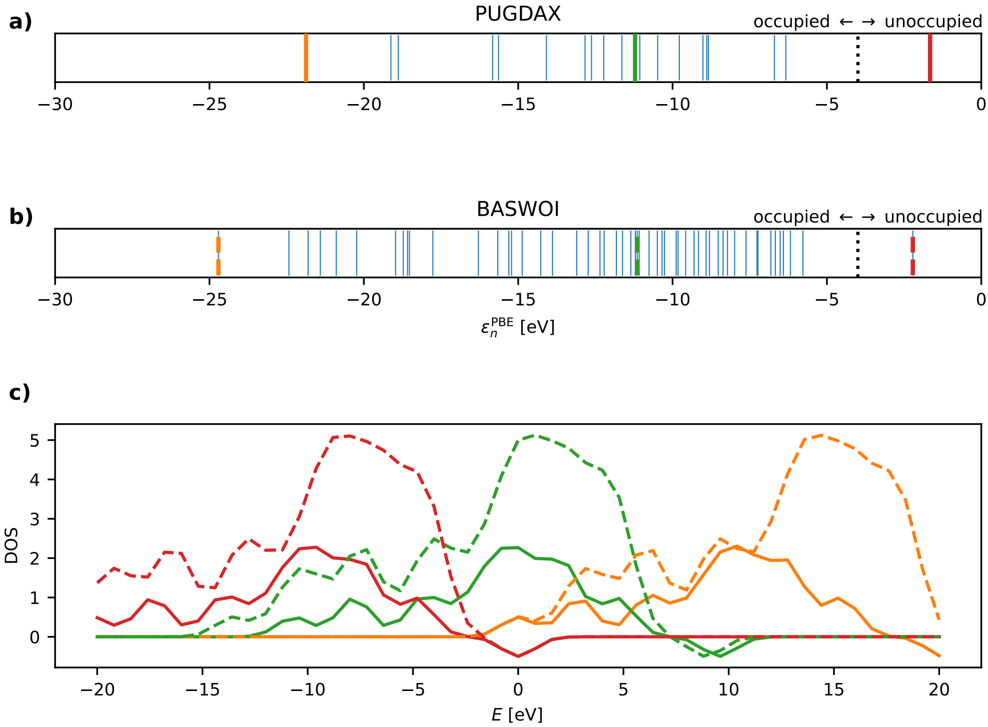


Figure 5.14: Examples of DOS fingerprints for two molecules with reference codes PUGDAX and BASWOI. a) and b) show the PBE energies of the two molecules with 3 states highlighted for each molecule. c) shows the corresponding DOS fingerprints of the six states.

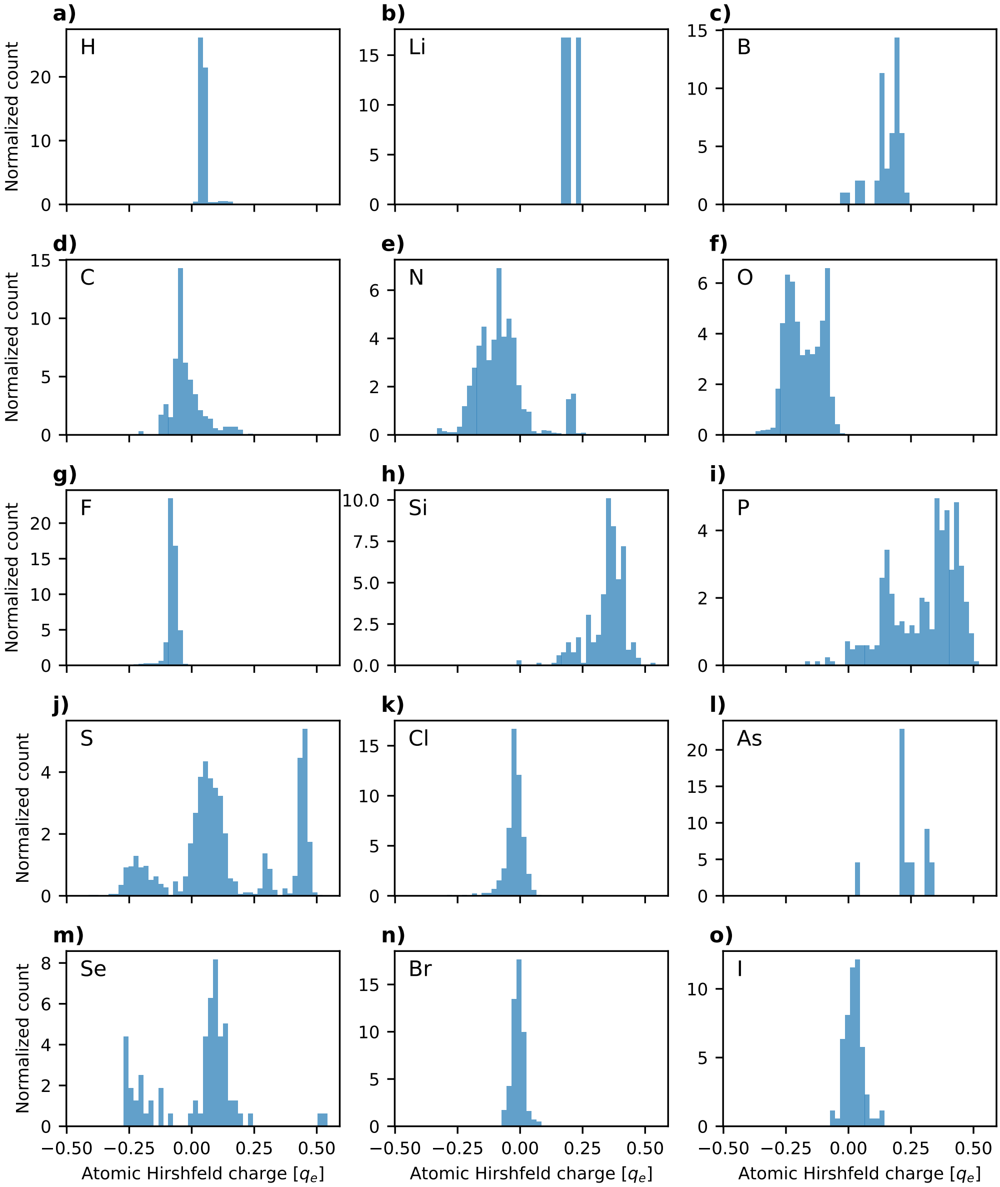


Figure 5.15: Histograms of Hirshfeld charges for the chemical elements present in the molecules in the OE62 database. Some of the elements have somewhat distinct charge distributions.

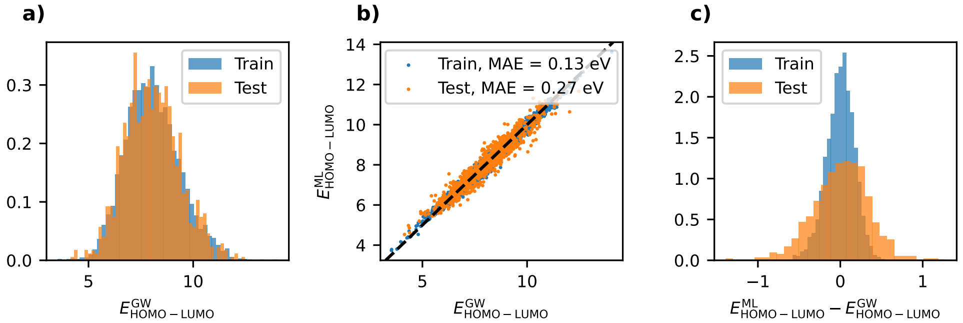


Figure 5.16: Results for the ML prediction of G_0W_0 band gaps for the 5k molecules. a) shows the distributions of G_0W_0 band gaps for the train and test set. b) shows the predicted vs. true G_0W_0 band gaps and c) shows the prediction residuals.

CHAPTER 6

Electron-phonon coupling and dynamical stability

The work presented in this chapter is a continuation of the work outlined in Publication [II] where the dynamical stability is predicted with a machine learning model trained using the RAD-PDOS fingerprint.

The dynamical stability is derived from a phonon calculation, but here the focus will be on the corresponding coupling between phonons and electrons with the electron-phonon matrix elements between states i and j:

$$g_{ij}^{u} \sim \langle i| \nabla_u V_{\text{eff}}(\boldsymbol{r}) |j\rangle \tag{6.1}$$

where $V_{\text{eff}}(\boldsymbol{r})$ is the effective potential and the gradient ∇_u is with respect to atomic displacements u.

In practice, the gradient of the effective potential is calculated using a finite difference method by displacing all atoms in the unit cell in both directions of all three Cartesian axes, i.e. for a system with N atoms in the unit cell, $6N$ DFT calculations are needed to calculate the matrix elements [72]. The work presented in this chapter is based on approximating the electron-phonon matrix elements to avoid performing the $6N$ DFT calculations. To calculate the full phonon band structure for arbitrary $\boldsymbol{q}$-vectors for periodic systems, the finite difference calculations are performed in a supercell, i.e. the primitive cell is repeated. This makes the computations of phonon band structures computationally expensive.

6.1 Atomic potentials

The first approximative step in this approach is to replace the true effective potential $V_{\text{eff}}(\boldsymbol{r})$ of a system with a potential setup by atomic contributions:

$$V_{\text{eff}}(\boldsymbol{r}) \approx \sum_{a} v_{\text{eff}}^{a}(\boldsymbol{r} - \boldsymbol{r}_a) = V_{\text{eff}}^{a}(\boldsymbol{r}) \tag{6.2}$$

where the atomic potentials $v_{\text{eff}}^{a}(\boldsymbol{r})$ are DFT effective potentials of the isolated atoms calculated once for all atoms. The sum runs over all atoms in the system, though for periodic systems it is limited to only the primitive cell and up to second nearest neighboring cells. This assumes that potentials from atoms further away are effectively zero inside the primitive cell.

In the following sections, two different sample systems are used for demonstrating the concepts. One is the periodic 2D material GaF_2 from C2DB, which is dynamically unstable based on a phonon calculation for a 2x2 supercell. The other system is an ethylene molecule (C_2H_4). Figure 6.1 and 6.2 shows examples of the potential setup from atomic contributions for the periodic system GaF_2 and C_2H_4, respectively. The top left panels show contour plots of the DFT effective potential in the xy-plane with the z-coordinated fixed at the center of the unit cell. The top center panels show the corresponding atomic potential and the top right panels show the difference between the two potentials $V^a_{\text{eff}}(\boldsymbol{r}) - V_{\text{eff}}(\boldsymbol{r})$. The bottom panels show the same information in the x-direction where both y- and z-coordinates are fixed at the center of the unit cell. It is seen that for both the periodic system and the molecule, the atomic potentials are qualitatively similar to the true potentials, though the differences can be quantitatively large. For GaF_2 in Figure 6.1, the difference is almost constant at least for the chosen hyperplane. This means that the gradients of the two potentials will be similar, which motivates the use of these atomic potentials as an approximation when calculating electron-phonon matrix elements.

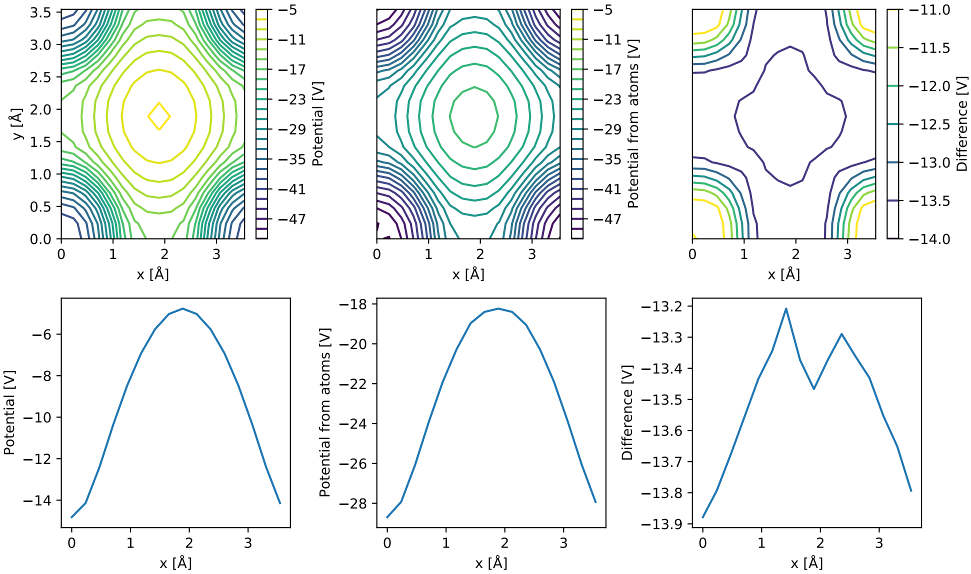


Figure 6.1: Plots showing potentials for GaF_2. The left plots show the DFT effective potential, the center plots show the atomic potential and the right plots show the difference between the two. The difference is nearly constant for this system.

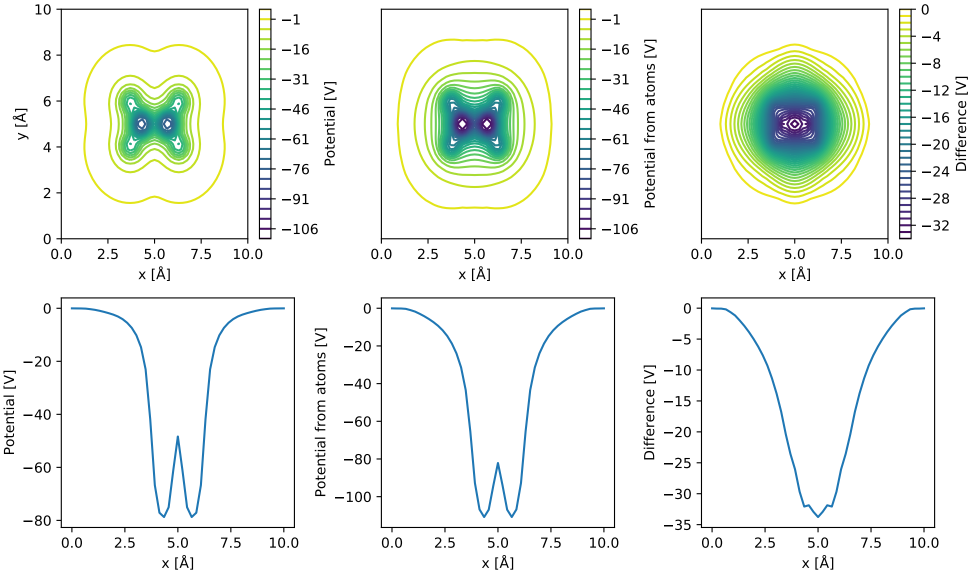


Figure 6.2: Plots showing potentials for C_2H_4. The left plots show the DFT effective potential, the center plots show the atomic potential and the right plots show the difference between the two. The DFT and atomic potentials are qualitatively similar, but the difference is quantitatively large especially close to the center of the molecule.

6.2 Electron-phonon matrix elements from atomic potentials

Approximated electron-phonon matrix elements can be calculated using the method to set up the potential from the atomic potentials. This is done using a finite difference method where each atom is displaced 0.01 Å along all directions. The difference is that instead of performing a DFT calculation to extract the effective potential for each displacement of atoms, the atomic potentials are used to calculate the atomic gradients of the effective potential, which is significantly faster.

In order to validate the method, the electron-phonon matrix elements are calculated using both the true effective potential and the summed atomic potentials. This is done for both GaF_2 and C_2H_4. The two systems are expanded by a factor of d, i.e. the cell and unit cell positions are multiplied by d, such that $d = 1$ corresponds to the equilibrium structure. This is done to evaluate the matrix elements in the limit where the systems are approaching isolated atoms. Figure 6.3 and 6.4 shows the matrix elements between the six first eigenstates of GaF_2 and C_2H_4, respectively, versus the expansion factor d. The dashed lines corresponds to the matrix elements calculated using the atomic potentials while the solid lines are matrix elements for the true potential. The matrix elements calculated using the atomic potentials are qualitatively similar to the true matrix elements. There is a tendency that the approximated matrix elements are quantitatively larger, but as the expansion factor is increased the two methods are approaching each other as expected.

6.3 Dynamical matrix from perturbation theory

The original thought was to use the approximated electron-phonon matrix elements to calculate the dynamical matrix using perturbation theory (see Section 3.3.1 for details). Dynamical stability is derived from the eigenvalues of the dynamical matrix. Negative eigenvalues of the dynamical matrix indicate that the total energy of the system can be reduced by displacing atoms along a certain mode. When calculating the dynamical matrix using perturbation theory, the change in the total energy is approximated as the change in eigenenergies of the occupied states of the system.

This approach of calculating the eigenvalues of the dynamical matrix using perturbation theory has been tested for 10 different materials from C2DB including both stable and unstable materials. Unfortunately, it was not possible to determine dynamical stability with this method. This is concluded since no correlation between the eigenvalues and the dynamical stability was found, i.e. the eigenvalues from the perturbation theory are wrong.

As a further investigation of why the approach of using perturbation theory to calculate changes in total energies fails, a small experiment is carried out. By calculating the total energy, the sum of eigenvalues of occupied states and their respective finite difference derivatives for the molecule C_2H_4 while varying the expansion factor d from 0.8 to 1.5, the perturbation theory approach is tested. This investigates the underlying assumption of using the changes in eigenvalues as a measure of the change in the total energy is examined.

Figure 6.5 shows the total energy and derivative in the top panels, and the sum of eigenvalues and derivative in the bottom panels. For the total energy, there is a clear minimum at $d = 1$ as expected, meaning that the derivative is zero at this point. The sum of eigenvalues of the occupied states is found to be monotonically increasing, i.e. the changes in eigenvalues cannot be used to measure the change in the total energy. For periodic 2D materials, the same conclusion is made based on similar experiments of a few materials.

6.4 Machine learning approaches to electron-phonon coupling

At this point, the conclusion is that the electron-phonon matrix elements can be approximated reasonably well by replacing the effective potential with a potential set up by atomic potentials. On the other hand, the matrix elements cannot be used to calculate the dynamical matrix and thereby the dynamical stability using perturbation theory. The next step is to use machine learning to utilize the information from the approximated electron-phonon matrix elements. This is done in two different steps;

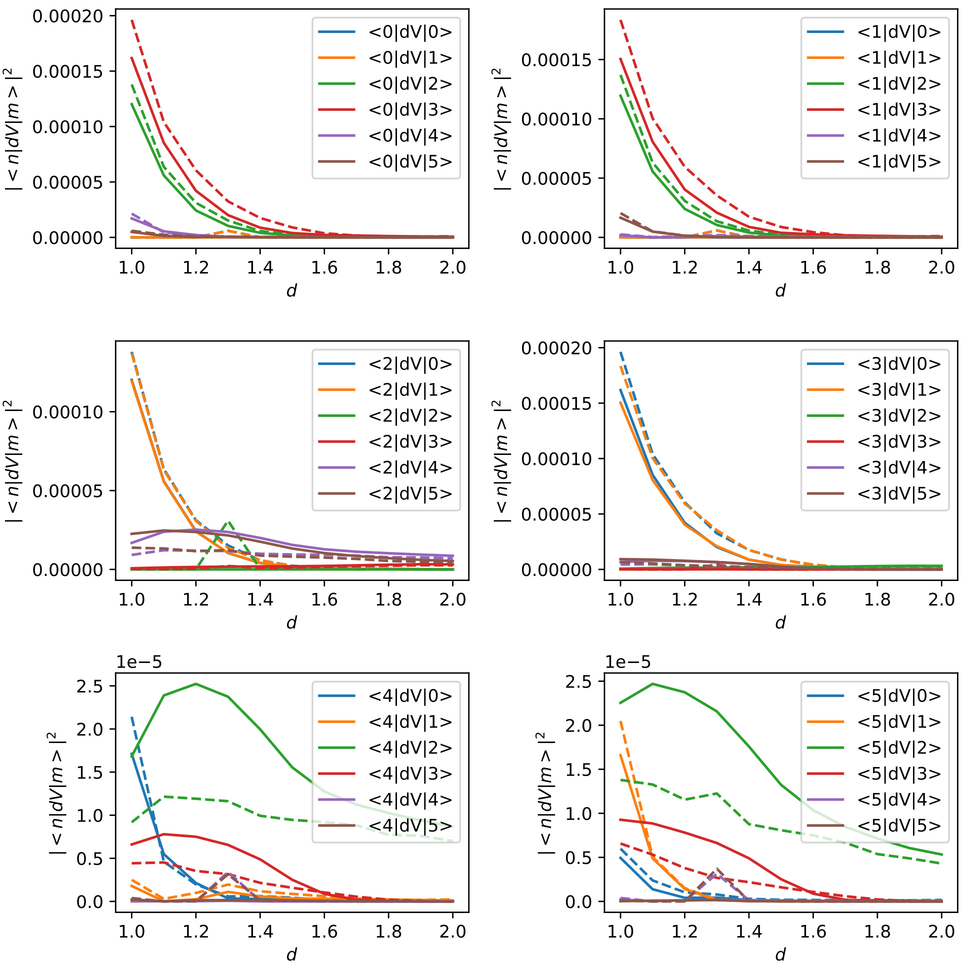


Figure 6.3: Electron-phonon matrix elements between the six lowest states of GaF_2 versus the expansion factor d. The solid lines refer to matrix elements calculated using the DFT potential while the dashed lines refer to matrix elements calculated using the atomic potentials. The two methods are qualitatively similar and approaching each other as the expansion factor is increased.

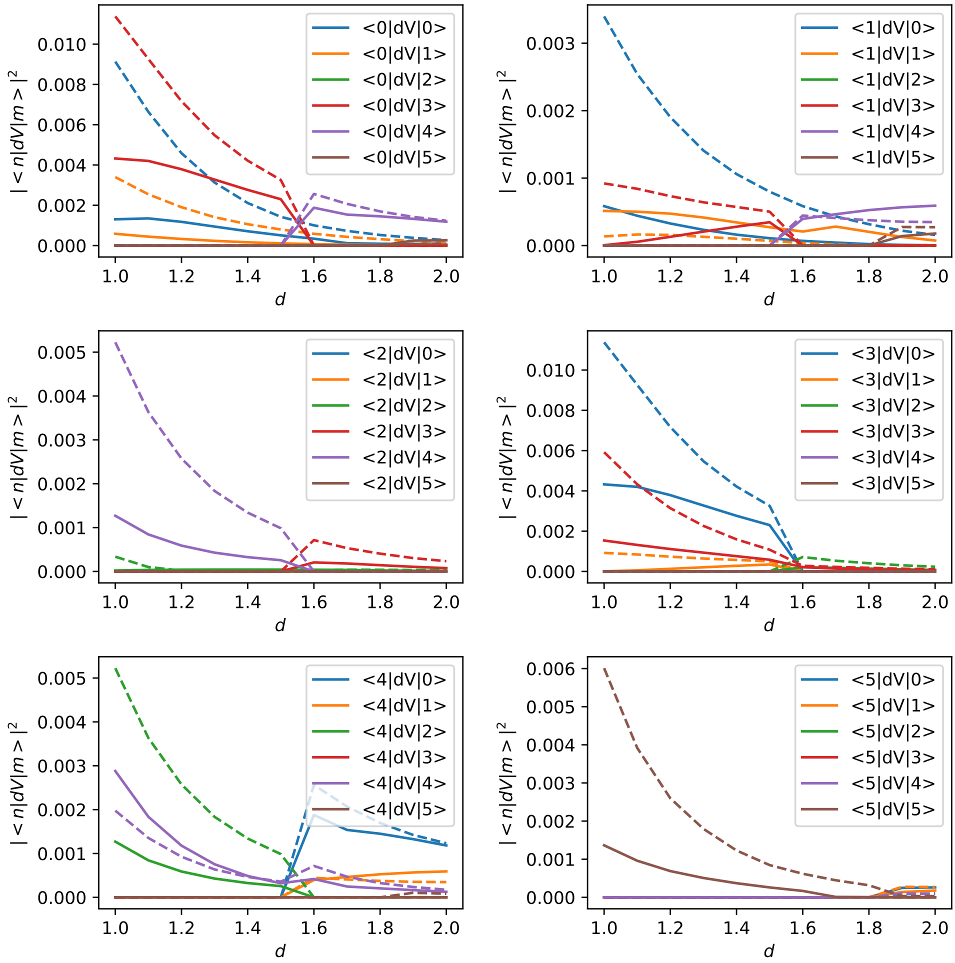


Figure 6.4: Electron-phonon matrix elements between the six lowest states of C_2H_4 versus the expansion factor d. The solid lines refer to matrix elements calculated using the DFT potential while the dashed lines refer to matrix elements calculated using the atomic potentials. The two methods are qualitatively similar and approaching each other as the expansion factor is increased.

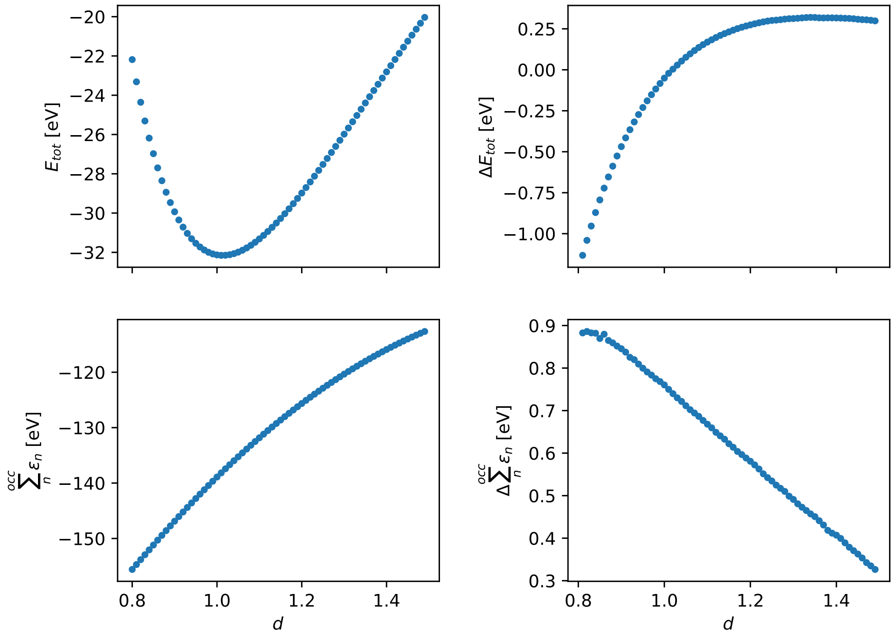


Figure 6.5: Analysis of total energy and sum of eigenenergies of occupied states as function of the expansion factor d. The two left panels show the total energy of C_2H_4 and the sum of eigenenergies while the right panels show the corresponding finite difference derivatives. The total energy is found to have a minimum at $d = 1$, but the sum of eigenenergies are monotonically increasing. This highlights that the sum of eigenenergies is not a good approximation to the total energy.

the first is to improve the approximated matrix elements by learning a mapping between the DFT effective potential and the atomic potential using a local fingerprint in real space. In this way, the potential is set up from the atomic potentials and then corrected by the machine learning model before calculating the electron-phonon matrix elements. The second approach is to use the matrix elements to generate an ENDOME fingerprint and then learn the dynamical stability using a classification model.

6.4.1 Learning the DFT effective potential

Since a DFT for the primitive cell is required to calculate the electron-phonon matrix elements, the true DFT effective potential is already available. The difference between the DFT effective potential and the potential set up by atomic potentials for the primitive cell is used as the target variable for a machine learning model. In this way, a relatively cheap mapping between the atomic potentials and the DFT effective potential is achieved. This mapping is then used to correct the potential and thereby improve the approximated electron-phonon matrix elements. A local many-body tensor representation (LMBTR) fingerprint is used to encode a local point in space for which the potential correction is predicted [73]. LMBTR is a modification of the regular MBTR where the $k = 1$ term is removed and the descriptor encodes 2 or 3-body interactions (distances, angles) between a point in space and the atoms in the system. For this purpose, only the 2-body distance interaction is used for the fingerprint. The distances is encoded using a grid of 25 points between 0 and 5 Å with Gaussian widths of 0.1 Å. This means that the number of features in the fingerprint is 25 times the number of atomic species in the system.

A feed-forward neural net with 2 hidden layers of 100 neurons each and ReLU activation functions is used as the machine learning algorithm. The fundamental requirement of the ML model is that it has to be continuous with respect to the input since the potential is a continuous function, which rules out e.g. decision tree based models. Also, the model should be able to handle large number of observations, since even for a single material the number of grid points can be large (> 20.000).

This approach of learning the mapping between the potentials is demonstrated for GaF_2. The data is split into a 50/50% train and test set by splitting the primitive cell along the x-axis. Figure 6.6 shows the distribution of the target variable, i.e. the difference between the potentials in the top left panel, while the top right shows the MAE of the train and test sets vs. the number of epochs. In the bottom left panel, the predicted values are plotted against the true values with a MAE of 0.004 Ha (0.11 eV). The bottom right shows the residuals, i.e. the difference between true and predicted values.

Another way to visualise the effect of the model is by plotting contours of the potentials. Figure 6.7 shows the DFT effective potential (top left), the atomic potential (top right), the atomic potential corrected by ML (bottom left) and the residuals between the DFT and the ML potentials (bottom right). The atomic potential is shifted down compared to the DFT potential, but the ML model corrects for this.

Since this method is in fact used to calculate the gradient of the potential wrt. atomic displacements using a finite difference method, it is interesting to see if the ML model is sensitive enough to catch the small changes in the potential when making a small displacement of an atom. Figure 6.8 shows the gradient when moving the atom located at $(x, y) = (0, 0)$. The left plot shows the gradient of the DFT effective potential, the center plot shows the gradient of the atomic potential and the right plot shows that of the potential corrected with the ML model. Even though the gradient of the atomic potential in this case is quite close to that of the DFT potential, the ML model is improving the gradient slightly.

To showcase the method further, the electron-phonon matrix elements are calculated using the gradients of the DFT potential, the atomic potentials and the potential corrected by the neural network for GaF_2. Figure 6.9 shows the approximated matrix elements vs. the true matrix elements (calculated with the DFT potential). Without the correction, the approximated matrix elements are typically larger than the true matrix elements. With the atomic potential corrected by the neural network, the

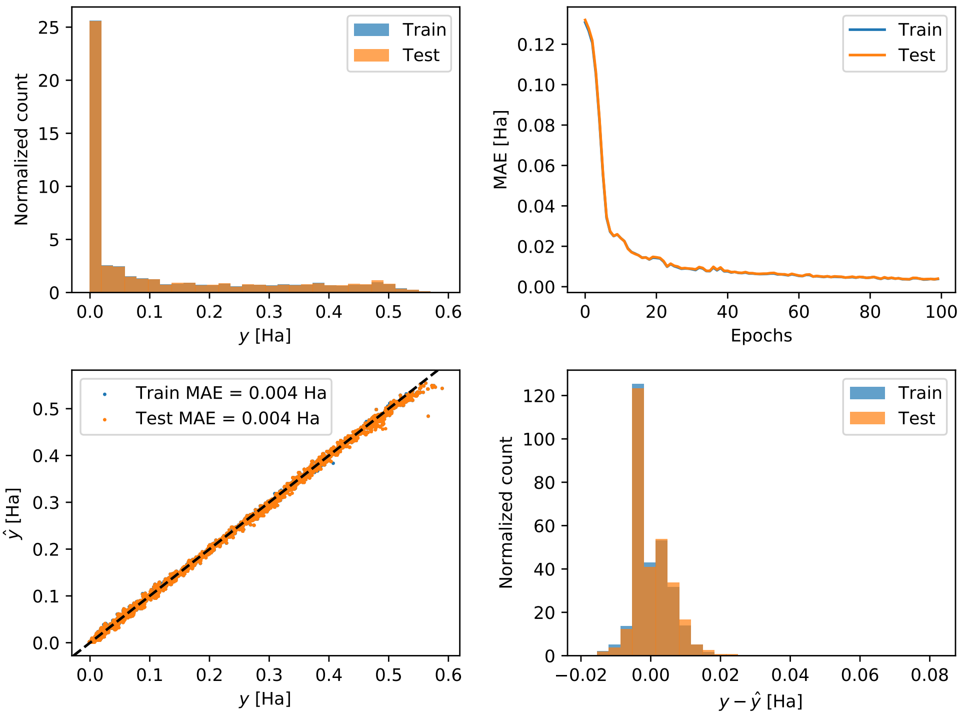


Figure 6.6: Results for the prediction of difference between the DFT potential and the atomic potential for GaF_2. The top left plot shows the distribution of the train and test set. The top right plot shows the learning curve when fitting the neural network, and the bottom left plot shows the predicted vs. true values for the final model. The bottom right plot shows the distribution of prediction residuals with a MAE of 0.004 Ha or 0.11 eV.

matrix elements are closer to the true matrix elements, and the RMSE is decreased from 37.0 $\frac{\mathrm{eV}^2}{\text{Å}^2}$ to 19.3 $\frac{\mathrm{eV}^2}{\text{Å}^2}$.

6.4.2 Preliminary results for electron-phonon ENDOME model

With the approximative electron-phonon matrix elements it is possible to construct an ENDOME fingerprint encoding a material as:

$$m_q(E_i, E_f) = \sum_{nmku} |\langle mk+q|\, \nabla_u V \,|nk\rangle|^2 G\left(E_i - (\varepsilon_{nk} - E_F); \delta_E\right) \\ \times G\left(E_f - (\varepsilon_{mk+q} - E_F); \delta_E\right) \tag{6.3}$$

Figure 6.10 shows examples of such fingerprints for GaF_2 at two different q-points. The electron-phonon ENDOME fingerprint along with the RAD-PDOS fingerprints are used to train a XGBoost classification model predicting dynamical stability as in Publication [II].

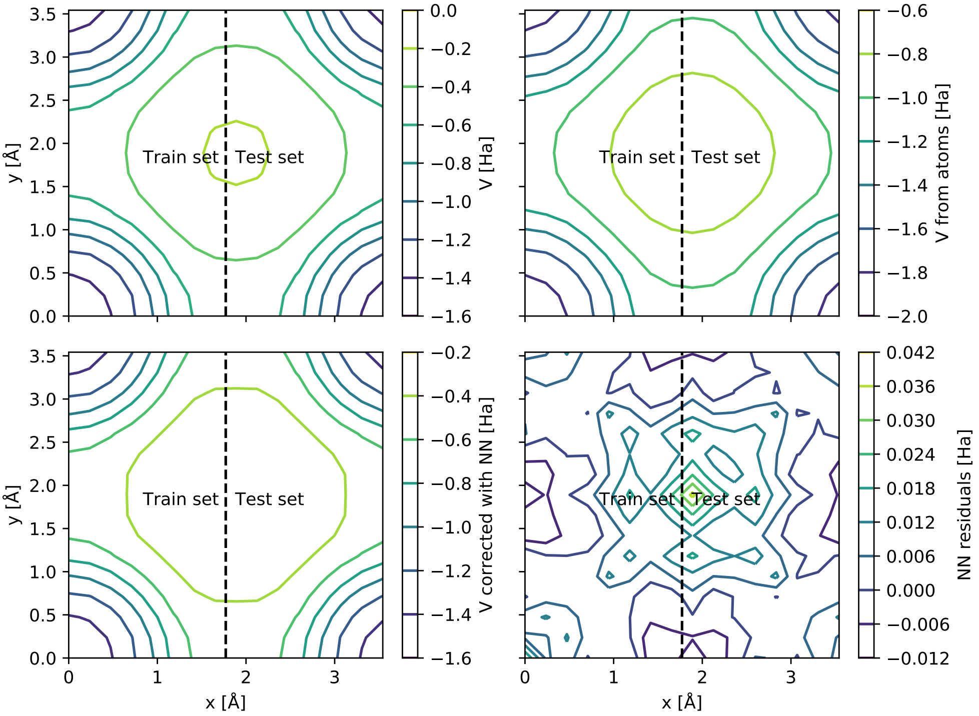


Figure 6.7: Contour plots of the potentials for GaF_2. The top left plot shows the DFT potential. The top right plot shows the atomic potential, which is shifted down in energy compared to the DFT potential. The bottom left plot shows the atomic potential corrected by the ML model and the bottom left shows the residual of the ML model.

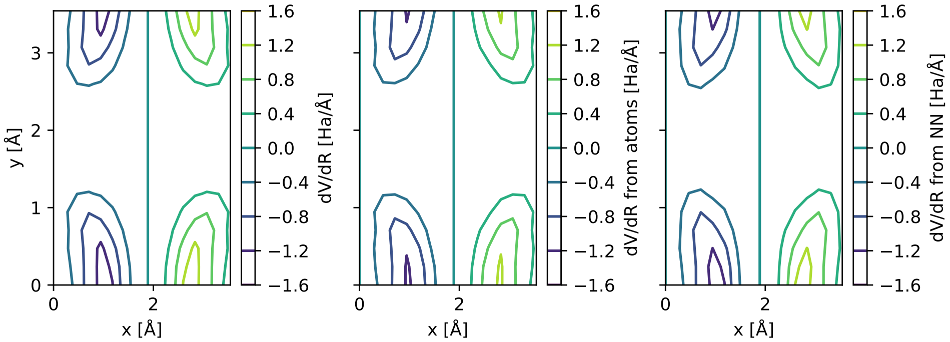


Figure 6.8: Plots showing the gradients of the DFT potential (left), atomic potential (center) and atomic potential corrected by the ML model (right) when moving an atom located at $(x, y) = (0, 0)$. Even though the gradient of atomic potential is already a good approximation in this case, the gradient is still improved slightly by the ML model.

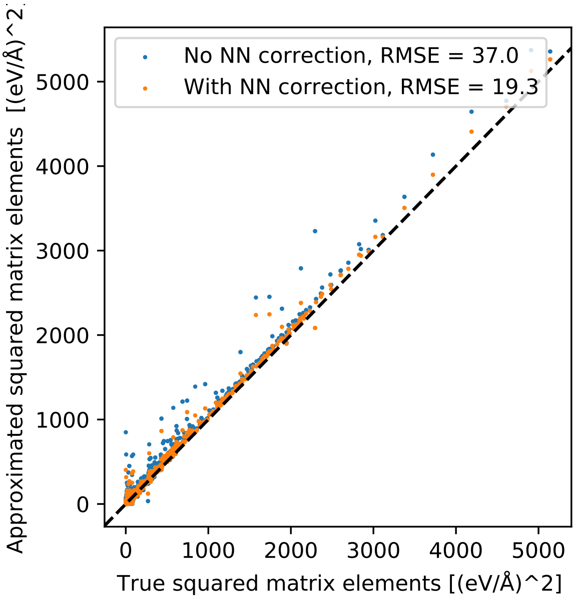


Figure 6.9: Parity plot showing the square of the approximated electron-phonon matrix elements vs. the true values calculated with DFT effective potential. Blue corresponds to matrix elements calculated with the atomic potentials while orange corresponds to the case where the atomic potentials are corrected using the trained neural network before calculating matrix elements. The RMSE is reduced from 37.0 to 19.3 $(\text{eV}/\text{Å})^2$ by using the neural net.

In the following, the preliminary results of using this fingerprint to predict dynamical stability are presented. Unfortunately, the project was not brought to conclusion before handing in this thesis, so a few things should be noted: First, the dataset for this model only contains 2352 materials from C2DB, which is >1000 less than in Publication [II]. This means that a direct comparison of the results is difficult. Secondly, the ENDOME fingerprint is actually based on the approximative electron-phonon matrix elements calculated without the correction of the potential using the neural net.

Figure 6.11 shows the preliminary results. The left panel shows the test ROC-AUC scores for three models based on 5-fold cross-validation. One model trained only with the RAD-PDOS fingerprint, one with only the electron-phonon ENDOME fingerprint and one trained with both fingerprints. The right panel shows the feature importance as evaluated by the model trained with both fingerprints. From the feature importance analysis, it is seen that the model actually uses a mixture of both fingerprints and not just the RAD-PDOS. The conclusion here is that the model trained with both RAD-PDOS and electron-phonon ENDOME fingerprints has a performance equal to the one trained only with RAD-PDOS within the uncertainties (AUC= 0.87). This means that the model does not gain performance by using the electron-phonon ENDOME fingerprint, even though the electron-phonon ENDOME show some predictive power itself by yielding an AUC of 0.67. There is reason to believe that the results can be improved, since the electron-phonon ENDOME fingerprint does not use the approximative electron-phonon matrix elements calculated without the correction of the potential using the neural net. Therefore, the hope is to gain some synergy between the electron-phonon ENDOME fingerprint and the RAD-PDOS fingerprint resulting in even better classifications of dynamical stability.

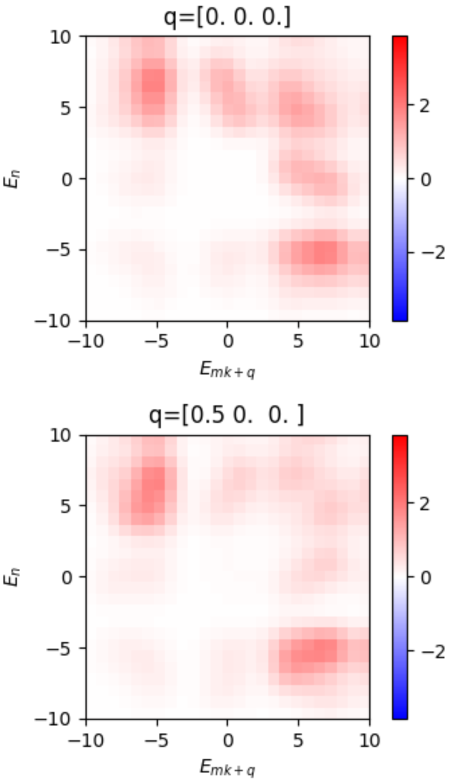


Figure 6.10: Examples of ENDOME fingerprints using the approximated electron-phonon matrix elements for GaF_2 for $q = [0, 0, 0]$ and $q = [0.5, 0, 0]$.

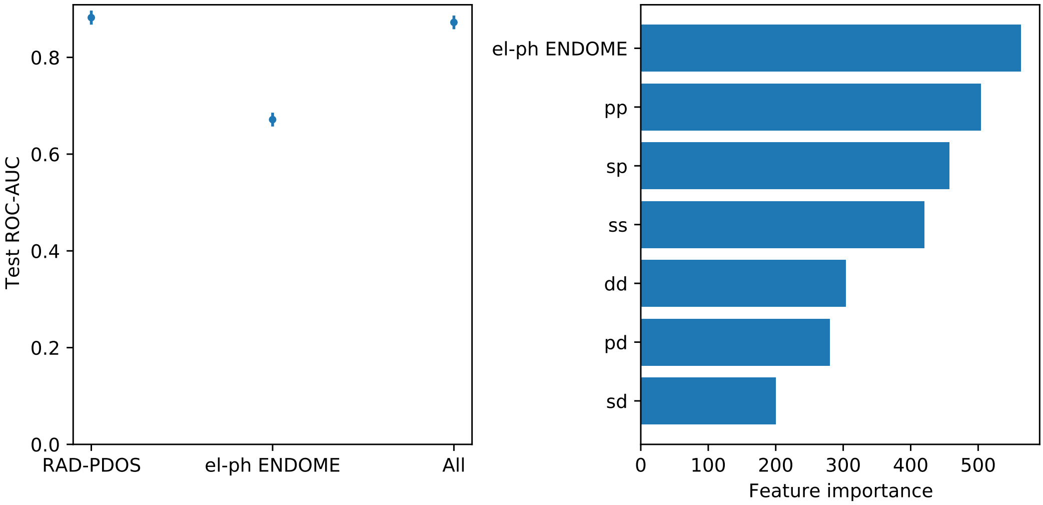


Figure 6.11: Preliminary results for the prediction of dynamical stability using the electron-phonon ENDOME fingerprint. The left panel shows the test ROC-AUC scores for models trained with the RAD-PDOS fingerprint, the electron-phonon ENDOME fingerprint and both fingerprints. The right panel show the feature importance of the model trained with both fingerprints.

CHAPTER 7

Conclusion

The application of machine learning methods in materials science is a fascinating research field. It combines theories of physics and quantum mechanics with computational concepts and mathematical algorithms from the field of machine learning. This Ph.D. project aimed to contribute to the active research by developing new methods useful for the computational materials science community.

After laying the theoretical foundation by reviewing the most fundamental concepts of machine learning and electronic structure methods applied in the project, the Computational 2D Materials Database (C2DB) was introduced as the primary dataset of the project. An initial machine learning application of C2DB was presented by combining unsupervised and supervised machine learning models. Based on a structural fingerprint, the materials were clustered using a k-means clustering method. Since the optimal number of clusters is ill-defined, the range from 1 to 10 clusters were examined. In the 10-cluster case, the clusters were found to differ in the distributions of selected material properties. The unsupervised approach was combined with supervised regression of the heat of formation and electronic band gap. By fitting Random Forest models for each cluster in the range from 1 to 10, it was found that increasing the number of clusters decreased the weighted mean absolute error of the clusters.

Next, the main contribution of this Ph.D. project was presented in terms of new fingerprint methods. The novelty of the fingerprint methods is that they encode an individual quantum state, and that they use information extracted from the electronic density and wavefunctions obtained by a DFT calculation. The energy decomposed operator matrix elements (ENDOME) fingerprint was constructed using matrix elements of quantum mechanical operators. Additionally, the radially decomposed projected density of states (RAD-PDOS) fingerprint was developed using projections of DFT wavefunctions onto atoms and angular orbitals. The ENDOME and RAD-PDOS fingerprints were applied in a machine learning model predicting the G_0W_0 correction energies to PBE eigenenergies for individual states using data from C2DB. The model was able to predict the correction energies with a MAE of 0.11 eV and G_0W_0 band gaps with a MAE of 0.15 eV. An analysis of the importance of different features in the fingerprint was performed. This showed that the fingerprint contains some redundant information but also that synergy effects between different features occurred.

The RAD-PDOS fingerprint was also used for machine learning the dynamical stability of 2D materials. The dynamical stability was based on phonon calculations, and a classification model was trained yielding a ROC-AUC score of 0.90. Additionally, the RAD-PDOS fingerprint was benchmarked against structural fingerprints across multiple properties from C2DB. This showed that the electronic fingerprint outperforms general structural fingerprints when used for machine learning different material properties.

Finally, the challenge of evaluating dynamical stability was revisited. This was done by developing approximative methods for calculating the electron-phonon coupling matrix elements. The key element in the approximation was replacing the DFT effective potential by a potential set up from atomic potentials. Calculating electron-phonon matrix elements for these approximated potentials yielded quantitatively good results. By training a neural network model to the differences between the DFT effective potential and the atomic potentials, the approximative method was further improved. When correcting the atomic potentials using the model before calculating the matrix elements, the error was reduced by a factor of ≈ 2.

7.1 Outlook

When doing research as a relatively inexperienced researcher, it is a challenge to end a Ph.D. project like this with a completely satisfied feeling, since there is always more that can be done. One key learning has been that it takes time to do things thoroughly, but it pays off in the end.

The work with the development of electronic fingerprints of individual states was left in a good state with Publication [I]. Naturally, the method could be further developed e.g. by testing other operators in ENDOME or constructing other fingerprints. From a personal point of view, it would be very interesting to test the method on other classes of materials than 2D materials, e.g. their bulk counterparts or other 3D crystals. This would also open for an exciting study of transfer learning, where transferring machine learning models trained on one class of materials would be applied on other classes. Another interesting application of transfer learning is transferring models between different properties, i.e. retraining a model pre-trained on a different property. This could be useful when training models on computationally expensive data such as *GW* band gaps, since a pre-trained model trained on cheaper data might be a good starting point, e.g. a model trained on cheaper band gaps such as PBE or even completely different properties like heat of formation.

Another logical follow-up study of the electronic fingerprints is a more thorough benchmark study. Publication [III] presented a small benchmark study comparing the RAD-PDOS to two structural fingerprints, but it could be interesting to test even more fingerprints. Also, the aspect of the choice of machine learning algorithm would be exciting to investigate. In other words, this would be a three-dimensional benchmark study comparing the use of various fingerprints in different machine learning algorithms across multiple material properties.

Finally, the electron-phonon project is left in an unfinished state. While the current method is capable of approximating the electron-phonon coupling matrix elements with reasonable accuracy, the actual connection to dynamical stability is still to be investigsated further. From a machine learning perspective, the electron-phonon ENDOME fingerprint should be improved in order to find the synergy effects with the RAD-PDOS fingerprint resulting in an improved result compared to Publication [II]. This also includes utilizing that in C2DB the dynamical matrix is calculated for individual q-points. This means that a classification model predicting the stability for a given q-point is achievable. In that way, the amount of data points can be increased which hopefully improves the model.

CHAPTER 8
Publications

8.1 Publication 1: Representing individual electronic states for machine learning GW band structures of 2D materials

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-28122-0 OPEN

Representing individual electronic states for machine learning GW band structures of 2D materials

Nikolaj Rørbæk Knøsgaard[1✉] & Kristian Sommer Thygesen[1]

Choosing optimal representation methods of atomic and electronic structures is essential when machine learning properties of materials. We address the problem of representing quantum states of electrons in a solid for the purpose of machine leaning state-specific electronic properties. Specifically, we construct a fingerprint based on energy decomposed operator matrix elements (ENDOME) and radially decomposed projected density of states (RAD-PDOS), which are both obtainable from a standard density functional theory (DFT) calculation. Using such fingerprints we train a gradient boosting model on a set of 46k G_0W_0 quasiparticle energies. The resulting model predicts the self-energy correction of states in materials not seen by the model with a mean absolute error of 0.14 eV. By including the material's calculated dielectric constant in the fingerprint the error can be further reduced by 30%, which we find is due to an enhanced ability to learn the correlation/screening part of the self-energy. Our work paves the way for accurate estimates of quasiparticle band structures at the cost of a standard DFT calculation.

[1] Computational Atomic-scale Materials Design (CAMD), Department of Physics, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark. ✉email: nirkn@dtu.dk

The electronic band structure is one of the most fundamental and important characteristics of a crystalline solid. It relates the quantum mechanical energy levels of an electron in the solid to its (crystal) momentum and provides the basis for describing and understanding a range of materials properties. As a consequence, the accurate prediction of electronic band structures represents a cornerstone problem of computational condensed matter physics.

Density functional theory (DFT)[1] with semi-local exchange-correlation functionals[2] is the standard method for solving the electronic structure problem of materials from first principles. However, the DFT single-particle energies do not in general provide an accurate model for the electronic band structure[3]. Instead, the gold standard for band structure calculations is represented by the GW self-energy method[4], which provides the true quasiparticle (QP) band structure, i.e., it goes beyond a mean-field description by explicitly accounting for exchange and many-body screening effects[5,6]. In ref. [7] the mean absolute error on the calculated bandgap relative to experimental references for a set of ten simple semiconductors and insulators was found to be 2.05 eV for DFT-LDA and 0.31 eV for non-self-consistent G_0W_0@LDA. Very similar results have been found in other studies[8,9]. The improved accuracy of the GW method comes at the price of a significantly more involved methodology and a much higher computational cost. In practice, this means that GW calculations are limited to small-scale studies of relatively simple materials.

Recently, machine learning (ML) has attracted widespread interest as a means to predict materials properties without performing expensive quantum mechanical calculations[10–15]. In the context of bandgap predictions, Zhou et al. trained a support vector machine on 3896 experimental bandgaps using a representation based only on elemental properties of the constituent atoms[16]. Rajan et al. used different regressions methods to predict bandgaps of MXene crystals using a training set of 76 G_0W_0 bandgaps and a representation encoding atomic and structural properties[17]. Liang et al. used a representation based on atomic ionicity descriptors to predict GW bandgaps of a set of 2D semiconductors[18]. In all these previous studies, the ML model was trained to predict the size of the bandgap rather than the full k-resolved band structure. Thereby, important information is missed including the type of the bandgap (direct or indirect), the curvature of the valence and conduction bands at the extrema points (effective masses), and the position and dispersion of other bands away from the bandgap. Predicting the full band structure directly from the atomic structure of the material is a daunting challenge that, although possible in principle, would require highly sophisticated ML models and immense amounts of training data.

Here we take a different approach, in which the output from a DFT calculation is taken as input to an ML model to predict the full GW band structure. The philosophy behind our approach is that standard DFT calculations are computationally very cheap, in particular, compared to GW, and although they do not directly produce the desired precision, they hold the gist of the material's genome and thus should provide an excellent starting point for accurate property predictions. In our scheme, the rich, but unmanageable, information contained in the DFT wave functions is encoded into low dimensional fingerprints via energy-resolved orbital projections and operator matrix elements. These state-specific electronic fingerprints provide a description of the local environment of a given electronic eigenstate in the infinite-dimensional Hilbert space and are thus analog to the well-known fingerprints used to describe atoms in chemical environments[19].

Using a data set of 286 G_0W_0 band structures of non-magnetic 2D semiconductors comprising a total of 46,000 $(\varepsilon_{nk}^{\mathrm{QP}}, k)$ pairs, we train a gradient boosting algorithm to predict the G_0W_0 correction of an eigenstate from its DFT fingerprint. The method achieves a mean absolute error (MAE) of 0.14 eV for individual band energies and 0.18 eV for the bandgap. These deviations are significantly smaller than the typical size of the G_0W_0 corrections and also lower than the accuracy of the G_0W_0 method itself. The model can be further and significantly improved by adding static electronic polarisability to the fingerprint. A SHAP feature analysis reveals that the inclusion of the polarisability allows the ML model to distinguish between materials with similar PBE band structures but different dielectric screening properties, which is directly related to the size of the GW correction.

We have used the resulting ML model to obtain G_0W_0 band structures for ~700 2D semiconductors from the Computational 2D Materials Database (C2DB)[20,21]. These materials are additional to the data set used in this study, and the band structures will be published on the C2DB web page[22].

Results

Figure 1a shows an example of a PBE (orange) and G_0W_0 (green) band structure for monolayer MoS_2 (note that spin–orbit interactions are not included throughout this work). It is clear that there are significant differences between the two descriptions. First of all, G_0W_0 yields a QP bandgap of 2.53 eV in good agreement with the experimental value of 2.5 eV[23] while PBE yields a significantly smaller bandgap of 1.58 eV. It can also be noted that unoccupied bands are shifted up in energy while occupied bands are shifted down. This is in fact a general trend across all the materials in the data set and it leads to a double peak in the histogram of G_0W_0 corrections with the peak of negative (positive) corrections corresponding to occupied

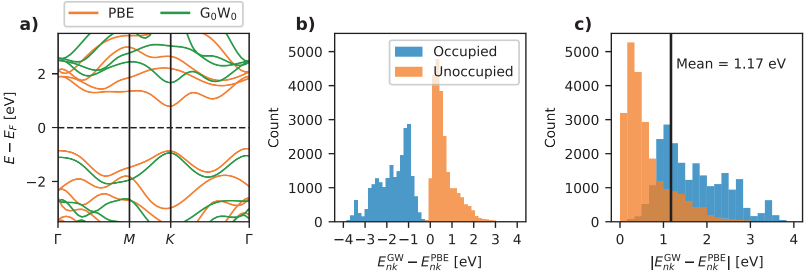


Fig. 1 G_0W_0 data. a Example of PBE and G_0W_0 band structures of monolayer MoS_2. The prediction target data is the difference in energy between the PBE and G_0W_0 energies. **b** Histogram of the G_0W_0 corrections for all states in all materials. **c** Histogram of the absolute values of the G_0W_0 corrections with a mean of 1.17 eV.

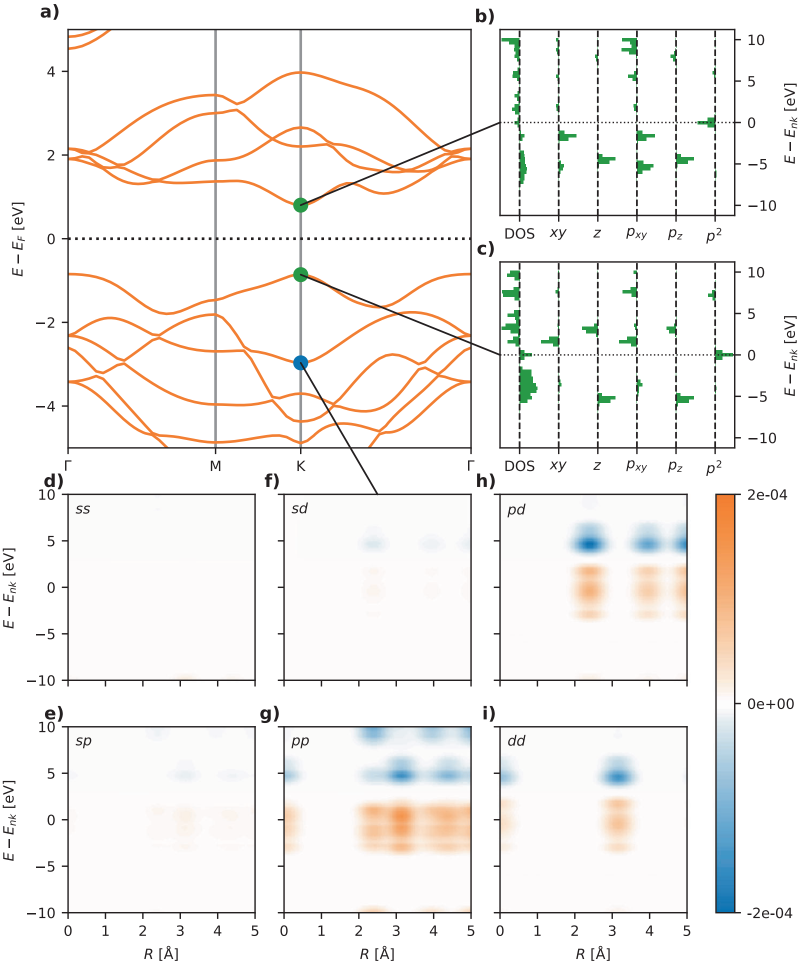


Fig. 2 Visualization of electronic state fingerprints for MoS_2. a Shows the PBE band structure. **b**, **c** Show ENDOME fingerprints of the conduction band minimum and valence band maximum states for the K-point. **d**–**i** Show six RAD-PDOS fingerprints for combinations of *s*, *p*, and *d* orbitals.

(empty) bands, see Fig. 1b. The absolute values of the G_0W_0 corrections range from 0 to 3 eV with an average value of 1.17 eV, see the histogram in Fig. 1c. Returning to the band diagram in panel (a) we further note that not all the bands are shifted by the same amount—even when disregarding the different signs for occupied/empty bands. Although for most materials, all the occupied bands experience similar, though material-specific, shifts and the same holds for the empty bands, there are several examples, like MoS_2, where this is not the case. Therefore, an accurate prediction of G_0W_0 corrections for general bands requires a representation that not only encodes the occupation of the state but also information about the energy and shape of the wave function and its relation to other relevant states of the crystal.

Electronic fingerprints. The ENDOME and RAD-PDOS representations, defined in the Methods Section, are attempts to generalize the notion of the local environment of an atom, which has been successfully employed to represent solids and molecules in machine learning studies, to the case of an electronic state. The ENDOME fingerprint represents the local environment of an energy eigenstate $|nk\rangle$ in terms of operator matrix elements between the state itself and other eigenstates of the crystal, $|\langle nk|\hat{A}|n'k'\rangle|^2$. These matrix elements are arranged on a grid as a function of the energy difference $\varepsilon_{nk} - \varepsilon_{n'k'}$, and their sign is used to encode the occupation of the final state $|n'k'\rangle$. With the ENDOME fingerprint, two states are thus considered similar if they have similar matrix elements with other states of similar relative energies. In this work, we include matrix elements for the position operator, momentum operator, and Laplacian operator. Since we exclusively consider 2D materials in the present work, the fingerprints are split into in-plane and out-of-plane components for the position operator (labeled *xy* and *z*, respectively) and the momentum operator (labeled p_{xy} and p_z). The RAD-PDOS fingerprint is a correlation function in energy and radial distance between the atomic orbital projections (onto angular momentum channels *s*, *p*, and *d*) of the reference eigenstate and all other eigenstates of the crystal. Figure 2 visualizes the two types of fingerprints for three different electronic states of MoS_2.

Any reasonable fingerprint should comply with certain general requirements[13] of which invariance and simplicity are the most

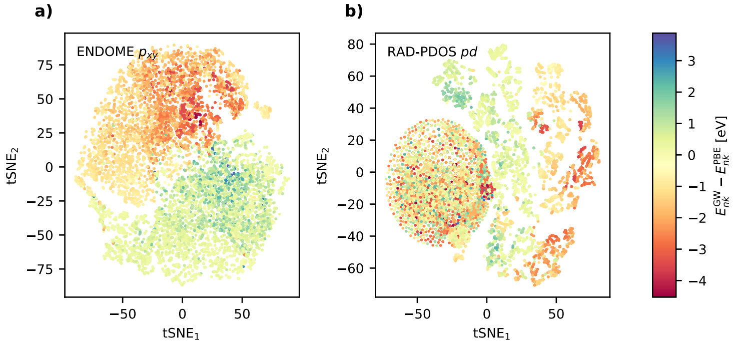


Fig. 3 tSNE visualizations of fingerprints. a tSNE components of ENDOME p_{xy}. **b** tSNE components of RAD-PDOS pd fingerprints color-coded with the GW corrections. For p_{xy}, states with similar GW corrections are also close in fingerprint space. In **b** a large amount of the states with both positive and negative GW corrections have similar distances in fingerprint space, corresponding to the materials without *d*-electrons where the RAD-PDOS *pd* fingerprint will be all zeros.

fundamental. In the present context, this means that the fingerprint should be invariant with respect to the choice of the unit cell (number of primitive cells, rotations, and translations), the gauge used for the Bloch wave functions and that it should be computationally cheap to generate compared to a full G_0W_0 calculation. Both the ENDOME and RAD-PDOS fingerprints clearly fulfill these requirements. Besides the invariance and simplicity conditions, the fingerprints should also be *unique* such that two different systems (here electronic states) are not mapped to the same fingerprint, and they should be *descriptive* such that systems with similar properties are close in fingerprint space. The interpretation and quantitative assessment of notions such as different systems and similar properties are obviously problem-dependent. This fact can make it difficult for problem independent fingerprints like the ENDOME and RAD-PDOS to meet these requirements in general. This is, however, not a principal problem, and can usually be solved by increasing the size of the training data set, at least as long as the fingerprints are complex and flexible enough to capture the variations in the considered systems that are relevant to the specific learning problem.

An impression of the descriptiveness of the fingerprints can be obtained from Fig. 3, which shows two-dimensional projections of the ENDOME-p_{xy} and RAD-PDOS-*dd* fingerprints using *t*-distributed stochastic neighbor embedding (tSNE) color-coded by the GW corrections. It is clear that data points, which are close in p_{xy}-space have similar GW corrections. The *pd* fingerprint is also descriptive for some data points, but there is also a large blob of data points that are indistinguishable in fingerprint space but have very different GW corrections. Not unexpectedly, these points correspond to the subset of materials without valence *d*-electrons, which results in all-zero *pd* fingerprint vectors. The tSNE plots for the other components of the ENDOME and RAD-PDOS fingerprints look similar.

State energies. To predict the state-specific G_0W_0 corrections to the PBE eigenvalues of 2D semiconductors, we use the XGBoost package[24] to build a machine learning model based on a gradient boosting algorithm for decision tree ensembles. The G_0W_0 data set was described and analysed in detail in ref. [25]. We split the data set into a training set of 228 randomly selected materials (37,851 electronic states) and a test set consisting of the remaining 58 materials (8766 electronic states). As an objective function, we use the mean absolute error (MAE) between the predicted and actual G_0W_0 corrections. The electronic states are represented by the ENDOME and RAD-PDOS fingerprints supplemented by a set of extra features consisting of the occupation of the state ($f_{nk} = 0, 1$), its distance to the Fermi energy ($\varepsilon_{nk} - E_F$), the PBE bandgap of the material (E_{gap}), and the static averaged in-plane and out-of-plane polarisabilities of the material ($\frac{1}{2}(\alpha_x + \alpha_y)$ and α_z). The averaged in-plane polarisability is used to ensure invariance of the feature with respect to rotations of the 2D material in the plane, which is important for materials with in-plane anisotropy. The effect of including the polarisabilities in the fingerprint has been analysed separately (see later discussion).

Table 1 Summary of results.

Methods	Target property MAE	
	Bandgap (eV)	State energies (eV)
G_0W_0 vs. experiment	≈ 0.3	N/A
PBE vs. G_0W_0	1.70	1.17
HSE06 vs. G_0W_0	0.85	0.47
PBE with ideal scissor-operator vs. G_0W_0	0	0.17
ML (8VB + 4CB) vs. G_0W_0	0.23, 0.18$^{(*)}$	0.14, 0.11$^{(*)}$
ML (VB + CB) vs. G_0W_0	0.18, 0.15$^{(*)}$	0.31, 0.22$^{(*)}$

The table shows the mean absolute error (MAE) on the bandgap and individual state energies for G_0W_0 versus experiments and different approximate methods versus G_0W_0, respectively. The MAE on state energies is always evaluated for the eight highest valence bands (VB) and four lowest conduction bands (CB). ML(X) refers to the test set MAE of the gradient boosting model after training on all bands (8VB+4CB) or only the highest valence and lowest conduction band (VB+CB), respectively. The values marked by (*) are obtained after training the model with the static polarisability of the materials included as extra features in the fingerprint.

The results of the model together with relevant baselines for assessing its performance are summarized in Table 1. The first row shows the estimated accuracy of our target G_0W_0 data relative to experiments based on previous reports in the literature[7–9]. Experimental data for individual band/state QP energies are scarce and subject to significant uncertainties, and thus do not represent a meaningful reference. The remaining rows of the Table show the mean absolute error (MAE) on the bandgap and individual state energies for different approximate methods versus G_0W_0. The MAE on state energies is evaluated over all the bands for which G_0W_0 data is available, namely the eight highest valence bands (VB) and four lowest conduction bands (CB). The second and third rows are straightforward comparisons of band energies from PBE and HSE06 with G_0W_0, respectively. The fourth row shows the MAE between G_0W_0 and

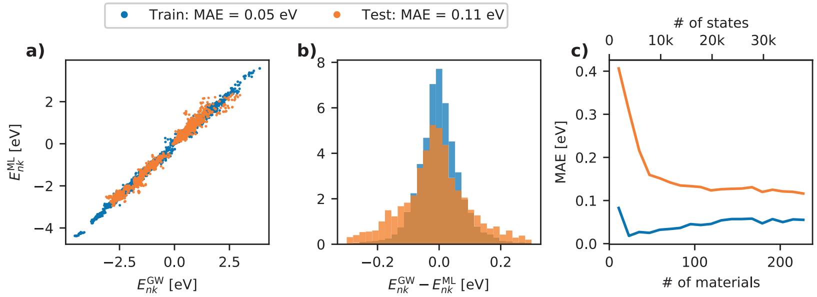


Fig. 4 Machine learning results. a Parity plot showing the ML predicted vs. true values of the GW correction for individual states for the train and test set. The MAEs of the train and test set are 0.05 and 0.11 eV, respectively. **b** Histograms of the prediction residuals of the train and test set. **c** Learning curve for the ML model showing validation MAE as a function of the number of materials/states in the training set.

PBE after the occupied and unoccupied PBE energies have been rigidly shifted (by applying a scissors operator) to match the valence band maximum (VBM) and conduction band minimum (CBM) of the G_0W_0 band structure. From this, it follows that the lowest possible MAE on individual band energies obtainable with a model trained to predict only the VBM and CBM energies is 0.17 eV. The last two rows of the table show the MAE on the test set obtained with the XGBoost model (see below for more details). Improved performance for the bandgap can be obtained by training the model only on the highest valence and lowest conduction band (last row); however, such a restriction on the training data reduces the prediction accuracy for bands further away from the bandgap. The numbers marked by (*) refer to the MAE obtained when the static polarisability of the materials is included in the fingerprint (see later discussion).

In the following, unless stated otherwise, results refer to the case where the model has been trained on all bands (8VB + 4CB) and with the static polarizabilities included in the fingerprint.

Figure 4a shows a parity plot of the predicted vs. true values for the train and test set. The evaluation yields MAEs of 0.05 and 0.11 eV for the train and test set, respectively. To test for the potential bias of the model, the residual distributions are plotted in Fig. 4b, showing that both the train and test set have residuals distributed evenly around 0 eV. To estimate the effect of adding more data to the train set, a learning curve is shown in Fig. 4c. The learning curve is calculated by continuously adding more materials to the training set while evaluating the performance on a constant test set. The test set MAE decreases significantly up to ≈50 materials after which the learning curve flattens considerably, although still presenting a slightly decreasing MAE. This suggests that a generalizable model can be trained using a rather limited number of materials, though it should be noted that overfitting issues decrease with the amount of materials in the training set. In general, it is difficult to assess whether the learning ability of the model is limited by the flexibility of the model/fingerprint or by the noise level in the data set. We do stress, however, that the numerical precision of the G_0W_0 corrections is not expected to be much better than 0.05 eV due to errors introduced by e.g., plane-wave extrapolation and linearisation of the self-energy, see ref. [25]. This could explain (part of) the finite prediction error of the model.

All MAEs reported in this paper were evaluated for a specific, randomly generated test set of 58 materials. We have verified that this test set is representative and fair by comparing it to MAEs obtained for 100 different random test sets, see Methods section.

The data used to train and evaluate the ML model represent states/energies evaluated at discrete uniformly distributed k-points of the Brillouin zone. However, the resulting ML model can of course be used to predict the G_0W_0 energy corrections of states at arbitrary k-points and thereby generate full, densely sampled band structures. Figure 5 shows examples of ML-generated band structures for PtO_2, SbClTe, GeS_2, and $CaCl_2$, which are all test set materials. For comparison, the PBE and the true discrete G_0W_0 energies are also shown. Overall, the ML bands closely interpolate the true G_0W_0 energies. In cases where the ML bands deviate, e.g. the conduction bands of $CaCl_2$, they still present a better description than PBE. Interestingly, the ML model is able to deviate from a scissors operator that would ascribe the same corrections to all occupied and all unoccupied bands, respectively. This is for example clear in the PtO_2 band structure where the four conduction bands are shifted by different amounts. We note that the single-point regression nature of the model, i.e., the fact that the model does not explicitly couple different k-points, can sometimes lead to weak and unphysical wiggles in the machine-learned band energies. These qualitative errors may be reduced by applying a smoothing function (e.g., a Gaussian filter) as post-processing of the ML energies across bands. This has been done for the plots in Fig. 5.

Bandgaps. The ML state energies can be translated into ML bandgaps by simply calculating the vertical difference between conduction band minimum and valence band maximum. Figure 6 shows parity plots of the predicted bandgaps vs. G_0W_0 bandgaps for an ML model trained on all bands and an ML model trained only on valence and conduction bands. Due to the discreteness of the original G_0W_0 data, the ML bandgap has been evaluated on the same states (discrete k-points) that define the G_0W_0 gap. The PBE and HSE06 data are also shown as baselines. Only data from the test set has been used for the comparison. The PBE and HSE06 functionals systematically underestimate the bandgaps leading to MAEs of 1.70 and 0.85 eV, respectively. The ML model trained on all bands achieves an MAE on the bandgap of 0.18 eV, while training the ML model only on valence and conduction bands reduces the bandgap MAE to 0.15 eV, but at the cost of increasing the MAE on the individual state energies across all bands from 0.11 to 0.22 eV.

While our ML model and fingerprints allow for the prediction of state-specific properties, such as individual band energies, it is of interest to compare its accuracy on bandgap predictions to alternative schemes reported in the literature. Lee and coworkers[26] used nonlinear support vector regression with fingerprints containing the Kohn-Sham bandgap obtained with both the PBE and the mBJ xc-functionals, together with a set of features describing the constituent chemical elements, to predict G_0W_0 bandgaps of inorganic bulk semiconductors. Using a

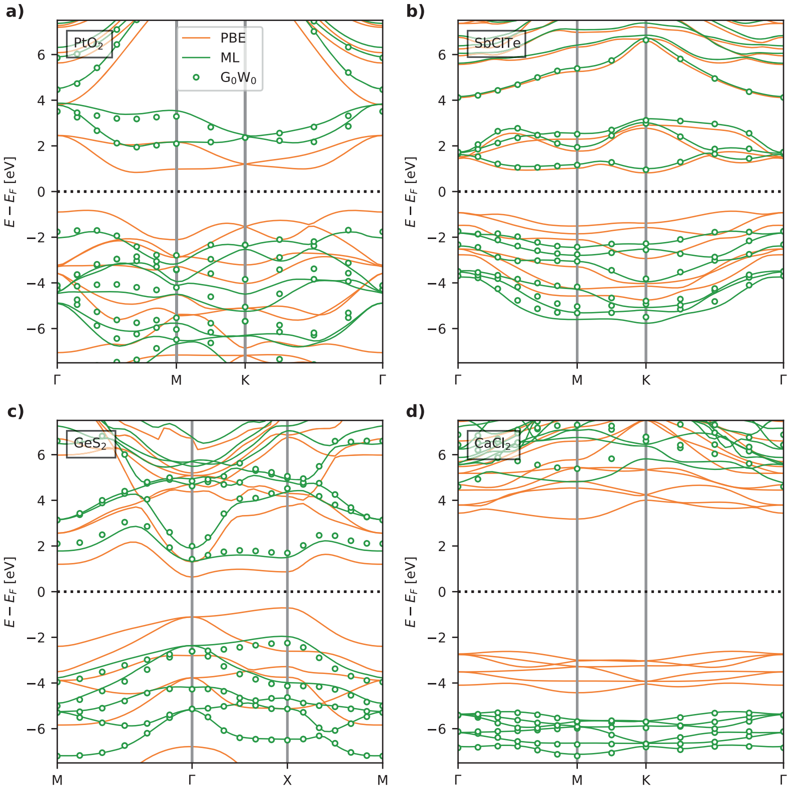


Fig. 5 Machine-learned band structures. Examples of band structures for four 2D materials from the test set. Both PBE and GW band structures are shown along with the ML predictions. The materials are selected to cover a wide range in the prediction accuracy of the test set. Band structures for PtO_2 (**a**), SbClTe (**b**), GeS_2 (**c**), and $CaCl_2$ (**d**).

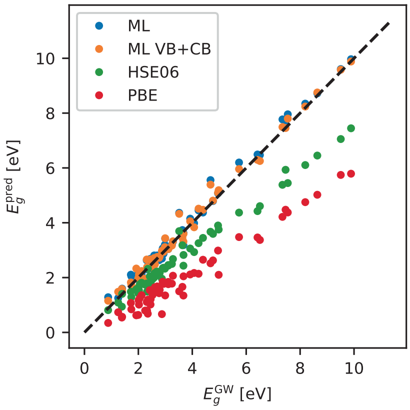


Fig. 6 Comparison of bandgaps. Parity plots for predicted bandgaps vs. GW bandgaps for PBE and HSE06 and two different ML models predicting GW corrections for either all bands (MAE = 0.18 eV) or only valence and conduction bands (MAE = 0.15 eV) which significantly outperform PBE and HSE06 with MAEs of 1.70 and 0.85 eV, respectively.

database of 270 G_0W_0 bandgaps, they obtained a root mean square error (RMSE) of 0.24 eV. Rajan et al. used a Gaussian process to predict G_0W_0 bandgaps of 2D MXene crystals with a fingerprint encoding atomic and structural properties of the MXenes[17]. Employing a training set of 76 G_0W_0 MXene bandgaps, they obtained an RMSE of 0.14 eV.

We stress that both the inorganic bulk semiconductors considered ref. [26] and, in particular, the MXene 2D crystals of ref. [17], represent more homogeneous sets of materials than the 2D crystals considered in the present work. Nevertheless, with an RMSE of 0.26 and 0.21 eV on the predicted G_0W_0 bandgap for the models trained on 8VB + 4CB and VB + CB, respectively, our general ML model with purely electronic fingerprints, is comparable in accuracy to the more system-specific ML models.

Additionally, by applying our ML model on ~700 semiconductors from C2DB we have found the bandgap to change nature (direct/indirect) in 12% of the materials when comparing the PBE and ML bandgaps. For these materials, 72% shift from direct to indirect gaps.

Effective masses. Since the ML model can be used to calculate G_0W_0 energies at any k-point grid, it is possible to use the method to calculate effective masses. Effective masses at the valence and conduction band extrema can be calculated by fitting a second-

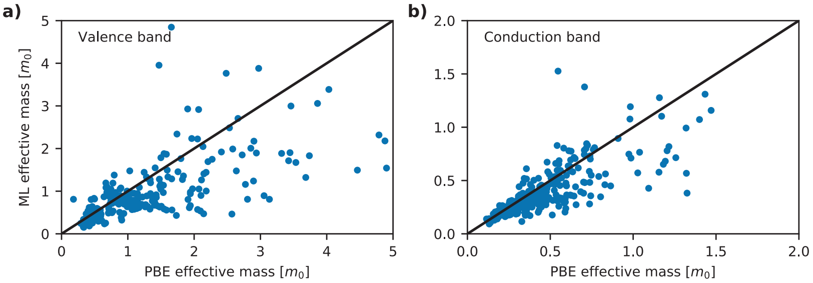


Fig. 7 Effective masses. Comparison of effective masses calculated using PBE and ML-G_0W_0 eigenvalues for valence and conduction band of ~800 materials. **a** Shows effective masses for the valence bands and **b** shows for the conduction bands. There seems to be a (weak) systematic trend for the ML model to predict smaller effective masses than PBE for both valence and conduction bands.

order polynomial to the energies at a densely sampled *k*-point grid centered around the band extrema[20,21]. This method is generally challenging with G_0W_0 due to the high computational cost of calculating the energies at sufficiently dense *k*-point grids, but using the ML model it is possible to achieve accurate estimates of the G_0W_0 effective masses.

Figure 7 shows effective masses calculated using PBE and ML energies for ≈330 materials using a *k*-point density of 55/Å^{-1} in a radius of 0.16 Å^{-1}.

The validity of the polynomial fit is evaluated using a mean absolute relative error (MARE) metric. The MARE is defined as the absolute difference between the parabolic fit and the actual ML-G_0W_0 band energies averaged over an energy range of 100 meV (from the band extremum) relative to the actual band energies averaged over the same energy range. The data shown in Fig. 7 includes only fits with MARE less than 10%.

Returning to Fig. 7 we note that the effective masses obtained with ML-G_0W_0 can deviate quite significantly from the PBE values. Specifically, the mean absolute deviation is $0.31m_0$ and $0.19m_0$ for valence and conduction bands, respectively, corresponding to relative deviations of 32 and 28%. We can also deduce that the ML-G_0W_0 method has a general tendency to yield smaller effective masses than PBE, although deviations from this trend occur relatively often.

Discussion

Feature importance. Often the evaluation of a machine learning model stops after considering the overall performance in terms of an objective function like the MAE. However, important insight may be gained by analysing how the model responds to different features in the input data. This is particularly important when devising new types of fingerprints. To extract information about the role of the different features composing the fingerprint vectors used in the present work, a feature importance analysis is performed using a feature subset hold-out method. The features are grouped at two different levels: The first level has four groups, namely the RAD-PDOS components, the ENDOME components, the extra features covering the PBE gap, occupation number, distance to the Fermi level, and finally the in-plane and out-of-plane polarizabilities. The second level breaks the RAD-PDOS and ENDOME components further down into their individual *ll*′ angular momentum blocks and operator matrix elements, respectively. The analysis is carried out in two complementary ways where a group of features is either used exclusively or dropped from the full fingerprint when training the ML model.

Figure 8 shows the test set MAE on individual state energies for the various feature groups with the all-feature baseline indicated by the vertical black line. Focusing first on panel (a), the analysis shows that both the RAD-PDOS and ENDOME perform well by themselves, though not as well as the full fingerprint. The extra features, in particular the polarisabilities, are unable to produce an accurate ML model. The poor performance of the polarisability-only feature is unsurprising as this feature is fully material-specific and not even able to distinguish between occupied and unoccupied states. Panel (b) shows the same analysis when the feature groups are broken further down. When used alone, the *pp*, *ss*, and *sp* components of the RAD-PDOS perform best followed by the various operator matrix elements of the ENDOME. An interesting observation is that at this level of feature grouping, almost any group of features can be dropped without increasing the MAE, except for the in-plane polarisability, α_{xy}, which results in a significant 27% increase of the MAE from 0.11 to 0.14 eV. This reveals a clear feature synergy since α_{xy} in itself does not have any predictive ability unless it is combined with other features (see below). In general, there seems to be some redundant information in the various fingerprint components since dropping any of the feature sets, at least at the second level of grouping, does not affect the test score by much. In some cases, the model might even gain performance when dropping some features (not visible on the scale of the plot). This suggests that a feature selection algorithm prior to the prediction algorithm might in general slightly improve the performance of the model. However, since gradient boosting algorithms like XGBoost already has some implicit feature selection in the training iterations, the improvement is not expected to be significant and is thus not considered here.

SHAP analysis. The role of the α_{xy} feature and its synergy with other features is further investigated using the general feature importance method SHAP, which is a game-theoretic approach to explain the output of any machine learning model[27]. SHAP builds an explanation model on top of an ML model which relates the output from the ML model to the importance of individual features for each predicted output. The SHAP values for a given feature can thus be interpreted as the direct effect of that feature on the model output, i.e. the difference between the model's prediction when used with and without that particular feature in the input. Figure 9a shows the SHAP values for α_{xy} as a function of α_{xy}. Only states from the test set are shown in Fig. 9, and the color code in panel (a) reflects the occupancy of the state. The plot shows a surprisingly clear trend: The SHAP values for occupied states increase consistently and monotonously for increasing α_{xy} while the opposite trend is seen for the empty states. In the following, we present a physical explanation for this observation.

The G_0W_0 correction can be split into two terms with distinctly different physical origin: $\Delta E_{nk}^{\mathrm{QP}} = (v_{nk}^{\mathrm{x}} - v_{nk}^{\mathrm{xc}}) + \Delta_{nk}^{\mathrm{scr}}$. The first term (in parenthesis) represents the difference between the local xc-potential (in this case the PBE potential) and the nonlocal exact exchange potential while the last term accounts for the interaction of

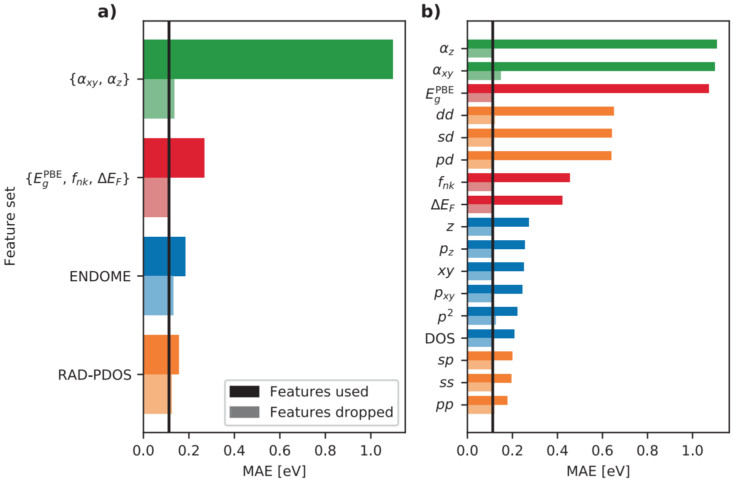


Fig. 8 Feature analysis of ML model. Solid bars refer to an ML model using only the specific features while the shaded bars are for an ML model without these features. **a** High-level feature group. **b** Low-level feature groups.

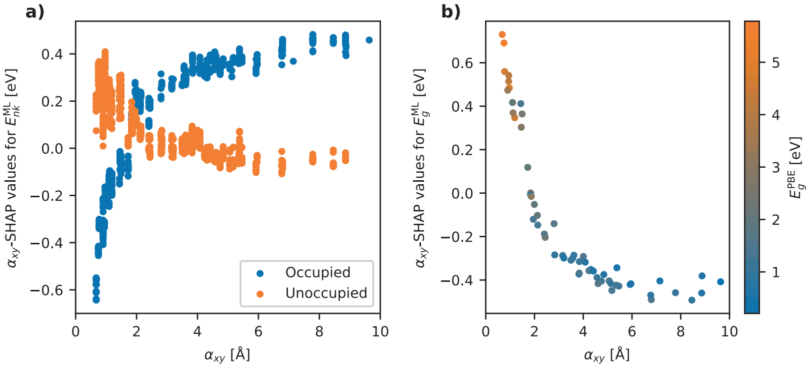


Fig. 9 SHAP analysis. a SHAP values for α_{xy} for the prediction of GW correction energies color-coded by occupancy. For materials with a low polarisability, the ML model predicts a more negative GW correction for the occupied states and a more positive correction for the unoccupied states. For materials with a high polarisability, the occupied states are predicted with a more positive correction when using the polarisability as a feature while the unoccupied states are only weakly affected. **b** SHAP values for α_{xy} for the prediction of bandgaps. This shows that the bandgap increases for materials with a low α_{xy} and decreases for high α_{xy} values.

the electron/hole with its own polarization cloud. The first term is typically negative for occupied states and positive for unoccupied states (Hartree–Fock typically opens the PBE gap), but its magnitude depends on the detailed shape of the wave functions of the system. In particular, this term can be quite different for different states of the same material. Moreover, one does not expect the size of this term to correlate with the material's static polarisability and thus it should not be captured by the α_{xy}-SHAP values. The second term is always positive for occupied states (hole quasiparticles) and negative for unoccupied states (electron quasiparticles) because the Coulomb interaction of the bare particle with its oppositely charged polarization cloud will always stabilize the quasiparticle, thus shifting occupied states up and empty states down in energy[28–30]. Now, the shape and size of the polarization cloud does not depend on the detailed shape of the wave function but is largely governed by the (microscopic) polarisability of the material. Therefore, on purely physical grounds, the static macroscopic polarisability, α_{xy}, is expected to provide a good descriptor for Δ_{nk}^{scr}: A large value of α_{xy} signals high screening ability of the material and therefore large QP polarization clouds, which in turn will yield a large Δ_{nk}^{scr} (with opposite signs for occupied/empty states). This is exactly what is seen in Fig. 9a. By subtracting the α_{xy}-SHAP values for the states at the CBM and VBM, we obtain the α_{xy}-SHAP values for the bandgap correction, see Fig. 9b. These show that the α_{xy} feature increases the bandgap in materials with low screening and decreases the bandgap in materials with high screening. Again, this is perfectly in line with the physical understanding of screening-induced renormalization of the bandgaps[28–30].

It can be noted that the α_{xy}-SHAP values for the state energies and bandgaps are significantly larger than the change in the MAE upon including/dropping α_{xy} from the feature set, see Fig. 8b. For example, the α_{xy}-SHAP values for the bandgap range from −0.50 to 0.70 eV while the MAE decreases by 0.03 eV when α_{xy} is included. This is due to the redundant information carried by the feature set. When the model is trained without α_{xy} as a feature, other features can, to a large extent, provide the same information. For example, the PBE bandgap alone correlates fairly well with α_{xy}. To test this hypothesis, we have carried out the same SHAP analysis for E_g^{PBE} on a model trained with and without α_{xy} in the feature set. The analysis shows that when α_{xy} is used to train the model, the E_g^{PBE}-SHAP values are fairly low (below ±0.1 eV) and do not show any clear trends. In contrast, when α_{xy} is not included in the fingerprint, the E_g^{PBE}-SHAP values are very similar to the α_{xy}-SHAP values shown in Fig. 9, although the values are slightly smaller and the trend less pronounced. This shows that in the absence of α_{xy} the model uses E_g^{PBE} to encode similar information. However, the model also finds that α_{xy} provides a better description of $\Delta_{nk}^{\mathrm{scr}}$ than does E_g^{PBE}, which is why the SHAP values of E_g^{PBE} are dwarfed by those of α_{xy} when both features are available for learning.

Summary. In summary, we have introduced two different methods to generate fingerprints of individual electronic states based on information available from a standard DFT ground-state calculation (eigenvalues and wave functions). The fingerprints were used to train a decision-tree-based ML model to predict the G_0W_0 corrections to the PBE band structure of a 2D semiconductor. The model achieves an MAE of 0.14 eV for individual state energies, which is reduced to 0.11 eV when the static polarisability is included in the fingerprint. For the bandgap, the MAE is 0.15–0.23 eV depending on whether the model is trained on all bands or only the valence/conduction bands and whether or not the static polarisability is included in the fingerprint. This level of precision is highly encouraging considering that the noise on the employed G_0W_0 data for individual state energies could be on the order of 0.05 eV and that the accuracy of the G_0W_0 method itself, when evaluated against experimental bandgaps, is about 0.3 eV. Since the bottleneck of the computations is the self-consistent DFT calculation (in particular the structural relaxation if performed), the method enables GW-quality band structures at the cost of a DFT calculation. Although the current work has focused on states in periodic 2D crystals, the methods can be straightforwardly used to fingerprint states in 3D crystals as well as non-periodic structures like molecules or surfaces. While the fingerprint methods can be used for e.g., 3D crystals, the ML model trained on 2D materials will not be transferable since some of the fingerprint components are divided into in-plane and out-of-plane parts. To use the full method of fingerprints and ML model for 3D crystals would require an ML model trained on a database of GW calculations of such systems.

Methods

This section describes the definition and generation of the Energy Decomposed Operator Matrix Elements (ENDOME) and Radially Decomposed Projected Density Of States (RAD-PDOS) fingerprints. In addition, the G_0W_0 band structure data set is presented along with a description of the employed machine learning model.

Electronic state fingerprints. The ENDOME fingerprint is based on operator matrix elements between electronic states (here assumed to be Bloch states of a periodic crystal)

$$A_{nk,n'k'} = |\langle nk|\hat{A}|n'k'\rangle|^2 \tag{1}$$

where $\hat{A}$ is some operator. For a reference state $|nk\rangle$ with energy ε_{nk}, the ENDOME fingerprint is defined as

$$m_{nk}^{A}(E) = \sum_{n'k'} A_{nk,n'k'} G\big(E - (\varepsilon_{nk} - \varepsilon_{n'k'}); \delta_E\big) \exp\big(-\alpha_E E\big) \mathrm{sign}(E_F - \varepsilon_{n'k'}), \tag{2}$$

where $G(x; \delta)$ is a Gaussian of width δ centered at $x = 0$. This function encodes the matrix element between the reference state and all other states at an energy distance of E from the reference state. In principle, any operator can be used to create fingerprints, but in this study, we include the position operators (x, y, z), the momentum operators (∇_x, ∇_y, ∇_z), and the Laplace operator (∇^2). These operators are all diagonal in the k index. In addition, we include the all-one matrix, $A_{nk,n'k'} = 1$, which essentially yields the density of states (DOS) translated to the energy of the reference state, ε_{nk}.

In practice, the function $m_{nk}^{A}(E)$ is represented on a uniformly spaced energy grid with 50 energy points from −10 to 10 eV around the reference state. Since we consider 2D materials, the in-plane (x and y) components of both the position and momentum operators are collected into a single fingerprint vector (i.e., $m_{nk}^{xy} = m_{nk}^{x} + m_{nk}^{y}$ and similarly for the momentum operator) while the out-of-plane z component is treated separately. For a given reference state, the ENDOME fingerprint thus consists of six 50-dimensional vectors resulting in a total of 300 features.

The RAD-PDOS encodes the electronic structure in terms of the density of states projected onto atomic orbitals. Specifically, a correlation function in energy and radial distance is defined as

$$\rho_{nk}^{\nu\nu'}(E,R) = \frac{1}{N_e}\sum_{n'k'aa'} \rho_{nk}^{a\nu}\rho_{n'k'}^{a'\nu'} G\big(R - |R_a - R_{a'}|; \delta_R\big) \exp\big(-\alpha_R R\big) G\big(E - (\varepsilon_{nk} - \varepsilon_{n'k'}); \delta_E\big) \times \exp\big(-\alpha_E E\big)\mathrm{sign}(E_F - \varepsilon_{n'k'}) \tag{3}$$

where N_e is the number of electrons in the system, a and a' denote atoms in the primitive unit cell and the entire crystal, respectively, and v and ν' denote atomic orbitals. The atomic projections are given by

$$\rho_{nk}^{a\nu} = |\langle \psi_{nk}|a\nu\rangle|^2 \tag{4}$$

The functions $\rho_{nk}^{\nu\nu'}(E,R)$ are represented on a uniform (E, R)-grid of size 25×20 spanning the intervals from −10 to 10 eV (centered around the reference energy ε_{nk}) and 0 to 5 Å, respectively. For the Gaussian smearing functions we use $\delta_E = 0.3$ eV and $\delta_R = 0.25$ Å, respectively. For a given state, the RAD-PDOS fingerprint consists of six 2D grids of 500 points each resulting in a total of 3000 features.

Figure 2 shows examples of ENDOME and RAD-PDOS fingerprints for three different states at the K-point of MoS_2. Note that some of the RAD-PDOS fingerprints are qualitatively similar (e.g., sp and pp) but the scales differ by about an order of magnitude. This is due to the fact that the density of states projected onto s and p orbitals have a similar dependence on energy.

The G_0W_0 data set. The data set comprises quasiparticle (QP) energies from 286 G_0W_0 band structures of non-magnetic 2D semiconductors covering 14 different crystal structures and 52 chemical elements. The QP energies have been obtained from plane-wave-based one-shot G_0W_0@PBE calculations with full frequency integration and were produced as a part of the Computational 2D Materials Database (C2DB)[20,21]. The data set has been described and analysed in detail in ref. [25].

The QP energies of the data set have been calculated under the standard assumption that the G_0W_0 self-energy can be treated within first-order perturbation theory and linearized around the non-interacting reference energy, $\omega = \varepsilon_{nk}$, leading to the expression

$$E_{nk}^{\mathrm{QP}} \approx \epsilon_{nk} + Z\mathrm{Re}\big[\langle \psi_{nk}|\Sigma(\epsilon_{nk})|\psi_{nk}\rangle\big] \tag{5}$$

where

$$Z = \left(1 - \left.\frac{\partial \Sigma}{\partial \omega}\right|_{\omega=\epsilon_{nk}}\right)^{-1} \tag{6}$$

is the QP weight and ψ_{nk} is the PBE wave function with eigenvalues ϵ_{nk}. In practice, the G_0W_0 correction to the PBE energies, $\Delta E_{nk}^{\mathrm{QP}} = E_{nk}^{\mathrm{QP}} - \epsilon_{nk}$, were used as targets for the machine learning model.

To ensure the highest data quality, the original data set was filtered such that only states with QP weights between 0.7 and 1.0 were kept. As shown in ref. [25] the MAE on the QP correction of such states due to the linearization of the QP equation is 0.04 eV.

Machine learning model. The choice of learning algorithm for a machine-learned model depends on different considerations such as the amount of training data available and the nature of the learning objective (regression/classification, discrete/continuous). The fingerprints presented here are not designed for a specific learning algorithm and can thus be used to train a wide range of algorithms. For this specific purpose of predicting G_0W_0 QP energies, several types of algorithms including tree-based ensemble methods, neural networks, and Gaussian process regression have been considered and tested. The machine learning model is built using a gradient boosting method from the XGBoost distribution based on decision trees in an ensemble[24]. The choice of XGBoost as a learning algorithm is based on its generality and good performance across multiple machine learning applications, the possibility to extract knowledge from single features, and the ability of training on large amounts of data. For this specific purpose, a neural network and a gaussian process regression method have also been tested resulting in similar prediction accuracy.

A train and test set is created using a random 80/20% split on the material level which results in a train set of 228 materials (37851 QP energies) and a test set of 58 materials (8766 QP energies). Hyperparameters of the learning algorithm (max depth = 5, learning rate = 0.15, and number of estimators = 60) are tuned using a grid search method with fivefold cross-validation of the 80% train set. The performance of the machine learning is based on the mean absolute error (MAE) of the 20% test set.

Since the test set size is only 58 materials, the test MAE might exhibit some test set dependence. To evaluate this effect, the entire process of splitting the data in 80/20% train/test set, training the model using fivefold cross-validation on the train set, and evaluating the MAE of the test set, has been repeated 100 times using different seeds for the random split. The distribution of the 100 test MAEs have a mean of 0.13 eV and a standard deviation of 0.02 eV. We note that the specific test set used for Table 1 yields an MAE within one standard deviation from the mean.

Since the XGBoost model is based on decision trees some small discontinuities in-band energies might be introduced by the model. When calculating effective masses using a harmonic fit on a much smaller energy scale than the full band structures it was necessary to use a neural network (feed-forward network with three hidden layers with 200 neurons and tanh activation functions) to ensure a more continuous output. This NN yielded a test MAE of 0.13 eV compared to the 0.11 eV of the XGBoost model.

Data availability

The structures of the materials used in this study have been deposited in C2DB[22] (https://doi.org/10.11583/DTU.14616660.v1). The data set generated for this study is available at https://gitlab.com/knosgaard/electronic-structure-fingerprints.

Code availability

The Python code used to compute the fingerprints can be found here https://gitlab.com/knosgaard/electronic-structure-fingerprints.

Received: 2 July 2021; Accepted: 21 December 2021;
Published online: 03 February 2022

References

1. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133 (1965).
2. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
3. Godby, R., Schlüter, M. & Sham, L. Accurate exchange-correlation potential for silicon and its discontinuity on addition of an electron. *Phys. Rev. Lett.* **56**, 2415 (1986).
4. Hedin, L. New method for calculating the one-particle green's function with application to the electron-gas problem. *Phys. Rev.* **139**, A796 (1965).
5. Golze, D., Dvorak, M. & Rinke, P. The gw compendium: a practical guide to theoretical photoemission spectroscopy. *Front. Chem.* **7**, 377 (2019).
6. Aryasetiawan, F. & Gunnarsson, O. The gw method. *Rep. Prog. Phys.* **61**, 237 (1998).
7. Hüser, F., Olsen, T. & Thygesen, K. S. Quasiparticle gw calculations for solids, molecules, and two-dimensional materials. *Phys. Rev. B* **87**, 235132 (2013).
8. Shishkin, M. & Kresse, G. Self-consistent GW calculations for semiconductors and insulators. *Phys. Rev. B* **75**, 235102 (2007).
9. Nabok, D., Gulans, A. & Draxl, C. Accurate all-electron G0W0 quasiparticle energies employing the full-potential augmented plane-wave method. *Phys. Rev. B* **94**, 035118 (2016).
10. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 1–7 (2016).
11. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
12. Rupp, M., Tkatchenko, A., Müller, K-R. & Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
13. Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quant. Chem.* **115**, 1094–1101 (2015).
14. Huo, H. & Rupp, M. Unified representation of molecules and crystals for machine learning. *arvix*:1704.06439 (2018).
15. Jørgensen, P. B. et al. Machine learning-based screening of complex molecules for polymer solar cells. *J. Chem. Phys.* **148**, 241735 (2018).
16. Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* **9**, 1668–1673 (2018).
17. Rajan, A. C. et al. Machine-learning-assisted accurate band gap predictions of functionalized mxene. *Chem. Mater.* **30**, 4031–4038 (2018).
18. Liang, J. & Zhu, X. Phillips-inspired machine learning for band gap and exciton binding energy prediction. *J. Phys. Chem. Lett.* **10**, 5640–5646 (2019).
19. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chemical Phys.* **18**, 13754–13769 (2016).
20. Haastrup, S. et al. The computational 2d materials database: high-throughput modeling and discovery of atomically thin crystals. *2D Mater.* **5**, 042002 (2018).
21. Gjerding, M. et al. Recent progress of the computational 2d materials database (c2db). *2D Mater.* **8**, 044002 (2021).
22. *Computational 2D Materials Database* (*C2DB*). https://cmr.fysik.dtu.dk/c2db/c2db.html. Accessed: 2021-07-01.
23. Klots, A. et al. Probing excitonic states in suspended two-dimensional semiconductors by photocurrent spectroscopy. *Sci. Rep.* **4**, 1–7 (2014).
24. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. 785–794 (ACM, 2016).
25. Rasmussen, A., Deilmann, T. & Thygesen, K. S. Towards fully automated gw band structure calculations: what we can learn from 60.000 self-energy evaluations. *npj Comput. Mater.* **7**, 22 (2021).
26. Lee, J., Seko, A., Shitara, K., Nakayama, K. & Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B* **93**, 115104 (2016).
27. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (eds Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).
28. Neaton, J. B., Hybertsen, M. S. & Louie, S. G. Renormalization of molecular electronic levels at metal-molecule interfaces. *Phys. Rev. Lett.* **97**, 216405 (2006).
29. Garcia-Lastra, J. M., Rostgaard, C., Rubio, A. & Thygesen, K. S. Polarization-induced renormalization of molecular levels at metallic and semiconducting surfaces. *Phys. Rev. B* **80**, 245427 (2009).
30. Schmidt, P. S., Patrick, C. E. & Thygesen, K. S. Simple vertex correction improves gw band energies of bulk and two-dimensional crystals. *Phys. Rev. B* **96**, 205206 (2017).

Acknowledgements

The Center for Nanostructured Graphene (CNG) is sponsored by The Danish National Research Foundation (project DNRF103). We acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program Grant No. 773122 (LIMA) and Grant agreement No. 951786 (NOMAD CoE). K.S.T. is a Villum Investigator supported by VILLUM FONDEN (grant no. 37789).

Author contributions

N.R.K. and K.S.T. developed the initial concept. N.R.K. developed the Python code for computing the fingerprints and training the machine learning models. K.S.T. supervised the work and helped in the interpretation of the results. All authors modified and discussed the paper together.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-28122-0.

Correspondence and requests for materials should be addressed to Nikolaj Rørbæk Knøsgaard.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

8.2 Publication 2: Predicting and machine learning structural instabilities in 2D materials

Predicting and machine learning structural instabilities in 2D materials

Simone Manti[ξ,*], Mark Kamper Svendsen[ξ], Nikolaj R. Knøsgaard[ξ], Peder M. Lyngby[ξ], and Kristian S. Thygesen[ξ,†]

[ξ]CAMD, Computational Atomic-Scale Materials Design, Department of Physics, Technical University of Denmark, 2800 Kgs. Lyngby Denmark

[†]Center for Nanostructured Graphene (CNG), Department of Physics, Technical University of Denmark, DK - 2800 Kongens Lyngby, Denmark

[*]Corresponding author: smanti@fysik.dtu.dk

Abstract. We address the problem of predicting the zero temperature dynamical stability (DS) of a periodic crystal without computing its full phonon band structure. Using data for two-dimensional (2D) crystals, we first present statistical evidence that DS can be inferred with good reliability from the phonon frequencies at the center and boundary of the Brillouin zone (BZ). This analysis represents a validation of the DS test employed by the Computational 2D Materials Database (C2DB). For 137 dynamically unstable 2D crystals, we displace the atoms along an unstable mode and relax the structure. This procedure yields a dynamically stable crystal in 49 cases. The elementary properties of these new structures are characterised using the C2DB workflow, and it is found that their properties can differ significantly from those of the original unstable crystals, e.g. band gaps are opened by 0.3 eV on average. All the crystal structures and properties are available in the C2DB. Finally, we train a classification model on the DS data for 3295 2D materials in the C2DB using a representation encoding the electronic structure of the crystal. We obtain an excellent receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.89, showing that the classification model can drastically reduce computational efforts in high-throughput studies.

1. Introduction

Computational materials discovery aims at identifying novel materials for specific applications, often employing first principles methods such as density functional theory (DFT) [1]. The potential of a given material for the targeted application is usually evaluated based on elementary properties of the crystal, such as the electronic band gap, the optical absorption spectrum, or the magnetic order. Such properties can be highly sensitive to even small distortions of the lattice that reduce the symmetry of the crystal, and it is therefore important to develop efficient methods for identifying and accounting for such distortions.

Lattice distortions can be classified according to their periodicity relative to the primitive cell of the crystal. Local instabilities conserve the periodicity of the crystal, i.e. they do not enlarge the number of atoms in the primitive cell. Other distortions, known as charge density wave (CDW) [2], lead to an enlargement of the period of the crystal, which can be either commensurate or incommensurate with the high-symmetry phase. A universal microscopic theory of the CDW phase is still missing due to the many possible and intertwined driving mechanisms, e.g. electron-phonon interaction [3], Fermi surface nesting, or phonon-phonon interactions[4], which makes a precise clear-cut definition of the CDW phase difficult. In addition, the CDW state is sensitive to external effects such as temperature and doping[5]. As a testimony to the complexity of the problem, different models and concepts are used to describe the CDW phase depending on the dimensionality of the material [6, 7, 8, 9, 10].

The last few years have witnessed an increased interest in CDW states of two-dimensional (2D) materials. For example, CDW physics is believed to govern the transition from the trigonal prismatic T-phase to the lower symmetry T'-phase in monolayer MoS_2 [11] as well as the plethora of temperature dependent phases in monolayers of NbSe_2 [12, 13], TaS_2 [14, 15], TaSe_2 [16, 17], and TiSe_2 [18, 19]. In addition, a number of recent studies have investigated the possibility to control CDW phase transitions. For instance, the T-phase of monolayer MoS_2 can be stabilized by argon bombardment [20], exposure to electron beams [11], or Li-ion intercalation [21]. Similar results have been reported for MoTe_2 [22].

Regardless of the fundamental origin of possible lattice distortions, it remains of great practical importance to devise efficient schemes that makes it possible to verify whether or not a given structure is dynamically stable (DS), i.e. whether it represents a local minimum of the potential energy surface. Structures that are not DS are frequently generated in computational studies, e.g. when a structure is relaxed under symmetry constraints or the chosen unit cell is too small to accommodate the stable phase. Tests for DS are rarely performed in large-scale discovery studies, because there is no established way of doing it apart from calculating the full phonon band structure[23], which is a time-consuming task. At the same time, the importance of incorporating such tests is in fact unclear; that is, it is not known how much symmetry-breaking distortions generally influence the properties of a materials.

A straightforward strategy to generate potentially stable structures from dynamically unstable ones, is to displace the atoms along an unstable phonon mode using a supercell that can accommodate the distortion. This approach has previously been adopted to explore structural distortions in bulk perovskites [24, 25] and one-dimensional organometallic chains [26]. However, systematic studies of structural instabilities in 2D materials, have so far been lacking.

In this work, we perform a systematic study of structural distortions in thermodynamically stable 2D crystals from the Computational 2D Materials Database (C2DB)[27, 28] and explore a machine learning-based approache to DS classification. Throughout, we focus on the most common case of small-period, commensurate distortions that can be accommodated in a 2×2 repetition of the primitive cell of the high-symmetry phase. We shall refer to the test for the occurrence of such distortions as the Center and Boundary Phonon (CBP) protocol. The motivation behind the present work is fourfold: (i) To assess the reliability of the CBP protocol (which is currently used for DS classification by the C2DB). (ii) To elucidate the effect of symmetry-breaking distortions on the basic electronic properties of crystals. (iii) To obtain the DS phases of a set of dynamically unstable 2D materials that were originally generated by combinatorial lattice decoration, and make them available to the community via the C2DB. (iv) To explore the viability of a machine learning based classification scheme for predicting DS using input from a DFT calculation of the prospect high-symmetry phase.

The paper is structured as follows. In Section 2 we describe the CBP protocol. In Section 3 we first benchmark the CBP protocol against full phonon band structure calculations and evaluate its statistical success rate. For 137 dynamically unstable 2D materials, we further analyse how the small-period distortions that stabilise the materials influence their electronic properties. Section 4 concludes the paper.

2. Methodology

In this Section we briefly discuss the CBP protocol for testing the dynamical stability of a crystal and for generating distorted, dynamically stable stable crystal structures. We also describe the methodology and computational details of the phonon calculations.

2.1. The CBP protocol: Stability test

Given a material that has been relaxed in some unit cell (from hereon referred to as the primitive unit cell), the CBP protocol proceeds by evaluating the stiffness tensor of the material and the Hessian matrix of a supercell obtained by repeating the primitive cell 2×2 times. In the current work, the stiffness tensor is calculated as a finite difference of the stress under an applied strain, while the Hessian matrix is calculated as a finite difference of the forces on all the atoms of the 2×2 supercell under displacement of the atoms in one primitive unit cell (this is equivalent to calculating the phonons at the center and specific high symmetry points at boundary of the BZ of the primitive cell, see Fig. (3). Next, the stiffness tensor and the Hessian matrix are diagonalised, and the eigenvalues are used to infer a structural stability. A negative eigenvalue of the stiffness tensor indicates an instability of the lattice (the shape of the unit cell) while a negative eigenvalue of the 2×2 Hessian signals an instability of the atomic structure. The obvious question here, is whether it suffices to consider the Hessian of the 2×2 supercell, or equivalently consider the phonons at the BZ center and boundaries.

All phonon calculations were performed using the `asr.phonopy` recipe of the Atomic Simulation Recipes (ASR) [29], which makes use of the Atomic Simulation Environment (ASE)[30] and PHONOPY [31]. The DFT calculations were performed with the GPAW[32] code and the Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional [33]. The BZ was sampled on a uniform k-point mesh of density of 6.0 Å^2 and the plane wave cutoff was set to 800 eV. To evaluate the Hessian matrix, the small displacement method was used with a displacement size of 0.01 Å and forces were converged up to 10^{-4} eV/Å. To benchmark the CBP protocol, we compare to full phonon band structures. In these calculations, the size of the supercell is chosen such that the Hessian matrix includes interactions between pairs of atoms within a radius of at least 12 Å. (This implies that the supercell must contain a sphere of radius 12 Å).

We can distinguish three possible outcomes when comparing the CBP protocol against full phonon calculations (see Figure (1)), namely a true positive result, a true negative result, and a false positive result. We note that the case of a false negative is not possible, because a material that is unstable in a 2×2 cell is de facto unstable. The false positive case occurs when a material is stable in a 2×2 supercell, but unstable if allowed to distort in a larger cell. Our results show that such large-period distortions that do not show as distortions in a 2×2 cell, are relatively rare (see Section 3.1).

2.2. The CBP protocol: Structure generation

Here we outline a simple procedure to generate distorted and potentially stable structures from an initial dynamically unstable structure. The basic idea is to displace the atoms along an unstable phonon mode followed by a relaxation. In practice, the unstable mode is obtained as the eigen function corresponding to a negative eigenvalue of the Hessian matrix of the 2×2 supercell. The procedure is illustrated in Figure (2) for the well known T-T' phase transition of MoS_2 [11]. The left panel shows the atomic structure and phonon band structure of monolayer MoS_2 in the T-phase. Both the primitive unit cell (black) and the 2×2 supercell (orange) are indicated. The CBP method identifies an unstable mode at the BZ boundary (M point). After displacing the atoms along the unstable mode, a distorted structure is obtained, which after relaxation leads to the dynamically stable T'-phase of MoS_2 shown in the right panel.

In this work, we have applied the method systematically to 137 dynamically unstable 2D materials. The 137 monolayers were selected from the C2DB according to the following two criteria: First, to ensure that all materials are chemically "reasonable", only materials with a low formation energy were selected. Specifically, we require that $\Delta H_{\text{hull}} < 0.2$ eV/atom, where ΔH_{hull} is the energy above the convex hull defined by the most stable (possibly mixed) bulk phases of the relevant composition[34, 28]. Secondly, we consider only materials with exactly one unstable mode, i.e. one negative eigenvalue of the Hessian matrix at a given q-point.

The 137 dynamically unstable materials were displaced along the only unstable mode. The size of the displacement was chosen such that the maximum atomic displacement was exactly 0.1 Å. This displacement size was chosen based on the MoS_2 example discussed above, where it results in a minimal number of subsequent relaxation steps. A smaller value does not guarantee that the system leaves the saddle point, while a larger value creates a too large distortion resulting in additional relaxation steps. During relaxation the unit cell was allowed to change with no symmetry constraints and the relaxation was stopped when the forces on all atoms were below 0.01 eV / Å.

2.3. Machine learning method

Finally, we describe the machine learning approach that we will employ in an attempt to accelerate the pre-

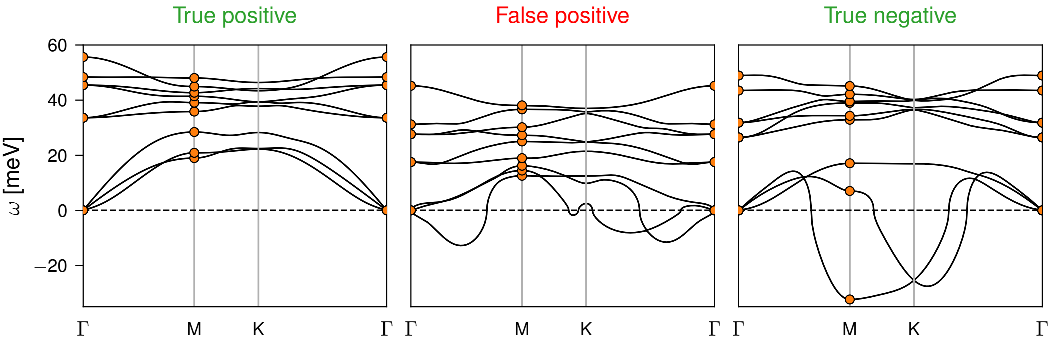


Figure 1. Phonon band structure for monolayer MoS_2 in the H-phase (left), NbSSe (middle), and MoS_2 in the T-phase (right). Note that imaginary phonon frequencies are represented by negative values. The CBP protocol (orange dots) is sufficient to conclude that a material is dynamically stable (unstable) in the situations depicted in the left (right) panels. In contrast, when the relevant distortion requires a supercell larger than a 2 × 2, and the phonon frequencies are real at the center and boundary of the BZ, the CBP protocol will result in a false positive result.

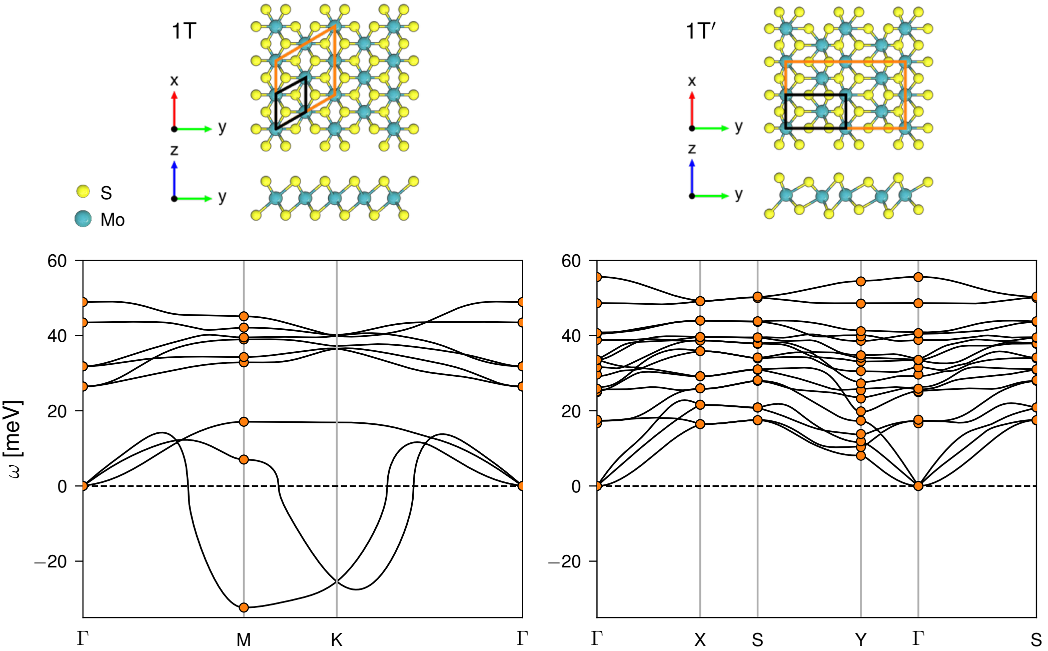


Figure 2. The CBP protocol captures the instability of the T-phase (left) of MoS_2. Both the primitive unit cell (black) and the 2 × 2 supercell (orange) are shown. Displacing the atoms along the unstable TA mode at the M-point ($\mathbf{q} = (\frac{1}{2}, 0)$), which can be accommodated in the 2 × 1 supercell, and subsequently relaxing the structure results in the dynamically stable T'-phase (right).

diction of dynamic instabilities. Our choice of machine learning algorithm is the library, XGBoost [35], due to its robustness and flexibility, while being a simpler model compared to neural network methods. XGBoost is a regularized high-performance implementation of gradient tree boosting, which makes predictions based on an ensemble of gradient boosted decision trees. The decision trees of the ensemble are grown sequentially while learning from the mistakes of the previous trees by minimizing the loss function through gradient descent. This loss function is regularized to reduce the complexity of the individual decision trees which reduces the risk of overfitting. In contrast, in the widely used decision tree ensemble model Random Forest, the decision trees are grown independently and without any regularization.

The dataset used is a subset of C2DB and consists of 3295 materials, which does not include the 137 distorted materials identified in the first part of the paper (as these will be used as a particularly challenging test case for the model performance). As input for the model we use the radially projected density of states (RAD-PDOS) material fingerprints. The RAD-PDOS starts from the wave functions projected onto the atomic orbitals (ν) of all the atoms (a) of the crystal, $\rho_{nk}^{a\nu} = |\langle\psi_{nk}|a\nu\rangle|^2$. For each state, these projections are then combined into a radially distributed orbital pair correlation function,

$$\rho_{nk}^{\nu\nu'}(R) = \sum_{aa'} \rho_{nk}^{a\nu}\rho_{nk}^{a'\nu'} G\left(R - |R_a - R_{a'}|; \delta_R\right) \times \exp\left(-\alpha_R R\right) \quad (1)$$

Finally, the radial functions are distributed on an energy grid,

$$\rho^{\nu\nu'}(R, E) = \sum_{nk} \rho_{nk}^{\nu\nu'}(R) G\left(E - (\varepsilon_{nk} - E_F); \delta_E\right) \times \exp\left(-\alpha_E R\right), \quad (2)$$

where $G(x;\delta)$ is a Gaussian of width δ centered at $x = 0$. For the materials in the dataset, the s, p, and d orbitals lead to six unique components of the RAD-PDOS fingerprint. The fingerprint involves some hyperparameters for which we use the values $E_{\min} = -10\text{eV}, E_{\max} = 10\text{eV}, N_E = 25, \delta_E = 0.3\text{eV}, \alpha_E = 0.2\text{eV}^{-1}, R_{\min} = 0, R_{\max} = 5\text{Å}, N_R = 20, \delta_R = 0.25\text{Å}, \alpha_R = 0.33\text{Å}^{-1}$.

In addition to the RAD-PDOS fingerprint, we consider a low-dimensional fingerprint consisting of six features, namely the PBE electronic band gap ($\varepsilon_{\text{gap}}^{\text{PBE}}$), crystal formation energy (ΔH), density of states at the Fermi level (DOS@E_F), energy above the convex hull (ΔH_{hull}), the total energy per atom in the unit cell and the total energy. The six-dimensional fingerprint is used to train a "baseline" ML model that we use to benchmark the performance of the ML model based on the more involved RAD-PDOS fingerprint. Common to all the features considered is that they are obtained from a single DFT calculation and thus are much faster to compute that the phonon frequencies.

The gradient boosting model introduces several hyperparameters such as depth of the trees, learning rate, minimum loss gain to perform a split and minimum weights in tree leafs. These parameters are optimized using Bayesian optimization where a Gaussian process is fitted to the mean test ROC-AUC of a 10-fold cross-validation.

The XGBoost classification model is in fact a logistic regression model, i.e. the output of the model is a number between 0 and 1 which is interpreted as a probability. In our case, 0 (1) refers to a dynamically stable (unstable) material.

3. Results

3.1. Assessment of the CBP protocol

To test the validity the CBP protocol, we have performed full phonon calculations for a set of 20 monolayers predicted as dynamically stable by the CBP protocol. The 20 materials were randomly selected from the C2DB and cover 7 different crystal structures. Out of the 20 materials 10 are metals and 10 are insulators/semiconductors. The calculated phonon band structures are reported in the supplementary material. For all materials, the phonon frequencies obtained with the CBP protocol equal the frequencies of the full phonon band structure at the q-points $\mathbf{q} \in \{(0,0), (\frac{1}{2},0), (0,\frac{1}{2}), (\frac{1}{2},\frac{1}{2})\}$. This is expected as the phonons at these q-points can be accommodated by the 2×2 supercell.

Within the set of 20 materials, we find three False-positive cases, namely CoTe_2, NbSSe, and TaTe_2. These materials exhibit unstable modes (imaginary frequencies or equivalently negative force constant eigenvalues) in the interior of the BZ (NbSSe and TaTe_2) or at the K-point (CoTe_2), while all phonon frequencies at the q-points covered by the CBP protocol, are real. This relatively low percentage of False-positives in our representative samples is consistent with the work by Mounet *et al.*[23] who computed the full phonon band structure of 258 monolayers predicted to be (easily) exfoliable from known bulk compounds. Applying the CBP protocol to their data yields 14 False-positive cases; half of these are transition metal dichalcogenides (TMDs) with Co, Nb or Ta.

We note that the small imaginary frequencies in the out of plane modes around the Γ-point seen in some of the phonon band structures are not distortions, but are rather due to the interpolation of the dynamical matrix. In particularly, these artifacts occur because of the broken crystal point-group symmetry in the force constant matrix and they will vanish if a larger supercell is used or the rotational sum rule is imposed [36].

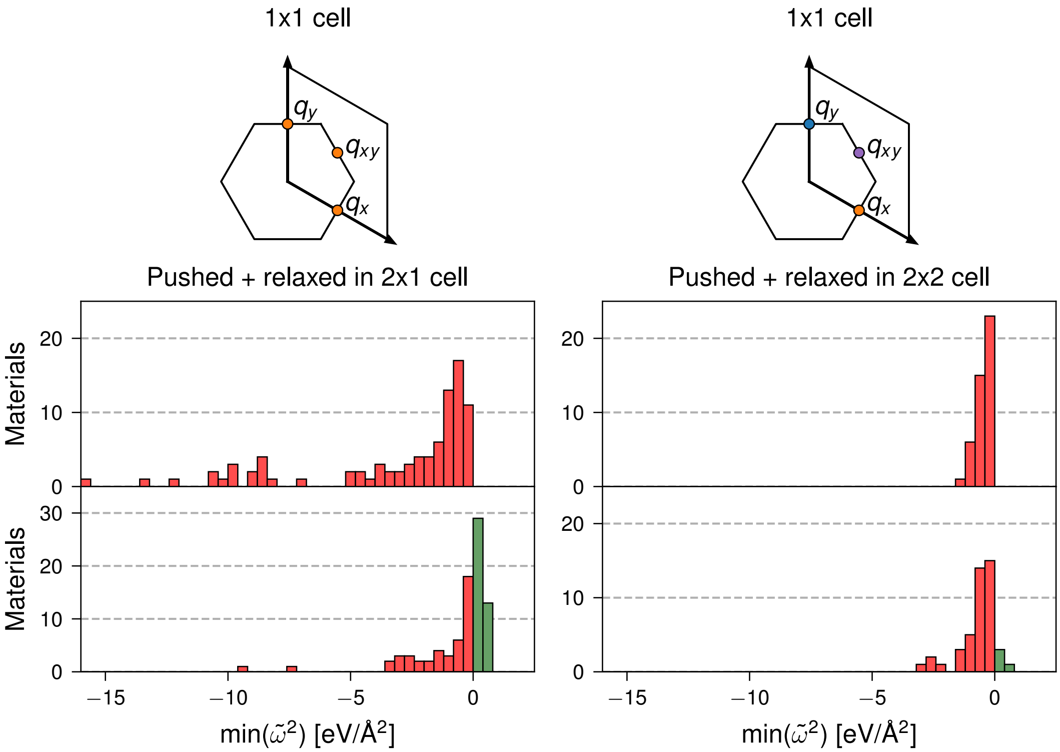


Figure 3. The 137 dynamically unstable 2D materials studied in this work can be divided into two groups depending on whether the negative eigenvalues of the Hessian matrix at $q = \{(\frac{1}{2}, 0), (0, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2})\}$ are equal (left panel) or different (right panel). For the first group of materials, displacing the atoms along the mode at $q = (\frac{1}{2}, 0)$ and relaxing in a 2×1 supercell, yields a dynamically stable structure in 43/91 cases. For the second group, displacing the atoms along $q = (\frac{1}{2}, \frac{1}{2})$ and relaxing in a 2×2 supercell, yields a dynamically stable structure in 6/46 cases.

3.2. Stable distorted monolayers

The 137 dynamically unstable materials, which were selected from the C2DB according to the criteria described in Section 2.2, can be divided into two groups depending on whether the eigenvalues of the Hessian at the wave vectors $q_x = (\frac{1}{2}, 0)$, $q_y = (0, \frac{1}{2})$ and $q_{xy} = (\frac{1}{2}, \frac{1}{2})$, are equal or not. Equality of the eigenvalues implies an isotropic Hessian. For such materials, we generate distorted structures by displacing the atoms along the unstable mode at $q_x = (\frac{1}{2}, 0)$, followed by relaxation in a 2×1 supercell. In the case of an anisotropic Hessian, the atoms were displaced along $q_{xy} = (\frac{1}{2}, \frac{1}{2})$ and relaxed in a 2×2 supercell.

After atomic displacement and subsequent relaxation, the CBP protocol was applied again to test for dynamical stability of the distorted structures. Histograms of the minimum eigenvalue of the Hessian matrix are shown in Figure (3) with the materials before and after atomic displacement shown in the upper and lower panels, respectively. Negative eigenvalues, corresponding to unstable materials, are shown in red while positive eigenvalues are shown in green. Out of the 137 unstable materials, 49 become dynamically stable (according to the CBP protocol). By far the highest success rate for generating stable crystals was found for the isotropic materials (left panel), where 43 out of 91 materials became stable while only 6 out of the 43 anisotropic materials became stable.

A wide range of elementary properties of the 49 distorted, dynamically stable materials were computed using the C2DB workflow (see Table (1) in [28] for a complete list of the properties). The atomic structures together with the calculated properties are available in the C2DB. Table (1) provides an overview of the symmetries, minimal Hessian eigenvalues, total energies, and electronic band gap of the 49 materials before and after the distortion.

Apart from the reduction in symmetry, the

Material	Space group - Wyckoff		ΔH_{hull} [eV/atom]		$\min(\tilde{\omega}^2)$ [eV/Å^2]		$\varepsilon_{\text{gap}}^{\text{PBE}}$ [eV]	
	before	after	before	after	before	after	before	after
$AgBr_2$	187-bi	1-a	0.05	0.00	-1.01	0.00	0.00	0.00
$AgCl_2$	164-bd	1-a	0.05	0.00	-2.60	0.06	0.00	0.00
AsClTe	156-ac	1-a	0.19	0.02	-0.51	0.25	1.29	1.48
CdBr	2-i	1-a	0.18	0.07	-1.00	0.03	0.00	1.28
CdCl	156-ab	1-a	0.17	0.04	-0.74	0.17	0.00	1.67
CoSe	164-bd	2-i	0.05	0.03	-4.82	0.09	0.00	0.00
CrBrCl	156-abc	1-a	0.11	0.06	-0.98	0.15	0.00	0.64
$CrBr_2$	164-bd	1-a	0.10	0.06	-0.59	0.07	0.00	0.49
$CrCl_2$	164-bd	1-a	0.11	0.05	-1.87	0.15	0.00	0.76
CrSSe	156-abc	1-a	0.15	0.09	-9.75	0.71	0.00	0.00
CrS_2	156-abc	2-i	0.18	0.05	-15.62	0.74	0.00	0.00
$CrSe_2$	156-abc	2-i	0.14	0.05	-12.17	0.77	0.00	0.00
$CrTe_2$	156-abc	2-i	0.02	0.01	-1.71	0.08	0.00	0.00
$CrPS_3$	3-1-a	1-a	0.09	0.03	-3.13	0.00	0.00	0.34
$FePSe_3$	3-1-a	1-a	0.13	0.12	-0.42	0.04	0.13	0.13
$FeSe_2$	187-bi	1-a	0.15	0.00	-2.04	0.00	0.00	0.00
HfBrCl	156-abc	1-a	0.14	0.03	-9.61	0.40	0.00	0.82
HfBrI	156-abc	1-a	0.22	0.05	-10.41	0.39	0.00	0.73
$HfBr_2$	164-bd	2-i	0.14	0.04	-10.21	0.41	0.00	0.8
$HfCl_2$	164-bd	2-i	0.14	0.04	-8.51	0.38	0.00	0.85
HgSe	156-ab	1-a	0.11	0.04	-0.83	0.04	0.08	0.37
HgTe	156-ab	1-a	0.11	0.04	-0.61	0.04	0.08	0.37
InTe	156-ab	1-a	0.18	0.10	-0.41	0.23	0.00	0.00
InBrSe	59-ab	2-i	0.03	0.02	-0.50	0.04	1.23	1.23
InSe	187-hi	6-ab	0.00	0.00	-0.28	0.01	1.39	1.39
MoSeTe	156-abc	1-a	0.20	0.08	-10.67	0.69	0.00	0.00
$MoTe_2$	164-bd	2-i	0.17	0.01	-13.23	0.67	0.00	0.00
NbS_2	187-bi	6-ab	0.00	0.00	-1.08	0.06	0.00	0.00
$NbTe_2$	187-bi	1-a	0.00	0.00	-0.37	0.51	0.00	0.00
PdI_2	164-bd	1-a	0.17	0.03	-0.56	0.12	0.00	0.59
RhI_2	164-bd	1-a	0.17	0.17	-0.64	0.04	0.00	0.00
RhO_2	164-bd	2-i	0.16	0.15	-2.19	0.65	0.00	0.00
$RhTe_2$	164-bd	1-a	0.11	0.07	-1.67	0.36	0.00	0.13
ScI_3	162-dk	1-a	0.00	0.00	-0.25	0.02	1.85	1.85
TiBrCl	156-abc	1-a	0.05	0.00	-9.15	0.52	0.00	0.29
$TiBr_2$	164-bd	1-a	0.06	0.04	-0.54	0.30	0.00	0.12
$TiCl_2$	164-bd	2-i	0.11	0.00	-9.99	0.53	0.00	0.32
TiO_2	164-bd	6-ab	0.14	0.12	-1.96	0.47	2.70	2.85
TiS_2	187-bi	4-a	0.14	0.14	-0.49	0.00	0.73	0.79
$TiPSe_3$	1-a	4-a	0.16	0.00	-1.24	0.00	0.00	0.00
VTe_2	164-bd	6-ab	0.02	0.00	-1.05	0.38	0.00	0.00
ZrBrCl	156-abc	1-a	0.12	0.02	-8.05	0.40	0.00	0.59
ZrBrI	156-abc	1-a	0.15	0.02	-8.89	0.38	0.00	0.48
$ZrBr_2$	164-bd	2-i	0.11	0.00	-8.65	0.47	0.00	0.59
ZrClI	156-abc	1-a	0.19	0.07	-8.69	0.26	0.00	0.48
$ZrCl_2$	164-bd	1-a	0.11	0.03	-7.13	0.46	0.00	0.60
ZrI_2	164-bd	2-i	0.14	0.00	-8.59	0.48	0.00	0.43
ZrS_2	187-bi	2-i	0.19	0.18	-0.80	0.15	0.96	1.13

Table 1. Some of the calculated properties of the subset of the 137 materials that became dynamically stable after displacing the atoms along an unstable phonon mode. The properties are shown before and after the distortion, i.e. for the original dynamically unstable structures and the final dynamically stable structures, respectively.

distortion also lowers the total energy of the materials. An important descriptor for the thermodynamic stability of a material is the energy above the convex hull, ΔH_{hull}. Figure (4) shows a plot of ΔH_{hull} before and after the distortion of the 49 materials. The reduction in energy upon distortion ranges from 0 to 0.2 eV/atom. In fact, several of the materials come very close to the convex hull and some even fall onto the hull, indicating their global thermodynamic stability (at $T = 0$ K) with respect to the reference bulk phases. We note that all DFT energies, including the reference bulk phases, were calculated using the PBE xc-functional, which does not account for van der Waals interactions. Accounting for the vdW interactions will downshift the energies of layered bulk phases and thus increase ΔH_{hull} for the monolayers slightly. This effect will, however, not influence the relative stability of the pristine and distorted monolayers, which is the main focus of the current work.

Another characteristic trend observed is the

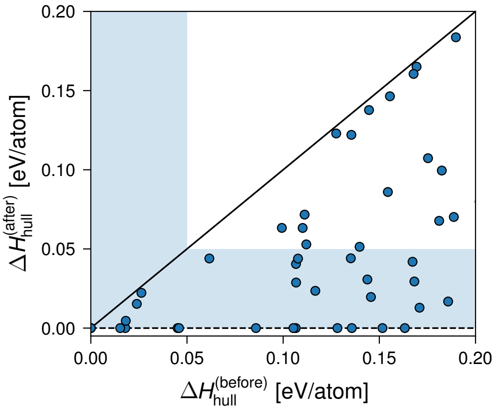


Figure 4. The energy above the convex hull for the 49 monolayers before and after distortion. Materials with a $\Delta H_{\rm hull}^{\rm (after)}$ close to zero are expected to be thermodynamically stable. The range up to 0.05 eV/atom above the convex hull has been indicated by a shaded blue region to visualise the importance of structural distortions for assessing the thermodynamic stability.

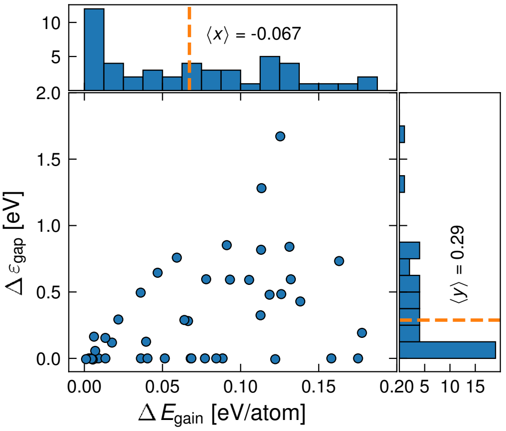


Figure 5. The average energy gain for the new stable materials is -0.067 eV / atom and the average gap opening is 0.29 eV.

opening/increase of the electronic band gap. The increase of the single-particle band gap is expected to be related to the total energy gained by making the distortion. Figure (5) shows the relation between the two quantities. Simplified models, for low dimensional systems and weak electron-phonon coupling, predict a proportionality between these two quantities [37]. From our results it is clear that there is no universal relationship between the change in band gap and total energy. In particular, several of the metals show large gain in total energy while the gap remains zero.

It is interesting that within a threshold of 0.1 eV 21 of the distorted and dynamically stable materials exhibit direct band gaps. Atomically thin direct band gap semiconductors are highly relevant as building blocks for opto-electronic or photonic devices, but only a hand full of such materials are known to date e.g. monolayers of transition metal dichalcogenides [38, 39] and black phosphorous [40]. As an example of a monolayer material that drastically changes from a metal to a direct band gap semiconductor upon distortion, we show the band structure of CdBr in Figure (6). The initial unstable metallic phase of the material becomes a dynamically stable upon distortion and opens a direct band gap of 1.28 eV at the C point.

3.3. Machine learning accelerated prediction of dynamical stability

We next attempt to accelerate the dynamic stability prediction using the machine learning model outlined in Section 2.3.

As an introductory exercise we consider the correlation between dynamical stability and six elementary materials properties, namely the energy above the convex hull, the PBE band gap, the DOS at the Fermi level, the total energy, the total energy per atom, and heat of formation. Figure 7 shows the distribution of these properties over the 3295 2D materials. The materials have been split into dynamically stable (blue) and dynamically unstable materials (orange), respectively. There is a clear correlation between dynamical stability and the first three materials properties shown in panels a-c. In particular, dynamically stable materials are closer to the convex hull, have larger band gap, and lower DOS at E_F as compared to dynamically unstable materials. The observed correlation with $\Delta H_{\rm hull}$ is consistent with previous findings based on phonon calculations[28]. In contrast, no or only weak correlation is found for the last three quantities in panels d-f. These six properties were used as a low-dimensional feature vector for training an XGBoost machine learning model that will serve as a baseline for an XGBoost model trained on the higher dimensional RAD-PDOS representation described in Sec. 2.3.

To evaluate the performance of our model we employ the receiver operating characteristic (ROC) curve. The ROC curve maps out the number of materials correctly predicted as unstable as a function of the number of materials incorrectly labelled as unstable, and it is calculated by varying the classification tolerance of the model. The area under the curve(AUC) is a measure of the performance of the classifier. Random guessing would amount to a linear

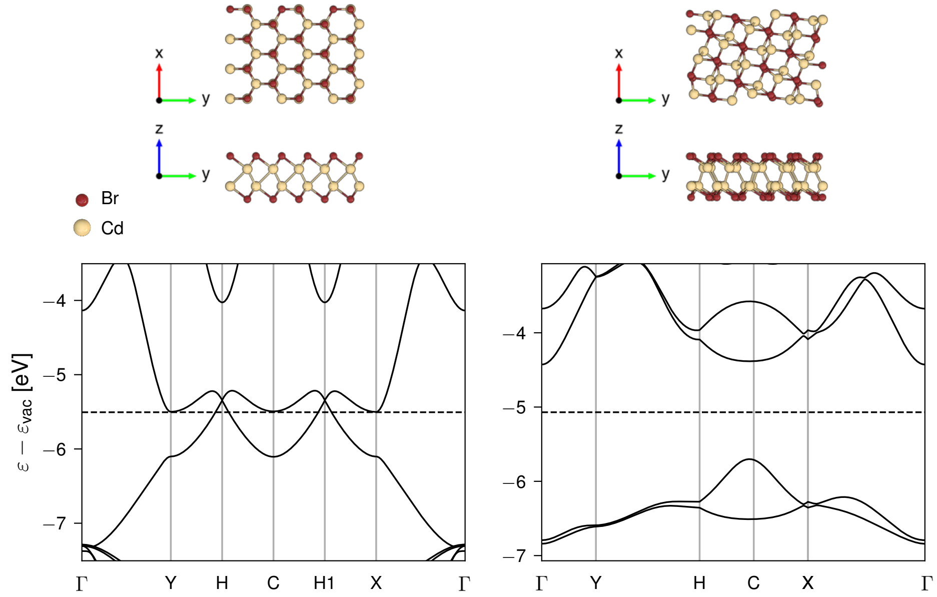


Figure 6. Extreme case of gap opening for the new stable material (CdBr) where the difference in the gap between the initial unstable metallic phase and the final structure is 1.28 eV.

ROC curve with unit slope, shown in Fig. 8 by the dashed grey line, and correspond to an AUC of 0.5 whereas a perfect classification model would have an AUC of 1. When calculating the ROC curve of our dynamic stability classifier we employ ten fold cross-validation (CV). This allows us to obtain a mean ROC curve and its standard deviation, which we then use to evaluate the performance of our model.

The results from the machine learning model is shown in Fig. 8. The mean ROC curve is shown in blue in Fig.8a; it achieves an excellent 10-fold CV AUC of 0.90 ± 0.01. This suggests that the XGBoost model is able to efficiently detect the dynamically unstable materials in the C2DB. We quantify the effect of the RAD-PDOS fingerprints by comparing the performance of the full model with a model trained on a reduced fingerprint based only on the six electronic properties: the PBE electronic band gap ($\varepsilon_{\text{gap}}^{\text{PBE}}$), crystal formation energy (ΔH), density of states at the Fermi level (DOS@E_F), energy above the convex hull (ΔH_{hull}), the total energy per atom in the unit cell and the total energy. We observe that the effect on including the RAD-PDOS in the fingerprint is statistically significant raising the AUC from 0.82 ± 0.01 to 0.90 ± 0.01. The relative impact of the RAD-PDOS fingerprints is shown in the feature importance evaluation in Fig. 8c. Here feature importance refers to how many times a feature is used to perform a split in the decision trees, and the feature importance have been summed for the six different components of the RAD-PDOS fingerprints, i.e. summing the radial distance and energy axes of the fingerprint. The vertical dashed black line shows the feature importance of random noise for reference. We observe that especially the RAD-PDOS *ss* fingerprint leads to many splits in the gradient boosted trees.

Because of the strong performance of the model, we envision that it can be deployed directly after the initial relaxation step of a high-throughput workflow to reduce the number of phonon calculations needed to remove the dynamically unstable materials. Depending on the number of stable candidates that one is willing to falsely label as unstable, it is possible to save a significant amount of phonon-calculations by pre-screening with the ML model. The willingness to sacrifice materials is controlled by the classification tolerance. The trade-off between the number of unstable materials removed and the number of stable materials lost is directly mapped out by the ROC curve. In Fig. 8b we have indicated the classification thresholds where we

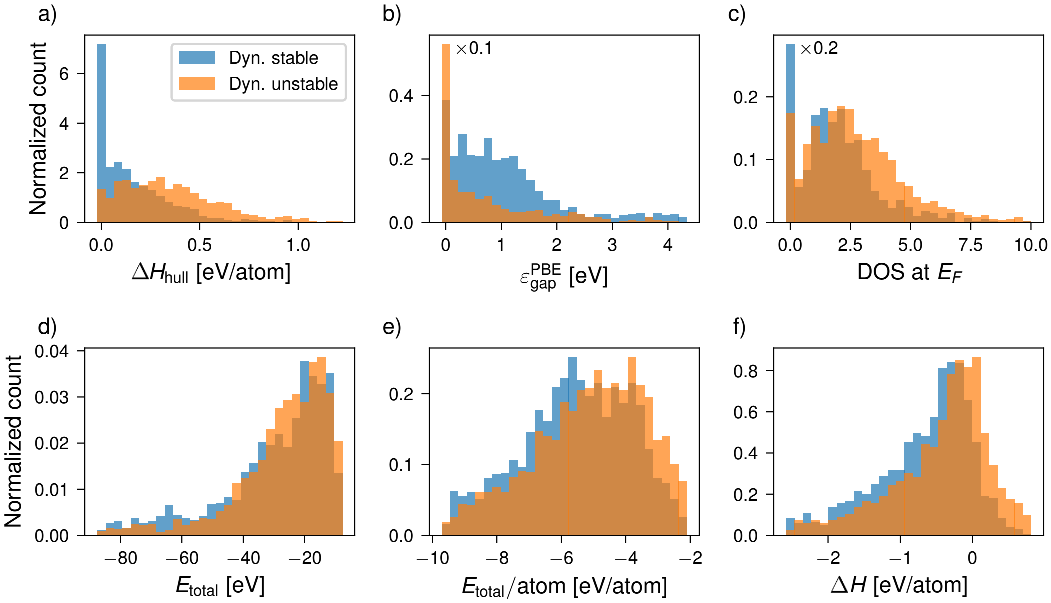


Figure 7. Histograms of electronic features for stable and unstable materials. a) shows the distribution of the energy above the convex hull for high and low stability predicted materials. The stable materials tend to be closer to the convex hull. b) shows that materials predicted to be stable more frequently have a PBE band gap larger than zero. c) shows the DOS at the Fermi level distributions, which supports that stable materials have lower DOS at E_F. d) is for the total energy while e) is for total energy per atom, which shows a slightly better seperation between stable and unstable materials as stable materials tend to have a slightly lower total energy per atom. In f) the unstable materials tend to have a slightly larger heat of formation. Note that for b) and c) the peaks at $x = 0$ for both stable and unstable materials have been scaled down by a factor of 10 and 5, respectively.

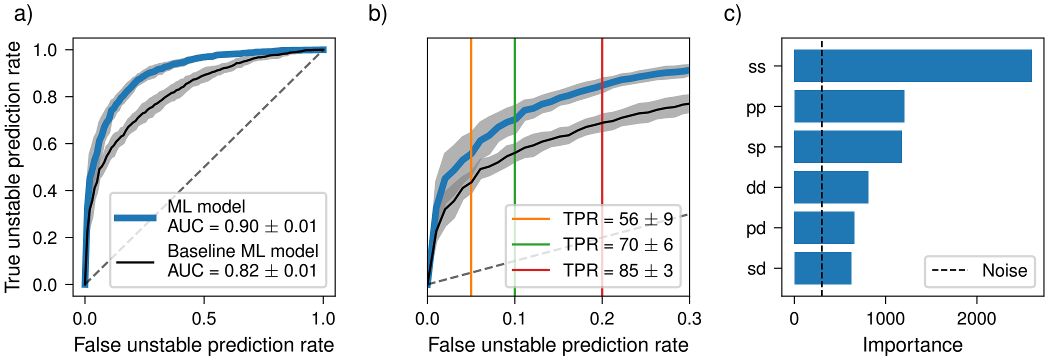


Figure 8. Machine learning results. a) shows the ROC curves for a machine learning model trained on RAD-PDOS fingerprints and a baseline model trained on electronic features from the C2DB. The AUC scores are 0.91 and 0.82, respectively. b) shows the ROC curves zoomed on low false prediction rates with different classification thresholds highlighted with vertical lines. By accepting 5% of the stable materials being falsely characterised as unstable, we can correctly label 56 ± 9% of the unstable materials. For 10% the true prediction rate is 70 ± 6% and for 20% it is 85 ± 3%. c) shows feature importances of the different RAD-PDOS fingerprints measured as how many times the fingerprints are used to perform a split in the ML model. The RAD-PDOS ss fingerprint is found to be the most important fingerprint for the ML model.

lose 5%, 10% and 20 % of the stable materials, and we observe that we can save 56±9%, 70±6% and 85±3% of the computations for the three thresholds, respectively.

As an additional test of the machine learning model we apply it to the set of the 137 dynamically unstable materials that were investigated using the CBP protocol in the first part of the paper. The dynamical stability of the materials is evaluated by the ML model both before and after being pushed along an unstable mode (recall that before the push all the 137 materials are unstable; after the push the subset of 49 materials listed in Table 1 become stable while the other materials remain unstable). It is found that before the push 56% of the unstable materials are labeled correctly. After the push, only 29% of the unstable materials are labelled as unstable while the accuracy of the stable materials are 72%. Overall, the ML model performs worse on this test set than on a randomly selected test set from the original data set. An obvious explanation is that the 137 materials were selected according to (i) low energy above the convex hull ($\Delta H_{\text{hull}} < 0.2eV/atom$) and (ii) dynamically unstable. As seen from Fig. 7a such materials are highly unusual and not well represented by the set of materials used to train the model.

4. Conclusions

In conclusion, we have performed a systematic study of structural instabilities in 2D materials. We have validated a simple protocol (here referred to as the CBP protocol) for identifying dynamical instabilities based on the frequency of phonons at the center and boundary of the BZ. The CBP protocol correctly classifies 2D materials as dynamically stable/unstable in 236 out of 250 cases[23] and is ideally suited for high-throughput studies where the computational cost of evaluating the full phonon band structure becomes prohibitive.

For 137 dynamically unstable monolayers with low formation energies, we displaced the atoms along an unstable phonon mode and relaxed the structure in a 2 × 1 or 2 × 2 supercell. This resulted in 49 distorted, dynamically stable monolayers. The success rate of obtaining a dynamically stable structure from this protocol was found to be significantly higher for materials with only one unstable phonon mode as compared to cases with several modes. In the latter case, the displacement vector is not unique and different choices generally lead to different, (dynamically unstable) structures. The 49 stable structures were fully characterised by an extensive computational property workflow and the results are available via the C2DB database. The properties of the distorted structures can deviate significantly from the original high symmetry structures, and we found only a weak, qualitative relation between the gain in total energy and band gap opening upon distortion.

Finally, we trained a machine learning classification model to predict the dynamical stability using a radially decomposed projected density of states (RAD-PDOS) representation as input and a gradient boosting decision tree ensemble method (XGBoost) as learning algorithm. The model achieves an excellent ROC-AUC score of 0.90 and lends itself to high-throughput assessment of dynamical stability.

5. Acknowledgments

The Center for Nanostructured Graphene (CNG) is sponsored by the Danish National Research Foundation, Project DNRF103. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement No 773122 (LIMA). K. S. T. is a Villum Investigator supported by VILLUM FONDEN (grant no. 37789).

References

[1] Kohn W and Sham L J 1965 *Phys. Rev.* **140**(4A) A1133–A1138

[2] Grüner G 1988 *Rev. Mod. Phys.* **60**(4) 1129–1181

[3] Weber F, Rosenkranz S, Castellan J P, Osborn R, Karapetrov G, Hott R, Heid R, Bohnen K P and Alatas A 2011 *Phys. Rev. Lett.* **107**(26) 266401 URL https://link.aps.org/doi/10.1103/PhysRevLett.107.266401

[4] Bianco R, Errea I, Monacelli L, Calandra M and Mauri F 2019 *Nano Letters* **19** 3098–3103 (*Preprint* https://doi.org/10.1021/acs.nanolett.9b00504)

[5] Zhou J S, Monacelli L, Bianco R, Errea I, Mauri F and Calandra M 2020 *Nano Letters* **20** 4809–4815 (*Preprint* https://doi.org/10.1021/acs.nanolett.0c00597)

[6] Leroux M, Le Tacon M, Calandra M, Cario L, Méasson M A, Diener P, Borrissenko E, Bosak A and Rodière P 2012 *Phys. Rev. B* **86**(15) 155125

[7] Bianco R, Errea I, Monacelli L, Calandra M and Mauri F 2019 *Nano Letters* **19** 3098–3103

[8] Calandra M, Mazin I I and Mauri F 2009 *Phys. Rev. B* **80**(24) 241108

[9] Zhu X, Cao Y, Zhang J, Plummer E W and Guo J 2015 *Proceedings of the National Academy of Sciences* **112** 2367–2371

[10] Johannes M and Mazin I 2008 *Physical Review B* **77** 165135

[11] Lin Y C, Dumcenco D O, Huang Y S and Suenaga K 2014 *Nature Nanotechnology* **9** 391–396 ISSN 1748-3395

[12] Xi X, Zhao L, Wang Z, Berger H, Forró L, Shan J and Mak K F 2015 *Nature Nanotechnology* **10** 765–769

[13] Ugeda M, Bradley A, Zhang Y, Onishi S, Chen Y, Ruan W, Ojeda-Aristizabal C, Ryu H, Edmonds M, Tsai H Z, Riss A, Mo S, Lee D, Zettl A, Hussain Z, Shen Z X and Crommie M 2016 *Nature Physics* **12** 92–97 ISSN 1745-2473

[14] Yang Y, Fang S, Fatemi V, Ruhman J, Navarro-Moratalla E, Watanabe K, Taniguchi T, Kaxiras E and Jarillo-Herrero P 2018 *Phys. Rev. B* **98**(3) 035203

[15] Yu Y, Yang F, Lu X F, Yan Y J, Cho Y H, Ma L, Niu X, Kim S, Son Y W, Feng D, Li S, Cheong S W, Chen X H and Zhang Y 2015 *Nature Nanotechnology* **10** 270–276 ISSN 1748-3395

[16] Ryu H, Chen Y, Kim H, Tsai H Z, Tang S, Jiang J, Liou F, Kahn S, Jia C, Omrani A A, Shim J H, Hussain Z, Shen Z X, Kim K, Min B I, Hwang C, Crommie M F and Mo S K 2018 *Nano Letters* **18** 689–694

[17] Ge Y and Liu A Y 2012 *Phys. Rev. B* **86**(10) 104101

[18] Sugawara K, Nakata Y, Shimizu R, Han P, Hitosugi T, Sato T and Takahashi T 2016 *ACS Nano* **10** 1341–1345

[19] Wang H, Chen Y, Duchamp M, Zeng Q, Wang X, Tsang S H, Li H, Jing L, Yu T, Teo E H T and Liu Z 2018 *Advanced Materials* **30** 1704382

[20] Zhu J, Wang Z, Yu H, Li N, Zhang J, Meng J, Liao M, Zhao J, Lu X, Du L, Yang R, Shi D, Jiang Y and Zhang G 2017 *Journal of the American Chemical Society* **139** 10216–10219

[21] Wang L, Xu Z, Wang W and Bai X 2014 *Journal of the American Chemical Society* **136** 6693–6697

[22] Krishnamoorthy A, Bassman L, Kalia R K, Nakano A, Shimojo F and Vashishta P 2018 *Nanoscale* **10**(6) 2742–2747

[23] Mounet N, Gibertini M, Schwaller P, Campi D, Merkys A, Marrazzo A, Sohier T, Castelli I E, Cepellotti A, Pizzi G and Marzari N 2018 *Nature Nanotechnology* **13** 246–252 ISSN 1748-3395

[24] Patrick C E, Jacobsen K W and Thygesen K S 2015 *Phys. Rev. B* **92**(20) 201205

[25] Yang R X, Skelton J M, da Silva E L, Frost J M and Walsh A 2020 *The Journal of Chemical Physics* **152** 024703

[26] Kayastha P and Ramakrishnan R 2021 *The Journal of Chemical Physics* **154** 061102

[27] Haastrup S, Strange M, Pandey M, Deilmann T, Schmidt P S, Hinsche N F, Gjerding M N, Torelli D, Larsen P M, Riis-Jensen A C, Gath J, Jacobsen K W, Mortensen J J, Olsen T and Thygesen K S 2018 *2D Materials* **5** 042002

[28] Gjerding M, Taghizadeh A, Rasmussen A, Ali S, Bertoldo F, Deilmann T, Holguin U, Knøsgaard N, Kruse M, Manti S *et al.* 2021 *2D Materials* **8** 044002

[29] Gjerding M, Skovhus T, Rasmussen A, Bertoldo F, Larsen A H, Mortensen J J and Thygesen K S 2021 *Computational Materials Science* **199** 110731

[30] Larsen A H, Mortensen J J, Blomqvist J, Castelli I E, Christensen R, Dułak M, Friis J, Groves M N, Hammer B, Hargus C *et al.* 2017 *Journal of Physics: Condensed Matter* **29** 273002

[31] Togo A and Tanaka I 2015 *Scripta Materialia* **108** 1 – 5 ISSN 1359-6462

[32] Enkovaara J, Rostgaard C, Mortensen J J, Chen J, Dułak M, Ferrighi L, Gavnholt J, Glinsvad C, Haikola V, Hansen H A, Kristoffersen H H, Kuisma M, Larsen A H, Lehtovaara L, Ljungberg M, Lopez-Acevedo O, Moses P G, Ojanen J, Olsen T, Petzold V, Romero N A, Stausholm-Møller J, Strange M, Tritsaris G A, Vanin M, Walter M, Hammer B, Häkkinen H, Madsen G K H, Nieminen R M, Nørskov J K, Puska M, Rantala T T, Schiøtz J, Thygesen K S and Jacobsen K W 2010 *Journal of Physics: Condensed Matter* **22** 253202 URL `https://doi.org/10.1088%2F0953-8984%2F22%2F25%2F253202`

[33] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77**(18) 3865–3868

[34] `https://cmr.fysik.dtu.dk/c2db/c2db.html`

[35] Chen T and Guestrin C 2016 Xgboost: A scalable tree boosting system *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '16 (New York, NY, USA: Association for Computing Machinery) p 785–794 ISBN 9781450342322 URL `https://doi.org/10.1145/2939672.2939785`

[36] Eriksson F, Fransson E and Erhart P *Advanced Theory and Simulations* **2** 1800184

[37] Rossnagel K 2011 *Journal of Physics: Condensed Matter* **23** 213001 URL `https://doi.org/10.1088/0953-8984/23/21/213001`

[38] Mak K F, Lee C, Hone J, Shan J and Heinz T F 2010 *Physical Review Letters* **105** 136805

[39] Manzeli S, Ovchinnikov D, Pasquier D, Yazyev O V and Kis A 2017 *Nature Reviews Materials* **2** 1–15

[40] Liu H, Neal A T, Zhu Z, Luo Z, Xu X, Tománek D and Ye P D 2014 *ACS Nano* **8** 4033–4041

8.3 Publication 3: Recent progress of the computational 2D materials database (C2DB)

Recent progress of the Computational 2D Materials Database (C2DB)

To cite this article: Morten Niklas Gjerding *et al* 2021 *2D Mater.* **8** 044002

2D Materials

RECEIVED
18 January 2021

REVISED
31 May 2021

ACCEPTED FOR PUBLICATION
30 June 2021

PUBLISHED
15 July 2021

PAPER

Recent progress of the Computational 2D Materials Database (C2DB)

Morten Niklas Gjerding[1], Alireza Taghizadeh[1,2], Asbjørn Rasmussen[1], Sajid Ali[1], Fabian Bertoldo[1], Thorsten Deilmann[3], Nikolaj Rørbæk Knøsgaard[1], Mads Kruse[1], Ask Hjorth Larsen[1], Simone Manti[1], Thomas Garm Pedersen[2], Urko Petralanda[1], Thorbjørn Skovhus[1], Mark Kamper Svendsen[1], Jens Jørgen Mortensen[1], Thomas Olsen[1] and Kristian Sommer Thygesen[1,*]

[1] Computational Atomic-scale Materials Design (CAMD), Department of Physics, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
[2] Department of Materials and Production, Aalborg University, 9220 Aalborg Øst, Denmark
[3] Institut für Festkörpertheorie, Westfälische Wilhelms-Universität Münster, 48149 Münster, Germany
[*] Author to whom any correspondence should be addressed.

E-mail: thygesen@fysik.dtu.dk

Keywords: 2D materials, high-throughput, ab-initio, database, density functional theory

Abstract

The Computational 2D Materials Database (C2DB) is a highly curated open database organising a wealth of computed properties for more than 4000 atomically thin two-dimensional (2D) materials. Here we report on new materials and properties that were added to the database since its first release in 2018. The set of new materials comprise several hundred monolayers exfoliated from experimentally known layered bulk materials, (homo)bilayers in various stacking configurations, native point defects in semiconducting monolayers, and chalcogen/halogen Janus monolayers. The new properties include exfoliation energies, Bader charges, spontaneous polarisations, Born charges, infrared polarisabilities, piezoelectric tensors, band topology invariants, exchange couplings, Raman spectra and second harmonic generation spectra. We also describe refinements of the employed material classification schemes, upgrades of the computational methodologies used for property evaluations, as well as significant enhancements of the data documentation and provenance. Finally, we explore the performance of Gaussian process-based regression for efficient prediction of mechanical and electronic materials properties. The combination of open access, detailed documentation, and extremely rich materials property data sets make the C2DB a unique resource that will advance the science of atomically thin materials.

1. Introduction

The discovery of new materials, or new properties of known materials, to meet a specific industrial or scientific requirement, is an exciting intellectual challenge of the utmost importance for our environment and economy. For example, the successful transition to a society based on sustainable energy sources and the realisation of quantum technologies (e.g. quantum computers and quantum communication) depend critically on new materials with novel functionalities. First-principles quantum mechanical calculations, e.g. based on density functional theory (DFT) [1], can predict the properties of materials with high accuracy even before they are made in the lab. They provide insight into mechanisms at the most fundamental (atomic and electronic) level and can pinpoint and calculate key properties that determine the performance of the material at the macroscopic level. Powered by high-performance computers, atomistic quantum calculations in combination with data science approaches, have the potential to revolutionise the way we discover and develop new materials.

Atomically thin, two-dimensional (2D) crystals represent a fascinating class of materials with exciting perspectives for both fundamental science and technology [2–5]. The family of 2D materials has been growing steadily over the past decade and counts about a hundred materials that have been realised in single-layer or few-layer form [6–10]. While some of these materials, including graphene, hexagonal boron nitride (hBN), and transition metal dichalcogenides (TMDCs), have been extensively studied, the majority have only been scarcely characterised and remain poorly understood. Computational studies indicate that around 1000 already known layered crystals have sufficiently weak interlayer (IL) bonding to allow the individual layers to be mechanically exfoliated [11, 12]. Supposedly, even more 2D materials could be realised beyond this set of already known crystals. Adding to this the possibility of stacking individual 2D layers (of the same or different kinds) into ultrathin van der Waals (vdW) crystals [13], and tuning the properties of such structures by varying the relative twist angle between adjacent layers [14, 15] or intercalating atoms into the vdW gap [16, 17], it is clear that the prospects of tailor made 2D materials are simply immense. To support experimental efforts and navigate the vast 2D materials space, first-principles calculations play a pivotal role. In particular, FAIR[5] [18] databases populated by high-throughput calculations can provide a convenient overview of known materials and point to new promising materials with desired (predicted) properties. Such databases are also a fundamental requirement for the successful introduction and deployment of artificial intelligence in materials science.

Many of the unique properties exhibited by 2D materials have their origin in quantum confinement and reduced dielectric screening. These effects tend to enhance many-body interactions and lead to profoundly new phenomena such as strongly bound excitons [19–21] with nonhydrogenic Rydberg series [22–24], phonons and plasmons with anomalous dispersion relations [25, 26], large dielectric band structure renormalisations [27, 28], unconventional Mott insulating and superconducting phases [14, 15], and high-temperature exciton condensates [29]. Recently, it has become clear that long range magnetic order can persist [30, 31] and (in-plane) ferroelectricity even be enhanced [32], in the single layer limit. In addition, first-principles studies of 2D crystals have revealed rich and abundant topological phases [33, 34]. The peculiar physics ruling the world of 2D materials entails that many of the conventional theories and concepts developed for bulk crystals break down or require special treatments when applied to 2D materials [26, 35, 36]. This means that computational studies must be performed with extra care, which in turn calls for well-organised and well-documented 2D property data sets that can form the basis for the development, benchmarking, and consolidation of physical theories and numerical implementations.

The Computational 2D Materials Database (C2DB) [6, 37] is a highly curated and fully open database containing elementary physical properties of around 4000 2D monolayer crystals. The data has been generated by automatic high-throughput calculations at the level of DFT and many-body perturbation theory as implemented in the GPAW [38, 39] electronic structure code. The computational workflow is constructed using the atomic simulation recipes (ASR) [40]—a recently developed Python framework for high-throughput materials modelling building on the atomic simulation environment (ASE) [41]—and managed/executed using the MyQueue task scheduler [42].

The C2DB differentiates itself from existing computational databases of bulk [43–45] and low-dimensional [11, 12, 46–50] materials, by the large number of physical properties available, see table 1. The use of beyond-DFT theories for excited state properties (GW band structures and Bethe–Salpeter equation (BSE) absorption for selected materials) and Berry-phase techniques for band topology and polarisation quantities (spontaneous polarisation, Born charges, piezoelectric tensors), are other unique features of the database.

The C2DB can be downloaded in its entirety or browsed and searched online. As a new feature, all data entries presented on the website are accompanied by a clickable help icon that presents a scientific documentation ('what does this piece of data describe?') and technical documentation ('how was this piece of data computed?'). This development enhances the usability of the database and improves the reproducibility and provenance of the data contained in C2DB. As another novelty it is possible to download all property data pertaining to a specific material or a specific type of property, e.g. the band gap, for all materials thus significantly improving data accessibility.

In this paper, we report on the significant C2DB developments that have taken place during the past two years. These developments can be roughly divided into four categories: (1) General updates of the workflow used to select, classify, and stability assess the materials. (2) Computational improvements for properties already described in the 2018 paper. (3) New properties. (4) New materials. The developments, described in four separate sections, cover both original work and review of previously published work. In addition, we have included some outlook discussions of ongoing work. In the last section we illustrate an application of statistical learning to predict properties directly from the atomic structure.

2. Selection, classification, and stability

Figure 1 illustrates the workflow behind the C2DB. In this section we describe the first part of the workflow

[5] FAIR data are data which meet principles of findability, accessibility, interoperability, and reusability.

Table 1. Properties calculated by the C2DB monolayer workflow. The computational method and the criteria used to decide whether the property should be evaluation for a given material is also shown. A '*' indicates that spin–orbit coupling (SOC) is included. All calculations are performed with the GPAW code using a plane wave basis except for the Raman calculations, which employ a double-zeta polarised basis of numerical atomic orbitals [51].

Property	Method	Criteria	Count
Bader charges	PBE	None	3809
Energy above convex hull	PBE	None	4044
Heat of formation	PBE	None	4044
Orbital projected band structure	PBE	None	2487
Out-of-plane dipole	PBE	None	4044
Phonons (Γ and BZ corners)	PBE	None	3865
Projected density of states	PBE	None	3332
Stiffness tensor	PBE	None	3968
Exchange couplings	PBE	Magnetic	538
Infrared polarisability	PBE	$E_{\text{gap}}^{\text{PBE}} > 0$	784
Second harmonic generation	PBE	$E_{\text{gap}}^{\text{PBE}} > 0$, non-magnetic, non-centrosymmetric	375
Electronic band structure PBE	PBE*	None	3496
Magnetic anisotropies	PBE*	Magnetic	823
Deformation potentials	PBE*	$E_{\text{gap}}^{\text{PBE}} > 0$	830
Effective masses	PBE*	$E_{\text{gap}}^{\text{PBE}} > 0$	1272
Fermi surface	PBE*	$E_{\text{gap}}^{\text{PBE}} = 0$	2505
Plasma frequency	PBE*	$E_{\text{gap}}^{\text{PBE}} = 0$	3144
Work function	PBE*	$E_{\text{gap}}^{\text{PBE}} = 0$	4044
Optical polarisability	RPA@PBE	None	3127
Electronic band structure	HSE06@PBE*	None	3155
Electronic band structure	G_0W_0@PBE*	$E_{\text{gap}}^{\text{PBE}} > 0$, $N_{\text{atoms}} < 5$	357
Born charges	PBE, Berry phase	$E_{\text{gap}}^{\text{PBE}} > 0$	639
Raman spectrum	PBE, LCAO basis set	Non-magnetic, dyn. stable	708
Piezoelectric tensor	PBE, Berry phase	$E_{\text{gap}}^{\text{PBE}}$, non-centrosym.	353
Optical absorbance	BSE@G_0W_0*	$E_{\text{gap}}^{\text{PBE}} > 0$, $N_{\text{atoms}} < 5$	378
Spontaneous polarisation	PBE, Berry phase	$E_{\text{gap}}^{\text{PBE}} > 0$, nearly centrosym. polar space group	151
Topological invariants	PBE*, Berry phase	$0 < E_{\text{gap}}^{\text{PBE}} < 0.3$ eV	242

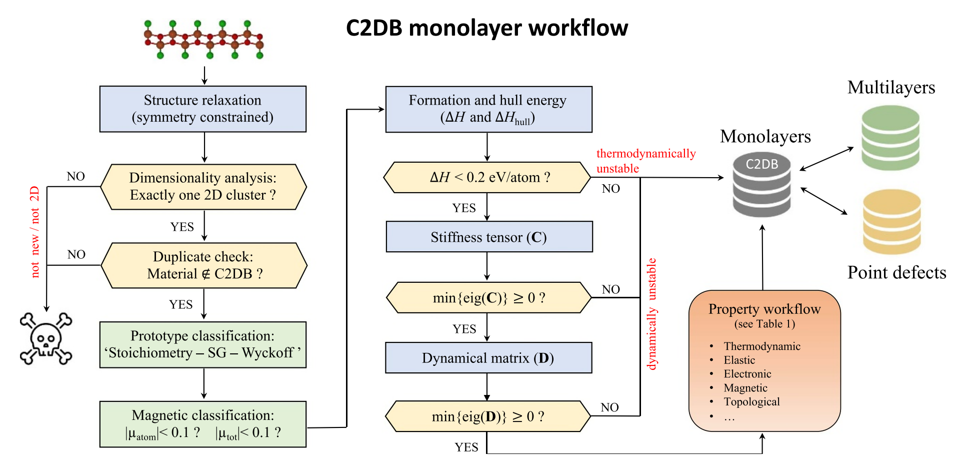

C2DB monolayer workflow
Structure relaxation (symmetry constrained)
Dimensionality analysis: Exactly one 2D cluster ?
NO
YES
Duplicate check: Material ∉ C2DB ?
NO
YES
not new / not 2D
Prototype classification: 'Stoichiometry – SG – Wyckoff'
Magnetic classification: $|\mu_{\text{atom}}| < 0.1$? $|\mu_{\text{tot}}| < 0.1$?
Formation and hull energy (ΔH and ΔH_{hull})
$\Delta H < 0.2$ eV/atom ?
NO
YES
thermodynamically unstable
Stiffness tensor (C)
min{eig(C)} ≥ 0 ?
NO
YES
Dynamical matrix (D)
min{eig(D)} ≥ 0 ?
NO
YES
dynamically unstable
Monolayers
C2DB
Multilayers
Point defects
Property workflow
(see Table 1)
Thermodynamic
Elastic
Electronic
Magnetic
Topological
…

Figure 1. The workflow behind the C2DB. After the structural relaxation, the dimensionality of the material is checked and it is verified that the material is not already present in the database. Next, the material is classified according to its chemical composition, crystal structure, and magnetic state. Finally, the thermodynamic and dynamic stabilities are assessed from the energy above the convex hull and the sign of the minimum eigenvalues of the dynamical matrix and stiffness tensor. Unstable materials are stored in the database; stable materials are subject to the property workflow. The C2DB monolayer database is interlinked with databases containing structures and properties of multilayer stacks and point defects in monolayers from the C2DB.

until the property calculations (red box), focusing on aspects related to selection criteria, classification, and stability assessment, that have been changed or updated since the 2018 paper.

2.1. Structure relaxation

Given a prospective 2D material, the first step is to carry out a structure optimisation. This calculation is performed with spin polarisation and with the symmetries of the original structure enforced. The latter is done to keep the highest level of control over the resulting structure by avoiding 'uncontrolled' symmetry breaking distortions. The prize to pay is a higher risk of generating dynamically unstable structures.

2.2. Selection: dimensionality analysis

A dimensionality analysis [52] is performed to identify and filter out materials that have disintegrated into non-2D structures during relaxation. Covalently bonded clusters are identified through an analysis of the connectivity of the structures where two atoms are considered to belong to the same cluster if their distance is less than some scaling of the sum of their covalent radii, i.e. $d < k(r_i^{\mathrm{cov}} + r_j^{\mathrm{cov}})$, where i and j are atomic indices. A scaling factor of $k = 1.35$ was determined empirically. Only structures that consist of a single 2D cluster after relaxation are further processed. Figure 2 shows three examples (graphene, $\mathrm{Ge_2Se_2}$, and $\mathrm{Pb_2O_6}$) of structures and their cluster dimensionalities before and after relaxation. All structures initially consist of a single 2D cluster, but upon relaxation $\mathrm{Ge_2Se_2}$ and $\mathrm{Pb_2O_6}$ disintegrate into two 2D clusters as well as one 2D and two 0D clusters, respectively. On the other hand, the relaxation of graphene decreases the in-plane lattice constant but does not affect the dimensionality. According to the criterion defined above only graphene will enter the database.

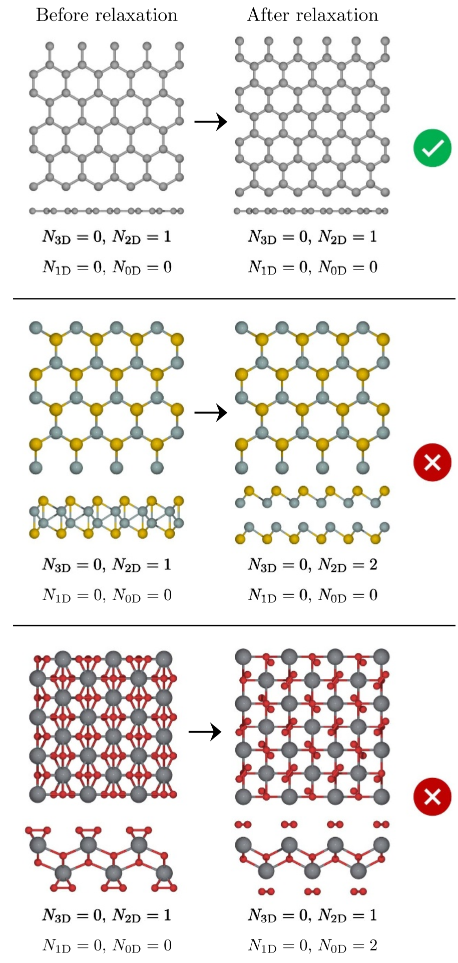

Before relaxation
After relaxation
$N_{3D} = 0,\ N_{2D} = 1$
$N_{1D} = 0,\ N_{0D} = 0$
$N_{3D} = 0,\ N_{2D} = 1$
$N_{1D} = 0,\ N_{0D} = 0$
✓
$N_{3D} = 0,\ N_{2D} = 1$
$N_{1D} = 0,\ N_{0D} = 0$
$N_{3D} = 0,\ N_{2D} = 2$
$N_{1D} = 0,\ N_{0D} = 0$
✗
$N_{3D} = 0,\ N_{2D} = 1$
$N_{1D} = 0,\ N_{0D} = 0$
$N_{3D} = 0,\ N_{2D} = 1$
$N_{1D} = 0,\ N_{0D} = 2$
✗

Figure 2. Three example structures from C2DB (top: graphene, middle: $\mathrm{Ge_2Se_2}$, bottom: $\mathrm{Pb_2O_6}$) with their respective cluster dimensionalities cluster before (left) and after (right) relaxation. The number $N_{x\mathrm{D}}$ denotes the number of clusters of dimensionality x. Note that the number of atoms of the structures depicted in the left and right columns can differ because the relaxation can change the lattice constants.

2.3. Selection: ranking similar structures

Maintaining a high-throughput database inevitably requires a strategy for comparing similar structures and ranking them according to their relevance. In particular, this is necessary in order to identify different representatives of the same material e.g. resulting from independent relaxations, and thereby avoid duplicate entries and redundant computations. The C2DB strategy to this end involves a combination of structure clustering and Pareto analysis.

First, a single-linkage clustering algorithm is used to group materials with identical reduced chemical formula and 'similar' atomic configurations. To quantify configuration similarity a slightly modified version of PyMatGen's [53] distance metric is employed where the cell volume normalisation is removed to make it applicable to 2D materials surrounded by vacuum. Roughly speaking, the metric measures the maximum distance an atom must be moved (in units of Å) in order to match the two atomic configurations. Two atomic configurations belong to the same cluster if their distance is below an empirically determined threshold of 0.3 Å.

At this point, the simplest strategy would be to remove all but the most stable compound within a cluster. However, this procedure would remove many high symmetry crystals for which a more stable distorted version exists. For example, the well known T-phase of $\mathrm{MoS_2}$ would be removed in favour of the more stable T′-phase. This is undesired as high-symmetry structures, even if dynamically unstable at $T = 0$, may provide useful information and might in fact become stabilised at higher temperatures [54]. Therefore, the general strategy adopted for the C2DB,

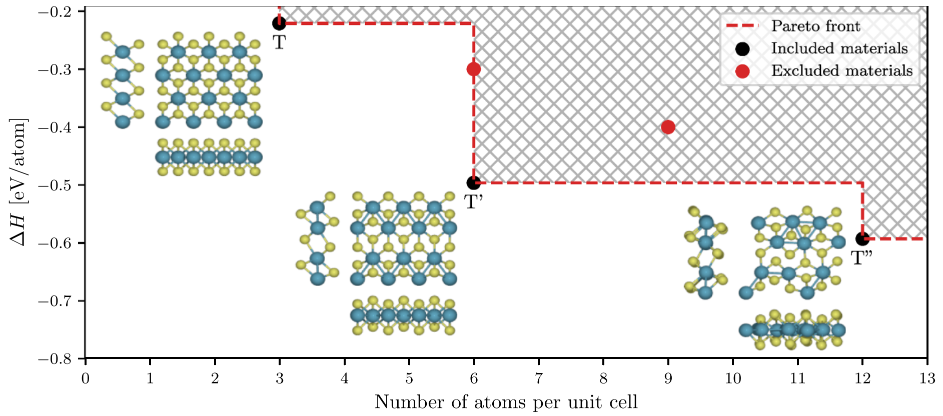


Figure 3. Illustration of the Pareto analysis used to filter out duplicates or irrelevant structures from the C2DB. All points represent materials with the same reduced chemical formula (in this case ReS_2) that belong to the same cluster defined by the structure metric. Only structures lying on the $(N, \Delta H)$-Pareto front are retained (black circles) while other materials are excluded (red circles). The philosophy behind the algorithm is to keep less stable materials if they contain fewer atoms per unit cell than more stable materials and thus represent structures of higher symmetry.

is to keep a material that is less stable than another material of the same cluster if it has fewer atoms in its primitive unit cell (and thus typically higher symmetry). Precisely, materials within a given cluster are kept only if they represent a defining point of the $(N, \Delta H)$-Pareto front, where N is the number of atoms in the unit cell and ΔH is the heat of formation. A graphical illustration of the Pareto analysis is shown in figure 3 for the case of ReS_2.

2.4. Classification: crystal structure

The original C2DB employed a *crystal prototype* classification scheme where specific materials were promoted to prototypes and used to label groups of materials with the same or very similar crystal structure. This approach was found to be difficult to maintain (as well as being non-transparent). Instead, materials are now classified according to their *crystal type* defined by the reduced stoichiometry, space group number, and the alphabetically sorted labels of the occupied Wyckoff positions. As an example, MoS_2 in the H-phase has the crystal type: AB2-187-bi.

2.5. Classification: magnetic state

In the new version of the C2DB, materials are classified according to their magnetic state as either *non-magnetic* or *magnetic*. A material is considered magnetic if any atom has a local magnetic moment greater than 0.1 μ_B.

In the original C2DB, the *magnetic* category was further subdivided into ferromagnetic (FM) and antiferromagnetic (AFM). But since the simplest antiferromagnetically ordered state typically does not represent the true ground state, all material entries with an AFM state have been removed from the C2DB and replaced by the material in its FM state. Although the latter is less stable, it represents a more well defined state of the material. Crucially, the nearest neighbour exchange couplings for all magnetic materials have been included in the C2DB (see section 5.8). This enables a more detailed and realistic description of the magnetic order via the Heisenberg model. In particular, the FM state of a material is not expected to represent the true magnetic ground if the exchange coupling $J < 0$.

2.6. Stability: thermodynamic

The heat of formation, ΔH, of a compound is defined as its energy per atom relative to its constituent elements in their standard states [55]. The thermodynamic stability of a compound is evaluated in terms of its energy above the *convex hull*, ΔH_{hull}, which gives the energy of the material relative to other competing phases of the same chemical composition, including mixed phases [6], see figure 4 for an example. Clearly, ΔH_{hull} depends on the pool of reference phases, which in turn defines the convex hull. The original C2DB employed a pool of reference phases comprised by 2807 elemental and binary bulk crystals from the convex hull of the Open Quantum Materials Database (OQMD) [55]. In the new version, this set has been extended by approximately 6783 ternary bulk compounds from the convex hull of OQMD, making a total of 9590 stable bulk reference compounds.

As a simple indicator for the thermodynamic stability of a material, the C2DB employs three labels (low, medium, high) as defined in table 2. These indicators are unchanged from the original version of the C2DB. In particular, the criterion $\Delta H_{\text{hull}} < 0.2$ eV atom^{-1}, defining the most stable category, was established based on an extensive analysis of 55 experimentally realised monolayer crystals [6].

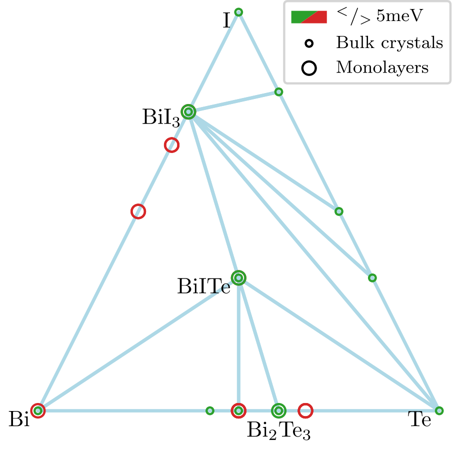


Figure 4. Convex hull diagram for (Bi,I,Te)-compounds. Green (red) colouring indicate materials that have a convex hull energy of less than (greater than) 5 meV. The monolayers BiI$_3$, Bi$_2$Te$_3$ and BiITe lie on the convex hull. The monolayers are degenerate with their layered bulk parent because the vdW interactions are not captured by the PBE xc-functional.

Table 2. Thermodynamic stability indicator assigned to all materials in the C2DB. ΔH and ΔH_{hull} denote the heat of formation and energy above the convex hull, respectively.

Thermodynamic stability indicator	Criterion (eV atom^{-1})
Low	$\Delta H > 0.2$
Medium	$\Delta H < 0.2$ and $\Delta H_{\text{hull}} > 0.2$
High	$\Delta H < 0.2$ and $\Delta H_{\text{hull}} < 0.2$

It should be emphasised that the energies of both monolayers and bulk reference crystals are calculated with the Perdew-Burke-Ernzerhof (PBE) xc-functional [56]. This implies that some inaccuracies must be expected, in particular for materials with strongly localised d-electrons, e.g. certain transition metal oxides, and materials for which dispersive interactions are important, e.g. layered van der Waals crystals. The latter implies that the energy of a monolayer and its layered bulk parent (if such exists in the pool of references) will have the same energy. For further details and discussions see reference [6].

2.7. Stability: dynamical

Dynamically stable materials are situated at a local minimum of the potential energy surface and are thus stable to small structural perturbations. Structures resulting from DFT relaxations can end up in saddle point configurations because of imposed symmetry constraints or an insufficient number of atoms in the unit cell.

In C2DB, the dynamical stability is assessed from the signs of the minimum eigenvalues of (1) the stiffness tensor (see section 3.1) and (2) the Γ-point Hessian matrix for a supercell containing 2 × 2 repetitions of the unit cell (the structure is not relaxed in the 2 × 2 supercell). If one of these minimal eigenvalues is negative the material is classified as dynamically unstable. This indicates that the energy can be reduced by displacing an atom and/or deforming the unit cell, respectively. The use of two categories for dynamical stability, i.e. stable/unstable, differs from the original version of the C2DB where an intermediate category was used for materials with negative but numerically small minimal eigenvalue of either the Hessian or stiffness tensors.

3. Improved property methodology

The new version of the C2DB has been generated using a significantly extended and improved workflow for property evaluations. This section focuses on improvements relating to properties that were already present in the original version of the C2DB while new properties are discussed in the next section.

3.1. Stiffness tensor

The stiffness tensor, C, is a rank-4 tensor that relates the stress of a material to the applied strain. In Mandel notation (a variant of Voigt notation) C is expressed as an $N \times N$ matrix relating the N independent components of the stress and strain tensors. For a 2D material $N = 3$ and the tensor takes the form:

$$\mathbf{C} = \begin{bmatrix} C_{xxxx} & C_{xxyy} & \sqrt{2}C_{xxxy} \\ C_{xxyy} & C_{yyyy} & \sqrt{2}C_{yyxy} \\ \sqrt{2}C_{xxxy} & \sqrt{2}C_{yyxy} & 2C_{xyxy} \end{bmatrix}, \qquad (1)$$

where the indices on the matrix elements refer to the rank-4 tensor. The factors multiplying the tensor elements account for their multiplicities in the full rank-4 tensor. In the C2DB workflow, C is calculated as a finite difference of the stress under an applied strain with full relaxation of atomic coordinates. A negative eigenvalue of C signals a dynamical instability, see section 2.7.

In the first version of the C2DB only the diagonal elements of the stiffness tensor were calculated. The new version also determines the shear components such that the full 3 × 3 stiffness tensor is now available. This improvement also leads to a more accurate assessment of dynamical stability [57].

3.2. Effective masses with parabolicity estimates

For all materials with a finite band gap the effective masses of electrons and holes are calculated for bands within 100 meV of the conduction band minimum and valence band maximum, respectively. The Hessian matrices at the band extrema (BE) are determined by fitting a second order polynomium to the PBE band structure including SOC, and the effective masses are obtained by subsequent diagonalisation of the Hessian. The main fitting-procedure is unaltered

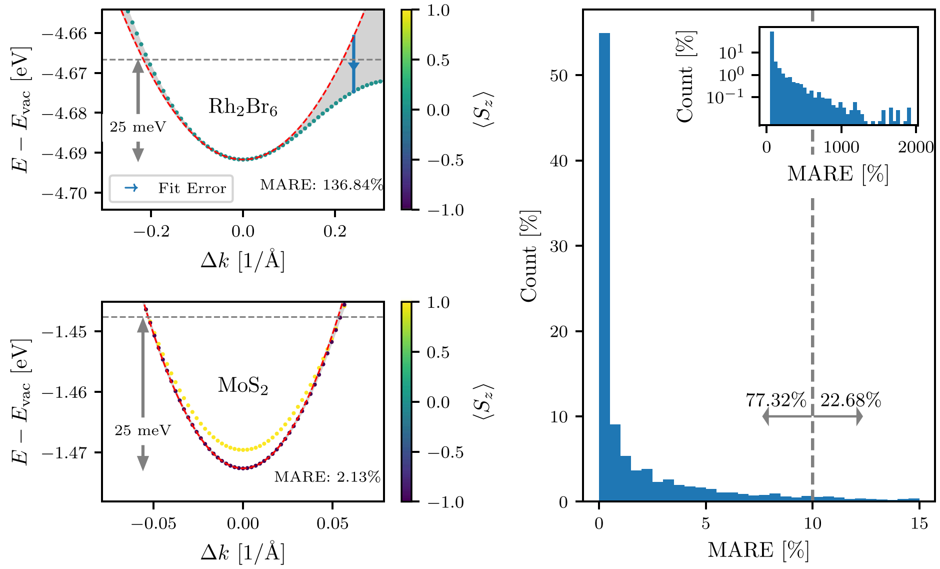


Figure 5. Left: The PBE band structures of Rh$_2$Br$_6$ and MoS$_2$ (coloured dots) in regions around the conduction band minimum. The dashed red line shows the fit made to estimate the effective masses of the lowest conduction band. The shaded grey region highlights the error between the fit and the true band structure. The mean absolute relative error (MARE) discussed in the main text is calculated for energies within 25 meV of the band minimum. For MoS$_2$ the fit is essentially on top of the band energies. Right: The distribution of the MARE of all effective mass fits in the C2DB. The inset shows the full distribution on a log scale. As mentioned in the main text, very large MAREs indicate that the band minimum/maximum was incorrectly identified by the algorithm and/or that the band is very flat. Only three materials have MAREs > 1000% but these each have several bands for which the fit fails.

from the first version of C2DB, but two important improvements have been made.

The first improvement consists in an additional k-mesh refinement step for better localisation of the BE in the Brillouin zone. After the location of the BE has been estimated based on a uniformly sampled band structure with k-point density of 12 Å, another one-shot calculation is performed with a denser k-mesh around the estimated BE positions. This ensures a more accurate and robust determination of the location of the BE, which can be important in cases with a small but still significant spin–orbit splitting or when the band is very flat or non-quadratic around the BE. The second refinement step is the same as in the first version of C2DB, i.e. the band energies are calculated on a highly dense k-mesh in a small disc around the BE, and the Hessian is obtained by fitting the band energies in the range up to 1 meV from the BE.

The second improvement is the calculation of the mean absolute relative error (MARE) of the polynomial fit in a 25 meV range from the BE. The value of 25 meV corresponds to the thermal energy at room temperature and is thus the relevant energy scale for many applications. To make the MARE independent of the absolute position of the band we calculate the average energy of the band over the 25 meV and compare the deviation of the fit to this energy scale. The MARE provides a useful measure of the parabolicity of the energy bands and thus the validity of the effective mass approximation over this energy scale.

Figure 5 shows two examples of band structures with the effective mass fits and corresponding fit errors indicated. Additionally, the distribution of MARE for all the effective mass fits in the C2DB are presented. Most materials have an insignificant MARE, but a few materials have very large errors. Materials with a MARE above a few hundreds of percentages fall into two classes. For some materials the algorithm does not correctly find the position of the BE. An example is Ti$_2$S$_2$ in the space group C2/m. For others, the fit and BE location are both correct, but the band flattens away from the BE which leads to a large MARE as is the case for Rh$_2$Br$_6$ shown in the figure or Cl$_2$Tl$_2$ in the space group P-1. In general a small MARE indicates a parabolic band while materials with large MARE should be handled on a case-by-case basis.

3.3. Orbital projected band structure

To facilitate a state-specific analysis of the PBE Kohn–Sham wave functions, an orbital projected band structure (PBS) is provided to complement the projected density of states (PDOS). In the PAW methodology, the all-electron wave functions are projected onto atomic orbitals inside the augmentation spheres centred at the position of each atom. The PBS resolves these atomic orbital contributions to the

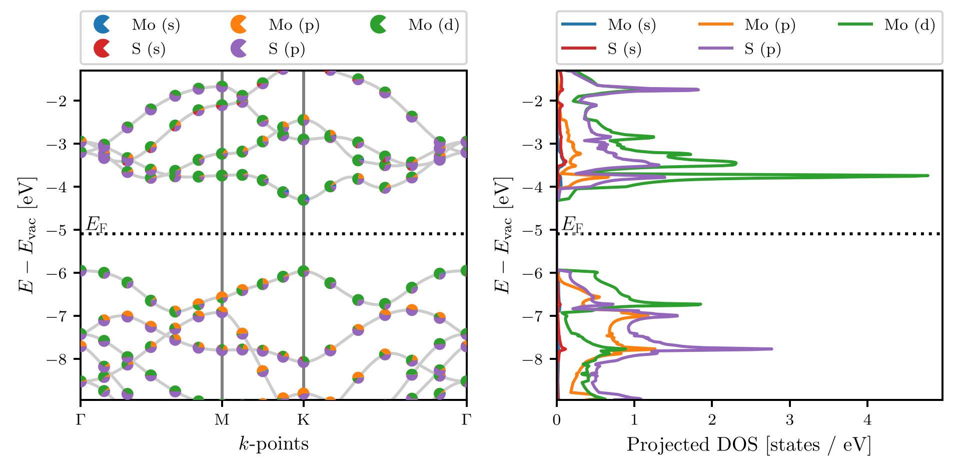


Figure 6. Orbital projected band structure and orbital projected density of states of MoS_2 in the H-phase. The pie chart symbols indicate the fractional atomic orbital character of the Kohn–Sham wave functions.

wave functions as a function of band and k-point whereas the PDOS resolves the atomic orbital character of the total density of states as a function of energy. The SOC is not included in the PBS or PDOS, as its effect is separately visualised by the spin-PBS also available in the C2DB.

As an example, figure 6 shows the PBS (left) and PDOS (right) of monolayer MoS_2 calculated with PBE. The relative orbital contribution to a given Bloch state is indicated by a pie chart symbol. In the present example, one can deduce from the PBS that even though Mo-p orbitals and S-p orbitals contribute roughly equally to the DOS in the valence band, the Mo-p orbital contributions are localised to a region in the BZ around the M-point, whereas the S-p orbitals contribute throughout the entire BZ.

3.4. Corrected G_0W_0 band structures

The C2DB contains G_0W_0 quasiparticle (QP) band structures of 370 monolayers covering 14 different crystal structures and 52 chemical elements. The details of these calculations can be found in the original C2DB paper [6]. A recent in-depth analysis of the 61.716 G_0W_0 data points making up the QP band structures led to several important conclusions relevant for high-throughput G_0W_0 calculations. In particular, it identified the linear QP approximation as a significant error source in standard G_0W_0 calculations and proposed an extremely simple correction scheme (the *empirical Z* (empZ) scheme), that reduces this error by a factor of two on average.

The empZ scheme divides the electronic states into two classes according to the size of the QP weight, Z. States with $Z \in [0.5, 1.0]$ are classified as QP consistent (QP-c) while states with $Z \notin [0.5, 1.0]$ are classified as QP inconsistent (QP-ic). With this definition, QP-c states will have at least half of their spectral weight in the QP peak. The distribution of the 60.000+ Z-values is shown in figure 7. It turns out that the linear approximation to the self-energy, which is the gist of the QP approximation, introduces significantly larger errors for QP-ic states than for QP-c states. Consequently, the empZ method replaces the calculated Z of QP-ic states with the mean of the Z-distribution, $Z_0 \approx 0.75$. This simple replacement reduces the average error of the linear approximation from 0.11 to 0.06 eV.

An illustration of the method applied to MoS_2 is shown in figure 7. The original uncorrected G_0W_0 band structure is shown in blue while the empZ corrected band structure is shown in orange. MoS_2 has only one QP-ic state in the third conduction band at the K-point. Due to a break-down of the QP approximation for this state, the G_0W_0 correction is greatly overestimated leading to a local discontinuity in the band structure. The replacement of Z by Z_0 for this particular state resolves the problem. All G_0W_0 band structures in the C2DB are now empZ corrected.

3.5. Optical absorbance

In the first version of the C2DB, the optical absorbance was obtained from the simple expression [6]

$$A(\omega) \approx \frac{\omega \mathrm{Im}\alpha^{2\mathrm{D}}(\omega)}{\epsilon_0 c}, \qquad (2)$$

where $\alpha^{2\mathrm{D}}$ is the long wavelength limit of the in-plane sheet polarisability density (note that the equation is written here in SI units). The sheet polarisability is related to the sheet conductivity via $\sigma^{2\mathrm{D}}(\omega) = -i\omega\alpha^{2\mathrm{D}}(\omega)$. The expression (2) assumes that the electric field inside the layer equals the incoming field (i.e. reflection is ignored), and hence, it may overestimate the absorbance.

In the new version, the absorbance is evaluated from $A = 1 - R - T$, where R and T are the reflected and transmitted powers of a plane wave at normal

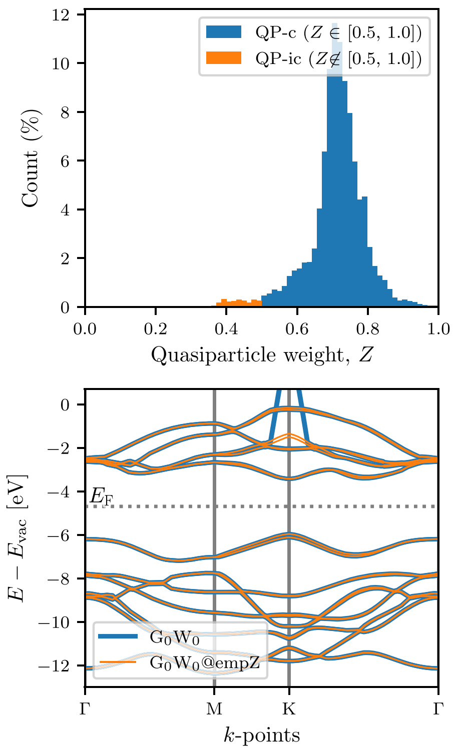


Figure 7. Top: Distribution of the 61 716 QP weights (Z) contained in the C2DB. The blue part of the distribution shows QP-consistent (QP-c) Z-values while the orange part shows QP-inconsistent (QP-ic) Z values. In general, the linear expansion of the self-energy performed when solving the QP equation works better for Z closer to 1. About 0.3% of the Z-values lie outside the interval from 0 to 1 and are not included in the distribution. Bottom: G_0W_0 band structure before (blue) and after (orange) applying the empZ correction, which replaces Z by the mean of the distribution for QP-ic states. In the case of MoS_2 only one state at K is QP-ic.

incidence, respectively. These can be obtained from the conventional transfer matrix method applied to a monolayer suspended in vacuum. The 2D material is here modelled as an infinitely thin layer with a sheet conductivity. Alternatively, it can be modelled as quasi-2D material of thickness d with a 'bulk' conductivity of $\sigma = \sigma^{2D}/d$ [58], but the two approaches yield very similar results, since the optical thickness of a 2D material is much smaller than the optical wavelength. Within this model, the expression for the absorbance of a suspended monolayer with the sheet conductivity σ^{2D} reads:

$$A(\omega) = \mathrm{Re}\left\{\sigma^{2D}(\omega)\eta_0\right\}\left|\frac{2}{2+\sigma^{2D}(\omega)\eta_0}\right|^2, \qquad (3)$$

where $\eta_0 = 1/(\epsilon_0 c) \approx 377\ \Omega$ is the vacuum impedance.

If the light–matter interaction is weak, i.e. $|\sigma^{2D}\eta_0| \ll 1$, equation (3) reduces to equation (2). Nonetheless, due the strong light–matter interaction in some 2D materials, this approximation is not reliable in general. In fact, it can be shown that the maximum possible absorption from equation (3) is 50%, which is known as the upper limit of light absorption in thin films [59]. This limit is not guaranteed by equation (2), which can even yield an absorbance above 100%.

As an example, figure 8 shows the absorption spectrum of monolayer MoS_2 for in- and out-of-plane polarised light as calculated with the exact equation (3) and the approximate equation (2), respectively. In all cases the sheet polarisability is obtained from the BSE to account for excitonic effects [6]. For weak light–matter interactions, e.g. for the z-polarised light, the two approaches agree quite well, but noticeable differences are observed in regions with stronger light–matter interaction.

4. New materials in the C2DB

In this section we discuss the most significant extensions of the C2DB in terms of new materials. The set of materials presented here is not complete, but represents the most important and/or well defined classes. The materials discussed in sections 4.1 and 4.2 (MXY Janus monolayers and monolayers extracted from experimental crystal structure databases) are already included in the C2DB. The materials described in sections 4.3 and 4.4 (homo-bilayers and monolayer point defect systems) will soon become available as separate C2DB-interlinked databases.

4.1. MXY Janus monolayers

The class of TMDC monolayers of the type MX_2 (where M is the transition metal and X is a chalcogen) exhibits a large variety of interesting and unique properties and has been widely discussed in the literature [60]. Recent experiments have shown that it is not only possible to synthesise different materials by changing the metal M or the chalcogen X, but also by exchanging the X on one side of the layer by another chalcogen (or halogen) [61–63]. This results in a class of 2D materials known as MXY Janus monolayers with broken mirror symmetry and finite out-of-plane dipole moments. The prototypical MXY crystal structures are shown in figure 9 for the case of MoSSe and BiTeI, which have both been experimentally realised [61–63]. Adopting the nomenclature from the TMDCs, the crystal structures are denoted as H- or T-phase, depending on whether X and Y atoms are vertically aligned or displaced, respectively.

In a recent work [64], the C2DB workflow was employed to scrutinise and classify the basic electronic and optical properties of 224 different MXY Janus monolayers. All data from the study is available in the C2DB. Here we provide a brief discussion of the Rashba physics in these materials and refer the

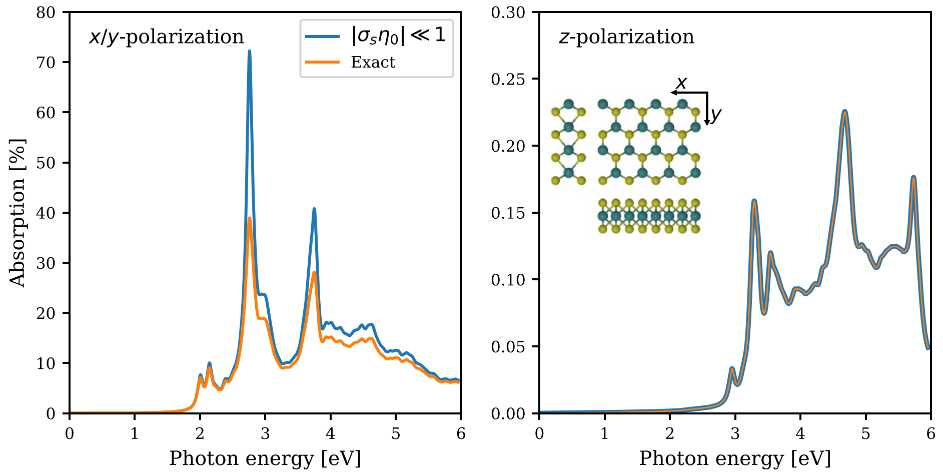


Figure 8. Optical absorption of standalone monolayer MoS_2 for *x*/*y*-polarisation (left) and *z*-polarisation (right) at normal incident in the BSE framework, obtained using equation (2) (blue) or equation (3) (orange). The crystal structure cross-sectional views are shown in the inset with the definition of directions.

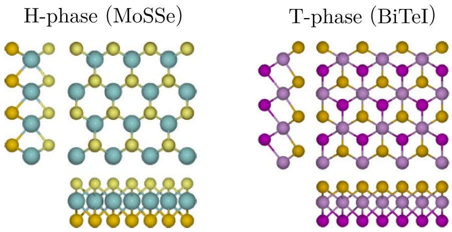


Figure 9. Atomic structure of the MXY Janus monolayers in the H-phase (left) and T-phase (right). The two prototype materials MoSSe and BiTeI are examples of experimentally realised monolayers adopting these crystal structures (not to scale).

interested reader to [64] for more details and analysis of other properties.

A key issue when considering hypothetical materials, i.e. materials not previously synthesised, is their stability. The experimentally synthesised MoSSe and BiTeI are both found to be dynamically stable and lie within 10 meV of the convex hull confirming their thermodynamic stability. Out of the 224 initial monolayers 93 are classified as stable according to the C2DB criteria (dynamically stable and $\Delta H_{\text{hull}} <$ $0.2\,\text{eV}\,\text{atom}^{-1}$). Out of the 93 stable materials, 70 exhibit a finite band gap when computed with the PBE xc-functional.

The Rashba effect is a momentum dependent splitting of the band energies of a 2D semiconductor in the vicinity of a band extremum arising due to the combined effect of spin–orbit interactions and a broken crystal symmetry in the direction perpendicular to the 2D plane. The simplest model used to describe the Rashba effect is a 2D electron gas in a perpendicular electric field (along the *z*-axis). Close to the band extremum, the energy of the two spin bands is described by the Rashba Hamiltonian [65, 66]:

$$H = \alpha_R(\boldsymbol{\sigma} \times \mathbf{k}) \cdot \hat{\mathbf{e}}_z, \qquad (4)$$

where $\boldsymbol{\sigma}$ is the vector of Pauli matrices, $\mathbf{k} = \mathbf{p}/\hbar$ is the wave number, and the Rashba parameter is proportional to the electric field strength, $\alpha_R \propto E_0$.

Although the Rashba Hamiltonian is only meant as a qualitative model, it is of interest to test its validity on the Janus monolayers. The electric field of the Rashba model is approximately given by $E_0 = \Delta V_{\text{vac}}/d$, where ΔV_{vac} is the shift in vacuum potential on the two sides of the layer (see left inset of figure 10) and d is the layer thickness. Assuming a similar thickness for all monolayers, the electric field is proportional to the potential shift. Not unexpected, the latter is found to correlate strongly with the difference in electronegativity of the X and Y atoms, see left panel of figure 10.

The Rashba energy, E_R, can be found by fitting $E(k) = \hbar^2k^2/2m^* \pm \alpha_R k$ to the band structure (see right inset of figure 10) and should scale with the electric field strength. However, as seen from the right panel of figure 10, there is no correlation between the two quantities. Hence we conclude that the simple Rashba model is completely inadequate and that the strength of the perpendicular electric field cannot be used to quantify the effect of spin–orbit interactions on band energies.

4.2. Monolayers from known layered bulk crystals

The C2DB has been extended with a number of monolayers that are likely exfoliable from experimentally known layered bulk compounds. Specifically, the Inorganic Crystal Structure Database (ICSD) [67] and Crystallography Open Database (COD) [68]

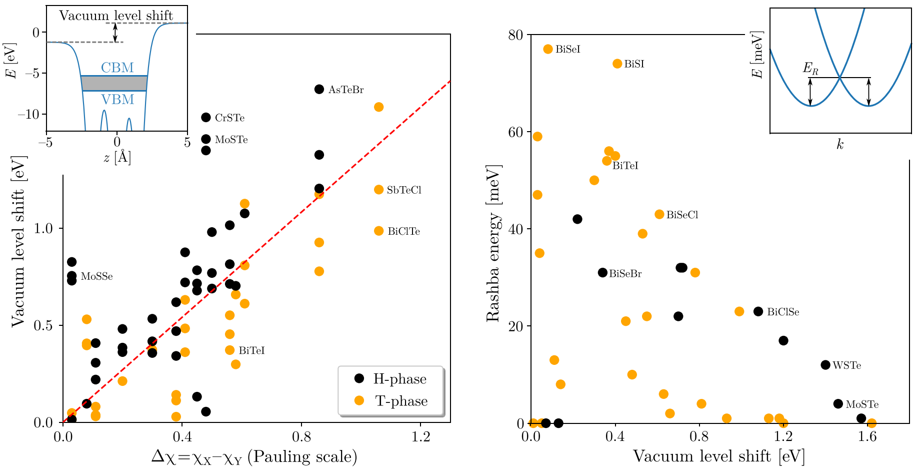


Figure 10. Left: Correlation between the electronegativity difference of X and Y in MXY Janus monolayers and the vacuum level shift across the layer. Right: Correlation between the Rashba energy and the vacuum level shift. Structures in the H-phase (e.g. MoSSe) are shown in black while structures in the T-phase (e.g. BiTeI) are shown in orange. The linear fit has the slope 1.35 eV/$\Delta\chi$ (Pauling scale). The insets show the definition of the vacuum level shift and the Rashba energy, respectively. Modified from [64].

have first been filtered for corrupted, duplicate and theoretical compounds, which reduce the initial set of 585.485 database entries to 167.767 unique materials. All of these have subsequently been assigned a 'dimensionality score' based on a purely geometrical descriptor. If the 2D score is larger than the sum of 0D, 1D and 3D scores we regard the material as being exfoliable and we extract the individual 2D components that comprise the material (see also section 2.2). We refer to the original work on the method for details [52] and note that similar approaches were applied in [11, 12] to identify potentially exfoliable monolayers from the ICSD and COD.

The search has been limited to bulk compounds containing less than six different elements and no rare earth elements. This reduces the set of relevant bulk materials to 2991. For all of these we extracted the 2D components containing less than 21 atoms in the unit cell, which were then relaxed and sorted for duplicates following the general C2DB workflow steps described in sections 2.1–2.3. At this point 781 materials remain. This set includes most known 2D materials and 207 of the 781 were already present in the C2DB prior to this addition. All the materials (including those that were already in C2DB) have been assigned an ICSD/COD identifier that refers to the parent bulk compound from which the 2D material was computationally exfoliated. We emphasise that we have not considered exfoliation energies in the analysis and a subset of these materials may thus be rather strongly bound and challenging to exfoliate even if the geometries indicate van der Waals bonded structures of the parent bulk compounds.

Figure 11 shows the distribution of energies above the convex hull for materials derived from parent structures in ICSD or COD as well as for the entire C2DB, which includes materials obtained from combinatorial lattice decoration as well. As expected, the materials derived from experimental bulk materials are situated rather close to the convex hull whereas those obtained from lattice decoration extend to energies far above the convex hull. It is also observed that a larger fraction of the experimentally derived materials are dynamically stable. There are, however, well known examples of van der Waals bonded structures where the monolayer undergoes a significant lattice distortion, which will manifest itself as a dynamical instability in the present context. For example, bulk MoS_2 exists in van der Waals bonded structures composed of either 2 H-MoS_2 or 1 T-MoS_2 layers, but a monolayer of the 1 T phase undergoes a structural deformation involving a doubling of the unit cell [69] and is thus categorised as dynamically unstable by the C2DB workflow. The dynamically stable materials derived from parent bulk structures in the ICSD and COD may serve as a useful subset of the C2DB that are likely to be exfoliable from known compounds and thus facilitate experimental verification. As a first application the subset has been used to search for magnetic 2D materials, which resulted in a total of 85 ferromagnets and 61 anti-ferromagnets [70].

4.3. Outlook: multilayers

The C2DB is concerned with the properties of covalently bonded monolayers (see discussion of dimensionality filtering in section 2.2). However, multilayer structures composed of two or more identical monolayers are equally interesting and often have properties that deviate from those of the monolayer. In fact, the synthesis of layered vdW structures with a

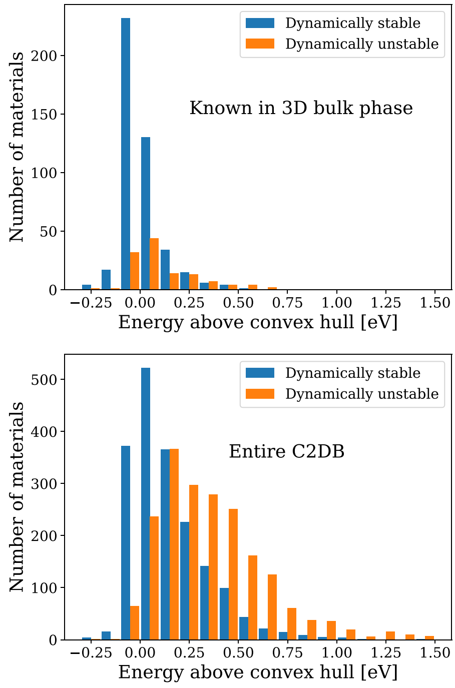


Figure 11. Distribution of energies above the convex hull for the 2D materials extracted from bulk compounds in ICSD and COD (top) and for the entire C2DB including those constructed from combinatorial lattice decoration (bottom). Dynamically stable materials are indicated in blue.

controllable number of layers represents an interesting avenue for atomic-scale materials design. Several examples of novel phenomena emerging in layered vdW structures have been demonstrated including direct-indirect band gap transitions in MoS_2 [71, 72], layer-parity selective Berry curvatures in few-layer WTe_2 [73], thickness-dependent magnetic order in CrI_3 [74, 75], and emergent ferroelectricity in bilayer hBN [76].

As a first step towards a systematic exploration of multilayer 2D structures, the C2DB has been used as basis for generating homobilayers in various stacking configurations and subsequently computing their properties following a modified version of the C2DB monolayer workflow. Specifically, the most stable monolayers (around 1000) are combined into bilayers by applying all possible transformations (unit cell preserving point group operations and translations) of one layer while keeping the other fixed. The candidate bilayers generated in this way are subject to a stability analysis, which evaluates the binding energy and optimal IL distance based on PBE-D3 [77] total energy calculations keeping the atoms of the monolayers fixed in their PBE relaxed geometry, see figures 12 and table 3.

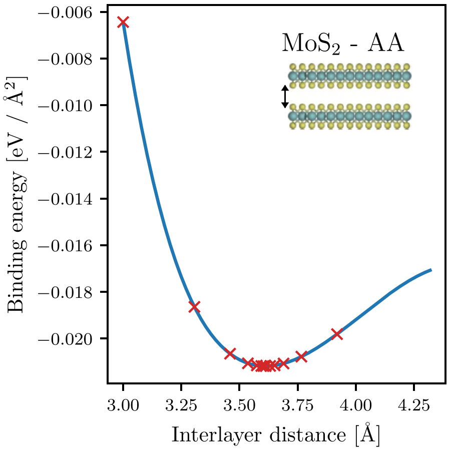


Figure 12. An illustration of the optimisation of the interlayer (IL) distance for MoS_2 in the AA stacking. The black crosses are the points sampled by the optimisation algorithm while the blue curve is a spline interpolation of the black crosses. The inset shows the MoS_2 AA stacking and the definition of the IL distance is indicated with a black double-sided arrow.

Table 3. Exfoliation energies for selected materials calculated with the PBE+D3 xc-functional as described in section 4.3 and compared with the DF2 and rVV10 results from [11]. The spacegroups are indicated in the column 'SG'. All numbers are in units of meV Å^{-2}.

Material	SG	PBE + D3	DF2	rVV10
MoS_2	P-6m2	28.9	21.6	28.8
$MoTe_2$	P-6m2	30.3	25.2	30.4
ZrNBr	Pmmn	18.5	10.5	18.5
C	P6/mmm	18.9	20.3	25.5
P	Pmna	21.9	38.4	30.7
BN	P-6m2	18.9	19.4	24.4
WTe_2	P-6m2	32.0	24.7	30.0
PbTe	P3m1	23.2	27.5	33.0

The calculated IL binding energies are generally in the range from a few to a hundred meV Å^{-2} and IL distances range from 1.5 to 3.8 Å. A scatter plot of preliminary binding energies and IL distances is shown in figure 13. The analysis of homobilayers provides an estimate of the energy required to peel a monolayer off a bulk structure. In particular, the binding energy for the most stable bilayer configuration provides a measure of the *exfoliation energy* of the monolayer. This key quantity is now available for all monolayers in the C2DB, see section 5.1.

4.4. Outlook: point defects

The C2DB is concerned with the properties of 2D materials in their pristine crystalline form. However, as is well known the perfect crystal is an idealised model of real materials, which always contain defects in smaller or larger amounts depending on the intrinsic materials properties and growth conditions. Crystal defects often have a negative impact on

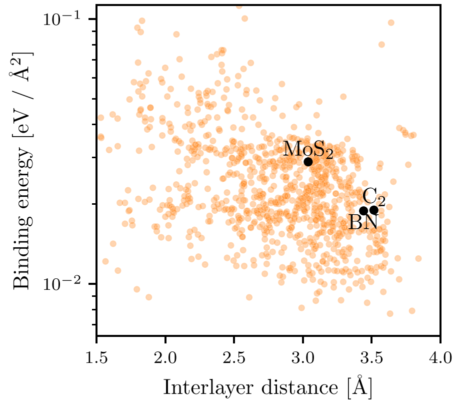


Figure 13. Scatter plot of the calculated interlayer distance and binding energies of (homo)bilayers of selected materials from C2DB. A few well known materials are highlighted: MoS$_2$, graphene (C$_2$), and hexagonal boron nitride (hBN). The bilayer binding energies provide an estimate of the monolayer exfoliation energies, see section 5.1.

physical properties, e.g. they lead to scattering and life time-reduction of charge carriers in semiconductors. However, there are also important situations where defects play a positive enabling role, e.g. in doping of semiconductors, as colour centres for photon emission [78, 79] or as active sites in catalysis.

To reduce the gap between the pristine model material and real experimentally accessible samples, a systematic evaluation of the basic properties of the simplest native point defects in a selected subset of monolayers from the C2DB has been initiated. The monolayers are selected based on the stability of the pristine crystal. Moreover, only non-magnetic semiconductors with a PBE band gap satisfying $E_{\mathrm{gap}} > 1$ eV are currently considered as such materials are candidates for quantum technology applications like single-photon sources and spin qubits. Following these selection criteria around 300 monolayers are identified and their vacancies and intrinsic substitutional defects are considered, yielding a total of about 1500 defect systems.

Each defect system is subject to the same workflow, which is briefly outlined below. To enable point defects to relax into their lowest energy configuration, the symmetry of the pristine host crystal is intentionally broken by the chosen supercell, see figure 14 (a). In order to minimise defect–defect interaction, supercells are furthermore chosen such that the minimum distance between periodic images of defects is larger than 15 Å. Unique point defects are created based on the analysis of equivalent Wyckoff positions for the host material. To illustrate some of the properties that will feature in the upcoming point defect database, we consider the specific example of monolayer CH$_2$Si.

First, the formation energy [80, 81] of a given defect is calculated from PBE total energies. Next, Slater–Janak transition state theory is used to obtain the charge transition levels [82, 83]. By combining these results, one obtains the formation energy of the defect in all possible charge states as a function of the Fermi level. An example of such a diagram is shown in figure 14 (b) for the case of the V$_{\mathrm{C}}$ and C$_{\mathrm{Si}}$ defects in monolayer CH$_2$Si. For each defect and each charge state, the PBE single-particle energy level diagram is calculated to provide a qualitative overview of the electronic structure. A symmetry analysis [84] is performed for the defect structure and the individual defect states lying inside the band gap. The energy level diagram of the neutral V$_{\mathrm{Si}}$ defect in CH$_2$Si is shown in figure 14 (c), where the defect states are labelled according to the irreducible representations of the C_s point group.

In general, excited electronic states can be modelled by solving the Kohn–Sham equations with non-Aufbau occupations. The excited-state solutions are saddle points of the Kohn–Sham energy functional, but common self-consistent field (SCF) approaches often struggle to find such solutions, especially when nearly degenerate states are involved. The calculation of excited states corresponding to transitions between localised states inside the band gap is therefore performed using an alternative method based on the direct optimisation (DO) of orbital rotations in combination with the maximum overlap method (MOM) [85]. This method ensures fast and robust convergence of the excited states, as compared to SCF. In figure 14 (d), the reorganisation energies for the ground and excited state, as well as the zero-phonon line (ZPL) energy are sketched. For the specific case of the Si vacancy in CH$_2$Si, the DO-MOM method yields $E_{\mathrm{ZPL}} = 3.84$ eV, $\lambda_{\mathrm{gs}}^{\mathrm{reorg}} = 0.11$ eV and $\lambda_{\mathrm{exc}}^{\mathrm{reorg}} = 0.16$ eV. For systems with large electron-phonon coupling (i.e. Huang–Rhys factor > 1) a one-dimensional approximation for displacements along the main phonon mode is used to produce the configuration coordinate diagram (see figure 14 (d)). In addition to the ZPL energies and reorganisation energies, the Huang–Rhys factors, photoluminescence spectrum from the 1D phonon model, hyperfine coupling and zero field splitting are calculated.

5. New properties in the C2DB

This section reports on new properties that have become available in the C2DB since the first release. The employed computational methodology is described in some detail and results are compared to the literature where relevant. In addition, some interesting property correlations are considered along with general discussions of the general significance and potential application of the available data.

5.1. Exfoliation energy

The exfoliation energy of a monolayer is estimated as the binding energy of its bilayer in the most stable

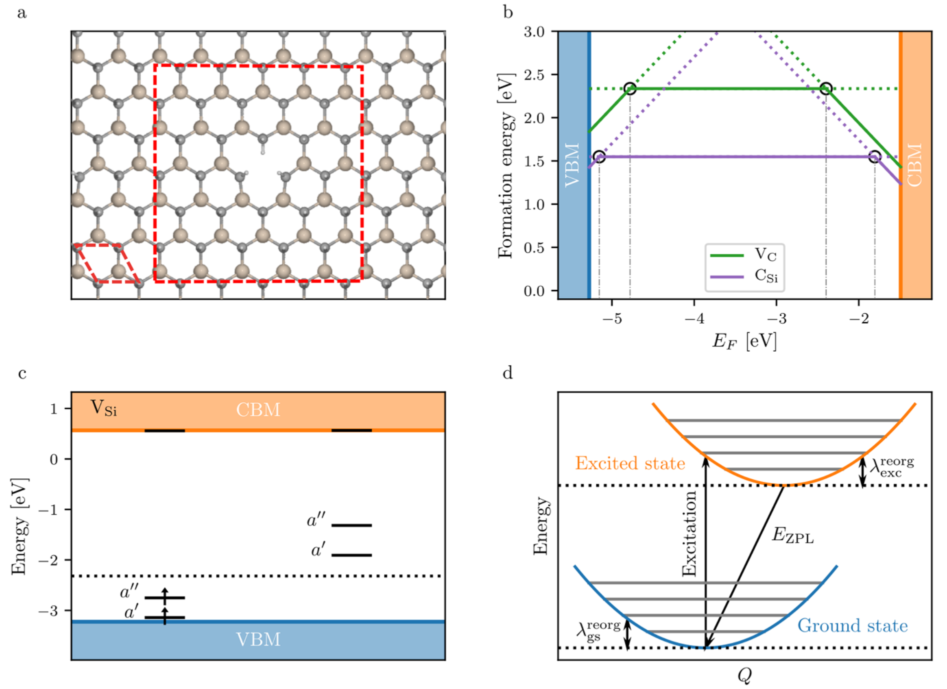


Figure 14. Overview of some of the properties included in the 2D defect database project for the example host material CH_2Si. (a) The supercell used to represent the defects (here a Si vacancy). The supercell is deliberately chosen to break the symmetry of the host crystal lattice. (b) Formation energies of a C vacancy (green) and C–Si substitutional defect (purple). (c) Energy and orbital symmetry of the localised single-particle states of the V_{Si} defect for both spin channels (left and right). The Fermi level is shown by the dotted line. (d) Schematic excited state configuration energy diagram. The transitions corresponding to the vertical absorption and the zero-phonon emission are indicated.

stacking configuration (see also section 4.3). The binding energy is calculated using the PBE + D3 xc-functional [86] with the atoms of both monolayers fixed in the PBE relaxed geometry. Table 3 compares exfoliation energies obtained in this way to values from Mounet *et al* [11] for a representative set of monolayers.

5.2. Bader charges

For all monolayers we calculate the net charge on the individual atoms using the Bader partitioning scheme [87]. The analysis is based purely on the electron density, which we calculate from the PAW pseudo density plus compensation charges using the PBE xc-functional. Details of the method and its implementation can be found in Tang *et al* [88]. In section 5.4 we compare and discuss the relation between Bader charges and Born charges.

5.3. Spontaneous polarisation

The spontaneous polarisation ($\mathbf{P}_s$) of a bulk material is defined as the charge displacement with respect to that of a reference centrosymmetric structure [89, 90]. Ferroelectric materials exhibit a finite value of $\mathbf{P}_s$ that may be switched by an applied external field and have attracted a large interest for a wide range of applications [91–93].

The spontaneous polarisation in bulk materials can be regarded as electric dipole moment per unit volume, but in contrast to the case of finite systems this quantity is ill-defined for periodic crystals [89]. Nevertheless, one can define the formal polarisation density:

$$\mathbf{P} = \frac{1}{2\pi}\frac{e}{V}\sum_l \phi_l \mathbf{a}_l, \tag{5}$$

where $\mathbf{a}_l$ (with $l \in \{1,2,3\}$) are the lattice vectors spanning the unit cell, V is the cell volume and e is the elementary charge. ϕ_l is the polarisation phase along the lattice vector defined by:

$$\phi_l = \sum_i Z_i \mathbf{b}_l \cdot \mathbf{u}_i - \phi_l^{\text{elec}}, \tag{6}$$

where $\mathbf{b}_l$ is the reciprocal lattice vector satisfying $\mathbf{b}_l \cdot \mathbf{R}_l = 2\pi$ and $\mathbf{u}_i$ is the position of nucleus i with charge eZ_i. The electronic contribution to the polarisation phase is defined as:

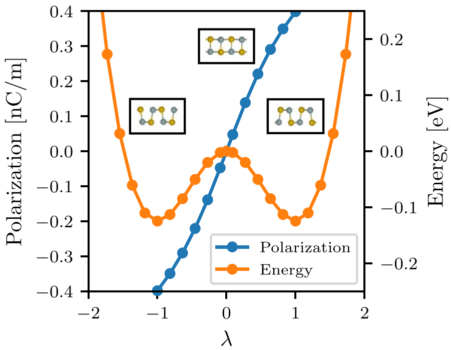


Figure 15. Depicted in the blue plot is the formal polarisation calculated along the adiabatic path for GeSe, using the methods described in the main text. The orange plot shows the energy potential along the path as well as outside. Figure inset: The structure of GeSe in the two non-centrosymmetric configurations corresponding to $-\mathbf{P}_\text{s}$ and $\mathbf{P}_\text{s}$ and the centrosymmetric configuration.

$$\phi_l^{\text{elec}} = \frac{1}{N_{k\perp\mathbf{b}_l}} \text{Im} \sum_{k\in \text{BZ}_{\perp\mathbf{b}_l}} \times \ln \prod_{j=0}^{N_{k\|\mathbf{b}_l}-1} \det_{occ} \left[\langle u_{n\mathbf{k}+j\delta\mathbf{k}} | u_{m\mathbf{k}+(j+1)\delta\mathbf{k}} \right], \tag{7}$$

where $\text{BZ}_{\perp\mathbf{b}_l} = \{\mathbf{k}|\mathbf{k}\cdot\mathbf{b}_l = 0\}$ is a plane of **k**-points orthogonal to $\mathbf{b}_l$, $\delta\mathbf{k}$ is the distance between neighbouring k-points in the $\mathbf{b}_l$ direction and $N_{k\|\mathbf{b}_l}$ ($N_{k\perp\mathbf{b}_l}$) is the number of **k**-points along (perpendicular to) the $\mathbf{b}_l$ direction. These expression generalise straightforwardly to 2D.

The formal polarisation is only well-defined modulo $e\mathbf{R}_n/V$ where $\mathbf{R}_n$ is any lattice vector. However, changes in polarisation are well defined and the spontaneous polarisation may thus be obtained by:

$$\mathbf{P}_\text{s} = \int_0^1 \frac{d\mathbf{P}(\lambda)}{d\lambda} d\lambda, \tag{8}$$

where λ is a dimensionless parameter that defines an adiabatic structural path connecting the polar phase ($\lambda = 1$) with a non-polar phase ($\lambda = 0$).

The methodology has been implemented in GPAW and used to calculate the spontaneous polarisation of all stable materials in the C2DB with a PBE band gap above 0.01 eV and a polar space group symmetry. For each material, the centrosymmetric phase with smallest atomic displacement from the polar phase is constructed and relaxed under the constraint of inversion symmetry. The adiabatic path connecting the two phases is then used to calculate the spontaneous polarisation using equations (5)–(8). An example of a calculation for GeSe is shown in figure 15 where the polarisation along the path connecting two equivalent polar phases via the centrosymmetric phase is shown together with the total energy. The spontaneous polarisation obtained from the path is 39.8 nC m^{-1} in good agreement with previous calculations [94].

5.4. Born charges

The Born charge of an atom a at position $\mathbf{u}_a$ in a solid is defined as:

$$Z_{ij}^a = \frac{V}{e} \frac{\partial P_i}{\partial u_{aj}} \bigg|_{E=0}. \tag{9}$$

It can be understood as an effective charge assigned to the atom to match the change in polarisation in direction i when its position is perturbed in direction j. Since the polarisation density and the atomic position are both vectors, the Born charge of an atom is a rank-2 tensor. The Born charge is calculated as a finite difference and relies on the Modern theory of polarisation [95] for the calculation of polarisation densities, see reference [96] for more details. The Born charge has been calculated for all stable materials in C2DB with a finite PBE band gap.

It is of interest to examine the relation between the Born charge and the Bader charge (see section 5.2). In materials with strong ionic bonds one would expect the charges to follow the atoms. On the other hand, in covalently bonded materials the hybridisation pattern and thus the charge distribution, depends on the atom positions in a complex way, and the idea of charges following the atom is expected to break down. In agreement with this idea, the (in-plane) Born charges in the strongly ionic hexagonal hBN ($\pm 2.71e$ for B and N, respectively) are in good agreement with the calculated Bader charges ($\pm 3.0e$). In contrast, (the in-plane) Born charges in MoS$_2$ ($-1.08e$ and $0.54e$ for Mo and S, respectively) deviate significantly from the Bader charges ($1.22e$ and $-0.61e$ for Mo and S, respectively). In fact, the values disagree even on the sign of the charges underlining the non-intuitive nature of the Born charges in covalently bonded materials.

Note that the out-of-plane Born charges never match the Bader charges, even for strongly ionic insulators, and are consistently smaller in value than the in-plane components. The smaller out-of-plane values are consistent with the generally smaller out-of-plane polarisability of 2D materials (for both electronic and phonon contributions) and agrees with the intuitive expectation that it is more difficult to polarise a 2D material in the out-of-plane direction as compared to the in-plane direction.

Figure 16 shows the average of the diagonal of the Born charge tensor, $\text{Tr}(Z^a)/3$, plotted against the Bader charges for all 585 materials in the C2DB for which the Born charges have been computed. The data points have been coloured according to the ionicity of the atom a defined as $I(a) = |\chi_a - \langle\chi\rangle|$, where χ_a and $\langle\chi\rangle$ are the Pauling electronegativity of atom a and the average electronegativity of all atoms in the

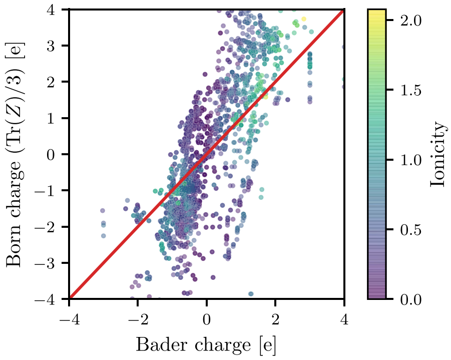


Figure 16. Born charges, $\mathrm{Tr}(Z)/3$, vs. Bader charges for 3025 atoms in the 585 materials for which the Born charges are calculated. The colors indicate the ionicity of the atoms (see main text).

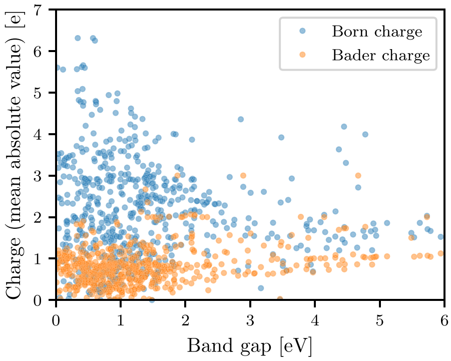


Figure 17. Bader and in-plane Born charges vs. band gap.

unit cell, respectively. The ionicity is thus a measure of the tendency of an atom to donate/accept charge relative to the average tendency of atoms in the material. It is clear from figure 16 that there is a larger propensity for the Born and Bader charges to match in materials with higher ionicity.

Figure 17 plots the average (in-plane) Born charge and the Bader charge versus the band gap. It is clear that large band gap materials typically exhibit integer Bader charges, whereas there is no clear correlation between the Born charge and the band gap.

5.5. Infrared polarisability

The original C2DB provided the frequency dependent polarisability computed in the random phase approximation (RPA) with inclusion of electronic interband and intraband (for metals) transitions [6]. However, phonons carrying a dipole moment (so-called IR active phonons) also contribute to the polarisability at frequencies comparable to the frequency of optical phonons. This response is described by the IR polarisability:

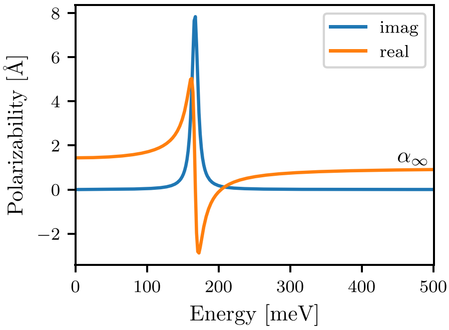


Figure 18. Total polarisability, including both electrons and phonons, of monolayer hBN in the infrared (IR) frequency regime. The resonance at around 180 meV is due to the Γ-point longitudinal optical phonon. At energies above all phonon frequencies (but below the band gap) the polarisability is approximately constant and equal to the static limit of the electronic polarisability, α_∞.

$$\alpha^{\mathrm{IR}}(\omega) = \frac{e^2}{A}\mathbf{Z}^T\mathbf{M}^{-1/2}\left(\sum_i \frac{\mathbf{d}_i\mathbf{d}_i^T}{\omega_i^2 - \omega^2 - i\gamma\omega}\right)\mathbf{M}^{-1/2}\mathbf{Z}, \tag{10}$$

where **Z** and **M** are matrix representations of the Born charges and atomic masses, ω_i^2 and d_i are eigenvectors and eigenvalues of the dynamical matrix, A is the in-plane cell area and γ is a broadening parameter representing the phonon lifetime and is set to 10 meV. The total polarisability is then the sum of the electronic polarisability and the IR polarisability.

The new C2DB includes the IR polarisability of all monolayers for which the Born charges have been calculated (stable materials with a finite band gap), see section (5.4). As an example, figure 18 shows the total polarisability of monolayer hexagonal hBN. For details on the calculation of the IR polarisability see reference [96].

5.6. Piezoelectric tensor

The piezoelectric effect is the accumulation of charges, or equivalently the formation of an electric polarisation, in a material in response to an applied mechanical stress or strain. It is an important material characteristic with numerous scientific and technological applications in sonar, microphones, accelerometers, ultrasonic transducers, energy conversion, etc [97, 98]. The change in polarisation originates from the movement of positive and negative charge centres as the material is deformed.

Piezoelectricity can be described by the (proper) piezoelectric tensor c_{ijk} with $i,j,k \in \{x,y,z\}$, given by [99]:

$$c_{ijk} = \frac{e}{2\pi V}\sum_l \frac{\partial \phi_l}{\partial \epsilon_{jk}} a_{li}, \tag{11}$$

which differs from equation (5) only by a derivative of the polarisation phase with respect to the strain tensor

Table 4. Comparison of computed piezoelectric tensor versus experimental values and previous calculations for hexagonal BN and a selected set of TMDCs (space group 187). All numbers are in units of nC/m. Experimental data for MoS_2 is obtained from [102].

Material	Exp.	Theory [101]	C2DB
BN	—	0.14	0.13
MoS_2	0.3	0.36	0.35
$MoSe_2$	—	0.39	0.38
$MoTe_2$	—	0.54	0.48
WS_2	—	0.25	0.24
WSe_2	—	0.27	0.26
WTe_2	—	0.34	0.34

ϵ_{jk}. Note that c_{ijk} does not depend on the chosen branch cut.

The piezoelectric tensor is a symmetric tensor with at most 18 independent components. Furthermore, the point group symmetry restricts the number of independent tensor elements and their relationships due to the well-known Neumann's principle [100]. For example, monolayer MoS_2 with point group D_{3h}, has only one non-vanishing independent element of c_{ijk}. Note that c_{ijk} vanishes identically for centrosymmetric materials. Using a finite-difference technique with a finite but small strain (1% in our case), equation (11) has been used to compute the proper piezoelectric tensor for all non-centrosymmetric materials in the C2DB with a finite band gap. Table 4 shows a comparison of the piezoelectric tensors in the C2DB with literature for a selected set of monolayer materials. Good agreement is obtained for all these materials.

5.7. Topological invariants

For all materials in the C2DB exhibiting a direct band gap below 1 eV, the k-space Berry phase spectrum of the occupied bands has been calculated from the PBE wave functions. Specifically, a particular k-point is written as $k_1\mathbf{b}_1 + k_2\mathbf{b}_2$ and the Berry phases $\gamma_n(k_2)$ of the occupied states on the path $k_1 = 0 \rightarrow k_1 = 1$ is calculated for each value of k_2. The connectivity of the Berry phase spectrum determines the topological properties of the 2D Bloch Hamiltonian [103, 104].

The calculated Berry phase spectra of the relevant materials are available for visual inspection on the C2DB webpage. Three different topological invariants have been extracted from these spectra and are reported in the C2DB: (1) The Chern number, C, takes an integer value and is well defined for any gapped 2D material. It determines the number of chiral edge states on any edge of the material. For any non-magnetic material the Chern number vanishes due to time-reversal symmetry. It is determined from the Berry phase spectrum as the number of crossings at any horizontal line in the spectrum. (2) The mirror Chern number, C_M, defined for gapped materials with a mirror plane in the atomic layer [105]. For such materials, all states may be chosen as mirror eigenstates with eigenvalues $\pm i$ and the Chern numbers $C_\pm$ can be defined for each mirror sector separately. For a material with vanishing Chern number, the mirror Chern number is defined as $C_M = (C_+ - C_-)/2$ and takes an integer value corresponding to the number of edge states on any mirror symmetry preserving edge. It is obtained from the Berry phase spectrum as the number of chiral crossings in each of the mirror sectors. (3) The Z_2 invariant, ν, which can take the values 0 and 1, is defined for materials with time-reversal symmetry. Materials with $\nu = 1$ are referred to as quantum spin Hall insulators and exhibit helical edge states at any time-reversal conserving edge. It is determined from the Berry phase spectrum as the number of crossing points modulus 2 at any horizontal line in the interval $k_2 \in [0, 1/2]$.

Figure 19 shows four representative Berry phase spectra corresponding to the three cases of non-vanishing C, C_M and ν as well as a trivial insulator. The four materials are: $OsCl_3$ (space group 147)—a Chern insulator with $C = 1$, $OsTe_2$ (space group 14)—a mirror crystalline insulator with $C_M = 2$, SbI (spacegroup 1)—a quantum spin Hall insulator with $\nu = 1$ and BiITe (spacegroup 156)—a trivial insulator. Note that a gap in the Berry phase spectrum always implies a trivial insulator.

In [106] the C2DB was screened for materials with non-trivial topology. At that point it was found that the database contained 7 Chern insulators, 21 mirror crystalline topological insulators and 48 quantum spin Hall insulators. However, that does not completely exhaust the the topological properties of materials in the C2DB. In particular, there may be materials that can be topologically classified based on crystalline symmetries other than the mirror plane of the layer. In addition, second order topological effects may be present in certain materials, which imply that flakes will exhibit topologically protected corner states. Again, the Berry phase spectra may be used to unravel the second order topology by means of nested Wilson loops [107].

5.8. Exchange coupling constants

The general C2DB workflow described in sections 2.1–2.3 will identify the FM ground state of a material and apply it as starting point for subsequent property calculations, whenever it is more stable than the spin-paired ground state. In reality, however, the FM state is not guaranteed to comprise the magnetic ground state. In fact, AFM states often have lower energy than the FM one, but in general it is non-trivial to obtain the true magnetic ground state. We have chosen to focus on the FM state due to its simplicity and because its atomic structure and stability are often very similar to those of other magnetic states. Whether or not the FM state is the true magnetic ground state is indicated by the nearest neighbour exchange coupling constant as described below.

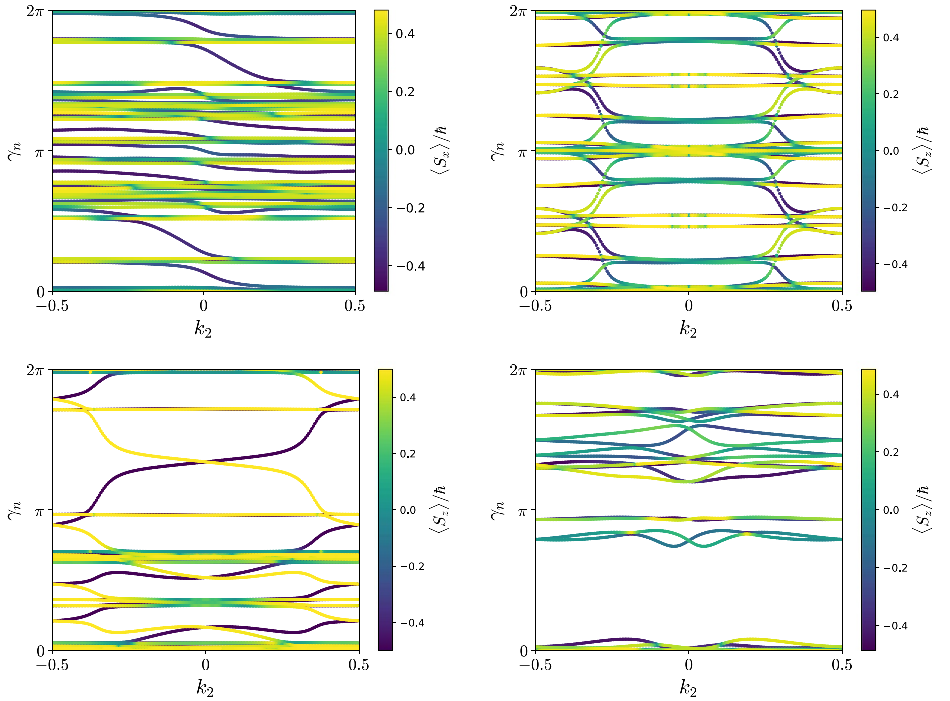


Figure 19. Berry phase spectra of the Chern insulator $OsCl_3$ (top left), the crystalline topological insulator $OsTe_2$ (top right), the quantum spin Hall insulator SbI (lower left) and the trivial insulator BiITe (lower right).

When investigating magnetic materials the thermodynamical properties (for example the critical temperatures for ordering) are of crucial interest. In two dimensions the Mermin–Wagner theorem [108] comprises an extreme example of the importance of thermal effects since it implies that magnetic order is only possible at $T = 0$ unless the spin-rotational symmetry is explicitly broken. The thermodynamic properties cannot be accessed directly by DFT. Consequently, magnetic models that capture the crucial features of magnetic interactions must be employed. For insulators, the Heisenberg model has proven highly successful in describing magnetic properties of solids in 3D as well as 2D [109]. It represents the magnetic degrees of freedom as a lattice of localised spins that interact through a set of exchange coupling constants. If the model is restricted to include only nearest neighbour exchange and assume magnetic isotropy in the plane, it reads:

$$H = -\frac{J}{2}\sum_{\langle ij\rangle}\mathbf{S}_i\cdot\mathbf{S}_j - \frac{\lambda}{2}\sum_{\langle ij\rangle}S_i^z S_j^z - A\sum_i\left(S_i^z\right)^2, \quad (12)$$

where J is the nearest neighbour exchange constant, λ is the nearest neighbour anisotropic exchange constant and A measures the strength of single-ion anisotropy. We also neglect off-diagonal exchange coupling constants that give rise to terms proportional to $S_i^x S_j^y$, $S_i^y S_j^z$ and $S_i^z S_j^x$. The out-of-plane direction has been chosen as z and $\langle ij\rangle$ implies that for each site i we sum over all nearest neighbour sites j. The parameters J, λ and A may be obtained from an energy mapping analysis involving four DFT calculations with different spin configurations [70, 110, 111]. The thermodynamic properties of the resulting 'first principles Heisenberg model' may subsequently be analysed with classical Monte Carlo simulations or renormalised spin wave theory [36, 112].

The C2DB provides the values of J, λ, and A as well as the number of nearest neighbours N_{nn} and the maximum eigenvalue of S_z (S), which is obtained from the total magnetic moment per atom in the FM ground state (rounded to nearest half-integer for metals). These key parameters facilitate easy post-processing analysis of thermal effects on the magnetic structure. In [113] such an analysis was applied to estimate the critical temperature of all FM materials in the C2DB based on a model expression for T_C and the parameters from equation (12).

For metals, the Heisenberg parameters available in C2DB should be used with care because the Heisenberg model is not expected to provide an accurate description of magnetic interactions in this case. Nevertheless, even for metals the sign and magnitude of the parameters provide an important qualitative measure of the magnetic interactions that may be used to screen and select materials for more detailed investigations of magnetic properties.

A negative value of J implies the existence of an AFM state with lower energy than the FM state used in C2DB. This parameter is thus crucial to consider when judging the stability and relevance of a material classified as magnetic in C2DB (see section 2.5). Figure 20 shows the distribution of exchange coupling constants (weighted by S^2) of the magnetic materials in the C2DB. The distribution is slightly skewed to the positive side indicating that FM order is more common than AFM order.

The origin of magnetic anisotropy may stem from either single-ion anisotropy or anisotropic exchange and it is in general difficult *a priori* to determine, which mechanism is most important. There is, however, a tendency in the literature to neglect anisotropic exchange terms in a Heisenberg model description of magnetism and focus solely on the single-ion anisotropy. In figure 20 we show a scatter plot of the anisotropy parameters A and λ for the FM materials ($J > 0$). The spread of the parameters indicate that the magnetic anisotropy is in general equally likely to originate from both mechanisms and neglecting anisotropic exchange is not advisable. For ferromagnets, the model (equation (12)) only exhibits magnetic order at finite temperatures if $A(2S-1)+\lambda N_{nn} > 0$ [113]. Neglecting anisotropic exchange thus excludes materials with $A < 0$ that satisfies $A(2S-1)+\lambda N_{nn} > 0$. This is in fact the case for 11 FM insulators and 31 FM metals in the C2DB.

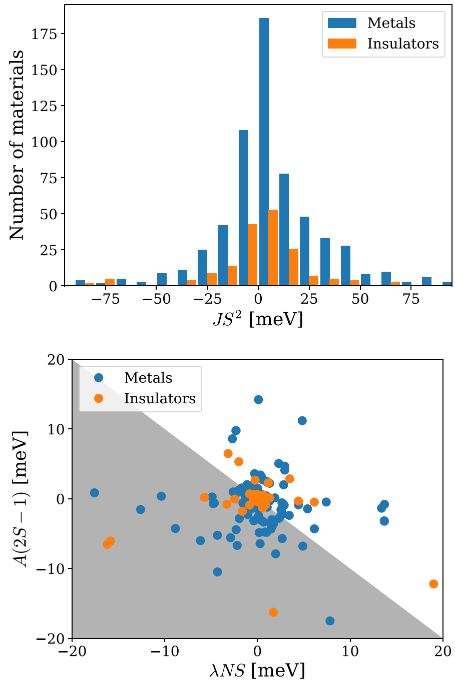


Figure 20. Top: Distribution of exchange coupling constants in C2DB. Bottom: Single-ion anisotropy A vs anisotropic exchange λ for ferromagnetic materials with $S > 1/2$. The shaded area indicates the part of parameter space where the model (equation (12)) does not yield an ordered state at finite temperatures.

5.9. Raman spectrum

Raman spectroscopy is an important technique used to probe the vibrational modes of a solid (or molecule) by means of inelastic scattering of light [114]. In fact, Raman spectroscopy is the dominant method for characterising 2D materials and can yield detailed information about chemical composition, crystal structure and layer thickness. There exist several different types of Raman spectroscopies that differ mainly by the number of photons and phonons involved in the scattering process [114]. The first-order Raman process, in which only a single phonon is involved, is the dominant scattering process in samples with low defect concentrations.

In a recent work, the first-order Raman spectra of 733 monolayer materials from the C2DB were calculated, and used as the basis for an automatic procedure for identifying a 2D material entirely from its experimental Raman spectrum [115]. The Raman spectrum is calculated using third-order perturbation theory to obtain the rate of scattering processes involving creation/annihilation of one phonon and two photons, see reference [115] for details. The light field is written as $\boldsymbol{\mathcal{F}}(t) = \mathcal{F}_{\text{in}}\mathbf{u}_{\text{in}}\exp(-i\omega_{\text{in}}t) + \mathcal{F}_{\text{out}}\mathbf{u}_{\text{out}}\exp(-i\omega_{\text{out}}t)$+c.c. where $\mathcal{F}_{\text{in/out}}$ and $\omega_{\text{in/out}}$ denote the amplitudes and frequencies of the input/output electromagnetic fields, respectively. In addition, $\mathbf{u}_{\text{in/out}} = \sum_i u^i_{\text{in/out}}\mathbf{e}_i$ are the corresponding polarisation vectors, where $\mathbf{e}_i$ denotes the unit vector along the i-direction with $i \in \{x,y,z\}$. Using this light field, the final expression for the Stokes Raman intensity involving scattering events by only one phonon reads [115]:

$$I(\omega) = I_0 \sum_{\nu} \frac{n_\nu + 1}{\omega_\nu} \left| \sum_{ij} u^i_{\text{in}} R^\nu_{ij} u^j_{\text{out}} \right|^2 \delta(\omega - \omega_\nu). \tag{13}$$

Here, I_0 is an unimportant constant (since Raman spectra are always reported normalised), and n_ν is obtained from the Bose–Einstein distribution, i.e. $n_\nu \equiv (\exp[\hbar\omega_\nu/k_BT]-1)^{-1}$ at temperature T for a Raman mode with energy $\hbar\omega_\nu$. Note that only phonons at the Brillouin zone center (with zero momentum) contribute to the one-phonon Raman processes due to momentum conservation. In equation (13), R^ν_{ij} is the Raman tensor for phonon mode ν, which involves electron–phonon and dipole matrix elements as well as the electronic transition energies and the incident excitation frequency. Equation (13) has been used to compute the Raman spectra of the 733 most stable, non-magnetic monolayers in C2DB for a range of excitation frequencies and polarisation configurations. Note that the Raman shift $\hbar\omega$ is typically expressed in cm^{-1} with

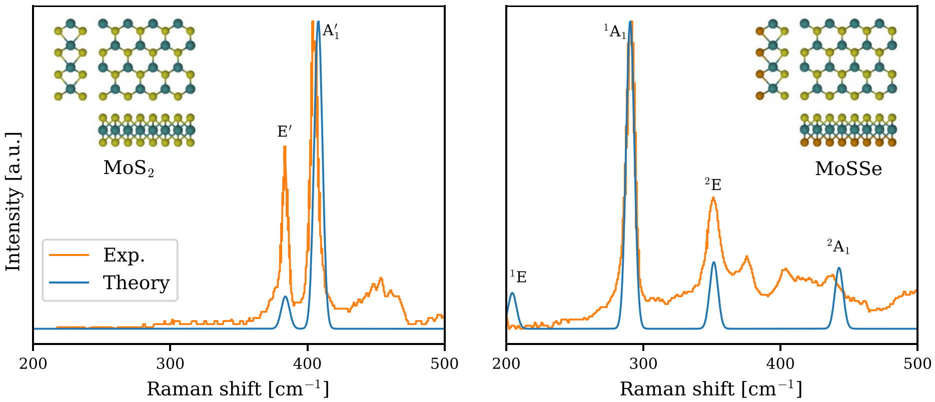


Figure 21. Comparison of the calculated and experimental (extracted from [62]) Raman spectrum of MoS_2 (left) and MoSSe (right). The excitation wavelength is 532 nm, and both the polarisation of both the incoming and outgoing photons are along the y-direction. The Raman peaks are labelled according to the irreducible representations of the corresponding vibrational modes. Adapted from [115].

1 meV equivalent to 8.0655 cm^{-1}. In addition, for generating the Raman spectra, we have used a Gaussian $[G(\omega) = (\sigma\sqrt{2\pi})^{-1}\exp(-\omega^2/2\sigma^2)]$ with a variance $\sigma = 3$ cm^{-1} to replace the Dirac delta function, which accounts for the inhomogeneous broadening of phonon modes.

As an example, figure 21 shows the calculated Raman spectrum of monolayer MoS_2 and the Janus monolayer MoSSe (see section 4.1). Experimental Raman spectra extracted from reference [62] are shown for comparison. For both materials, good agreement between theory and experiment is observed for the peak positions and relative amplitudes of the main peaks. The small deviations can presumably be attributed to substrate interactions and defects in the experimental samples as well as the neglect of excitonic effects in the calculations. The qualitative differences between the Raman spectra can be explained by the different point groups of the materials (C_{3v} and D_{3h}, respectively), see reference [115]. In particular, the lower symmetry of MoSSe results in a lower degeneracy of its vibrational modes leading to more peaks in the Raman spectrum.

Very recently, the Raman spectra computed from third order perturbation theory as described above, were supplemented by spectra obtained from the more conventional Kramers–Heisenberg–Dirac (KHD) approach. Within the KHD method, the Raman tensor is obtained as the derivative of the static electric polarisability (or equivalently, the susceptibility) along the vibrational normal modes [116, 117]:

$$R^{\nu}_{ij} = \sum_{\alpha l} \frac{\partial \chi^{(1)}_{ij}}{\partial r_{\alpha l}} \frac{v^{\nu}_{\alpha l}}{\sqrt{M_\alpha}}. \quad (14)$$

Here, $\chi^{(1)}_{ij}$ is the (first-order) susceptibility tensor, r_α and M_α are the position and atomic mass of atom α, respectively, and $v^{\nu}_{\alpha l}$ is the eigenmode of phonon ν. The two approaches, i.e. the KHD and third-order perturbation approach, can be shown to be equivalent [114], at least when local field effects can be ignored as is typically the case for 2D materials [35]. We have also confirmed this equivalence from our calculations. Furthermore, the computational cost of both methods is also similar [115]. However, the KHD approach typically converge faster with respect to both the number of bands and k-grid compared to the third-order perturbation method. This stems from the general fact that higher-order perturbation calculations converge slower with respect to k-grid and they require additional summations over a complete basis set (virtual states) and hence a larger number of bands [118]. Currently, Raman spectra from both approaches can be found at the C2DB website.

5.10. Second harmonics generation

Nonlinear optical (NLO) phenomena such as harmonic generation, Kerr, and Pockels effects are of great technological importance for lasers, frequency converters, modulators, etc. In addition, NLO spectroscopy has been extensively employed to obtain insight into materials properties [119] that are not accessible by e.g. linear optical spectroscopy. Among numerous nonlinear processes, second-harmonic generation (SHG) has been widely used for generating new frequencies in lasers as well as identifying crystal orientations and symmetries.

Recently, the SHG spectrum was calculated for 375 non-magnetic, non-centrosymmetric semiconducting monolayers of the C2DB, and multiple 2D materials with giant optical nonlinearities were identified [120]. In the SHG process, two incident photons at frequency ω generate an emitted photon at frequency of 2ω. Assume that a mono-harmonic electric

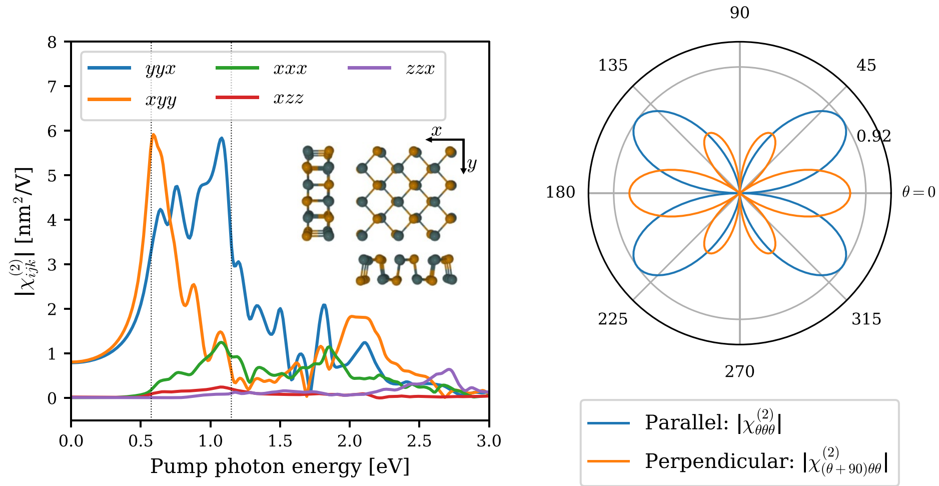


Figure 22. (Left panel) SHG spectra of monolayer Ge$_2$Se$_2$, where only non-vanishing independent tensor elements are shown. The vertical dashed lines mark $\hbar\omega = E_g/2$ and $\hbar\omega = E_g$, respectively. The crystal structure of Ge$_2$Se$_2$ structure is shown in the inset. (Right panel) The rotational anisotropy of the static ($\omega = 0$) SHG signal for parallel (blue) and perpendicular (red) polarisation configurations with θ defined with respect to the crystal x-axis.

field written $\boldsymbol{\mathcal{F}}(t) = \sum_i \mathcal{F}_i \mathbf{e}_i e^{-i\omega t}$+c.c. is incident on the material, where $\mathbf{e}_i$ denotes the unit vector along direction $i \in \{x, y, z\}$. The electric field induces a SHG polarisation density $\mathbf{P}^{(2)}$, which can be obtained from the quadratic susceptibility tensor $\chi^{(2)}_{ijk}$,

$$P^{(2)}_i(t) = \epsilon_0 \sum_{jk} \chi^{(2)}_{ijk}(\omega, \omega)\mathcal{F}_j\mathcal{F}_j e^{-2i\omega t} + \text{c.c.}, \quad (15)$$

where ε_0 denotes the vacuum permittivity. $\chi^{(2)}_{ijk}$ is a symmetric (due to intrinsic permutation symmetry i.e. $\chi^{(2)}_{ijk} = \chi^{(2)}_{ijk}$) rank-3 tensor with at most 18 independent elements. Furthermore, similar to the piezoelectric tensor, the point group symmetry reduces the number of independent tensor elements.

In the C2DB, the quadratic susceptibility is calculated using density matrices and perturbation theory [118, 121] with the involved transition dipole matrix elements and band energies obtained from DFT. The use of DFT single-particle orbitals implies that excitonic effects are not accounted for. The number of empty bands included in the sum over bands was set to three times the number of occupied bands. The width of the Fermi–Dirac occupation factor was set to $k_BT = 50$ meV, and a line-shape broadening of $\eta = 50$ meV was used in all spectra. Furthermore, time-reversal symmetry was imposed in order to reduce the $\mathbf{k}$-integrals to half the BZ. For various 2D crystal classes, it was verified by explicit calculation that the quadratic tensor elements fulfil the expected symmetries, e.g. that they all vanish identically for centrosymmetric crystals.

As an example, the calculated SHG spectra for monolayer Ge$_2$Se$_2$ is shown in figure 22 (left panel). Monolayer Ge$_2$Se$_2$ has five independent tensor elements, $\chi^{(2)}_{xxx}$, $\chi^{(2)}_{xyy}$, $\chi^{(2)}_{xzz}$, $\chi^{(2)}_{yyx} = \chi^{(2)}_{yxy}$, and $\chi^{(2)}_{zzx} = \chi^{(2)}_{zxz}$, since it is a group-IV dichalcogenide with an orthorhombic crystal structure (space group 31 and point group C_{2v}). Note that, similar to the linear susceptibility, the bulk quadratic susceptibility (with SI units of m V^{-1}) is ill-defined for 2D materials (since the volume is ambiguous) [120]. Instead, the unambiguous *sheet* quadratic susceptibility (with SI units of m^2 V^{-1}) is evaluated. In addition to the frequency-dependent SHG spectrum, the angular dependence of the static ($\omega = 0$) SHG intensity at normal incidence for parallel and perpendicular polarisations (relative to the incident electric field) is calculated, see figure 22 (right panel). Such angular resolved SHG spectroscopy has been widely used for determining the crystal orientation of 2D materials. The calculated SHG spectra for all non-vanishing inequivalent polarisation configurations and their angular dependence, are available in the C2DB.

Since C2DB has already gathered various material properties of numerous 2D materials, it provides a unique opportunity to investigate interrelations between different material properties. For example, the strong dependence of the quadratic optical response on the electronic band gap was demonstrated on basis of the C2DB data [120]. As another example of a useful correlation, the static quadratic susceptibility is plotted versus the static linear susceptibility for 67 TMDCs (with formula MX$_2$, space group 187) in figure 23. Note that for materials with several independent tensor elements, only the largest is shown. There is a very clear correlation between the two quantities. This is not unexpected as both

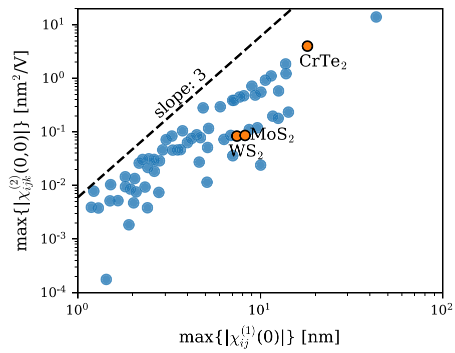


Figure 23. Scatter plot (double log scale) of the static sheet quadratic susceptibility $|\chi^{(2)}_{ijk}|$ versus the static sheet linear susceptibility $|\chi^{(1)}_{ij}|$ for 67 TMDCs (with chemical formula MX$_2$ and space group 187). A few well known materials are highlighted.

the linear and quadratic optical responses are functions of the transition dipole moments and transition energies. More interestingly, the strength of the quadratic response seems to a very good approximation to be given by a universal constant times the linear susceptibility to the power of three (ignoring polarisation indices), i.e.

$$\chi^{(2)}(0,0) \approx A\chi^{(1)}(0)^3, \qquad (16)$$

where A is only weakly material dependent. Note that this scaling law is also known in classical optics as semi-empirical Miller's rule for non-resonant quadratic responses [122], which states that the second order electric susceptibility is proportional to the product of the first-order susceptibilities at the three frequencies involved.

6. Machine learning properties

In recent years, material scientists have shown great interest in exploiting the use of machine learning (ML) techniques for predicting materials properties and guiding the search for new materials. ML is the scientific study of algorithms and statistical models that computer systems can use to perform a specific task without using explicit instructions but instead relying on patterns and inference. Within the domain of materials science, one of the most frequent problems is the mapping from atomic configuration to material property, which can be used e.g. to screen large material spaces in search of optimal candidates for specific applications [123, 124].

In the ML literature, the mathematical representation of the input observations is often referred to as a fingerprint. Any fingerprint must satisfy a number of general requirements [125]. In particular, a fingerprint must be:

(a) *Complete:* The fingerprint should incorporate all the relevant input for the underlying problem, i.e. materials with different properties should have different fingerprints.
(b) *Compact:* The fingerprint should contain no or a minimal number of features redundant to the underlying problem. This includes being invariant to rotations, translations and other transformations that leave the properties of the system invariant.
(c) *Descriptive:* Materials with similar target values should have similar fingerprints.
(d) *Simple:* The fingerprint should be efficient to evaluate. In the present context, this means that calculating the fingerprint should be significantly faster than calculating the target property.

Several types of atomic-level materials fingerprints have been proposed in the literature, including general purpose fingerprints based on atomistic properties [126, 127] possibly encoding information about the atomic structure, i.e. atomic positions [125, 128, 129], and specialised fingerprints tailored for specific applications (materials/properties) [130, 131].

The aim of this section is to demonstrate how the C2DB may be utilised for ML-based prediction of general materials properties. Moreover, the study serves to illustrate the important role of the fingerprint for such problems. The 2D materials are represented using three different fingerprints: two popular structural fingerprints and a more advanced fingerprint that encodes information about the electronic structure via the PDOS. The target properties include the HSE06 band gap, the PBE heat of formation (ΔH), the exciton binding energy (E_B) obtained from the many-body BSE, the in-plane static polarisability calculated in the RPA averaged over the x and y polarisation directions ($\langle\alpha_i\rangle$), and the in-plane Voigt modulus ($\langle C_{ii}\rangle$) defined as $\frac{1}{4}(C_{11}+C_{22}+2C_{12})$, where C_{ij} is a component of the elastic stiffness tensor in Mandel notation.

To introduce the data, figure 24 shows pair-plots of the dual-property relations of these properties. The plots in the diagonal show the single-property histograms, whereas the off-diagonals show dual-property scatter plots below the diagonal and histograms above the diagonal. Clearly, there are only weak correlations between most of the properties, with the largest degree of correlation observed between the HSE06 gap and exciton binding energy. The lack of strong correlations motivates the use of ML for predicting the properties.

The prediction models are build using the Ewald sum matrix and many-body tensor representation (MBTR) as structural fingerprints. The Ewald fingerprint is a version of the simple Coulomb matrix fingerprint [128] modified to periodic systems [125]. The MBTR encodes first, second and third order

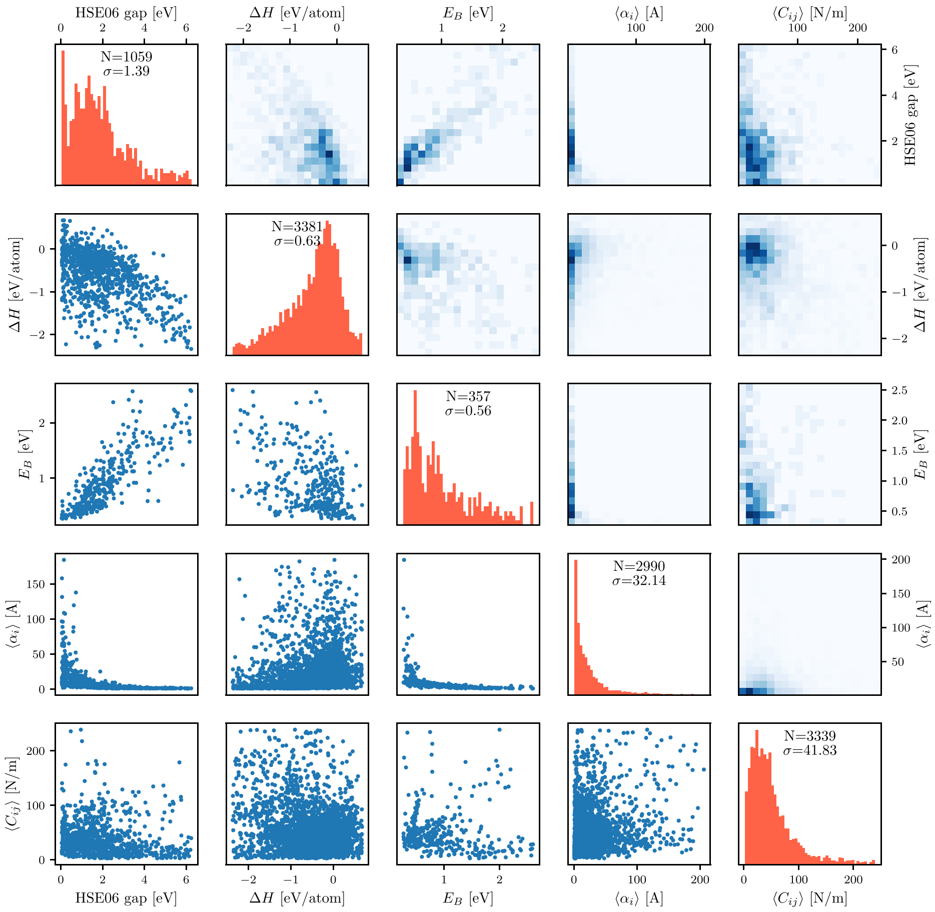


Figure 24. Pair-plot of selected properties from C2DB. The diagonal contains the single property histograms. Below the diagonal are two-property scatter plots showing the correlation between properties and above the diagonal are two-property histograms. properties include the HSE06 band gap, the PBE heat of formation (ΔH), the exciton binding energy (E_B) calculated from the BSE, the in-plane static polarisability calculated in the RPA and averaged over the x and y polarisation directions ($\langle \alpha_i \rangle$), and the in-plane Voigt modulus ($\langle C_{ii} \rangle$) defined as $\frac{1}{4}(C_{11} + C_{22} + 2C_{12})$, where C_{ij} is a component of the elastic stiffness tensor.

terms like atomic numbers, distances and angles between atoms in the system [129]. As an alternative to the structural fingerprints, a representation based on the PBE PDOS is also tested. This fingerprint[6] encodes the coupling between the PDOS at different atomic orbitals in both energy and real space. It is defined as:

$$\rho_{\nu\nu'}(E,R) = \sum_{a \in \text{cell}} \sum_{a'} \rho_{a\nu}(E)\rho_{a'\nu'}(E)G \times (R - |R_a - R_{a'}|), \tag{17}$$

where G is a Gaussian smearing function, a denotes the atoms, ν denotes atomic orbitals, and the PDOS is given by:

$$\rho_{a\nu}(E) = \sum_n |\langle \psi_n | a\nu \rangle|^2 G(E - \epsilon_n), \tag{18}$$

where n runs over all eigenstates of the system. Since this fingerprint requires a DFT-PBE calculation to be performed, additional features derivable from the DFT calculation can be added to the fingerprint. In this study, the PDOS fingerprint is amended by the PBE band gap. The latter can in principle be extracted from the PDOS, but its explicit inclusion has been found to improve the performance of the model.

A Gaussian process regression using a simple Gaussian kernel with a noise component is used as learning algorithm. The models are trained using 5-fold cross validation on a training set consisting of 80% of the materials with the remaining 20% held aside as test data. Prior to training the model, the input space is reduced to 50 features using principal component analysis (PCA). This step is necessary to reduce the huge number of features in the MBTR fingerprint to a manageable size. Although this is not required for the Ewald and PDOS fingerprints,

[6] Details will be published elsewhere.

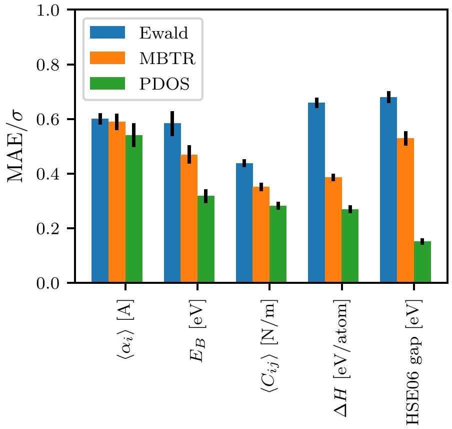


Figure 25. Prediction scores (MAE normalised to standard deviation of property values) for the test sets of selected properties using a Gaussian process regression.

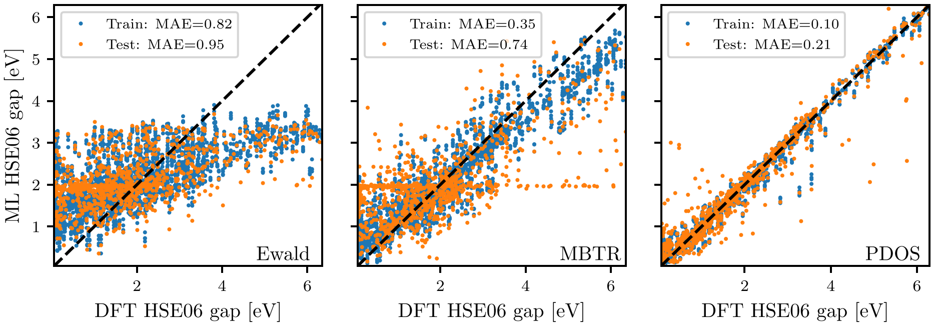


Figure 26. ML predicted HSE06 gap values vs. true values for Ewald, MBTR and PDOS fingerprints with MAE's for train and test set included. The PDOS is found to perform significantly better for the prediction of HSE06 gap.

we perform the same feature reduction in all cases. The optimal number of features depends on the choice of fingerprint, target property and learning algorithm, but for consistency 50 PCA components are used for all fingerprints and properties in this study.

Figure 25 shows the prediction scores obtained for the five properties using the three different fingerprints. The employed prediction score is the mean absolute error of the test set normalised by the standard deviation of the property values (standard deviations are annotated in the diagonal plots in figure 24). In general, the PDOS fingerprint outperforms the structural fingerprints. The difference between prediction scores is smallest for the static polarisability $\langle\alpha_i\rangle$ and largest for the HSE06 gap. It should be stressed that although the evaluation of the PBE-PDOS fingerprint is significantly more time consuming than the evaluation of the structural fingerprints, it is still much faster than the evaluation of all the target properties. Moreover, structural fingerprints require the atomic structure, which in turns requires a DFT structure optimisation (unless the structure is available by other means).

The HSE06 band gap shows the largest sensitivity to the employed fingerprint. To elaborate on the HSE06 results, figure 26 shows the band gap predicted using each of the three different fingerprints plotted against the true band gap. The mean absolute errors on the test set is 0.95 and 0.74 eV for Ewald and MBTR fingerprints, respectively, while the PDOS significantly outperforms the other fingerprints with a test MAE of only 0.21 eV. This improvement in prediction accuracy is partly due to the presence of the PBE gap in the PDOS fingerprint. However, our analysis shows that the pure PDOS fingerprint without the PBE gap still outperforms the structural fingerprints. Using only the PBE gap as feature results in a test MAE of 0.28 eV.

The current results show that the precision of ML-based predictions are highly dependent on the type of target property and the chosen material representation. For some properties, the mapping between atomic structure and property is easier to learn while others might require more/deeper information, e.g. in terms of electronic structure fingerprints. Our results clearly demonstrate the potential of encoding electronic structure information into the material fingerprint, and we anticipate more work on this relevant and exciting topic in the future.

7. Summary and outlook

We have documented a number of extensions and improvements of the C2DB made in the period 2018–2020. The new developments include: (1) A refined and more stringent workflow for filtering prospective 2D materials and classifying them according to their crystal structure, magnetic state and stability. (2) Improvements of the methodology used to compute certain challenging properties such as the full stiffness tensor, effective masses, G_0W_0 band structures, and optical absorption spectra. (3) New materials including 216 MXY Janus monolayers and 574 monolayers exfoliated from experimentally known bulk crystals. In addition, ongoing efforts to systematically obtain and characterise bilayers in all possible stacking configurations as well as point defects in the semiconducting monolayers, have been described. (4) New properties including exfoliation energies, spontaneous polarisations, Bader charges, piezoelectric tensors, IR polarisabilities, topological invariants, magnetic exchange couplings, Raman spectra, and SHG spectra. It should be stressed that the C2DB will continue to grow as new structures and properties are being added, and thus the present paper should not be seen as a final report on the C2DB but rather a snapshot of its current state.

In addition to the above mentioned improvements relating to data quantity and quality, the C2DB has been endowed with a comprehensive documentation layer. In particular, all data presented on the C2DB website are now accompanied by an information field that explains the meaning and representation (if applicable) of the data and details how it was calculated thus making the data easier to understand, reproduce, and deploy.

The C2DB has been produced using the ASR in combination with the GPAW electronic structure code and the MyQueue task and workflow scheduling system. The ASR is a newly developed Python-based framework designed for high-throughput materials computations. The highly flexible and modular nature of the ASR and its strong coupling to the well established community-driven ASE project, makes it a versatile framework for both high- and low-throughput materials simulation projects. The ASR and the C2DB-ASR workflow are distributed as open source code. A detailed documentation of the ASR will be published elsewhere.

While the C2DB itself is solely concerned with the properties of perfect monolayer crystals, ongoing efforts focus on the systematic characterisation of homo-bilayer structures as well as point defects in monolayers. The data resulting from these and other similar projects will be published as separate, independent databases, but will be directly interlinked with the C2DB making it possible to switch between them in a completely seamless fashion. These developments will significantly broaden the scope and usability of the C2DB+ (+ stands for associated databases) that will help theoreticians and experimentalists to navigate one of the most vibrant and rapidly expanding research fields at the crossroads of condensed matter physics, photonics, nanotechnology, and chemistry.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.11583/DTU.14616660.

Acknowledgments

The Center for Nanostructured Graphene (CNG) is sponsored by the Danish National Research Foundation, Project DNRF103. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program Grant Agreement No. 773122 (LIMA) and Grant Agreement No. 951786 (NOMAD CoE). T D acknowledges financial support from the German Research Foundation (DFG Projects No. DE 2749/2-1).

ORCID iDs

Morten Niklas Gjerding https://orcid.org/0000-0002-5256-660X
Alireza Taghizadeh https://orcid.org/0000-0003-0876-9538
Asbjørn Rasmussen https://orcid.org/0000-0001-7110-9255
Sajid Ali https://orcid.org/0000-0001-7865-2664
Fabian Bertoldo https://orcid.org/0000-0002-1219-8689
Thorsten Deilmann https://orcid.org/0000-0003-4165-2446
Nikolaj Rørbæk Knøsgaard https://orcid.org/0000-0003-3709-5464
Mads Kruse https://orcid.org/0000-0002-0599-5110
Ask Hjorth Larsen https://orcid.org/0000-0001-5267-6852
Simone Manti https://orcid.org/0000-0003-3770-0863
Thomas Garm Pedersen https://orcid.org/0000-0002-9466-6190
Urko Petralanda https://orcid.org/0000-0003-0226-0028

Thorbjørn Skovhus https://orcid.org/0000-0001-5215-6419
Mark Kamper Svendsen https://orcid.org/0000-0001-9718-849X
Jens Jørgen Mortensen https://orcid.org/0000-0001-5090-6706
Thomas Olsen https://orcid.org/0000-0001-6256-9284
Kristian Sommer Thygesen https://orcid.org/0000-0001-5197-214X

References

[1] Kohn W and Sham L J 1965 *Phys. Rev.* **140** A1133
[2] Schwierz F 2010 *Nat. Nanotechnol.* **5** 487
[3] Novoselov K, Mishchenko A, Carvalho A and Castro Neto A 2016 *Science* **353** 6298
[4] Ferrari A C *et al* 2015 *Nanoscale* **7** 4598–810
[5] Bhimanapati G R *et al* 2015 *ACS Nano* **9** 11509–39
[6] Haastrup S *et al* 2018 *2D Mater.* **5** 042002
[7] Shivayogimath A *et al* 2019 *Nat. Commun.* **10** 1–7
[8] Zhou J *et al* 2018 *Nature* **556** 355–9
[9] Anasori B, Lukatskaya M R and Gogotsi Y 2017 *Nat. Rev. Mater.* **2** 1–17
[10] Dou L *et al* 2015 *Science* **349** 1518–21
[11] Mounet N *et al* 2018 *Nat. Nanotechnol.* **13** 246–52
[12] Ashton M, Paul J, Sinnott S B and Hennig R G 2017 *Phys. Rev. Lett.* **118** 106101
[13] Geim A K and Grigorieva I V 2013 *Nature* **499** 419–25
[14] Cao Y, Fatemi V, Fang S, Watanabe K, Taniguchi T, Kaxiras E and Jarillo-Herrero P 2018 *Nature* **556** 43–50
[15] Bistritzer R and MacDonald A H 2011 *Proc. Natl Acad. Sci.* **108** 12233–7
[16] Zhao X *et al* 2020 *Nature* **581** 171–7
[17] Wan J, Lacey S D, Dai J, Bao W, Fuhrer M S and Hu L 2016 *Chem. Soc. Rev.* **45** 6742–65
[18] Wilkinson M D *et al* 2016 *Sci. Data* **3** 1–9
[19] Wirtz L, Marini A and Rubio A 2006 *Phys. Rev. Lett.* **96** 126104
[20] Cudazzo P, Tokatly I V and Rubio A 2011 *Phys. Rev.* B **84** 085406
[21] Klots A *et al* 2014 *Sci. Rep.* **4** 6608
[22] Chernikov A, Berkelbach T C, Hill H M, Rigosi A, Li Y, Aslan O B, Reichman D R, Hybertsen M S and Heinz T F 2014 *Phys. Rev. Lett.* **113** 076802
[23] Olsen T, Latini S, Rasmussen F and Thygesen K S 2016 *Phys. Rev. Lett.* **116** 056401
[24] Riis-Jensen A C, Gjerding M N, Russo S and Thygesen K S 2020 *Phys. Rev.* B **102** 201402
[25] Felipe H, Xian L, Rubio A and Louie S G 2020 *Nat. Commun.* **11** 1–10
[26] Sohier T, Gibertini M, Calandra M, Mauri F and Marzari N 2017 *Nano Lett.* **17** 3758–63
[27] Ugeda M M *et al* 2014 *Nat. Mater.* **13** 1091–5
[28] Winther K T and Thygesen K S 2017 *2D Mater.* **4** 025059
[29] Wang Z, Rhodes D A, Watanabe K, Taniguchi T, Hone J C, Shan J and Mak K F 2019 *Nature* **574** 76–80
[30] Gong C *et al* 2017 *Nature* **546** 265–9
[31] Huang B *et al* 2017 *Nature* **546** 270–3
[32] Chang K *et al* 2016 *Science* **353** 274–8
[33] Olsen T, Andersen E, Okugawa T, Torelli D, Deilmann T and Thygesen K S 2019 *Phys. Rev. Mater.* **3** 024005
[34] Marrazzo A, Gibertini M, Campi D, Mounet N and Marzari N 2019 *Nano Lett.* **19** 8431–40
[35] Thygesen K S 2017 *2D Mater.* **4** 022004
[36] Torelli D and Olsen T 2018 *2D Mater.* **6** 015028
[37] Rasmussen F A and Thygesen K S 2015 *J. Phys. Chem.* C **119** 13169–83
[38] Enkovaara J *et al* 2010 *J. Phys.: Condens. Matter.* **22** 253202
[39] Mortensen J J, Hansen L B and Jacobsen K W 2005 *Phys. Rev.* B **71** 035109
[40] Gjerding M, Skovhus T, Rasmussen A, Bertoldo F, Larsen A H, Mortensen J J and Thygesen K S 2021 Atomic simulation recipes—a python framework and library for automated workflows (arXiv:2104.13431)
[41] Larsen A H *et al* 2017 *J. Phys.: Condens. Matter.* **29** 273002
[42] Mortensen J J, Gjerding M and Thygesen K S 2020 *J. Open Source Softw.* **5** 1844
[43] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 *JOM* **65** 1501–9
[44] Jain A *et al* 2013 *APL Mater.* **1** 011002
[45] Curtarolo S *et al* 2012 *Comput. Mater. Sci.* **58** 218–26
[46] Ataca C, Sahin H and Ciraci S 2012 *J. Phys. Chem.* C **116** 8983–99
[47] Lebègue S, Björkman T, Klintenberg M, Nieminen R M and Eriksson O 2013 *Phys. Rev.* X **3** 031002
[48] Kormányos A, Burkard G, Gmitra M, Fabian J, Zólyomi V, Drummond N D and Fal'ko V 2015 *2D Mater.* **2** 022001
[49] Zhou J *et al* 2019 *Sci. Data* **6** 1–10
[50] Choudhary K, Kalish I, Beams R and Tavazza F 2017 *Sci. Rep.* **7** 1–16
[51] Larsen A H, Vanin M, Mortensen J J, Thygesen K S and Jacobsen K W 2009 *Phys. Rev.* B **80** 195112
[52] Larsen P M, Pandey M, Strange M and Jacobsen K W 2019 *Phys. Rev. Mater.* **3** 034003
[53] Ong S P *et al* 2013 *Comput. Mater. Sci.* **68** 314–19
[54] Patrick C E, Jacobsen K W and Thygesen K S 2015 *Phys. Rev.* B **92** 201205
[55] Kirklin S, Saal J E, Meredig B, Thompson A, Doak J W, Aykol M, Rühl S and Wolverton C 2015 *npj Computat. Mater.* **1** 1–15
[56] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865–8
[57] Maździarz M 2019 *2D Mater.* **6** 048001
[58] Li Y and Heinz T F 2018 *2D Mater.* **5** 025021
[59] Hadley L N and Dennison D 1947 *J. Opt. Soc. Am.* **37** 451–65
[60] Wang G, Chernikov A, Glazov M M, Heinz T F, Marie X, Amand T and Urbaszek B 2018 *Rev. Mod. Phys.* **90** 021001
[61] Lu A Y *et al* 2017 *Nat. Nanotechnol.* **12** 744–9
[62] Zhang J *et al* 2017 *ACS Nano* **11** 8192–8
[63] Fülöp B *et al* 2018 *2D Mater.* **5** 031013
[64] Riis-Jensen A C, Deilmann T, Olsen T and Thygesen K S 2019 *ACS Nano* **13** 13354
[65] Bychkov Y A and Rashba E I 1984 *J. Phys. C: Solid State Phys.* **17** 6039
[66] Petersen L and Hedegård P 2000 *Surf. Sci.* **459** 49–56
[67] Bergerhoff G, Brown I and Allen F *et al* 1987 *Int. Union Crystallogr., Chester* **360** 77–95
[68] Gražulis S *et al* 2012 *Nucleic Acids Res.* **40** D420–7
[69] Qian X, Liu J, Fu L and Li J 2014 *Science* **346** 1344–7
[70] Torelli D, Moustafa H, Jacobsen K W and Olsen T 2020 *npj Comput. Mater.* **6** 158
[71] Mak K F, Lee C, Hone J, Shan J and Heinz T F 2010 *Phys. Rev. Lett.* **105** 136805
[72] Splendiani A, Sun L, Zhang Y, Li T, Kim J, Chim C Y, Galli G and Wang F 2010 *Nano Lett.* **10** 1271–5
[73] Xiao J *et al* 2020 *Nat. Phys.* **16** 1028–34
[74] Sivadas N, Okamoto S, Xu X, Fennie C J and Xiao D 2018 *Nano Lett.* **18** 7658–64
[75] Liu Y, Wu L, Tong X, Li J, Tao J, Zhu Y and Petrovic C 2019 *Sci. Rep.* **9** 1–8
[76] Yasuda K, Wang X, Watanabe K, Taniguchi T and Jarillo-Herrero P 2020 (arXiv:2010.06600)
[77] Grimme S, Antony J, Ehrlich S and Krieg H 2010 *J. Chem. Phys.* **132** 154104
[78] Northup T and Blatt R 2014 *Nat. Photon.* **8** 356–63
[79] O'Brien J, Furusawa A and Vuckovic J 2009 Photonic quantum technologies nat *Photonics* **3** 687
[80] Zhang S and Northrup J E 1991 *Phys. Rev. Lett.* **67** 2339

[81] Van de Walle C G, Laks D, Neumark G and Pantelides S 1993 *Phys. Rev.* B **47** 9425

[82] Janak J F 1978 *Phys. Rev.* B **18** 7165

[83] Pandey M, Rasmussen F A, Kuhar K, Olsen T, Jacobsen K W and Thygesen K S 2016 *Nano Lett.* **16** 2234–9

[84] Kaappa S, Malola S and Häkkinen H 2018 *J. Phys. Chem.* A **122** 8576–84

[85] Levi G, Ivanov A V and Jonsson H 2020 *Faraday Discuss.* **224** 448-66

[86] Grimme S, Antony J, Ehrlich S and Krieg H 2010 *J. Chem. Phys.* **132** 154104

[87] Bader R F W 1990 *Atoms in Molecules: A Quantum Theory (The Int. Series of Monographs on Chemistry* vol 22) (Oxford: Clarendon)

[88] Tang W, Sanville E and Henkelman G 2009 *J. Phys.: Condens. Matter.* **21** 084204

[89] Resta R 1992 *Ferroelectrics* **136** 51–5

[90] King-Smith R D and Vanderbilt D 1993 *Phys. Rev.* B **47** 3

[91] Zhang S and Yu F 2011 *J. Am. Ceram. Soc.* **94** 3153–70

[92] Maeder M D, Damjanovic D and Setter N 2004 *J. Electroceram.* **13** 385–92

[93] Scott J F 2000 *Ferroelectric Memories* vol 3 (Berlin: Springer)

[94] Rangel T, Fregoso B M, Mendoza B S, Morimoto T, Moore J E and Neaton J B 2017 *Phys. Rev. Lett.* **119** 067402

[95] Resta R and Vanderbilt D 2007 Theory of Polarization: A Modern Approach *Phys. Ferroelectr.* vol 105 (Berlin: Springer) pp 31–68

[96] Gjerding M N, Cavalcante L S R, Chaves A and Thygesen K S 2020 *J. Phys. Chem.* C **124** 11609–16

[97] Ye Z G 2008 *Handbook of Advanced Dielectric, Piezoelectric and Ferroelectric Materials: Synthesis, Properties and Applications* (Amsterdam: Elsevier)

[98] Ogawa T 2016 *Piezoelectric Materials* (Croatia: InTech)

[99] Vanderbilt D 1999 *J. Phys. Chem. Solids* **61** 147–51

[100] Authier A 2003 *Int. Tables for Crystallography: Volume D: Physical Properties of Crystals* (Dordrecht: Springer)

[101] Duerloo K A N, Ong M T and Reed E J 2012 *J. Phys. Chem. Lett.* **3** 2871–6

[102] Zhu H *et al* 2015 *Nat. Nanotechnol.* **10** 151–5

[103] Taherinejad M, Garrity K F and Vanderbilt D 2014 *Phys. Rev.* B **89** 115102

[104] Olsen T 2016 *Phys. Rev.* B **94** 235106

[105] Fu L 2011 *Phys. Rev. Lett.* **106** 106802

[106] Olsen T, Andersen E, Okugawa T, Torelli D, Deilmann T and Thygesen K S 2019 *Phys. Rev. Mater.* **3** 024005

[107] Benalcazar W A, Bernevig B A and Hughes T L 2017 *Phys. Rev.* B **96** 245115

[108] Mermin N D and Wagner H 1966 *Phys. Rev. Lett.* **17** 1133–6

[109] Olsen T 2019 *MRS Commun.* **9** 1142–50

[110] Olsen T 2017 *Phys. Rev.* B **96** 125143

[111] Torelli D and Olsen T 2020 *J. Phys.: Condens. Matter.* **32** 335802

[112] Lado J L and Fernández-Rossier J 2017 *2D Mater.* **4** 035002

[113] Torelli D, Thygesen K S and Olsen T 2019 *2D Mater.* **6** 045018

[114] Long D A 2002 *The Raman Effect: A Unified Treatment of the Theory of Raman Scattering by Molecules* (Chichester: Wiley)

[115] Taghizadeh A, Leffers U, Pedersen T G and Thygesen K S 2020 *Nat. Commun.* **11** 3011

[116] Lee S Y and Heller E J 1979 *J. Chem. Phys.* **71** 4777

[117] Umari P and Pasquarello A 2003 *J. Phys. Condens. Matter* **15** S1547–52

[118] Taghizadeh A, Hipolito F and Pedersen T G 2017 *Phys. Rev.* B **96** 195413

[119] Prylepa A *et al* 2018 *J. Phys. D: Appl. Phys* **51** 043001

[120] Taghizadeh A, Thygesen K S and Pedersen T G 2021 *ACS Nano* **15** 7155

[121] Aversa C and Sipe J E 1995 *Phys. Rev.* B **52** 14636–45

[122] Miller R C 1964 *Appl. Phys. Lett.* **5** 17–19

[123] Schmidt J, Marques M R G, Botti S and Marques M A L 2019 *Computat. Mater.* **5** 83

[124] Zhuo Y, Mansouri Tehrani A and Brgoch J 2018 *J. Phys. Chem. Lett.* **9** 1668–73

[125] Faber F, Lindmaa A, von Lilienfeld O A and Armiento R 2015 *Int. J. Quantum Chem.* **115** 1094–101

[126] Ward L, Agrawal A, Choudhary A and Wolverton C 2016 *Computat. Mater.* **2** 1–7

[127] Ghiringhelli L M, Vybiral J, Levchenko S V, Draxl C and Scheffler M 2015 *Phys. Rev. Lett.* **114** 105503

[128] Rupp M, Tkatchenko A, Müller K R and Von Lilienfeld A 2012 *Phys. Rev. Lett.* **108** 058301

[129] Huo H and Rupp M 2018 Unified representation of molecules and crystals for machine learning (arXiv:1704.06439)

[130] Jorgensen P B, Mesta M, Shil S, García Lastra J M, Jacobsen K W, Thygesen K S and Schmidt M N 2018 *J. Chem. Phys.* **148** 241735

[131] Rajan A C, Mishra A, Satsangi S, Vaish R, Mizuseki H, Lee K R and Singh A K 2018 *Chem. Mater.* **30** 4031–8

Bibliography

[1] 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 (2013).

[2] G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, and A. Fazzio, "From DFT to machine learning: recent approaches to materials science–a review," Journal of Physics: Materials **2**, 032001 (2019).

[3] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, "A high-throughput infrastructure for density functional theory calculations," Computational Materials Science **50**, 2295 (2011).

[4] D. J. Griffiths and D. F. Schroeter, *Introduction to quantum mechanics*, 3rd edition (Cambridge University Press, 2018).

[5] L. E. Ballentine, *Quantum mechanics*, 2nd (WORLD SCIENTIFIC, 2014).

[6] H. Bruus and K. Flensberg, *Many-body quantum field theory in condensed matter physics: an introduction*, English (Oxford University Press, United Kingdom, 2003).

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

[8] J. Alzubi, A. Nayyar, and A. Kumar, "Machine learning from theory to algorithms: an overview," Journal of Physics: Conference Series **1142**, 012012 (2018).

[9] 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 Computational Materials **5**, `10.1038/s41524-019-0221-0` (2019).

[10] 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).

[11] H. Huo and M. Rupp, *Unified representation of molecules and crystals for machine learning*, 2017.

[12] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, "The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies," npj Computational Materials **1**, `10.1038/npjcompumats.2015.10` (2015).

[13] A. J. Cohen, P. Mori-Sánchez, and W. Yang, "Challenges for density functional theory," Chemical Reviews **112**, 289 (2011).

[14] P. Verma and D. G. Truhlar, “Status and challenges of density functional theory,” Trends in Chemistry **2**, Special Issue - Laying Groundwork for the Future, 302 (2020).

[15] J. Kohanoff, *Electronic structure calculations for solids and molecules: theory and computational methods* (Cambridge University Press, 2006).

[16] 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 Materials **1**, 011002 (2013).

[17] T. J. Jacobsson, A. Hultqvist, A. García-Fernández, A. Anand, A. Al-Ashouri, A. Hagfeldt, A. Crovetto, A. Abate, A. G. Ricciardulli, A. Vijayan, A. Kulkarni, A. Y. Anderson, B. P. Darwich, B. Yang, B. L. Coles, C. A. R. Perini, C. Rehermann, D. Ramirez, D. Fairen-Jimenez, D. D. Girolamo, D. Jia, E. Avila, E. J. Juarez-Perez, F. Baumann, F. Mathies, G. S. A. González, G. Boschloo, G. Nasti, G. Paramasivam, G. Martínez-Denegri, H. Näsström, H. Michaels, H. Köbler, H. Wu, I. Benesperi, M. I. Dar, I. B. Pehlivan, I. E. Gould, J. N. Vagott, J. Dagar, J. Kettle, J. Yang, J. Li, J. A. Smith, J. Pascual, J. J. Jerónimo-Rendón, J. F. Montoya, J.-P. Correa-Baena, J. Qiu, J. Wang, K. Sveinbjörnsson, K. Hirselandt, K. Dey, K. Frohna, L. Mathies, L. A. Castriotta, M. H. Aldamasy, M. Vasquez-Montoya, M. A. Ruiz-Preciado, M. A. Flatken, M. V. Khenkin, M. Grischek, M. Kedia, M. Saliba, M. Anaya, M. Veldhoen, N. Arora, O. Shargaieva, O. Maus, O. S. Game, O. Yudilevich, P. Fassl, Q. Zhou, R. Betancur, R. Munir, R. Patidar, S. D. Stranks, S. Alam, S. Kar, T. Unold, T. Abzieher, T. Edvinsson, T. W. David, U. W. Paetzold, W. Zia, W. Fu, W. Zuo, V. R. F. Schröder, W. Tress, X. Zhang, Y.-H. Chiang, Z. Iqbal, Z. Xie, and E. Unger, “An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles,” Nature Energy **7**, 107 (2021).

[18] F. Ricci, W. Chen, U. Aydemir, G. J. Snyder, G.-M. Rignanese, A. Jain, and G. Hautier, “An ab initio electronic transport database for inorganic materials,” Scientific Data **4**, `10.1038/sdata.2017.85` (2017).

[19] K. S. Novoselov, A. Mishchenko, A. Carvalho, and A. H. C. Neto, “2d materials and van der waals heterostructures,” Science **353**, `10.1126/science.aac9439` (2016).

[20] A. Gupta, T. Sakthivel, and S. Seal, “Recent development in 2d materials beyond graphene,” Progress in Materials Science **73**, 44 (2015).

[21] S. Li, Y. Liu, D. Chen, Y. Jiang, Z. Nie, and F. Pan, “Encoding the atomic structure for machine learning in materials science,” WIREs Computational Molecular Science **12**, `10.1002/wcms.1558` (2021).

[22] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, “Crystal structure representations for machine learning models of formation energies,” International Journal of Quantum Chemistry **115**, 1094 (2015).

[23] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. von Lilienfeld, "Prediction errors of molecular machine learning models lower than hybrid DFT error," Journal of Chemical Theory and Computation **13**, 5255 (2017).

[24] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," npj Computational Materials **2**, 10.1038/npjcompumats.2016.28 (2016).

[25] T. Xie and J. C. Grossman, "Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties," Physical Review Letters **120**, 145301 (2018).

[26] G. G. C. Peterson and J. Brgoch, "Materials discovery through machine learning formation energy," Journal of Physics: Energy **3**, 022002 (2021).

[27] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, "Graph networks as a universal machine learning framework for molecules and crystals," Chemistry of Materials **31**, 3564 (2019).

[28] G. R. Schleder, C. M. Acosta, and A. Fazzio, "Exploring two-dimensional materials thermodynamic stability via machine learning," ACS Applied Materials & Interfaces **12**, 20149 (2019).

[29] B. Olsthoorn, R. M. Geilhufe, S. S. Borysov, and A. V. Balatsky, "Band gap prediction for large organic crystal structures with machine learning," Advanced Quantum Technologies **2**, 1900023 (2019).

[30] Y. Zhuo, A. M. Tehrani, and J. Brgoch, "Predicting the band gaps of inorganic solids by machine learning," The Journal of Physical Chemistry Letters **9**, 1668 (2018).

[31] A. C. Rajan, A. Mishra, S. Satsangi, R. Vaish, H. Mizuseki, K.-R. Lee, and A. K. Singh, "Machine-learning-assisted accurate band gap predictions of functionalized MXene," Chemistry of Materials **30**, 4031 (2018).

[32] D. Wolpert and W. Macready, "No free lunch theorems for optimization," IEEE Transactions on Evolutionary Computation **1**, 67 (1997).

[33] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," The bulletin of mathematical biophysics **5**, 115 (1943).

[34] L. Breiman, "Bagging predictors," Machine Learning **24**, 123 (1996).

[35] T. K. Ho, "Random decision forests," in Proceedings of 3rd international conference on document analysis and recognition, Vol. 1 (1995), 278–282 vol.1.

[36] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, KDD '16 (2016), pages 785–794.

[37] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," The Annals of Statistics **29**, 1189 (2001).

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

[39] R. Cohn and E. Holm, "Unsupervised machine learning via transfer learning and k-means clustering to classify materials image data," Integrating Materials and Manufacturing Innovation **10**, 231 (2021).

[40] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu, "A comprehensive survey of clustering algorithms: state-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," Engineering Applications of Artificial Intelligence **110**, 104743 (2022).

[41] R. Ahuja, A. Chug, S. Gupta, P. Ahuja, and S. Kohli, "Classification and clustering algorithms of machine learning with their applications," in *Nature-inspired computation in data mining and machine learning* (Springer International Publishing, September 2019), pages 225–248.

[42] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," Journal of Machine Learning Research **9**, 2579 (2008).

[43] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning* (Springer New York, 2009).

[44] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in neural information processing systems, Vol. 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (2017).

[45] B. Parsaeifard, D. S. De, A. S. Christensen, F. A. Faber, E. Kocer, S. De, J. Behler, O. A. von Lilienfeld, and S. Goedecker, "An assessment of the structural resolution of various fingerprints commonly used in machine learning," Machine Learning: Science and Technology **2**, 015018 (2021).

[46] R. Batra, H. D. Tran, C. Kim, J. Chapman, L. Chen, A. Chandrasekaran, and R. Ramprasad, "General atomic neighborhood fingerprint for machine learning-based methods," The Journal of Physical Chemistry C **123**, 15859 (2019).

[47] M. F. Langer, A. Goeßmann, and M. Rupp, "Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning," npj Computational Materials **8**, `10.1038/s41524-022-00721-x` (2022).

[48] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," Phys. Chem. Chem. Phys. **18**, 13754 (2016).

[49] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, "DScribe: Library of descriptors for machine learning in materials science," Computer Physics Communications **247**, 106949 (2020).

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

[51] E. Fermi, "Un metodo statistico per la determinazione di alcune priorieta dell'atome," Rend. Accad. Naz. Lincei **6**, 32 (1927).

[52] L. H. Thomas, "The calculation of atomic fields," Mathematical Proceedings of the Cambridge Philosophical Society **23**, 542 (1927).

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

[54] J. P. Perdew, "Density-functional approximation for the correlation energy of the inhomogeneous electron gas," Phys. Rev. B **33**, 8822 (1986).

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

[56] J. Heyd, G. E. Scuseria, and M. Ernzerhof, "Hybrid functionals based on a screened coulomb potential," The Journal of Chemical Physics **118**, 8207 (2003).

[57] A. Seidl, A. Görling, P. Vogl, J. A. Majewski, and M. Levy, "Generalized kohn-sham schemes and the band-gap problem," Physical Review B **53**, 3764 (1996).

[58] J. P. Perdew, "Density functional theory and the band gap problem," International Journal of Quantum Chemistry **28**, 497 (1985).

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

[60] D. Golze, M. Dvorak, and P. Rinke, "The gw compendium: a practical guide to theoretical photoemission spectroscopy," Frontiers in Chemistry **7**, `10.3389/fchem.2019.00377` (2019).

[61] G. v. Minnigerode, "Göran grimvall, e. p. wohlfahrt (eds.): the electron-phonon interaction in metals, vol. 16, aus: selected topics in solid state physics. north holland publishing company, amsterdam, new york, oxford 1981. 304 seiten, dfl 125,—," Berichte der Bunsengesellschaft für physikalische Chemie **87**, 453 (1983).

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

[63] M. N. Gjerding, A. Taghizadeh, A. Rasmussen, S. Ali, F. Bertoldo, T. Deilmann, N. R. Knøsgaard, M. Kruse, A. H. Larsen, S. Manti, T. G. Pedersen, U. Petralanda, T. Skovhus, M. K. Svendsen, J. J. Mortensen, T. Olsen, and K. S. Thygesen, "Recent progress of the computational 2d materials database (c2db)," 2D Materials **8**, 044002 (2021).

[64] J. J. Mortensen, L. B. Hansen, and K. W. Jacobsen, "Real-space grid implementation of the projector augmented wave method," Phys. Rev. B **71**, 035109 (2005).

[65] J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. A. Hansen, H. H. Kristoffersen, M. Kuisma, A. H. Larsen, L. Lehtovaara, M. Ljungberg, O. Lopez-Acevedo, P. G. Moses, J. Ojanen, T. Olsen, V. Petzold, N. A. Romero, J. Stausholm-Møller, M. Strange, G. A. Tritsaris, M. Vanin, M. Walter, B. Hammer, H. Häkkinen, G. K. H. Madsen, R. M. Nieminen, J. K. Nørskov, M. Puska, T. T. Rantala, J. Schiøtz, K. S. Thygesen, and K. W. Jacobsen, "Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method," Journal of Physics: Condensed Matter **22**, 253202 (2010).

[66] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, "The atomic simulation environment—a python library for working with atoms," Journal of Physics: Condensed Matter **29**, 273002 (2017).

[67] M. Gjerding, T. Skovhus, A. Rasmussen, F. Bertoldo, A. H. Larsen, J. J. Mortensen, and K. S. Thygesen, "Atomic simulation recipes: a python framework and library for automated workflows," Computational Materials Science **199**, 110731 (2021).

[68] N. R. Knøsgaard and K. S. Thygesen, "Representing individual electronic states for machine learning GW band structures of 2D materials," Nat. Commun. **13**, 1 (2022).

[69] S. Manti, M. K. Svendsen, N. R. Knøsgaard, P. M. Lyngby, and K. S. Thygesen, "Predicting and machine learning structural instabilities in 2D materials," arXiv preprint arXiv:2201.08091 (2022).

[70] A. Stuke, C. Kunkel, D. Golze, M. Todorović, J. T. Margraf, K. Reuter, P. Rinke, and H. Oberhofer, "Atomic structures and orbital energies of 61,489 crystal-forming organic molecules," Scientific Data **7**, 58 (2020).

[71] F. L. Hirshfeld, "Bonded-atom fragments for describing molecular charge densities," Theoretica chimica acta **44**, 129 (1977).

[72] D. Alfè, "Phon: a program to calculate phonons using the small displacement method," Computer Physics Communications **180**, 40 YEARS OF CPC: A celebratory issue focused on quality software for high performance, grid and novel computing architectures, 2622 (2009).

[73] *Local many-body tensor representation*, `https://singroup.github.io/dscribe/0.3.x/tutorials/lmbtr.html`, Accessed: 2022-11-30.

We are constantly looking for new materials. We need these new materials to develop new technologies and to improve existing technologies such as constructing more efficient solar cells and longer lasting batteries, and to replace materials depending on limited resources.

So how do we find new materials with the right properties? In computational materials science, the laboratory is a computer. Here we can design new materials and calculate their properties. The materials we are interested in are atomic-scale materials, and more specifically two-dimensional crystals. On this scale, the applicable laws of nature are those of quantum mechanics, and the relevant properties are quantum mechanical properties calculated using first-principles electronic structure methods. This has been done in many years resulting in large databases of materials and their properties, and these are the foundation of the next step in materials science; use machine learning models to reduce the computational costs of calculating material properties. Machine learning models are mathematical models learning automatically from experience in form of data.

This thesis presents developments and applications of machine learning methods for computational materials science. First, novel methods for representing individual quantum states as input features to machine learning models are introduced. The representation methods utilize information from the electronic structure and quantum mechanical wavefunctions of the materials. These methods are then applied in a machine learning model predicting accurate electronic band structures and band gaps of two-dimensional materials, which are important properties for e.g. photovoltaic applications such as solar cells.

Additionally, the concept of dynamical stability of 2D materials is studied. Dynamical stability tells whether it is favorable for a material to dislocate from the equilibrium structure, and it is an important check when doing computational studies of materials. Approximative methods for determining the dynamical stability is developed using simple atomic potentials and machine learning methods.

Technical University of Denmark
DTU Physics
Department of Physics

Fysikvej 311
2800 Kongens Lyngby, Denmark
Phone: +45 4525 3344
info@fysik.dtu.dk