Classification of materials based on similarity measures

Dissertation

zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

im Fach Physik
Spezialisierung: Theoretische Physik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin
von
M.Sc. Martin Kuban

Präsidentin der Humboldt Universität zu Berlin
Prof. Dr. Julia von Blumenthal
Dekanin der Mathematisch-Naturwissenschaftlichen Fakultät
Prof. Dr. Caren Tischendorf

Abstract

The discovery and characterization of novel materials are crucial for the development of new technology. Finding suitable materials for specific applications, however, is challenging due to the diverse and sometimes conflicting requirements for their properties. The decreasing cost of computing material properties and the recent development of data infrastructures have drastically increased the amount of available materials data. Being computed for various purposes, the available data employ different physical approximations and numerical parameters. This heterogeneity poses significant challenges in integrating and comparing data from different sources.

In this thesis, we make use of descriptors and metrics to quantitatively evaluate the similarity between different materials, represented by individual calculations. To achieve this task, we developed a computational framework that allows users to compose and manage datasets, specify and compute different descriptors and metrics, compute similarity matrices, and use methods of unsupervised machine learning. We furthermore present a spectral fingerprint, *i.e.*, a novel descriptor that encodes spectra as binary-valued raster images, allowing us to compare the similarity of different quantities, such as the electronic density-of-states, or optical absorption spectra.

We apply our methodology to assess the quality of materials data and explore large data-spaces. We demonstrate with various examples that the spectral fingerprint can be used to quantitatively describe the differences between theoretical results obtained with different physical approximations or numerical parameters, or results stemming from independent experiments. By applying our methods to larger data sets, we identify and visualize the correlations between the precision of computational results and the relevant numerical parameters. This also allows us to find calculations based on different parameters that show very similar results. To explore large data spaces, we conduct similarity searches on materials data, which reveal unexpected similarities between materials with different compositions. Furthermore, we use a clustering algorithm to find sets of materials with similar electronic structure. We identify and rationalize the main mechanisms leading to these similarities. Importantly, we find outliers that cannot be explained by simple rules. Finally, we compare the results of clustering with different similarity measures, showcasing correlations between them.

Zusammenfassung

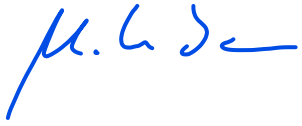
Die Entdeckung und Charakterisierung neuer Materialien sind von entscheidender Bedeutung für die Entwicklung neuer Technologien. Das Finden geeigneter Materialien stellt jedoch aufgrund der diversen und teils widersprüchlichen Anforderungen an ihre Eigenschaften eine Herausforderung dar. Aufgrund der sinkenden Kosten der Berechnung von Materialeigenschaften und der Entwicklung von Dateninfrastrukturen hat sich die Menge der verfügbaren Daten drastisch erhöht. Da sie für verschiedene Zwecke berechnet wurden, wurden diese Daten mit verschiedenen physikalischen Näherungen und numerischen Parametern erzeugt. Diese Heterogenität stellt eine signifikante Herausforderung dar, wenn Daten aus verschiedenen Quellen verglichen und gemeinsam verwendet werden sollen.

In dieser Dissertation verwenden wir Deskriptoren und Metriken, um die Ähnlichkeit von verschiedenen Materialien quantitativ zu evaluieren, deren Eigenschaften mittels unterschiedlicher Berechnungen bestimmt wurden. Zu diesem Zweck haben wir ein Softwareframework entwickelt, das es Nutzer:innnen erlaubt, Datensätze zusammenzustellen und zu verwalten, verschiedene Deskriptoren und Metriken zu definieren und zu berechnen und Methoden des *Unsupervised Learning* zu verwenden. Darüber hinaus stellen wir einen spektralen Fingerabdruck als neuartigen Deskriptor vor, der Spektren als binäre Rasterbilder darstellt und es erlaubt, die Ähnlichkeit verschiedener spektraler Größen, wie die elektronische Zustandsdichte oder optische Absorbtionsspektren, zu vergleichen.

Wir verwenden unsere Methodik, um die Qualität von Materialdaten zu beurteilen und große Datenräume zu erkunden. Für ersteres demonstrieren wir an verschiedenen Beispielen, dass der spektrale Fingerabdruck verwendet werden kann, um die Unterschiede zwischen theoretischen Ergebnissen, die mit unterschiedlichen physikalischen Näherungen und numerischen Parametern durchgeführt wurden, oder Ergebnissen die aus unterschiedlichen Experimenten stammen, quantitativ zu beschreiben. Danach identifizieren und visualisieren wir Korrelationen zwischen der Genauigkeit von Rechenergebnissen und den relevanten numerischen Parametern, indem wir unsere Methode auf größere Datensätze anwenden. Das erlaubt es uns, Rechnungen zu finden, die unterschiedliche Parameter verwenden und dennoch sehr ähnliche Ergebnisse erzielen. Danach fokussieren wir uns auf die Erkundung großer Datenräume. Dafür führen wir Ähnlichkeitssuchen auf Materialdaten durch, welche unerwartete Ähnlichkeiten zwischen Materialien offenlegen. Des weiteren verwenden wir einen Clustering-Algorithmus, um Sets von Materialien mit ähnlicher elektronischer Struktur zu finden. Wir identifizieren und rationalisieren die Hauptmechanismen, die zu diesen Ähnlichkeiten führen. Dabei finden wir Ausnahmefälle, die sich nicht durch einfache Regeln erklären lassen. Schließlich vergleichen wir die Ergebisse die duch Clustering mit verschiedenen Ähnlichkeitsmaßen erreicht wurden und zeigen Korrelationen zwischen ihnen auf.

Ich erkläre, dass ich die Dissertation selbständig und nur unter Verwendung der von mir gemäß § 7 Abs. 3 der Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät, veröffentlicht im Amtlichen Mitteilungsblatt der Humboldt-Universität zu Berlin Nr. 42/2018 am 11.07.2018 angegebenen Hilfsmittel angefertigt habe.

Berlin, den 20. Dezember 2024



Gewidmet meiner Familie

und allen die ich dazu zähle.

Contents

List of Acronyms

API	Application Programming Interface
BSE	Bethe-Salpeter Equation
C2DB	Computational 2D-materials Database
CPU	Central Processing Unit
DFT	Density-Functional Theory
DOI	Digital Object Identifier
DOS	Density Of States
EOS	Equation Of State
GGA	Generalized Gradient Approximation
HF	Hartree-Fock
HPC	High Performace Computing
HT	High Throughput
h-BN	hexagonal Boron Nitride
KS	Kohn-Sham
LDA	Local Density Approximation
MD	Molecular Dynamics
ML	Machine Learning
MP	Materials Project
MPI	Message Passing Interface
OQMD	Open Quantum Materials Database
PBE	Perdew, Burke, and Ernzerhof
PCA	Principle Component Analysis
PDOS	Projected Density of States
PLMF	Property-Labeled Materials Fragments
PP	Pseudopotential

PTE	Periodic Table of Elements
SG	Space Group
SOAP	Smooth Overlap of Atomic Positions
TM	Transition Metal
TMDC	Transition-Metal Dichalcogenides
VASP	Vienna *Ab-initio* Simulation Package
XC	Exchange-Correlation

1 Introduction

Materials are a fundamental component of technological advancement. Consequently, the discovery of novel materials is of great significance for a vast range of applications. In light of the ongoing energy and climate crises, there is an urgent demand for innovative materials to address pressing challenges, such as the generation of electricity from renewable sources, and its transmission and storage. The requirements for these materials are diverse, and in some cases conflicting. For example, hard materials tend to be brittle, which limits their spectrum of applications. Similarly, the currently most promising solar cell materials, hybrid perovskites, contain toxic elements, or disintegrate under light exposure. Addressing these challenges requires materials that optimally combine desired properties. Unfortunately, the time span for developing novel materials, *i.e.*, the time from their discovery to their application on the industrial scale, are long –typically 20 years[1]– and cost intensive. To speed up this process, data-driven methods have been established recently as a new paradigm of research.[2–4]

The workhorse of computational material science to calculate materials properties from first principles, *i.e.*, without empirical assumptions, is density functional theory (DFT). In short, it allows the computation of the total energy of a many-electron system in terms of the electron density. This approach is, in principle exact, however, its practical implementation requires the use of approximations and computational parameters. Both introduce deviations of the computed quantities from the ideal exact result.

Advances in computational hardware and numerical algorithms have drastically reduced the cost of performing DFT calculations. This enabled performing *high-throughput* experiments,[2] *i.e.*, running hundreds and thousands of these calculations, covering large parts of the configurational space of materials[5] (at least those

comprised of a few elements only). The large amounts of data produced in this way are a key resource for data-driven analysis[4] and are made available to the scientific community through large databases.[5–9] However, providers of different databases use different approximations and parameters for their calculations, making it difficult to combine data from different databases.[10] Moreover, even if datasets are based on consistent parameters, the sheer amount of available data requires novel techniques to learn from them.

Similarity is a core concept in materials design. Evaluating the similarity of different compounds allows one to infer their properties without explicitly computing or measuring them. However, similarity is often evaluated only qualitatively, because the materials science community has defined only few quantitative metrics to measure similarity, and lacks tools to apply these metrics on a large scale. In this work, we address this issue by showing how quantifying the similarity of materials allows to address two main challenges of materials science, *i.e.*, data quality assessment and big-data analysis.

This thesis is structured as follows: We introduce the theoretical background to our work in Chap. 2. There we provide a brief introduction into DFT in Sec. 2.1, followed by an overview of data management and databases in computational materials science in Sec. 2.2. We review the machine learning methods most relevant to this thesis in Sec. 2.3 and introduce descriptors used in computational materials science in Sec. 2.4. Chapter 3 presents the methodology that we contribute, including a general framework for similarity analysis in Sec. 3.1, a descriptor for spectral properties in Sec. 3.2, an application-specific clustering algorithm in Sec. 3.3, and additional descriptors for the analysis of clusters in Sec. 3.4. The following chapter 4 presents the applications of similarity to data quality assessment. First, in Sec. 4.1, we show how the use of similarity measures can help to identify in which cases materials data are (dis)similar. Then, in Sec. 4.2, we show how to find calculations in large datasets that have similar numerical precision. Finally, a summary of our findings is presented in Sec. 4.3. Then, in Chap. 5, we show how similarity measures can guide materials research. In Sec. 5.1, we apply similarity searches to materials data and study how the descriptor used affects the results. In Sec. 5.2, we use unsupervised machine learning to find all sets of similar materials

in a database of 2D materials and show how different similarity measures correlate with each other. We summarize our findings in Sec. 5.3. Finally, we discuss the impact and limitations of our work in Chap. 6, conclude all findings in Chap. 7, and provide a general outlook in Chap. 8.

2 Background

2.1 Density functional theory

The properties of materials can be computed from first principles, also called *ab initio*, by determining the quantum-mechanical ground state of the system, *i.e.*, the wave-function corresponding to the lowest eigenvalue that satisfies the many-body Schrödinger equation. Generally, to compute the properties of a solid, it is necessary to treat both nuclei and electrons as part of the same system. However, due to their higher mass, it is assumed that the motion of nuclei is much slower, allowing for a separation of variables known as the Born-Oppenheimer approximation. This allows to solve the ionic and the electronic part of the problem independently. In the latter, the nuclei act as a static potential for the electrons. The influence of the electrons on the nuclei can then be deduced from the solutions of the electronic problem. Solving the many-body Schrödinger equation of the electronic system is still difficult, because it depends on the coordinates of all N electrons. To approach this challenge, in Density-Functional Theory (DFT), an electronic system is expressed in terms of the electronic density, which depends only on three degrees of freedom.

In order to describe the properties of an inhomogeneous electron gas immersed in an external potential $v(\boldsymbol{r})$, Hohenberg and Kohn showed[11] that its total energy E can be expressed in terms of a universal functional $F[n]$ of the electron density $n(\boldsymbol{r})$ and the external potential:

$$E[n] = F[n] + \int d\boldsymbol{r} v(\boldsymbol{r}) n(\boldsymbol{r}). \tag{2.1}$$

In search for a practical realization of this formula, Kohn and Sham[12] found that

the interacting electron problem can be mapped exactly onto a fictitious system of non-interacting electrons, called the Kohn-Sham (KS) system. In this system, the electron density is represented by:

$$n = \sum_{i=1}^{N} |\Psi_i(\boldsymbol{r})|^2, \tag{2.2}$$

where $\Psi_i(\boldsymbol{r})$ are the so-called KS orbitals, obtained as solutions of single-particle Schrödinger equations,

$$\left(-\frac{1}{2}\nabla^2 + v_{\text{eff}}[n](\boldsymbol{r})\right)\Psi_i(\boldsymbol{r}) = \epsilon_i \Psi_i(\boldsymbol{r}), \tag{2.3}$$

where ϵ_i are the respective energy eigenvalues, and $v_{\text{eff}}[n](\boldsymbol{r})$ is the KS potential. This effective potential $v_{\text{eff}}[n](\boldsymbol{r})$ is the sum of three terms:

$$v_{\text{eff}} = v_{\text{ext}} + v_{\text{H}} + v_{\text{xc}}, \tag{2.4}$$

namely the external potential v_{ext}, which accounts for the Coulomb interaction with the nuclei and external fields, the Hartree potential v_{H}, which accounts for the electrostatic effects of the electron charge density, and the XC potential v_{xc}, which accounts for the quantum-mechanical part of the electron-electron interaction. The latter consists of two contributions, exchange and correlation. v_{ext} is known and v_{H} can be expressed analytically in terms of the electron density, but such expression does not exist for v_{xc}. Thus it needs to be approximated. The choice of this approximation can have strong impact on the accuracy (see also Sec. 2.2.3) of the computed properties. The most basic approximation is the local density approximation (LDA),[12] where the XC contribution E_{XC} to total energy is given by:

$$E_{\text{XC}}^{\text{LDA}}[n] = \int d\boldsymbol{r}\; n(\boldsymbol{r})\; \varepsilon_{XC}^{\text{hom}}(n(\boldsymbol{r})), \tag{2.5}$$

where $\varepsilon_{XC}^{\text{hom}}(n)$ is the XC energy density of a homogeneous electron gas with density n. For practical applications, $\varepsilon_{XC}^{\text{hom}}(n)$ is given in a parameterized, analytical form.[13] Semi-local extensions, called generalized-gradient approximations (GGAs),

improve (in many cases) the accuracy of computed physical quantities by taking the gradient of the electron density into account. The most commonly used parameterization of this functional has been suggested by Perdew, Burke, and Ernzerhof (PBE).[14] Further improvements of XC functional include meta-GGA functionals, which also consider the kinetic energy density, and hybrid functionals, which include a proportion of exact exchange computed within Hartree-Fock (HF) theory. The amount of HF exchange is a parameter, which can have large effects on the results.[15] A complete description of these methods is beyond the scope of this work. We discuss the impact of hybrid functionals on the electronic structure in Sec. 4.1.

Since the potential $v_{\text{eff}}[n](\boldsymbol{r})$ depends on the electron density n, the equation has to be solved in a self-consistent way. To do so, first, n is approximated, *e.g.*, from the electron densities of isolated atoms. In practical calculations, the KS orbitals $\Psi_i(\boldsymbol{r})$ are represented using basis functions ϕ_j, *i.e.*,

$$\Psi_i(\boldsymbol{r}) \approx \sum_j C_{ij}\phi_j(\boldsymbol{r}). \tag{2.6}$$

Using these basis functions, Eq. 2.3 allows to recast the KS problem as a (generalized) matrix eigenvalue problem with the Hamiltonian matrix H, the overlap matrix S, and the expansion coefficient vectors $\boldsymbol{C}_i$:

$$H\boldsymbol{C}_i = \varepsilon_i S\boldsymbol{C}_i. \tag{2.7}$$

This equation which can be solved by matrix diagonalization. The latter can be performed efficiently by current computer hardware. The thus obtained eigenvectors are the expansion coefficients representing an electronic orbital $\Psi_i(\boldsymbol{r})$, and the corresponding energy eigenvalue ϵ_i is the energy of the electronic state i. Using Eq. 2.2, a new electron density is calculated, allowing to set up a new H. This process is repeated until the density converges, *i.e.*, do not change (or change insignificantly) w.r.t. to the last iteration.

Different types of basis functions can be used to represent the KS orbitals in Eq. 2.6. In theory, different expansions should lead to the same result. In practice,

however, a finite set of basis functions must be used, which introduces deviations from the exact result. In calculations for periodic systems, another important computational parameter is the $\boldsymbol{k}$-point sampling, which represents the discretization of the Brillouin zone (*i.e.*, the smallest unit cell in momentum space). Besides these important examples, many other parameters are used in any implementation of the KS formalism and methods beyond. Their influence on the results of a calculation is called *precision*, the *accuracy* determines the difference to experimental results,[16] *e.g.*, introduced by the XC functional. The quality of DFT data will be discussed in more detail in Sec. 2.2.3. We show how the number of basis functions, the $\boldsymbol{k}$-sampling, and other settings influence the electronic structure in Chapter 4.

The electronic structure is characterized by the electronic density-of-states (DOS) and the band structure (BS). The former describes how many electronic states are available for a given energy:

$$DOS(E) = \sum_i \delta(\varepsilon_i - E), \tag{2.8}$$

where $\delta(x)$ is the Dirac delta, and ε_i are the KS eigenvalues obtained by Eq. 2.3. The contribution of an atomic orbital j to the total DOS, the projected DOS (PDOS), is computed by projecting the KS orbitals onto a atomic-like orbital basis ϕ_j:

$$PDOS_j(E) = \sum_i |\langle \Psi_i, \phi_j \rangle|^2 \, \delta(\varepsilon_i - E). \tag{2.9}$$

In order to obtain a smooth DOS from this discrete distribution, $\delta(x)$ is usually broadened using a Gaussian distribution function, *i.e.*, $\delta(x) \rightarrow \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$, with the smearing parameter σ.

The DOS is an integrated quantity, containing the electronic states for all $\boldsymbol{k}$points. To get more information about the dispersion of electronic states along a specific path, *i.e.*, along selected high-symmetry path in $\boldsymbol{k}$-space, the BS can be computed. Thus, it describes which energies (or states) an electron with momentum $\boldsymbol{k}$ can occupy in a unit cell, the so-called energy bands.

In this work, we deal with the electronic structure of different materials. This

requires the introduction of some terms, which we briefly recall here (for more details, see Ref. 17):

- *Fermi energy*: Highest occupied electronic state
- *Band gap*: Region in the BS and DS where no electronic states are available
- *Valence bands*: Occupied bands below the band gap
- *Conduction bands*: Unoccupied bands above the band gap
- *Metal*: A material that has bands crossing the Fermi energy
- *Semiconductor*: A material were the Fermi energy lies in a band gap

The electronic structure of a crystal is strongly influenced by the atomic positions and the unit-cell volume. Thus, to obtain accurate results, the geometry must be optimized by minimizing the total energy. The relationship between the unit-cell volume and the total energy is described by the equation of state (EOS). For this, often the Birch-Murnaghan function[18] is used, which provides information about the equilibrium volume, the pressure, and the bulk modulus.

There are numerous implementations of the DFT formalism. Here, we focus on three popular examples, each representing a different numerical approach to the KS problem. Results obtained with these codes are discussed later in this work. The first example is `VASP`,[19,20] which uses plane waves as basis functions. Using plane waves is numerically very convenient, however, the rapidly varying wave-functions of core electrons are not well represented using this ansatz. Therefore, many basis functions are required to represent these core states. To make the problem tractable, plane-wave codes make use of pseudo-potentials (PP). By approximating the contributions of core electrons to the effective potential the number of required plane waves can be drastically reduced. There are many different PPs available,[21] and different choices can impact the precision and accuracy (see Sec. 2.2.3) of the results.

In a different approach, numeric atom-centered orbitals (NAO) are used as basis functions. An example for this ansatz can be found in the DFT code `FHI-aims`.[22] NAOs are centered at the atomic positions. Their functional form, consisting of the product of radial functions and spherical harmonics, allows to efficiently

account for all electrons, including the rapid oscillations of the core-electron wavefunctions. Thus, the amount of basis functions required to obtain precise results is small. Furthermore, due to the local nature of NAOs, numerical integrals can be performed efficiently, allowing to compute the properties of isolated molecules, surfaces, and very large systems.[22,23]

The last example presented here is the code **exciting**.[24] It uses a basis constructed from augmented plane-waves and local orbitals (LAPW+lo). In this approach, the space is partitioned into so-called muffin-tin (MT) spheres around the nuclei, and the interstitial, *i.e.*, the remaining space in the unit cell. In the interstitial, the electronic wave functions are smooth and can be described efficiently using plane waves.[24] In the MT spheres, atom-centered basis functions using products of spherical harmonics and atomic-like radial functions are used to represent the electronic wave functions. Local orbitals, which are similarly defined within the MT sphere, but strictly 0 in the interstitial, are added to increase the flexibility of the basis. Beyond the electronic ground state, **exciting** puts a focus on the calculation of excited states[24] and is known to achieve micro-Hartree precision.[25]

2.2 Data management in materials science

Due to the steady increase of computational power and availability of computational resources and the usage of efficient numerical methods the cost for performing *ab initio* calculations of material properties continues to decrease. As a consequence, the large materials science community generates more and more data, representing a rich source of information. This gives the opportunity to make use of big data-driven methods[3,4,26] to obtain scientific insight from large quantities of materials data, rather than from specialized studies on small datasets. Providing a large pool of data requires the development of an extensive data infrastructure to ensure its availability and the correctness of the data. These requirements are not exclusive to the materials-science community, but are common to most of scientific disciplines. Many of the challenges that arise around the development of data infrastructures can be summarized as the *four V challenge*[4] of big data. These are *variety*, meaning the heterogeneity in the type of data that is processed, *veracity*,

the precision of the data, *volume*, the amount of data, and *velocity*, the rate at which new data is added.

Ultimately, the goal of providing a research data infrastructure is to improve the reproducibility of scientific results and speed up the research process. The requirements to achieve these goals have been formalized as the FAIR principles of scientific data management.[27] In the following, we discuss these and emphazise the challenges that arise specifically for materials-science data. Then, we review how major providers of materials-science data address these challenges, and present ongoing concerns w.r.t. data quality.

2.2.1 FAIR principles of data management

The FAIR principles[27] of scientific data management were formulated in 2016 and have since gained increasing attention. *FAIR* is an acronym: Data should be Findable, Accessible, Interoperable, and Re-usable.

Findability requires data to be labelled with a globally unique and persistent identifier, described with rich metadata, and registered or indexed by a resource that can be searched.[27] Currently, most database providers (see, *e.g.*, Sec. 2.2.2) provide unique identifiers for their entries. Some, *e.g.*, NOMAD,[9,28] additionally allow to assign a DOI[1] to a dataset, providing users with a way to cite the data in publications. Beyond the technical challenges, enabling scientists to find *relevant* data is much more involved. This is partly because even defining the requirements that a material should fulfil is challenging. For example, a researcher may be interested in finding new materials to replace a component in a device. Obviously, they can search for materials that have the desired properties, *e.g.*, the required value of the electronic band gap. However, there are other requirements for using the material in a device, such as its stability under certain conditions, or the toxicity of its constituents. Formalizing such requirements requires extensive expert knowledge as well as very flexible search interfaces.

Accessibilty requires that data can be retrieved via an open and free standardized

[1] Digital Object Identifier, `https://doi.org`

communication protocol that can be used regardless of the resources or physical location of the user. Furthermore, it requires metadata to be available even when the data are no longer available.[27] Most materials-science databases implement this using Application Programming Interfaces (API), which allow to access the data in a programmatic way. The protocols for searching and accessing the data are made available online. Notably, the OPTIMADE consortium[2], a joint effort of 24 database providers, provides a standardised API specification,[29,30] which gives access to –at the time of writing this manuscript– 29 databases.

Interoperability mandates the use of a formalized language for representing (meta)data.[27] This concerns, *e.g.*, file formats, but also requires that the schema used within these data files are disclosed to the public. An example for an interoperable data schema is the NOMAD MetaInfo,[31] which uses the `json`[32] file format and is fully documented[3]. In the context of materials science, interoperability also means which data can be combined *in a meaningful way*, regardless of the method that was used to obtain them. The results of DFT, *e.g.*, obtained with different codes, should be in principle identical. In the limit of highly converged calculations, it could be shown that the results are indeed very similar[164]. However, many of the data that are available in large databases are not highly converged, depending on the purpose of their creation. Thus, the interoperabilty of these data present a significant challenge, especially with respect ot the data-hungry nature of modern machine-learning models[33]

Reusability requires that data includes rich metadata, which describe the data. Furthermore, the provenance of the data must be disclosed.[27] Licensing issues can be met by requiring that users publish their data under an open-source license. Thus, by default, all data on the platform will be open source. Ensuring that data provenance is available is the first step in enabling that data can be reused in a different context. For computational results, data provenance can be addressed by storing, *e.g.*, all input and output files of a DFT calculation, as well the version of the software that was used.

[2] See also: `https://www.optimade.org/`
[3] See, *e.g.*, `https://nomad-lab.eu/prod/v1/gui/analyze/metainfo`.
[4] For a more detailed discussion, we refer to Sec. 2.2.3.

2.2.2 Databases of materials science data

A significant amount of the currently available computational materials data was produced in *high-throughput* (HT) investigations. HT is an approach for exploring many materials in view of certain properties. To do so, large databases of the thermodynamical and electronic properties of experimentally observed and not (yet) existing materials have been generated.[2] Searches for new materials are often performed in a combinatorial way, *i.e.*, crystal structures of know compounds are systematically decorated with different species to calculate the properties of the new materials. The HT approach, however, goes beyond that and includes that data are stored in a systematic way and that advanced data analytics is used for the selection of new candidate materials.[2] To achieve this, HT searches rely on software frameworks for setting up new calculations, running DFT codes, handling errors, and extracting results.[2]

The rapid increase of data produced in such HT searches in the past years has several reasons.[34] First, DFT codes are increasingly supporting the automatic computation of more and more properties, also making the usage of codes easier for non-expert users. This simplifies the setup of large-scale computation of advanced material properties, which are of interest for many applications. Second, the cost of numerical approaches has decreased due to the increased availability of computational resources. Third, the availability of open-source HT frameworks[35–38] and post-processing tools[39,40] allows users to set up workflows and extract scientific results with little development effort.

In the following, we review several material databases, focusing on those that are used in Chapters 4 and 5 of this work. The selection provided here is by far not exhaustive, but is rather used to illustrate the variety of different approaches, their similarities, and differences.

ICSD

The Inorganic Crystal Structure Database (ICSD) contains manually curated entries of compounds, such as ceramics, minerals, and metals. These entries stem

from the scientific literature, the earliest articles dating back to 1915.[41] The ICSD was started in 1990, today it contains more than 300,000 characterized crystal structures. That means that the atomic coordinates and the composition of the compound are known. Originally, the ICSD contained exclusively experimentally observed structures. Recently, structures from theoretical publications have also been included. The ICSD is often used for finding crystal structure prototypes as a starting point for combinatorial searches, or to populate HT databases (see below). Access to the ICSD requires the purchase of a license, however, many computational databases mark, wherever possible, to which ICSD entry each database entry corresponds.

AFLOW

The HT database AFLOW[6] contains over 3.5 million entries[42] of material properties computed with DFT. For the data generation, the eponymous computational HT framework `AFLOW`[35,43] was used. The latter is a monolithic open-source computer program[5], implemented in the C++ programming language. It supports tasks like crystal-structure generation and manipulation and symmetry and structure analysis, and it includes post-processing tools. Different components of the software are accessible through a web interface at `https://aflow.org`, allowing users to avoid installing the complete `AFLOW` program.[42] Notably, the AFLOW ecosystem provides a library of 1,100 crystallographic structure prototypes, which can be used for combinatorial HT searches.[44–46]

The AFLOW repository contains a variety of different material properties, including phase diagrams, symmetry related properties, energetics, band gaps, and magnetic moments.[6] Properties of the electronic structure can be analysed using BS and orbital-projected DOS spectra. Furthermore, the original and relaxed crystal structure can be downloaded. The consistency and interoperability of the results is supported by a standard set of parameters for the DFT code `VASP`, which is used for all calculations.[47] However, the data are organized in so-called *catalogs*,[42] grouping data that was created for a specific purpose, such as computing

[5] A software monolith contains different modules and functionalities, which are part of a the same codebase and can only accessed through the main program.

the properties of materials from the ICSD, or combinatorial searches, using slightly different approximations. For example, for the former, the Hubbard correction for the XC functional (DFT+U) was used to obtain accurate results, whereas for the latter it is omitted to obtain accurate and comparable formation enthalpies.[42]

OQMD

The Open Quantum Materials Database (OQMD) is a DFT-based HT materials database focused on thermodynamic stability.[5] As of 2015, it contained about $30,000$ compounds from the ICSD, extended by $\sim 260,000$ novel, predicted crystal structures found by combinatorial HT searches.[38] More recently, about half a million of compounds have been reported.[48] All calculations are performed using `VASP`. To study the thermodynamic stability, the accuracy of the formation energies is vital. Therefore, the chemical potentials used to compute formation energies are corrected using experimental data.[38] Earlier, good agreement between experiment and theory has been found,[5] and the remaining uncertainty in accuracy was suspected to be based on the comparatively low precision of experimental references.[38]

An important feature of the OQMD is the automatic construction of multi-component phase diagrams. These can be used to assess the thermodynamic stability of materials by comparing their energies to the energies of many other materials at the same time. This is achieved by formulating the generation of a multi-component phase diagram as a linear algebra problem.[49,50] This technique was also applied to study reaction pathways of materials with potential applications for hydrogen storage.[49] Later, the approach was extended to compute a phase stability network of 21,000 stable materials from the OQMD.[48]

Exceptionally, the entire OQMD can be downloaded as a single file from the website, `https://oqmd.org/download/`.[38]

Materials Project

The Materials Project[7] (MP) is a multi-institution collaboration as part of the US Department of Energy Office of Basic Energy Sciences.[51] The MP maintains an open-access HT database of more than 150,000 materials and molecules, see `https://next-gen.materialsproject.org/`, calculated with `VASP`. The MP has a large user base with 40,000 registered users[6]. For the majority of materials, the PBE XC functional is employed. For transition metal oxides and sulfides, DFT+U is used.

The MP furthermore provides and maintains a large suite of open-source software, including the analysis and structure manipulation library `pymatgen`,[40] the workflow management tool `Fireworks`,[37] and the workflow library `atomate`.[52] On its website, the MP provides *Apps*[7], *i.e.*, online applications that allow to combine properties from different materials, to analyze and visualize them.[51] These apps can be used, *e.g.*, to generate phase diagrams from MP data. Further apps are for crystal structure analysis and for screening for materials with specific properties.

C2DB

The Computational 2D Materials Database (C2DB),[8,53] is an HT database of atomically thin systems, computed with the DFT code `GPAW`.[54] A large fraction of its entries is generated using a combinatorial approach, based on several structural prototypes, including Xane (e.g., graphane), Xene (e.g., graphene), MXY Janus (e.g., MoSSe), and TMDCs (e.g., MoS2). Recently, more stable structure have been added, which were found using a novel ML-based approach.[55] Currently, the C2DB contains 16,789 structures, that are available at `https://c2db.fysik.dtu.dk/`. Among these, 4,194 are labeled as *dynamically stable*, with an energy of maximal 0.2 eV above the convex hull[8]. The remaining materials are metastable, *i.e.*, they are not likely to be synthesized, but, under ideal conditions, *e.g.*, by controlling temperature and pressure, it may still be possible.

[6] Note that the MP requires registration to access the database.

[7] See `https://next-gen.materialsproject.org/apps`.

[8] The convex hull is formed by all compounds that have the lowest energy of formation for their specific atomic composition.

The properties computed for stable materials include the heat of formation, the stiffness tensor, phonons to assess the dynamic stability, magnetic properties, and the optical polarizabiliy. Furthermore, the data includes band structures, orbital-projected PDOS, and (if applicable) band gaps.

NOMAD

The NOMAD data infrastructure[4,9,28] follows a different approach. Conceptualized as an open materials-science platform, it allows users to upload and share their data. In addition to the uploads from hundreds of individual users, NOMAD also contains the data created by different HT projects,[9] totaling in $\sim$ 13 million identified entries (see `https://nomad-lab.eu/prod/v1/gui/search/entries`), which can consist of several single-point calculations. Among these $\sim$ 52% stem from AFLOW, $\sim$ 5% from the OQMD, and another $\sim$ 2% originate from the MP. Recently, NOMAD has been expanded to include experimental data. All data in NOMAD are fully open access under the Creative Commons Attribution 4.0 License and can be downloaded without registration.

To achieve a FAIR data infrastructure, NOMAD presents the data in different ways.[9] From the very beginning, it accepted raw computational data, *i.e.*, the input and output files of DFT calculations. NOMAD supports the entire electronic-structure community by accepting results from more than 50 different DFT codes. The repository aims at data sharing, providing long-term storage of the files for at least 10 years. Each DFT code has its own data formats, file types, and even units used to represent physical quantities. That means that these data are not interoperable. To address this challenge, the NOMAD software[56] parses and normalizes these files, and stores them in unified form. To achieve this, NOMAD relies on a domain-specific metadata schema, called the *MetaInfo*,[31] which can be found at `https://nomad-lab.eu/prod/v1/gui/analyze/metainfo`. Using this schema, the information contained in the raw files can be represented in an interoperable way. To lower the access barrier for non-expert users, NOMAD created the *Encyclopedia*, providing an aggregated view of the data, *i.e.*, showing the results of all calculations for the same material. Even if it is currently

not actively further developed, users can use the Encyclopedia inspect the crystal structure, electronic and vibrational properties, as well as an overview of different methodologies employed to obtain the data.[28] It allows users to directly compare the spread of results obtained by different calculations and see the impact of different methodologies on physical properties. The NOMAD infrastructure also maintains the *AI toolkit*,[57] a collection of data analysis and AI tools available in web-based interactive notebooks. They act both as tutorials to learn data analytics and machine learning techniques (see Sec. 2.3) as well as a starting point for users to explore the data of the NOMAD Archive using state-of-the-art data processing methods.

2.2.3 High throughput calculations and data quality

The correctness of DFT implementations[10,16,58] and data contained in HT materials databases[34,59] are subject of recent investigations. In the following we review these studies.

In Sec. 2.1, we have briefly touched on the conceptual differences between the terms *precision* and *accuracy*.[16] To recall, the former refers to numerical precision. *i.e.*, deviations in computed quantities due to details of implementations and the usage of certain computational parameters. Conversely, accuracy is used to identify systematic deviations introduced by methodologies and approximations, *e.g.*, the XC functional. It can be determined only in comparison to higher-level theory or experiment. In a large collaborative effort,[16] the precision of 15 different DFT codes and 40 different pseudo-potentials (PP) was evaluated, using the so-called Δ factor as a metric. The latter is defined as the differences between the areas below the EOS curves (see Sec. 2.1) obtained by two different codes[58] for a single material. By comparing this score, averaged across a set of 71 elemental solids, it was revealed that, for highly converged calculations, the precision of DFT codes is below 0.5 meV/atom for about 10 codes. Notably, several codes have improved significantly compared to early implementations. More recent studies, using a larger test set[59] have further emphasized that codes based on PPs show lower precision.

The results in these benchmarks were obtained using very stringent numerical settings. This is not representative of the data contained in HT databases.[10] Thus, comparing or combining data from different datasets can lead to uncontrollable uncertainties. Therefore, Carbogono *et al.*[10] systematically investigated the precision of results from four codes w.r.t. their highest converged settings. This was achieved by comparing the total energies of these highly converged calculations to those with less stringent numerical settings. It was found that for three codes, *i.e.*, **exciting**, FHI-aims, and GPAW, as expected, the error decreases steadily with increasing convergence parameters. For VASP, this relationship is more complex due to an error reduction scheme employed by the code. Furthermore, the authors pointed out that the achieved precision can vary considerably between different elemental species, their atomic coordinations, and different methodologies.[10] It was also shown that the precision of the energies of binary alloys can be approximated well using a stichiometric average of the respective errors for each elemental solids, using the same numerical settings.

In a later study, a large-scale comparison between the data contained in the AFLOW, OQMD and Materials Project databases revealed that formation energies and volumes of materials show generally better agreement than electronic band gaps and magnetic moments.[34] The differences in energies were shown to stem mostly from the usage of different elemental reference states and post-calculation corrections, *e.g.*, by those employed by the OQMD to improve the accuracy of DFT results.[38] The usage of different PPs was found to have a larger impact on band gaps and magnetization.[34]

To address the challenges related to the quality of data obtained in HT experiments, several approaches have been proposed. General consensus is that more verification studies are necessary, specifically including more XC functionals[16,59] and a variety of properties. Bosoni *et al.* emphasize that, to quantify and understand differences in calculations, novel metrics are required, which should rely on quantities that can be determined experimentally.[59] How to better control the precision of the data produced in HT workflows is actively discussed in the literature. Bosoni *et al.* suggest to perform convergence tests automatically in HT workflows.[59] Conversely, Hegde *et al.* point out the computational cost of such

approach and focus on quantifying the uncertainties employing existing data.[34] Also, Carbogno *et al.*, propose the prediction of uncertainties of HT data based on statistical models.[10] The predictive power of such models can be improved by using more elaborate machine-learning models.[60]

2.3 Machine learning

The popularity and success of machine learning (ML), and artificial intelligence (AI) in general, in materials research has been increasing for more than 10 years.[26,61–63] In the following section, we introduce the basic concepts of ML, in order to put this work into the general context and to provide an overview of the methods applied in this work. Note that we do not introduce specific popular methods, such as neural networks, as they go too far beyond the scope of our work. The scientific literature contains a large variety of introductions to the topic, *e.g.*, in Ref. 64. We conclude this section by providing a brief review of some applications of ML to materials research.

2.3.1 Supervised learning

In supervised learning tasks, also called learning from a teacher, refers to learning tasks where the value of a *dependent*, or *target*, variable y is predicted from an *independent* variable $\boldsymbol{x}$.[64] The latter is typically a vector and its components are called *features*. Denoting the predicted value as $\hat{y}$, the goal is to find a function, or *model*, f,

$$f_\beta(\boldsymbol{x}_i) = \hat{y}_i, \tag{2.10}$$

such that the prediction error between the target and the predicted variable is minimized. The subscript $\boldsymbol{\beta}$ denotes parameters of the model. The prediction error is evaluated using a *loss function* $\mathcal{L}$, typically the root mean squared error

(RMSE):

$$RMSE(\hat{y}, y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{2.11}$$

where N denotes the total number of samples. To find the optimal set of parameters, the loss function is minimized w.r.t. the model parameters:

$$\boldsymbol{\beta} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ \mathcal{L}(y, f_{\boldsymbol{\beta}}(\boldsymbol{x})). \tag{2.12}$$

This process is also called *fitting*. All samples that are used for fitting the model are called the *training set*. The *generalizibility* of a model describes its capability to make correct predictions for unseen data.[64] Most models also contain *hyperparameters*, which are parameters controlling, *e.g.*, the number of model parameters $\boldsymbol{\beta}$ or their maximal value, which are not learned from the training set, but are selected *a priori*. To choose the values of these hyperparameters, the loss is evaluated repeatedly for a *validation set*, *i.e.*, samples that are not used for the fitting process. The hyperparameters that lead to the lowest loss for the validation set are considered optimal. The final performance is reported by computing a suitable loss function for the *test* or *hold-out set*, *i.e.*, samples, that have not been used before.

2.3.2 Unsupervised learning

In unsupervised learning, there is no dependent variable: the goal is to directly deduce the characteristics of the data.[64] It encompasses a variety of different techniques, such as finding association rules that describe the correlations between samples, finding a low-dimensional subspace which maximizes the variance of the data with principle component analysis (PCA), or finding sets of samples that are similar to each other with clustering. In this work, we focus on the third example.

Clusters are obtained by applying a clustering algorithm to the data at hand. Numerous of such algorithms have been published in the scientific literature, which can be classified by different criteria. A comprehensive review is available, *e.g.*, in

Ref. 65. The input to clustering algorithms are, similar to supervised learning, the variables $\boldsymbol{x}_i$ describing each sample i. From these, based on various assumptions about the distribution of the data, the algorithm assigns *cluster labels* to each sample i. All samples with the same label belong to the same cluster and are called *members* of the cluster. We identify the *size* of a clusters as the number of its members. Some algorithms allow some samples not to be assigned to any cluster. These samples are called *orphans*.

While in supervised ML the optimal solution is defined as the minimum of the loss function, in unsupervised ML there is no direct measure of success.[64] Therefore, it is required to explain the results of clustering[65] qualitatively or through additional metrics.

2.3.3 Applications in materials science

ML can be used for many purposes in materials science. Here, we give a brief overview of the recent applications, following the very comprehensive review of Ref. 26.

A persistent challenge in materials research is the discovery of novel materials. Related questions are, for example: Will a given stoichiometric combination of elements form a stable crystal? What crystal symmetries will the resulting compound have? What mechanical, electronic, and optical properties will it have? Supervised ML can be used to address many of these questions. The stability of materials can be studied by training ML models on the energy of formation of existing, *i.e.*, experimentally observed, materials, and then predicting it for new compounds. This approach is, however, limited in its effectiveness due to the relatively small amount of available experimental data. Conversely, predicting the energy of formation from theoretical data alone requires knowledge of all competing compounds for a given stoichiometry, which ultimately limits the accuracy of such predictions.

Electronic and optical properties of materials can also be addressed with ML. A typical target variable for this task is the electronic band gap, which is important for many applications, such as electronics or photovoltaics, but requires expen-

sive calculations, *e.g.*, using non-local XC functionals or many-body perturbation theory to obtain accurate results. Thus, the acquisition of a sufficient amount of training data is challenging. One way to mitigate this challenge is using the less accurate band gaps obtained by DFT as a feature, and learning the differences to results obtained by more accurate calculations. However, the availability of highly accurate data remains a bottleneck.

To deal with small datasets, active learning approaches can be employed. Here, the model, or an estimate of its uncertainty, is used to predict which training data points would lead to an improvement of the model. These are then acquired and used to retrain the model, leading to rapidly increasing performance.

As a final example, we focus on molecular dynamics (MD) simulations. In MD, the movements of atoms or molecules are studied. This is often done using force fields, *i.e.*, mathematical models that describe the forces acting on individual atoms. Instead of using classical models, these force fields can be based on ML trained on DFT data, drastically improving the accuracy of the predictions.

2.4 Descriptors

The examples presented in the previous section clearly show the that ML approaches can be used a variety of applications. However, so far, we have not discussed an important aspect of ML for materials science, *i.e.*, the representation of materials as features $\boldsymbol{x}$ (see Eq. 2.10). To compute material properties using *ab initio* methods such as DFT (see Sec. 2.1), only the unit cell, the positions of the atoms, and their atomic numbers need to be known.[66] Due to the periodicity of the crystal lattice, however, an infinite number of representations is possible, which leave intrinsic properties of the system invariant. This poses a challenge to ML methods, which would have to learn to account for these invariances,[66] potentially requiring an unaffordable amount of training data. This problem can be avoided using *descriptors*, which represent the materials in a form that is already invariant to transformations, *i.e.*, leave intrinsic properties unchanged. The properties of ideal descriptors are often discussed in the scientific literature:[66–70]

1. The numerical value of a descriptor should allow to uniquely identify a sample.
2. (Dis)similar materials should have (dis)similar descriptor values.
3. Small changes in the material should lead to small changes in the descriptor.
4. The computation of the descriptor should be significantly cheaper than computing the material property directly using *ab initio*methods.
5. The descriptor should be as compact as possible.
6. Descriptors of the atomic structure should be invariant to rotations, spacial translations, and permutation of atomic indices.

Some points found in single references are not included here, and different authors weigh the importance of each property of descriptors differently, depending to the task at hand. Another favorable property of descriptors is a constant size, *i.e.*, the numerical representation of the descriptor should have the same size (*e.g.*, length of the feature vector $\boldsymbol{x}_i$) for every possible material. Not all descriptors fulfill by design all desired properties, such as those listed above. In many cases, issues arising from this can be mitigated by adapting the original design of the descriptor, or by adding additional post-processing steps.[66] We note, however, that the requirements listed above are often postulated in an *a-prori* manner, and the importance of fulfilling them is not shown on materials datasets.

In the following, we review different descriptors used in materials science, distinguishing three different general approaches.

2.4.1 Feature vectors

Feature vectors are sets of physically meaningful parameters, which are provided in a tabular form. To illustrate the concept, we make use of an example from Ref. 71, which describes a general framework for generating feature vectors based on the composition of a material. The largest fraction of such features are based on atomic features, such as the atomic number or the covalent radius. The features of individual atoms are combined by mathematical operations such as the mean

weighted by the composition, average deviation from the mean, or the minimum and maximum value. These descriptors have been successfully used to identify potential candidates for solar-cell materials, and to find metallic glass alloys.[71] One downside of this approach is that such feature vectors are postulated *ad hoc*, based on physical intuition. Another disadvantage is that correlations of the descriptor with the ML target property cannot be controlled, especially, the descriptor cannot be optimized to improve to predictive power of models trained using these descriptors. To mitigate this problem, Ghiringhelli *et al.*[70] proposed an ansatz based on symbolic regression: To obtain features with higher predictive power, they are generated by combining so-called primary features using algebraic operators. Typical primary features are the elemental electron affinity or the covalent radius, and the algebraic operators that are used include linear operations, such as addition, as well as non-linear ones, such as the square root or exponential functions. This process can be repeated using features obtained in previous iterations, until a sufficient level of complexity is reached. Due to the combinatorial nature of this approach, the total number of features becomes very large. From this large set of features, a suitable subset is selected, *i.e.*, the smallest subset that provides the best predictive power on a training dataset. Using this approach, they were, for example, able to train a model that predicts the (small) energy differences between rock-salt and zinc-blende structures for a set of 82 binary materials. This approach allowed to classify with high accuracy in which of these crystal structure a binary material is more stable. Issues with linearly dependent features could be solved later using the SISSO approach.[72]

2.4.2 Structural descriptors

Structural descriptors are used to represent the atomic structure of materials. An important distinction can be made between *local* and *global* descriptors.[66] The former are used for local environments, *e.g.*, the coordination of an atom, given by its next nearest neighbors. Global descriptors are defined for the whole structure, *e.g.*, all atoms in a unit cell at once. In most cases, global descriptors can be obtained from local ones by averaging over all local descriptors found in a compound, or by using kernel methods for matching sets of local descriptors.[73]

Oganov and Valle presented a so-called fingerprint function,[67,68] based on the distribution of pairwise atomic distances in a solid. Their approach is related to the radial distribution function (RDF),[68] and the structure factor used in diffraction experiments.[67] The effectiveness of this approach was demonstrated by computing the descriptor values for various randomized crystal structures. It was shown that the similarity between the descriptor values correlates with the the degree of order in the randomized structures, their energy at constant volume, and a measure of quasi-entropy of the system.[67] The descriptor was also applied to identify duplicates in an evolutionary algorithm, where they cluster the found structures based on their descriptors, and use the cluster centroid as a representative for all members. This allows expert users to use the evolutionary algorithm more effectively.[68] A more generalized version of this approach to crystal structure representation is the many-body tensor representation (MBTR).[74] Here, the values of so-called geometry functions, *i.e.*, functions of properties like the atomic number, pair distances, or angles between atoms, are computed and concatenated to obtain a vector representation of the material.

A different approach was taken by Bartók *et al.*:[75] In order to fit potential energy surfaces used for modeling interatomic potentials (see also Sec. 2.3.3), a representation of the bispectrum,[69,75] *i.e.*, a statistical function that is commonly used in signal processing and telecommunication, of the local atomic density[76] is used. Later, a more generalized formulation of this descriptor was introduced,[69] based on the power spectrum computed from the expansion of the atomic density in spherical harmonics. This descriptor, called smooth overlap of atomic positions (SOAP),[69] combined with so-called matching kernels[73] allows to describe the similarity of atomic systems of arbitrary size, as demonstrated by compiling *e.g.*, maps of different phases and clusters of carbon atoms. These representations suffer from a combinatorial explosion of coefficients (*i.e.*, the descriptor length) if many atomic species are included, which can be mitigated by compression techniques.[77]

Alternatively, materials can be represented as fragments, or building blocks, of their crystal structure.[78] Numerically, this can be achieved by comparing a material to a pre-compiled list of structural motifs. By denoting the number of occurrences of each pattern in the structure, weighted by the stoichiometric ratio of their

constituents, a representative feature vector can be obtained. Similar approaches were established as standard in other scientific fields, *e.g.*, drug discovery[79] (see also Sec. 2.5). Building upon this concept, property-labeled materials fragments (PLMF),[80] are constructed by determining the atomic connectivity graph, represented by the adjacency matrix. The adjacency matrix encodes the environment of atoms as 1 (0), if two atoms are connected (not connected) by a chemical bond and is used to identify neighboring atoms in a crystal structure. The individual entries of the descriptor vector are then constructed by summing over the differences of various atomic, experimentally obtained, and derived properties between neighboring atoms in a fragment.

2.4.3 Electronic-structure descriptors

All descriptors introduced above are derived from atomic, elemental, or structural properties of the material. These are typical input parameters of *ab initio* methods or can be found in tables. The electronic structure, however, can be only determined by performing a calculation. As such, the electronic structure is not a suitable descriptor for predicting ground-state properties of materials. However, descriptors based on it have been used to compile *material cartograms*,[78] building predictive models for the electronic DOS,[81] and predict electronic states obtained with higher levels of theory.[82]

For the first example, two different representations were used: Isayev *et al.*[78] constructed a representation of the electronic DOS by encoding it point-wise in the energy range between -10 and 10 eV as a series of 256 real numbers. Similarly, the electronic BS was represented by discretizing the energy of electronic states along the high symmetry paths in the Brillouin zone. The number of states falling into each bin was then denoted in the feature vector. For the second example, *i.e.*, training ML models to predict the DOS, another pointwise representation of the DOS was used.[81] Such representations were claimed to be inefficient, as they potentially require many sampling points of the DOS to efficiently train the ML models. Furthermore, employing loss functions derived from this representation has been shown to be indifferent to spectral features with small overlap. To mitigate these

problems, one can reduce the degrees of freedom by principal-component analysis (PCA), effectively smoothing the DOS. Alternatively, a representation based on the cumulative distribution function of the DOS was proposed, which showed an increased sensitivity of the loss function to non-overlapping spectral features. In the third example, a high-dimensional descriptor based on the PDOS (see Sec. 2.1), followed by PCA dimensionality reduction, has been proposed.[53] It was used to to predict electronic band-gaps at the accuracy level of hybrid XC functionals, using the PBE PDOS as input. Later, a similar descriptor was used for the prediction of electronic states obtained by post-DFT many-body methods.[82]

In Sec. 3.2, we present a spectral fingerprint that can be used to describe the DOS of materials, but also other spectral quantities. A variety of use cases are presented in Chapters 4 and 5.

2.5 Similarity

The concept of similarity and its applications are well researched topics in other scientific fields, *e.g.*, medicinal chemistry and drug design.[79] There, *molecular similarity*, *i.e.*, the usage of similarity measures for molecules, plays an important role in the discovery of new drugs in the pharmaceutical industry.[83,84] An interesting parallel that can be drawn is that databases of chemical compounds are used extensively, both in public and private research. While this has not been the case for materials science in the past, the recent introduction of large, open materials databases (see Sec. 2.2.2) has opened up new opportunities for data-driven research. To facilitate the advantage that research in molecular similarity has due to its early start, in the following, we review the literature, summarize the key concepts, and relate them to materials research.

The advance of similarity-based studies started in the 1990s with the formulation of the similarity property principle (SPP),[79,83] which states that similar compounds have similar properties. While this may seem obvious at first glance, the matter is actually more complex: Obtaining a clear definition of similarity and properly accounting for it, turned out to be a challenging task. In general,

finding a *quantitative structure-property relationship* (QSPR) is described as the combination of three steps:[79,83] selection of a descriptor, weighting of its features, and selection of a suitable similarity score.

Analogous to descriptors for materials, in molecular similarity they are used to represent the molecular structure in a numerical format. The most computationally efficient, and therefore, in many cases favored representations are based on bit vectors, *i.e.*, vectors, where every entry can either be 0 or 1.[83,85] Weighting of the descriptor features allows to emphasize specific structural patters or to incorporate the number of occurrences of a structural feature.[83] similarity scores are used to compute the similarity between two descriptors.

The key property of compounds that can be used as new drugs is their (biological) activity.[79,85,86] In *similarity searches*, a known (active) molecule, also called a *lead*, is used as a reference to scan database entries, aiming to find compounds with similar structural features, which, according to the SPP, may be active as well. One goal of this analysis is to find active molecules that are "dissimilar enough" to justify patenting the molecule.[79,85] Furthermore, similarity measures are used to increase the diversity of molecular datasets, e.g., by selecting subsets of larger databases such that the dissimilarity between the selected entries is maximized.[83]

However, despite the success of molecular similarity, several limitations and shortcomings are known. Most importantly, similarity is subjective,[79] and can only be defined within a certain context.[85] This means, that even if a similarity search with a specific descriptor yields good results for a specific application, it does not mean that the same methodology will be equally successful for another application. Rather, descriptors and similarity scores have to be selected on a case-by-case basis.[85] This also means a general threshold S_{thres}, such that all compounds with a similarity $S > S_{\text{thres}}$ to a reference compound have similar properties to it, does not exist.[79] This is especially important when interpreting the results of similarity searches. Furthermore, it is possible that small differences in the descriptor lead to large differences in the biological activity. Finally, many similarity measures are known to have an asymmetry based on the size of the descriptors, *i.e.*, larger molecules tend to score higher in similarity rankings.[79]

Materials discovery faces similar challenges as drug design: Determining material properties through experiment or accurate calculations is expensive. Therefore, predicting material properties based on their similarity to known compounds[61] is desirable. Furthermore, materials databases are large, to an extent that extracting knowledge from them requires the usage of data- analytics tools. However, as mentioned in Chapter 1, the materials science community has, to this point, not adapted similarity-based methods to face these challenges.

For materials, similarity searches can be used in the same way as they are used for molecules. This allows one to effectively search materials with desired properties in large databases, increasing the findability of scientific results. Additionally, quantifying the similarity between individual calculations or measurements offers new opportunities to address the challenge of interoperability of materials data. In the following chapters, we establish the required terms and methodology, and apply them to a variety of different use cases. While doing so, we learn from the successes and pitfalls of molecular similarity and carefully adapt, apply, and extend already established and well-researched methods.

3 Novel methods for similarity analysis

While it may be intuitively clear to a domain expert to decide whether two materials are similar on a qualitative basis, quantifying that similarity is challenging. Part of this challenge is that many physical quantities that characterize a material are arrays, such as the electronic DOS. This means that comparing them, even qualitatively, requires to add context, such as the energy region in which the spectra are (dis)similar. But even for scalar quantities, such as the electronic band gap, it is not possible to give a definitive answer, since whether a band gap is too large or too small is defined by the application that the material is used for. To address this challenge, we present a flexible framework of methods that allows for the quantification of the similarity of materials, based on individual components and tools that can be combined for a specific purpose. In order to do so, technical challenges need to be overcome, concerning, *e.g.*, collecting and storing datasets, or finding a suitable descriptor for spectra. We explain these challenges below and introduce the tools to address them.

The following section 3.1 introduces `MADAS`,[87] a framework for materials data similarity analysis, implemented in the programming language Python. We describe the core modules of the program as well as their interplay. Section 3.2 presents a spectral fingerprint, which can be used to compute the similarity between two spectra. Section 3.3 introduces a threshold-based clustering algorithm for the analysis of materials similarity. Finally, we present additional fingerprints in Sec. 3.4, which are used to interpret the results of clustering.

3.1 Similarity analysis with MADAS

Similarity analysis is a data-driven approach, making it essential to access and manage (large) materials datasets. In this section, we describe how this can be achieved using using the `MADAS` framework, following our publication Ref. 87.

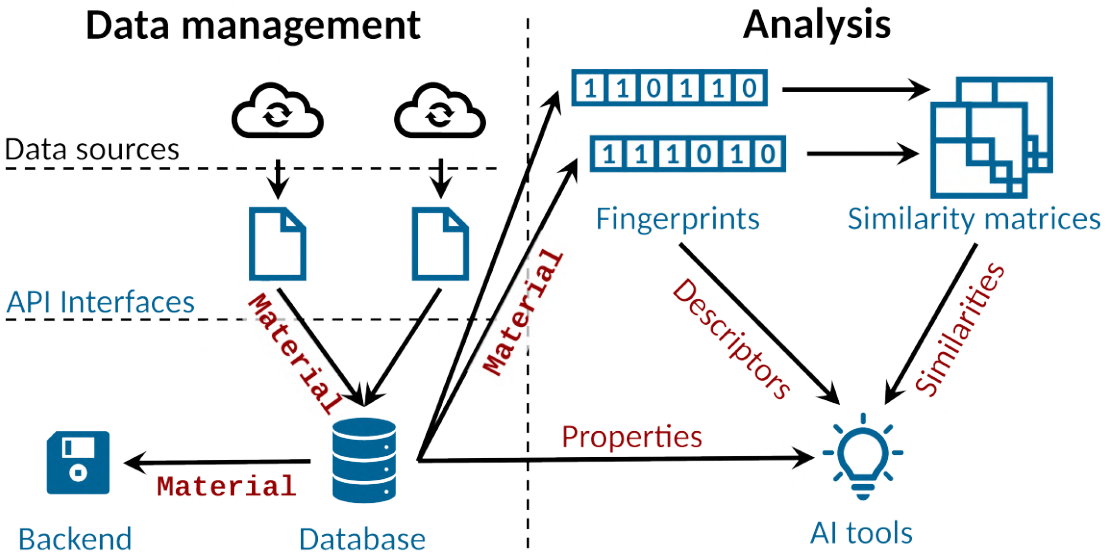


Figure 3.1: Workflow of similarity analysis with `MADAS`. *Data* are obtained from one ore more *sources*, using *API interfaces*. Using `Material` objects, data are stored in a *Database*. File management is handled by a *Backend*. From database entries, *Fingerprints* are computed, which are in turn the input for *Similarity matrices*. Combining data, descriptors and similarity scores, *AI tools* are used for analysis. Symbols represent components of `MADAS`, arrows represent the data flow. Elements in blue color are implemented in Python modules, red labels indicate the type of data that are exchanged. Figure from.[87]

An example of how similarity-analysis workflows can be realized using `MADAS` is shown in Fig. 3.1. The workflow starts in the top left corner with *data sources*, *i.e.*, online databases or local files, which are used to compile datasets. To access these data, an *API interface*, *i.e.*, an interface connecting the Application Programming Interfaces (API) of the data source to `MADAS`, is required. It is responsible for querying the external database and returning `Material` objects. The latter are used to unify the communication between different components of `MADAS`.

`Material` objects are stored in the *Database*, which, in turn, allows for reading from and writing to the *Backend* and generating and storing the *Fingerprint*s. Fingerprints are used as input for the calculation of *Similarity matrices*. Finally, *AI tools* can combine data from the database, fingerprints, and similarity matrices, to analyze the data. Below, we discuss the individual components in detail.

`MADAS` is published as open-source software, available at `https://github.com/kubanmar/madas` and via the Python package index (PyPI) at `https://pypi.org/project/madas`. An extensive documentation, including examples and tutorials, can be found online at `https://madas.readthedocs.io`.

3.1.1 Data download and management

As mentioned above, the first step of data analysis is the collection of the data. To allow users access to data in an efficient way, providers of online databases maintain web APIs. These are programmatic access points to the contents of the databases, allowing to perform tasks such as searching and download of data. Data can be retrieved from the these APIs by using a unique identifier of individual calculations or measurements, or by using a set of search terms, called *queries*, to retrieve even large quantities of data at once. While this functionality is maintained by most of the data providers, the individual implementation, including, *e.g.*, the definition of keywords and metadata, may differ significantly between databases. This poses a challenge to the data-collection task, since each database requires specific knowledge of its implementation to access the data. Aware of this issue, the providers of the largest databases jointly developed the OPTIMADE API,[29,30] which is a standard that describes how various materials properties can be made available to users. It is supported by an increasing amount of data providers. However, the development of such standards is slow, and the providers' own APIs may be more expressive and give access to more data. One drawback of non-standard APIs is that they may not provide data consistently, and providers can change their APIs over time without notice. To address this challenge, `MADAS` employs a Python class for implementing and updating interfaces to these APIs, called `APIClass`. It is equipped with a common naming schema and data model

and can be used as a standardized template for accessing data from different sources. The output of an `APIClass` should be given using `Material` objects[1], allowing them to seamlessly integrate with the other components of `MADAS`.

Besides that APIs may change over time, also the data are not guaranteed to be immutable. The Materials Project, *e.g.*, provides a version history of the data at `https://docs.materialsproject.org/changes/database-versions`. According to the MPs deprecation policy, this data are still available, but may not be found through the search interface. Therefore, in order to support reproducibility of scientific results (or at least traceability, in case erroneous data led to false results), data used for analysis often need to be stored locally. To do so, a database should be used, as database software is optimized towards fast and convenient access to the data, and provides methods to protect against data loss through consistency checks. Thus, consistently storing the data that enter the analysis pipeline in a database greatly enhances reproducibility of results. Furthermore, storing the data locally reduces overhead from repeated downloads from the same source. In `MADAS`, the database is implemented in a `MaterialsDatabase` class. It acts as an anchor point between the API interfaces, the file handling, and the subsequent data analysis, as shown in the left and right sides of Fig. 3.1, respectively. The file handling is managed by a `Backend` class, which writes and reads `Material` objects to and from database files. The logical distinction between the `MaterialsDatabase` and the `Backend` allows to implement interfaces to different databases, such as SQLite (`https://www.sqlite.org/index.html`), in different `Backend` classes, without the need to adapt the code that interfaces with the data analysis classes and methods.

[1] Python, as a dynamically typed programming language, does not allow to verify that a function returns the correct data type before the code is executed. However, `MADAS` supports the effort by using type hints in the source code and providing tutorials.

```
1  from madas import MaterialsDatabase
2  from madas.apis.NOMAD_web_API import API
3  from madas.backend import ASEBackend
4
5  api = API()
6
7  backend = ASEBackend(
8    filename = "materials_database.db",
9    filepath = "data")
10
11 db = MaterialsDatabase(
12   api = api,
13   backend = backend)
14
15 db.add_material("8Ax_8kdhSkpHy4qYb2wzqGH2Cnem")
```

Listing 3.1: Setup of a `MaterialsDatabase` with `MADAS`.

Listing 3.1 shows how a `MaterialsDatabase` can be used to download data using the NOMAD Archive API[2]. In line 5, an `API` object is created, which is used to access data from NOMAD. In line 7, a `Backend` object is initialized. As keyword arguments, the name of the database file and relative path are defined in lines 8 and 9, respectively. In line 11, the `MaterialsDatabase` is initialized, taking the previously created `API` and `Backend` as keyword arguments. In line 15, a material is added to the database, *i.e.*, the data is downloaded from NOMAD, using its entry id[3], via the `API` and stored in a database file called *materials_database.db*.

[2]For more information on the NOMAD API, visit `https://nomad-lab.eu/prod/v1/staging/docs/tutorial/access_api.html` and `https://nomad-lab.eu/prod/v1/gui/analyze/apis`

[3]`https://nomad-lab.eu/entry/id/8Ax_8kdhSkpHy4qYb2wzqGH2Cnem`

3.1.2 Fingerprinting

The collected data can be analyzed, as shown on the right side of Fig. 3.1. As discussed in Sec. 2.4, in order to represent materials and their properties for data analytics and machine learning, a suitable descriptor of the data must be used. Here, focusing on similarity analysis, the data are represented using material *fingerprints*, which we define as the combination of a descriptor d (see also 2.4) and a similarity score S that assigns the similarity S_{ij} to each pair of descriptors (d_i, d_j). The similarity score ranges between 0 and 1, where $S = 0$ ($S = 1$) defines that the two descriptors (d_i, d_j) are maximally different (identical). The choice of the similarity score is arbitrary in general, however, depending on the research question, functions with specific properties may be required. For most applications, symmetric scores ($S(d_i, d_j) = S(d_j, d_i)$) are beneficial. Asymmetric scores, such as the Tversky[88] coefficient, as used in drug discovery (see Sec. 2.5), have, to our knowledge, not found applications in materials science so far. The complement $D(d_i, d_j) = 1 - S(d_i, d_j)$ of a similarity score defines a measure of distance between two descriptors. It is convenient that these measures fulfil the following properties of a metric:[86]

1. $D(d_i, d_j) \geq 0$ and $D(d_i, d_i) = 0$ for all d_i, d_j
2. $D(d_i, d_j) = D(d_j, d_i)$
3. $D(d_i, d_j) \leq D(d_i, d_k) + D(d_k, d_j)$
4. $d_i \neq d_j \rightarrow D(d_i, d_j) > 0$.

The third property, *i.e.*, fulfilling the *triangle inequality*, is important for several applications, such as clustering (see also Sec. 3.3, where this is used), because it supports the interpretability and consistency of the results.

It is useful to distinguish between different fingerprint *types* and *parameterizations*. Different types refers to using different descriptors, *e.g.*, SOAP or MBTR (both introduced in Sec. 2.4), to represent the atomic structure of a material. While, for this example, the former describes the material as a power spectrum derived from the atomic density, the latter uses broadened geometry functions of the crystal geometry. Naturally, these two representations are not compatible. Ad-

ditionally, most descriptors also use different numerical settings in the definition of the descriptor.[66] Such settings could, *e.g.*, be the cutoff of radial basis functions in SOAP,[69] or the choice of geometry functions in MBTR.[74] Fingerprints with different parameterizations cannot –in general– be used to compute meaningful similarity scores. However, the development of fingerprint types that allow for computing the similarity in these cases is possible and may be useful in some scenarios.

In `MADAS`, fingerprints are implemented using a `Fingerprint` class. It defines all relevant methods to integrate with the rest of the `MADAS` framework and provides a template for quickly developing new fingerprints. By convention, two methods must be implemented: `calculate` and `from_material`. The former takes data as arguments, does the required transformation, and stores them in the `Fingerprint`s `data` attribute. Such data can be, *e.g.*, the atomic structure encoded as an ASE `Atoms` object, and a possible transformation operating on it is computing a SOAP vector. The `from_material` method is used to calculate the fingerprint directly from a `MADAS Material` object. To do so, it retrieves the required data from a `Materials` object and passes it to the `calculate` method. Furthermore, methods to retrieve the fingerprint data in the `json` format[32] and to serialize and deserialize the fingerprint are implemented. This allows to store the fingerprints both in a `MADAS MaterialsDatabase` (see above) or in a text file, if necessary.

```
1 from madas import Fingerprint
2 from madas.fingerprints.DOS_fingerprint import
      DOSFingerprint
3
4 fp1 = Fingerprint("DOS").calculate(energy_values_1,
      dos_values_1)
5 fp2 = DOSFingerprint().calculate(energy_values_2,
      dos_values_2)
6
7 sim = fp1.get_similarity(fp2)
8 many_sims = fp1.get_similarities([fp1, fp2])
```

Listing 3.2: Generation and usage of `Fingerprints` with `MADAS`.

Listing 3.2 shows the initialization and calculation of DOS fingerprints (see also Sec. 3.2 using `MADAS`. Built-in fingerprints, *i.e.*, those that are defined within `MADAS` and not created by a user, can be imported directly (line 2), or initialized through the generic `Fingerprint` class (line 4). `Fingerprints` can also be calculated directly from `MADAS` `Material` objects. Lines 7 and 8 show how the similarity to one or more fingerprints can be calculated.

3.1.3 Similarity matrices

The matrix containing all pairwise similarities between members of a set of fingerprints is called *similarity matrix*. For a symmetric similarity score, this matrix is also symmetric. `MADAS` implements several classes that can be used for many tasks associated with similarity matrices, including to calculate, manipulate, and store them, or to retrieve individual entries or sub-matrices. First, we introduce the general `SimilarityMatrix` class. It is used for square matrices, which have the similarity between the same materials in their columns and rows. The column and row indices correspond to the same fingerprints, therefore the diagonal elements represent the similarity of a fingerprint to itself. The calculation of symmetric ma-

trices can be sped up by a factor of two, by calculating only the upper (or lower) triangular part of the matrix. Additionally, the calculation of the matrix can be parallelized: Since the rows of the matrix are independent of each other, they can be computed as individual tasks on different cores of a CPU. Additional features are, among others, retrieving the most similar materials for a given reference, aligning two or more matrices to ensure that their rows and columns represent the same materials, or retrieval of all unique entries. These functions have proven to be valuable tools for understanding the relations between materials in a dataset and can be used for a variety of analysis and visualization tasks.

```
1  from madas import SimilarityMatrix
2
3  simat = SimilarityMatrix().calculate(fingerprint_list)
4
5  simat.save(
6    filename = 'similarity_matrix.npy',
7    filepath: str = '.')
8
9  sub_matrix = simat.get_sub_matrix(list_of_mids)
10 matrix_row_1 = simat[mid]
11 matrix_row_2 = simat[0]
12 matrix_entry = simat.get_entry(mid1, mid2)
```

Listing 3.3: Calculation and usage of `SimilarityMatrix` objects with `MADAS`.

Listing 3.3 presents some of the methods provided by the `SimilarityMatrix` class in `MADAS`. The matrix is calculated from a list of fingerprints in line 3. In line 5, it is saved to a file called *similarity_matrix.npy* (line 6) in the current directory (line 7). A sub-matrix can be obtained by providing a list of material IDs (`mids`, line 9). Matrix rows can be accessed either by the respective `mid` (line 10) or matrix index (line 11). Individual matrix entries can be obtained using the `mids` of the fingerprints they were calculated for (line 12).

A more specialized type of matrix is the *overlap similarity matrix*, implemented in a `OverlapSimilarityMatrix` class. Its rows and columns correspond to the fingerprints of different materials, *i.e.*, the entries of an overlap matrix O with similarity score S, for sets A and B of fingerprints f, with $f_i \in A$, $f_j \in B$, and $A \cap B = \emptyset$, are $O_{ij} = S(f_i, f_j)$. The `OverlapSimilariyMatrix` can be used, *e.g.*, for testing the predictions of ML models. The training data of a ML model (see Sec. 2.3) can be used to generate one set of fingerprints, say, the column index of the matrix. Then, these fingerprints are used to calculate the similarities to the test data, *i.e.*, a dataset that was not used in the training of the ML model. The values of the matrix then show how (dis-)similar the training set is to the test set. This can give insight into the expected error of the model: If the similarity of the training data to the test data is low, the model is more likely to perform badly.

Computing similarity matrices can be resource-intensive, both in terms of CPU and memory, since the number of entries to compute and store grows quadratically with the number of fingerprints. To reduce the compute time for (very) large matrices, additional parallelization can be used by dividing matrices into several sub-matrices. These sub-matrices can be computed on individual CPUs without shared memory, such as the nodes of a high-performance compute (HPC) cluster. This distributed calculation of entries is implemented in `MADAS` for symmetric matrices in a class called `BatchedSimilarityMatrix`.

Figure 3.2 shows how the calculation of a matrix of 5 fingerprints, *i.e.*, 25 entries, can be split into 4 individual processes. This is achieved by assigning the independent blocks of the matrix (labeled (a)-(f)) to different tasks. Each task can be executed by, *e.g.*, a different CPU / MPI task on a HPC system. The implementation of this functionality makes use of both `SimilarityMatrix` and `OverlapSimilarityMatrix` objects to represent the individual matrix blocks. In Fig. 3.2, matrix blocks (f), (e), and (c) are symmetric matrices, and (a), (b), and (d) are overlap matrices. To reduce the memory footprint of the computation, only the fingerprints necessary for computing a sub-matrix are kept in memory. This is realized using the `serialize` function of each `Fingerprint` object (see Sec. 3.1.2 above) by writing the serialized fingerprints to files corresponding to each submatrix. For the example in Fig. 3.2, this means that three fingerprint

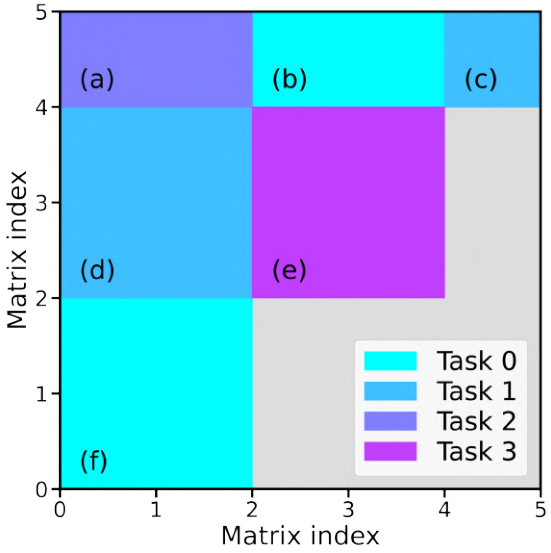


Figure 3.2: Illustration of matrix batching for parallel computation of entries for a symmetric matrix. In this example, 4 independent tasks compute parts of the matrix. The color indicates which task computes which part of the matrix. The gray area represents duplicate entries that are not computed. The alphabetic labels are described in the text.

files are written: for indices 0-2, required by blocks (a), (d), and (f), for indices 2-4, required by blocks (b), (d), and (e), and for indices 4-5, required by blocks (a), (b), and (c).

3.2 Spectral fingerprints

Many material properties are expressed in terms of spectra. A spectrum represents a mapping between an *independent variable*, *e.g.*, the energy, and a *dependent variable*, *e.g.*, the number of electronic states at that energy. Examples are, among others, the electronic or vibrational DOS, or optical absorption spectra. Computing the similarity between spectra is challenging because the values of the dependent variable can only be meaningfully compared if the values of the independent variable are the same. For example, it must be verified that the compared spectra represent the same energy region. The challenge becomes even bigger because spectra are given as discrete arrays of values, rather than continuous functions. It

is possible that two spectra map the same region of independent variables, but using different sampling points. Then the dependent variables of the spectra cannot be compared either.

To address this challenge, we develop a spectral fingerprint that can be used to encode any spectrum as a binary-valued 2D image. It is computed using a non-uniform transformation, which has the advantage that the fingerprint can be adapted to focus on specific ranges of the independent variable. This is achieved by defining a so-called *feature region*, in which the spectrum is represented with high resolution. Outside of the feature region, the resolution of the representation can be reduced. Our implementation mitigates the common drawback of electronic-structure descriptors found in the scientific literature (see Sec. 2.4.3) where all contributions to the spectrum are weighted equally. This means that they cannot account for the fact that different energy regions are associated with distinct physical phenomena. The optical properties of semiconductors, for example, are largely influenced by the size of the band gap. For metals, the region around the Fermi energy is most relevant. Although the usage of PCA for dimensionality reduction, as used in the literature[53,81] can introduce a reweighting of spectral features, the importance of a specific energy region cannot be tuned manually to focus on a specific research question. Rather, it is determined by the variance of the data that is used for generating the descriptors.

The following explanation of the fingerprint generation and similarity metric closely follows our publications, Ref. 89 and 33.

3.2.1 Fingerprint generation

Figure 3.3 shows the generation of a spectral fingerprint from an electronic DOS. As a first step, if required, the spectrum can be shifted by an energy, $\Delta\varepsilon$, to define the reference energy, $\varepsilon_{\mathrm{ref}}$, of the feature region. Here, the spectrum is centered at the Fermi energy, $\varepsilon_{\mathrm{ref}} = E_{\mathrm{Fermi}} = 0$ eV. Then, the spectrum is integrated over an even number of intervals, N_ε, to obtain a histogram, $\{\rho_i\}$:

$$\rho_i = \int_{\varepsilon_i}^{\varepsilon_{i+1}} \rho(\varepsilon) d\varepsilon, \tag{3.1}$$

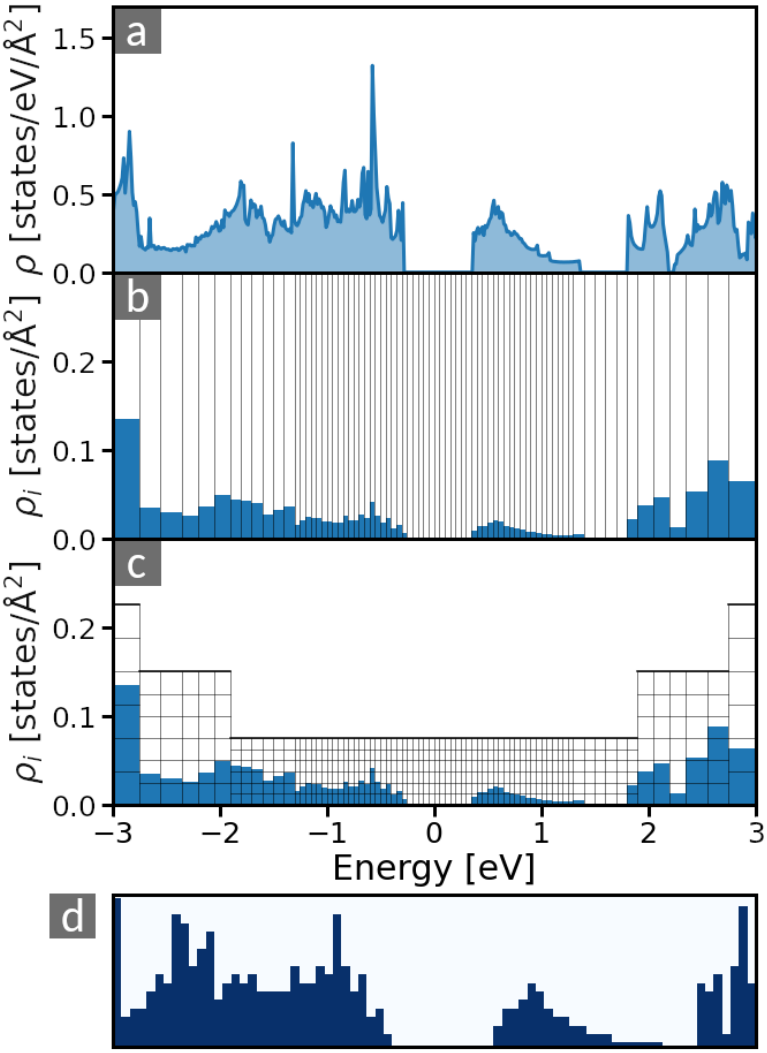


Figure 3.3: Generation of spectral fingerprints from the electronic DOS. The DOS, $\rho(E)$, of a material is numerically integrated over energy intervals $[\varepsilon_i, \varepsilon_i + \Delta\varepsilon_i)$ (See Eq. 3.1). The resulting histogram (b) is subsequently discretized using a grid, and transformed to a raster image (d) representing the cells of this grid. The binary valued fingerprint vector is obtained by concatenating the columns of the raster image. Each dark (light) pixel corresponds to a value of 1 (0) in the fingerprint vector. To increase visibility, in panel (c) only every fifth discretization step is shown. Furthermore, we use a coarse grid with $N_\rho = 30$, $\rho_{\min} = 0.075$, and $\rho_{\max} = 0.825$. Figure from Ref. 89.

with $i \in [-N_\varepsilon/2,\, N_\varepsilon/2]$, $i \in \mathbb{Z}$, $\varepsilon_0 = 0$, $\varepsilon_{i+1} = \varepsilon_i + \Delta\varepsilon_i$ for $i \geq 0$, and $\varepsilon_{-i} = -\varepsilon_i$. The (non-uniform) integration intervals are defined as

$$\Delta\varepsilon_i = n(\varepsilon_i, W, N)\, \Delta\varepsilon_{min}, \tag{3.2}$$

where the parameter $\Delta\varepsilon_{min}$ is the minimal integration width and

$$n(\varepsilon, W, N) = \lfloor g(\varepsilon, W)N + 1 \rfloor \in [1, N]\,. \tag{3.3}$$

Here, $\lfloor \cdot \rfloor$ denotes the 'round down' operator and $g(\varepsilon, W)$ is defined as

$$g(\varepsilon, W) = (1 - \exp(-\varepsilon^2/2W^2)). \tag{3.4}$$

$N \in \mathbb{N}$ $(N > 1)$ determines the maximum interval width $\Delta\varepsilon_{max} = N\Delta\varepsilon_{min}$. The parameter W is used to define the width of the feature region: For $\varepsilon = 0$, $\Delta\varepsilon_i$ equals $\Delta\varepsilon_{min}$, for $|\varepsilon| > W$, it approaches $\Delta\varepsilon_{max}$. Using this functional form, the integration intervals, indicated by vertical lines in Fig. 3.3 (b), are smaller in the feature region $|\varepsilon| < W$, resulting in a finer discretization, and therefore a more detailed representation of the spectrum. The energy cutoff values $\varepsilon_{\min}$ and $\varepsilon_{\max}$ can be used to limit the resulting histogram to a specific energy range. The histogram is then overlaid by a grid as shown in Fig. 3.3 (c). For every column i of the histogram, the grid contains N_ρ intervals of height

$$\Delta\rho_i = n(\varepsilon_i, W_H, N_H)\, \Delta\rho_{min}. \tag{3.5}$$

The parameters W_H, N_H, and $\Delta\rho_{min}$ are used analogous to W, N, and $\Delta\varepsilon_{min}$ above: Close to $\varepsilon = 0$, the height of the grid is small $\Delta\rho_i = \Delta\rho_{min}$, resulting in a fine representation of the height of the histogram, while it approaches $\Delta\rho_{max} = N_H\Delta\rho_{min}$ for $|\varepsilon| > W_H$, reducing the accuracy of the representation. The grid is used to obtain a raster graphic of pixels: The number of "filled" pixels in column i is determined by

$$\min\left(\left\lfloor \frac{\rho_i}{\Delta\rho_i} \right\rfloor, N_\rho\right), \tag{3.6}$$

resulting in the 2D raster image shown in panel (d) of Fig. 3.3. It contains $N_\varepsilon \times N_\rho$ pixels, enumerated by an index α. By concatenating the columns of the image, it is transformed into a binary-encoded vector $\boldsymbol{d} = (f_1, ..., f_{N_\varepsilon \times N_\rho})$, where the component $f_\alpha = 1$ represents that the pixel α is completely filled and 0 otherwise.

3.2.2 Similarity metrics

As a similarity metric $S(\boldsymbol{d}_i, \boldsymbol{d}_j)$ for the DOS fingerprints $\boldsymbol{d}_i$ and $\boldsymbol{d}_j$, we use the Tanimoto coefficient (Tc),[90] defined as:

$$S(\boldsymbol{d}_i, \boldsymbol{d}_j) = \frac{\boldsymbol{d}_i \cdot \boldsymbol{d}_j}{|\boldsymbol{d}_i|^2 + |\boldsymbol{d}_j|^2 - \boldsymbol{d}_i \cdot \boldsymbol{d}_i}. \tag{3.7}$$

$S(\boldsymbol{d}_i, \boldsymbol{d}_j)$ can be interpreted as the intersection of the areas covered by the raster images $\boldsymbol{d}_i$ and $\boldsymbol{d}_j$, divided by their union. As a similarity score (see also Sec. 3.1.2), S can take real values in the range $[0, 1]$. It is equal to 1 (0) if $\boldsymbol{d}_i$ and $\boldsymbol{d}_j$ are identical (have no overlap). Note that the values of $S(\boldsymbol{d}_i, \boldsymbol{d}_j)$ do not map the overlap of the fingerprints linearly to the similarity score, *i.e.*, they do not generally describe the percentage of the overlap between the two spectra. We can illustrate their relationship by considering two spectra of equal area A. In this case, the overlapping area is given by $A \cdot 2S/(1+S)$. Using this equation reveals that a value of $S = 0.5$ corresponds to an overlap of 2/3 of the areas. To illustrate this metric, Fig. 3.4 exemplifies the process of calculating the similarity. The left panels show the DOS obtained by two calculations[10] with different numbers of basis functions (N_b, see also Sec. 2.1), but otherwise identical settings. They are converted into spectral fingerprints (middle panels) using the procedure described above. The right panel shows how the two fingerprints overlap: Red and blue correspond to the areas covered by the individual fingerprints, purple color indicates the area covered by both fingerprints. This results in a similarity S of 0.77.

To further illustrate the metric, Fig. 3.5 shows the DOS of four different materials (left panel) from the C2DB (see Sec. 2.2.2) and their respective similarities in a similarity matrix (right panel). Comparing the quantitative results in the matrix to the qualitative differences between the DOS, we can see that C_2 (graphene,

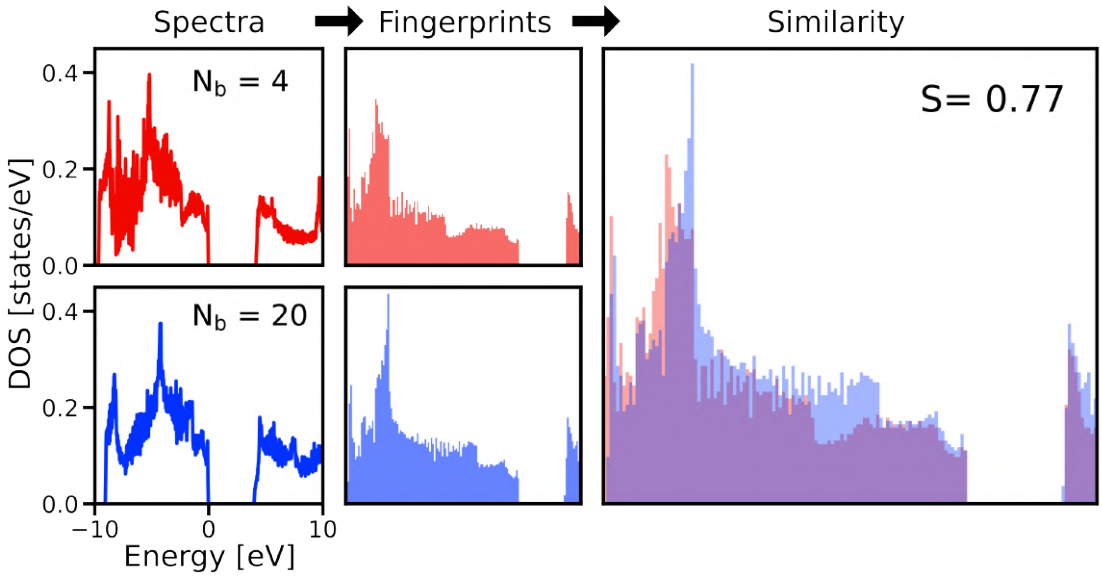


Figure 3.4: Workflow for the computing the similarity of DOS spectra. Two spectra (left panels), obtained by DFT (see Sec. 2.1) with different basis-set size N_b, are first converted to fingerprints (middle panels). Then, the similarity S between them is calculated (right panel). Figure from Ref. 33.

blue) has much fewer available states in the considered energy region than the other examples. Primarily for this reason, the similarity score with respect to all other materials is low, *i.e.*, $S \leq 0.14$. MoS_2 (orange) and WMo_3S_8 (green) show a high similarity score of $S = 0.84$, since both shape and magnitude are similar. Comparing them to FeO_2 (red) shows that, besides the similar shapes of the DOS for $|\varepsilon| > 1eV$, the presence of the band gap in MoS_2 and WMo_3S_8 leads to a low similarity of $S = 0.4$. In Sec. 5.1.2 we discuss the influence of the feature region of the fingerprint on the similarity score.

In the literature, many different similarity scores are reported for binary-valued descriptors.[79,86,91] In general, but specifically for our application, using the Tanimoto coefficient has several advantages. As discussed earlier, the similarity is highly interpretable, due to its relation to the overlap of the two spectra. Furthermore, unlike, *e.g.*, the Dice coefficient,[86] the Tanimoto coefficient obeys the triangle inequality, making it more suitable for applications such as clustering (see

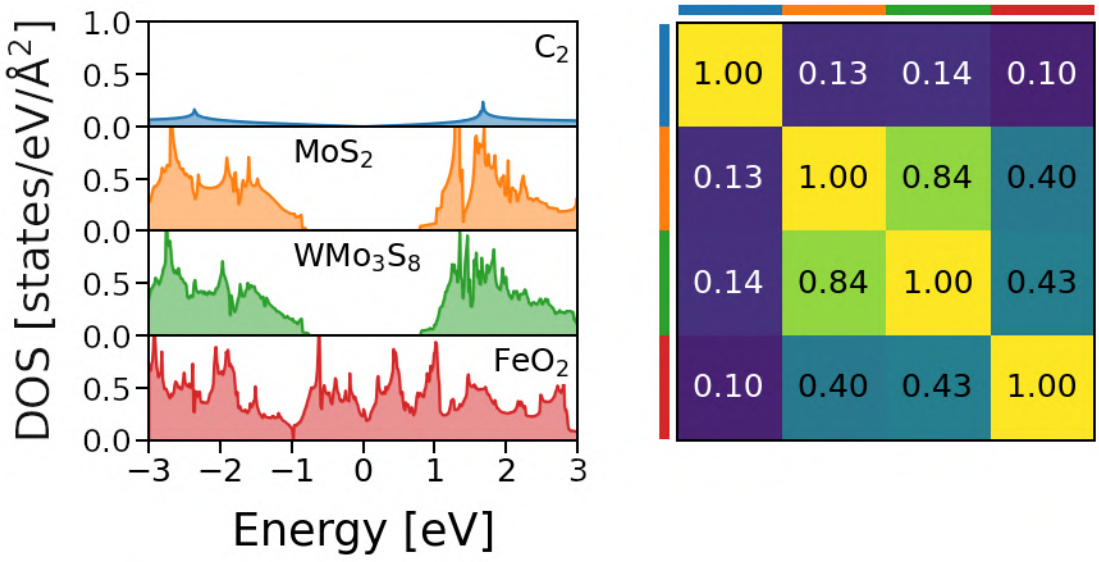


Figure 3.5: Similarity scores between DOS spectra exemplified on four different materials. The left panel presents the DOS of graphene (C_2), MoS_2, WMo_3S_8, and FeO_2, where the Fermi level is located at $E = 0$ eV. The right panel shows the corresponding similarity matrix, where the rows and columns follow the same order and are color coded. Figure from Ref. 89.

Sec. 3.3). In contrast to, *e.g.*, the cosine similarity, which requires the calculation of a square root, Tc is computationally cheap and can be computed efficiently using only binary operations on fingerprint vectors and bit counts. We note that for molecular similarity searches, it has been shown that the Tanimoto coefficient is a reasonable choice.[92]

3.2.3 Implementation

A Python implementation of the spectral fingerprint is available at GitHub (`https://github.com/kubanmar/dos-fingerprints`). It features an object model (`DOSFingerprint` and `Grid` classes) and is extensively documented in the code. Furthermore, it provides functions to visualize the fingerprints and grids. At the time of writing of this thesis, it is in version 2.0, implementing not only the non-uniform grid described above, but also uniform and user-defined grids. The description of parameters fol-

lows closely the description shown here and in Ref. 89. The package is well tested and will be extended and updated in the future, as new use cases arise.

The spectral fingerprints are also integrated within `MADAS`, optimized for the electronic DOS, in the `DOSFingerprint` class.

3.3 Clustering based on minimal similarity

Being able to compare the similarities between individual data points naturally raises the question whether the dataset contains an overarching pattern or structure when similar data are grouped together. Clustering algorithms, as introduced in Sec. 2.3.2, can be employed to search for an answer to this question. However, since every algorithm is based on assumptions about the structure of the data, no general algorithm exists that can be used for every problem.

Many clustering algorithms, such as the popular k-means algorithm,[64] use the number of clusters as the main parameter. Often, for materials-science data, this number cannot be estimated. This is due to the heterogeneity of the data: For example, the electronic structure of materials consisting of the same elements and having similar crystal structures, can be largely different (see, *e.g.*, the example of In_2S_2 in Sec. 5.2.1). Thus, the distribution of the data, measured by, *e.g.*, its chemical diversity, cannot provide enough information about the expected number of clusters.

Other clustering algorithms make use of a similarities or distances between data points to find clusters. We define the cluster radius r_c as the largest distance between two members of a cluster,

$$r_c = 1 - S_{\min}, \tag{3.8}$$

where $S_{\min}$ denotes the minimal similarity between any two cluster members. A large cluster radius means that two cluster members can be very dissimilar. A small radius means that each member is highly similar to every other member. We call clusters with a small radius *compact* clusters in the following. In many

clustering algorithms, a data point is considered a member of a cluster if it is similar enough to at least one other member. Thus, the resulting clusters can have very large cluster radii, *i.e.*, they are not compact, and their properties can be very different. We therefore need a clustering algorithm that can find compact clusters.

On a technical note, the implementations of many popular clustering algorithms provide only a limited number of metrics to measure similarities or distances between data points. This allows them to be optimized towards numerical efficiency. However, customized similarity scores cannot be used with these implementations.

We therefore devise a simple clustering method, that yields all compact clusters of all materials that are more similar than a given threshold S_{thres} to their the *cluster center*, also called the *centroid*, *i.e.*, the cluster member with the highest average similarity to all other members. To do so, we require that the similarity score S obeys the triangle inequality for similarity scores, derived from the triangle inequality of their complements (1-S),[86]

$$S(\boldsymbol{d}_i, \boldsymbol{d}_j) \geq S(\boldsymbol{d}_i, \boldsymbol{d}_k) + S(\boldsymbol{d}_k, \boldsymbol{d}_j) - 1. \tag{3.9}$$

From this, it follows that any two members of the cluster that are more similar to the centroid than S_{thres}, will have a similarity of at least $2S_{\text{thres}} - 1$ to each other.

The algorithm uses a similarity matrix as input and can be described in a few steps: For each row of the matrix, the number of entries that are more similar than S_{thres}, termed *similar entries*, are identified. If no such row can be found, the algorithm stops because all possible clusters are found. From all rows, the one with the highest number of similar entries is selected. These are considered a cluster and the centroid is identified by the index of the matrix row. The row index of all cluster members (including the centroid) are noted together with their cluster label. Afterwards, the rows and columns of all cluster members are removed from the matrix. Then, the algorithm repeats the process from the beginning. When two centroids have the same number of similar entries and share any of them, the cluster with the highest average similarity is selected. The rows that do not belong to any cluster are considered *orphans* and receive the cluster label -1. Finally, the assignment of clusters can be read from the list of cluster labels, sorted by the

row index of the input matrix.

This clustering algorithm is also implemented as a standalone Python program, published as open-source software at `https://github.com/kubanmar/similarity_threshold_clusterer`.

Within `MADAS` (see Sec. 3.1), clustering algorithms can be applied to similarity matrices. For this purpose, `MADAS` implements a `SimilarityMatrixClusterer` object, which can interface all algorithms using the `scikit-learn`[93] API[4]. It allows for easy handling of the results returned by the clustering algorithm. Among other features, it can be used to relate the cluster labels to the material IDs used in the database or the rows and columns of the similarity matrix, or retrieving the similarity matrix sorted by cluster labels.

3.4 Additional fingerprints

Complementing the spectral fingerprint introduced in Sec. 3.2, we employ additional similarity measures. In Sec. 5.2.1, we use simple and highly interpretable fingerprints to effectively filter and analyse clusters based on the electronic structure. Furthermore, in Sec. 5.2.2, we employ a fingerprint of the atomic structure, based on a species-agnostic version of the SOAP descriptor.[69]

Periodic table of elements fingerprint

The electronic structure of a material close to the Fermi energy can often be understood in terms of the configuration of the valence electrons of its constituent atoms. Information about this configuration can be obtained by considering the column number in the Periodic Table of Elements (PTE). We therefore devise a descriptor based on this information by averaging this column number over the species of all atoms in the unit cell:

$$\overline{c}_m = \frac{1}{N} \sum_i^N c_{im}, \tag{3.10}$$

[4] See, *e.g.*, `https://scikit-learn.org/stable/api/sklearn.cluster.html`

where i runs over all N atoms in the unit cell of material m, and c_{im} denotes their column in the PTE. This descriptor is called *PTE descriptor* in the following. Due to its simplicity, it can be computed easily for any set of atoms.

To measure the similarity between two materials (m_j, m_k) based on the PTE descriptor, we compute their l_1 norm, also called Manhattan norm, and transform it to a similarity score using:

$$S(m_j, m_k) = \frac{1}{1 + ||\overline{c}_{m_j} - \overline{c}_{m_k}||_1}. \tag{3.11}$$

The similarity based on the PTE descriptor reflects the assumption that elements from the same column of the PTE behave electronically similar. This descriptor and similarity metric are implemented in `MADAS` in the `PTEFingerprint` class.

Symmetry-based fingerprint

The symmetries of the crystal lattice have a large influence on all material properties. Therefore, it can be expected that materials with similar symmetries share certain properties. Based on this assumption, we consider a similarity measure based on the space group (SG) of the material. The goal is to identify two materials as similar even in cases where the symmetry is broken, *e.g.*, due to chemical disorder. To achieve this, we remove all information about the species from the structure. In practice, this is done by replacing all atoms in a unit cell by a single species. Then, we determine the SG of the lattice using the software package `spglib`.[94] We use a tolerance of `symprec` $= 1 \times 10^{-1}$ to account for effects of crystal-structure relaxation. In the following, we call the thus obtained number the *SG descriptor*.

Based on this SG descriptor, we implement a symmetry-based fingerprint. First, the SG number is assigned a vector that encodes its symmetry operations in a one-hot representation. Then, the Tanimoto coefficient (see Eq. 3.7) is used as a metric. In `MADAS`, this is implemented in the `SYMFingerprint` class.

Atomic structure fingerprint

Besides the symmetry, we also encode the atomic structure by using the SOAP descriptor (see Sec. 2.4.2) as implemented in the `dscribe`[66] library. However, to reduce the size of the resulting SOAP vector and to allow comparison of the atomic positions independent of the atomic species, we use a species-agnostic version of the descriptor. This is realized by setting the atomic species of all atoms in the unit cell to the same element. Then the SOAP vectors, averaged over atomic sites, are computed using `dscribe`.

As a similarity metric we use the pairwise Gaussian kernel of `scikit-learn`.[93] These fingerprints are currently not included in the `MADAS` source code, but can be obtained from GitHub[5].

[5] `https://github.com/kubanmar/madas-examples/blob/master/notebooks/analyze_similarity_correlations.ipynb`

4 Data quality assessment

The first application of similarity measures that we will focus on is the assessment of data quality. Both in theory and experiment, it is assumed that there is an exact value for each considered physical quantity. However, particular experimental conditions or computational parameters may yield values that differ from the exact ones. For example, in DFT calculations (see Sec. 2.1), the KS wavefunctions cannot be correctly represented when too few basis functions are used. If the number of basis functions is increased, initially, large changes in the computed property can be expected, since the correct ground state is increasingly well represented. When enough basis functions are used, the result converges, *i.e.*, additional basis functions do not contribute significantly. The closer a computed or measured property is to the exact, or alternatively, the converged value, the higher the quality of the data. Here, similarity measures provide a "natural" way to frame this problem: Even if the exact value of a property is not know, the degree of convergence can be estimated by quantifying the differences between data obtained with, *e.g.*, different numerical settings, or experimental conditions. This can be achieved by designing fingerprints that represent the quantities under study, and verifying that with increasing numerical parameters the results indeed become more similar. Moreover, given a large enough set of calculations, we can find out which sets of numerical settings and approximations yield similar results, and select the data that are converged based on our observations. Therefore, using similarity provides a framework that offers high interpretability for finding interoperable[1] datasets.

The first section of this chapter, Sec. 4.1, presents a variety of examples of how specialized fingerprints can be used to quantify similarities and differences between data. Different examples of the impact of methods, approximations, and structural

[1] See also Sec. 2.2.1.

and numerical parameters on unit-cell volumes, electronic structure, computed and experimental optical spectra, are shown. Section 4.2 shows how interoperable datasets can be found by sorting the data either with respect to their numerical settings, or based on the mean similarity to the rest of the dataset.

4.1 Quantification of dissimilarities

Being able to show when properties are similar is equivalent to being able to tell when they differ. This is especially useful when we want to study what impact an approximation or numerical parameter (or a specific method of measuring a physical quantity) have on the results of a calculation (or experiment). The first step of understanding these differences on a quantitative level is explicitly showing that results obtained with different methods yield only low similarity scores.

In the following, we first illustrate the veracity of materials data on the example of the volumes of NaCl crystal structures, obtained from different open materials databases. We find, in agreement with the scientific literature, significant differences between material properties reported in these databases. We then proceed to showcase differences in the electronic DOSs that come from using various approximations. On the example of SiC, we show how a rigid shift of the conduction bands, introduced by the G_0W_0 approximation, can be quantified. For PbI_2, we show the impact of hybrid XC functionals and spin-orbit coupling on the DOS. We then show how the optical spectra of h-BN are influenced by the $\boldsymbol{k}$-point sampling. Finally, we highlight the applicability of our spectral fingerprint to experimental spectra.

4.1.1 Data from different online databases

Computational HT databases employ different workflows and parameter sets to obtain their results. That means that the data from different databases are not necessarily interoperable, as discussed in Sec. 2.2.3. Indeed, building mixed training sets for ML models from different HT databases leads to systematic prediction

errors.[33] We showcase these discrepancies by using `MADAS` (see Sec. 3.1) to download data from three large DFT databases, *i.e.*, AFLOW,[6] Materials Project[7] (MP), and OQMD.[5] Further details about these databases are presented in Sec. 2.2.2. All of them use the same DFT code, `VASP`,[20] and employ GGA in the PBE parametrization for the XC functional.

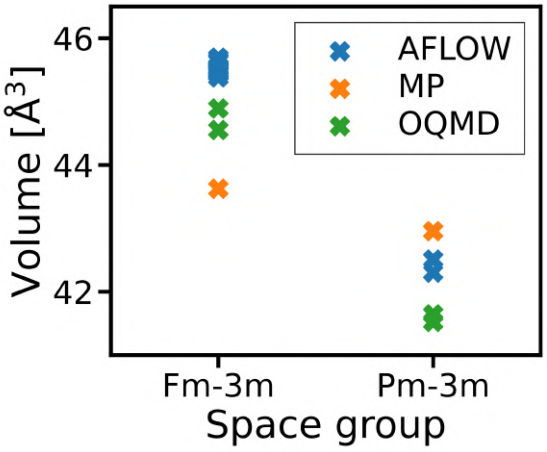


Figure 4.1: Volumes of two different space groups of NaCl, obtained from DFT calculations. The data stem from the databases AFLOW,[6] Materials Project,[7] and OQMD[5] and were collected using `MADAS` (see Sec. 3.1.1). Figure from Ref. 87.

Figure 4.1 shows the volumes of two different two-atomic NaCl unit cells, with space groups Fm-3m and Pm-3m, respectively. They were relaxed in terms of unit-cell volumes and atomic positions. We verified that the structures are symmetrically equivalent, using the method described in Ref. 95 and implemented in ASE,[39] version v3.22.1. Although the same computational approach was used in all three cases, the volumes differ by up to 2Å^3. While the MP database reports the lowest volume for the face-centered structure, it shows the highest volume for the primitive lattice. This shows that the differences in data quality not only introduce a constant difference between the volumes, but can influence the relative differences between them.

The results presented in Ref. 34 (see also Sec. 2.2.3) suggest that these discrepancies stem from differences in the plane-wave cutoff and different relaxation schemes. While the plane-wave cutoff is rather easy to interpret and higher cutoff

values can be considered to give more precise results, the benefits of one relaxation schema over the other are not obvious. Unified protocols for structure relaxation have been proposed recently.[96] Despite the fact that the proposed implementation relies on the usage of the workflow and provenance tracking system `AiiDA`,[36] which may not be suitable for every problem or material, it does not address the issue of interoperability for already existing data.

4.1.2 Electronic structure

Different approximations and numerical settings, called the *setup* in the following, can have a strong influence on the accuracy and precision of DFT results, respectively. The optimal setup is material dependent, requiring costly convergence tests for each new material studied. In contrast, HT databases employ the same set of parameters for all calculations, leading to the deviations described above and in Sec. 2.2.3. This unsatisfactory situation can be mitigated by understanding which materials or materials classes require which computational setups. This classification may be achieved by measuring the (dis)similarities between materials. The electronic structure of a material is a key result of a DFT calculation (see also Sec. 2.1). As such, it is a good candidate for monitoring the convergence of a calculation and the impact of different approximations and levels of theory.

To do so, we make use of spectral fingerprints (see Sec. 3.2) to encode the electronic DOS of calculations employing different approximations and levels of theory.[97] To specifically address differences in different energy regions, we use the cutoff parameter of the fingerprints to restrict them to the occupied bands (-10 eV $< \varepsilon <$ 0 eV) or unoccupied bands (0 eV $< \varepsilon <$ 10 eV). Additionally, we consider the whole energy range (-10 eV $< \varepsilon <$ 10 eV).

Figure 4.2 presents the DOS of SiC[98] stemming from two calculations uploaded to NOMAD. One was performed with DFT using the LDA XC functional (yellow), the other with the G_0W_0 approximation[99] (green). The latter is method within many-body perturbation theory, which provides corrections to be applied to the KS eigenvalues. These corrections significantly improve the accuracy of computed electronic band gaps. Here, the LDA and G_0W_0 results were obtained using the

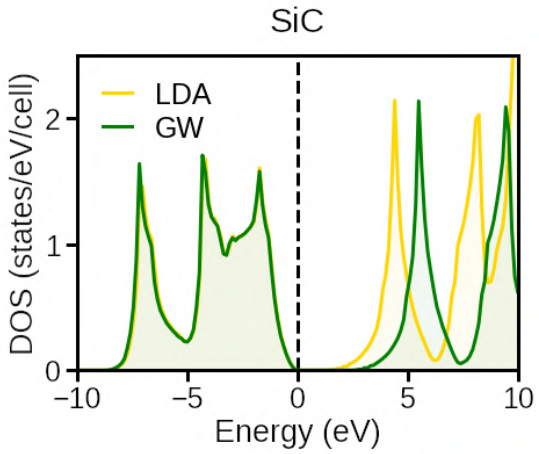


Figure 4.2: DOS of SiC[98] calculated using the LDA XC functional (yellow) and the G_0W_0 approximation (green). Considering the whole energy range, the similarity score is moderate (Tc = 0.66). The high overlap of the DOSs in the valence bands is reflected in their high similarity (Tc = 0.96), whereas the misalignment in the conduction bands results in a low Tc of 0.27. The dashed vertical line indicates the Fermi energy. Figure adapted from Ref. 97.

exciting code (see Sec. 2.1) based on the same numerical settings. Differences between both DOSs can therefore be attributed solely to the method.

Computing the similarity score for the whole energy range reveals a moderate value of Tc = 0.66. This can be easily understood by a qualitative comparison of the results: Below the band gap at $E_{\text{Fermi}} = 0$ eV, both DOSs overlap almost perfectly. The unoccupied bands, however, are subject to a rigid shift towards higher energies when the G_0W_0 approximation is used, opening the band gap from $\sim$ 1.28 eV (LDA) to $\sim$ 2.17 eV (G_0W_0). For comparison, the experimental value of the band gap is $\sim$ 2.42 eV[100] at a temperature of 2 K. By confining the DOS fingerprints to the valence region, we can quantitatively describe their overlap, leading to a high similarity of Tc = 0.96 in the occupied bands. Inspecting only the conduction bands, the shift towards higher energies for the G_0W_0 calculation is reflected in the small similarity score of Tc = 0.27. This observation allows us to outline an automatized way to detect how the usage of the G_0W_0 method affects the electronic structure: In a first step, the similarity is computed over the entire

energy region. If the similarity score is high, the effect of considering quasi-particle effects in the electronic structure is low. If it is moderate or low, we compute Tc separately for the valence and conduction bands. If the similarity is high in the valence region, we can focus our analysis on the conduction bands. Whether the differences in the conduction bands stem from a rigid shift can be tested by, *e.g.*, artificially shifting the G_0W_0 DOS by the energy difference between the band gaps. If the similarity score is high after applying this shift, G_0W_0 introduced a rigid shift. If the similarity is still moderate, the shape of DOSs differs and qualitative analysis is required.

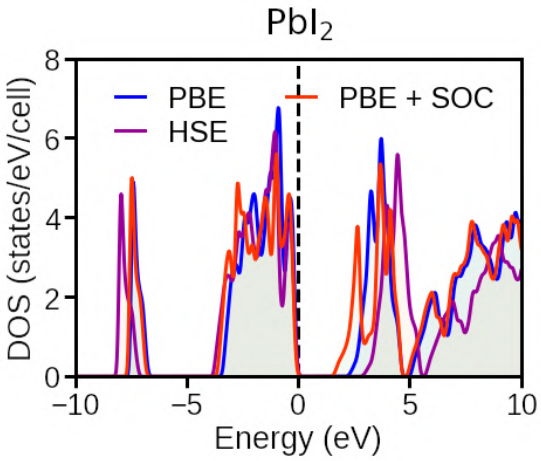


Figure 4.3: Impact of XC functional (PBE vs HSE06, Tc = 0.60) and SOC (PBE vs PBE+SOC, Tc = 0.71) on the DOS of PbI_2.[15] The dashed vertical line indicates the Fermi energy. Figure adapted from Ref. 97.

Figure 4.3 shows the DOS of PbI_2[101] for three different scenarios. It is computed with PBE, with and without the effects of spin-orbit coupling (SOC). SOC adds a correction that describes the relativistic effects coming from the interaction of the spin of an electron with its angular momentum. The third calculation is based on the HSE functional. In the present example, PBE is the lowest level of theory, so we consider it as a baseline to show to show how more accurate approximations affect the DOS. Focusing first on the valence bands, we find that PBE and HSE agree well close to the Fermi level ($E_{\text{Fermi}} = 0$ eV), however, going to lower energies, the HSE bands are subject to shifts towards lower energies. This is especially visible for the isolated bands at ~ -7 eV and leads to a moderate Tc of 0.73.

In the conduction bands, we see that HSE introduces a rigid shift towards higher energies, reflected in the low Tc of 0.45. Comparing PBE and PBE+SOC, we see a qualitatively different picture: While the position of the isolated bands at ~ -7 eV is the same, the shape of the DOS differs considerably in the upper valence bands (-4 eV $< \varepsilon < 0$ eV), leading to a moderate similarity score of Tc = 0.75. Close to the Fermi level ($0 \geq \varepsilon \geq 5$ eV), the shape of the DOS is different, more specifically, the peak at ~ 3 eV in the PBE DOS is shifted towards lower energies for PBE+SOC. Consequently, the conduction band minimum (CBM) of the PBE+SOC calculation is located at lower energies. At higher energies ($\varepsilon \geq 5$ eV), the DOSs overlap well. This leads to a moderate similarity in the conduction bands of Tc = 0.67. Considering the similarity over the whole energy range, the effect of HSE is greater, as reflected by the lower similarity of Tc = 0.60, than the effect of PBE+SOC (Tc = 0.71) compared to PBE.

Notably, the PBE band gap of PbI_2 is rather close to experiment,[102–104] however for the wrong reason. It can be understood by a cancellation of the shifts towards lower and higher energies of the CBM introduced by including SOC and exact exchange (HSE), respectively.

In both examples in this section, SiC and PbI_2, the impact of different approximations to the electronic structure can be captured by spectral fingerprints. By automatizing our analysis, as sketched for SiC, large amounts of data can be scanned. This would allow the characterization of materials based on the effect that given approximations have on the electronic structure. Based on this characterization, the computational effort for retrieving accurate results can be reduced, and interesting classes of materials for future research can be identified.

4.1.3 Optical absorption spectra

The application of spectral fingerprints is not limited to the electronic DOS. They can also be used for spectroscopy, as exemplified here for optical spectra obtained by many-body perturbation theory. The spectra that are analyzed in the following were computed using the **`exciting`** code. The imaginary part of the frequency-dependent macroscopic dielectric tensor is represented by a summation

of Lorentzian functions, centered at the eigenvalues of the Bethe-Salpeter equation (BSE).[105] The BSE accounts for excitonic effects, *i.e.*, bound states formed by electrons and holes. They play an essential role in the absorption behavior of semiconductors and insulators.

For computing the fingerprints, we use a fine, uniform grid, *i.e.*, $N = 1$, with $\Delta\varepsilon_{\min} = 0.02$ eV, $\Delta\rho_{\min} = 3 \times 10^{-3}$, $N_\rho = 256$, and restrict it to the energy range of 4 to 8 eV. Using a uniform grid means to give equal weight to all parts of the spectra. This is useful in cases when highly converged results, identical over the whole energy range, are required. Therefore, defining a feature region of the fingerprint would add an unnecessary bias. The following example closely follows Ref. 97. The data can be found in Ref. 106.

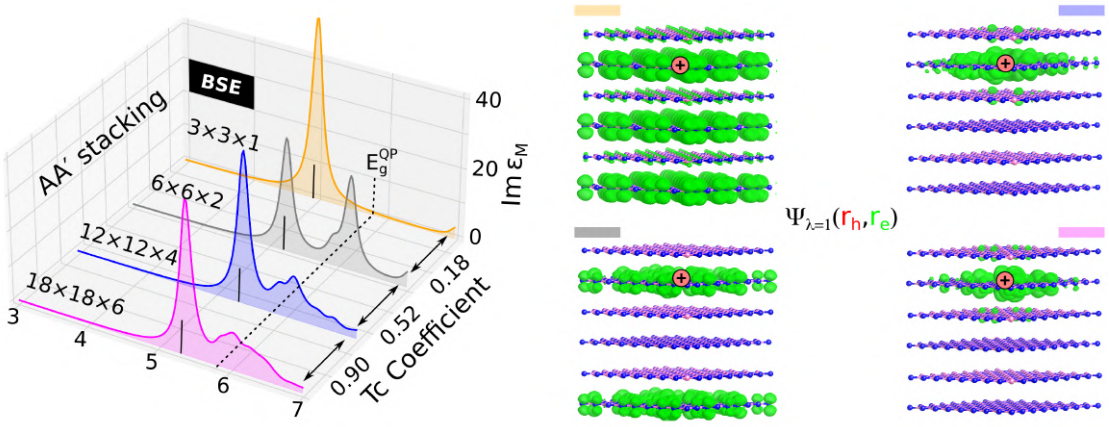


Figure 4.4: Optical spectra of h-BN. Left: Imaginary part of the dielectric function obtained with different $\boldsymbol{k}$-point grids. The similarity, measured by the Tanimoto coefficient Tc, between two adjacent curves is denoted on the right. The vertical dashed line indicates the direct quasi-particle band gap, the solid lines mark the absorption onsets. Right: Electron distribution of the electron-hole wavefunction $\Psi_{\lambda=1}(\boldsymbol{r}_h, \boldsymbol{r}_e)$, with the hole position fixed at the red dot, for different $\boldsymbol{k}$-samplings, indicated by the color of the spectra in the left panel. Adopted with slight modifications from Ref. 97.

The left panel of Fig. 4.4 illustrates how the optical absorption spectra of hexagonal boron nitride (h-BN) with A'A stacking vary depending on the $\boldsymbol{k}$-grid employed in the calculations. The spectra exhibit an absorption threshold (solid ver-

tical lines) significantly below the band gap (dashed line), indicating pronounced excitonic effects.[107] When the $\boldsymbol{k}$-point sampling is improved from the coarsest ($3 \times 3 \times 1$, orange) to the densest ($18 \times 18 \times 6$, magenta), the excitonic binding energy decreases, as indicated by the shift of absorption onset towards higher energies. A decreased binding energy can be associated with a increased delocalization of the exciton. The latter can be studied by visualizing the excitonic wavefunction, *i.e.*, the wavefunction describing the electron-hole pair forming an excition. It is displayed in the right panel of Fig. 4.4 for the same $\boldsymbol{k}$-meshes. In order to be able to visualize this quantity, which depends on the coordinates of both the electron and hole, the position of the hole is fixed on the h-BN layer and the electron distribution is shown. Comparing the wavefunction for different $\boldsymbol{k}$-grids, suggests that the exciton becomes more localized with increasing $\boldsymbol{k}$-sampling. This is, however, not the case and needs clarification. This apparent contradiction to the behavior of the absorption onset can be understood by considering the impact of $\boldsymbol{k}$-points on the calculation. Looking, for instance, at the behavior in out-of-plane direction, the reason for having the same electron distribution in every second plane is simply explained by the fact that the unit cell contains two such planes. By restricting ourselves to one $\boldsymbol{k}$-point in this direction, we obtain a replica of the same wavefunction in every other plane. The same applies to the in-plane directions where we also observe an apparent "delocalization". Only by considering a denser $\boldsymbol{k}$-grid, the localized character of the exciton becomes clear.

Using spectral fingerprints to represent the optical spectra, we can quantitatively describe their convergence behavior. The similarity score between the $3 \times 3 \times 1$ and $6 \times 6 \times 2$ results is $\mathrm{Tc} = 0.18$, increasing to $\mathrm{Tc} = 0.52$ between $6 \times 6 \times 2$, and $12 \times 12 \times 4$ and to $\mathrm{Tc} = 0.90$ between $12 \times 12 \times 4$ and $18 \times 18 \times 6$. The latter two spectra are almost identical in the onset region and only slightly differ at higher energies. This gradual increase in similarity indirectly reflects the behavior of the exciton wavefunction.

Keeping the computational cost of such calculations in mind, the usage of similarity measures presents itself as an excellent opportunity for quantifying the level of convergence w.r.t. computational parameters, automatizing the convergence process, and potentially reducing the number of required calculations.

4.1.4 Experimental spectra

The application of spectral fingerprints is not limited to data obtained by theory. We can apply the same methodology to experimental data to quantify the (dis)similarity of measurements. In this case, interoperability can be an even bigger issue than for computational results. One reason is that the same physical property can be measured by different methods. For example, the dielectric function can be obtained by, *e.g.*, optical absorption or reflection spectroscopy, ellipsometry, or electron-loss spectroscopy. While this variety of methods is advantageous, as it allows reducing any bias that could possibly be introduced by individual probes, all of these measurements yield the property after some transformations or modeling. This makes it harder to determine the origin of deviations, since it is not always clear if they should be attributed to the method, the transformations, or other, previously not considered influences, such as the resolution the detectors, or the temperature of the sample.

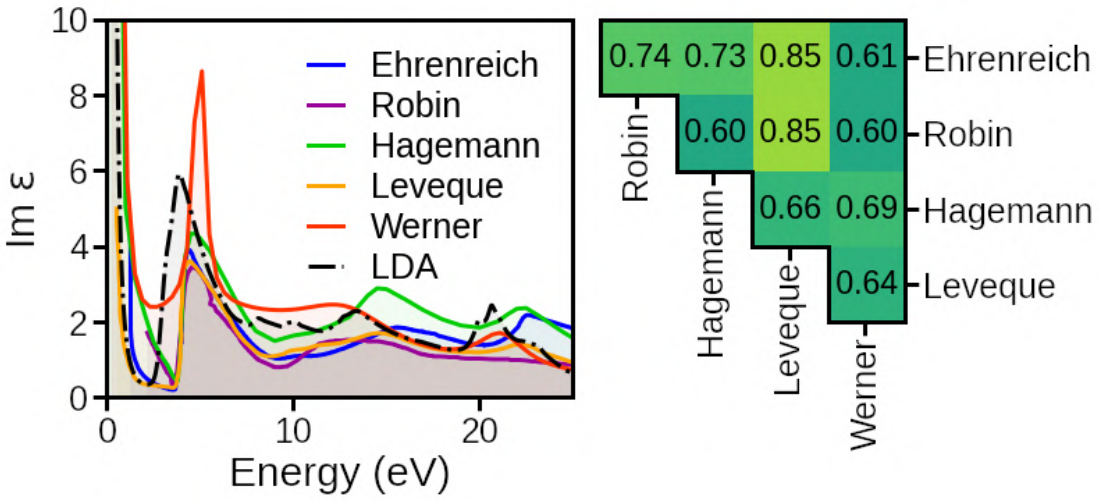


Figure 4.5: Left: Comparison of optical absorption spectra of elemental silver, stemming from different sources and measurements.[108–112] For comparison, a calculation within the independent-particle approximation based on a DFT calculations using the LDA XC functional is shown. Right: Corresponding similarity matrix, obtained using spectral fingerprints. Due to symmetry, only the upper off-diagonal elements of the matrix are shown. Figure from.[97]

In the following example,[97] we compare the optical absorption spectra of the elemental solid silver. They are shown in the left panel of Fig. 4.5, obtained by different sources and measurements.[108–112] The right panel presents the corresponding similarity scores, computed with spectral fingerprints using uniform grids with $N = 1$, $\Delta\varepsilon_{\min} = 0.02$ eV, $\Delta\rho_{\min} = 1 \times 10^{-3}$, $N_\rho = 256$, and a cutoff of 2.2 to 25 eV.

The absorption onset of the spectra at ~ 5 eV is almost identical in all measurements. However, the peak positions vary up to several eV and their heights roughly by a factor of two. Especially at higher energies, the differences in the shape of the features are larger. The best agreement can be seen between the data from Leveque[111] with both Ehrenreich[108] and Robin[109] [2]. This is well reflected in their comparatively large similarity scores of Tc = 0.85. The largest deviations can be seen between the measurements of Robin, Hagemann[110] and Werner *et al.*,[112] leading to low similarity scores of Tc = 0.6.

We furthermore compare the measured spectra to theoretical results. The latter were computed within the independent-particle approximation, using the LDA XC functional.[112] The similarity of this predicted spectrum to the experimental data of Robin[109] yields a moderate Tc of 0.63. Overall, they agree best with the experimental data of Werner *et al.*, Tc = 0.78.[112]

The variety of the experimental data shown here is undoubtedly large, and without detailed information about the respective samples, the experimental methods, and even environmental conditions, *i.e.*, the metadata, as required by the principles of FAIR data management (see Sec. 2.2.1), no conclusions about the data quality can be made. Clearly, it remains difficult to obtain large quantities of high-quality experimental data. Similarity measures can help to identify which data show the largest deviation. Furthermore, comparing experimental data with different levels of theory, allows to not only increase the accuracy of theoretical results, but also to identify which level of theory is required to reproduce experimental findings.

[2] Note that the highest similarity in this case does not mean that these results are the most reliable (see Ref. 112 for the scientific discussion).

4.2 Finding optimal parameter sets

In the previous section we have shown how similarity scores can be used to quantify the effects of different theoretical and experimental methods. This information by itself, however, does not tell which parameters or approximations give a consistent answer. The next step, therefore, is to explore if converged settings can be extracted from large sets of data. To illustrate this, we use data from NOMAD[9] (see also Sec.2.2.2) in the following. The specific dataset[113] was computed using the DFT code `FHI-aims`[22] as part of a systematic study of the impact of computational parameters on DFT results.[10] For this study, a grid search of numerical settings and approximations was performed, *i.e.*, all combinations of a fixed range of numerical settings, such as the number of $\boldsymbol{k}$-points used for Brillouin zone sampling or the basis set size, or approximations, such as the relativistic treatment and the exchange-correlation functional have been used to compute the same material.

In the following, we present two examples from this dataset. In the first example, we show that sorting the similarity matrix with respect to the convergence parameters reveals sets of calculations that are highly similar, meaning that the results of these calculations are interoperable. In the second example, we sort a similarity matrix based on the mean similarity of each entry and show that interoperable calculations can be found without full knowledge of the convergence parameters.

4.2.1 Sorting by numerical settings

We use the DOS of h-BN computed at the experimentally determined equilibrium volume. To compare the various results, we use spectral fingerprints of the DOS. The parameters are set to $\varepsilon_{\text{ref}} = -2$ eV, $\Delta\varepsilon_{\text{min}} = 0.05$ eV, $\Delta\varepsilon_{\text{max}} = 1.05$ eV, $\Delta\rho_{\text{min}} = 0.5$, $\Delta\rho_{\text{max}} = 5.5$, $N_\rho = 2048$, $W = 7$ eV, and the cutoff is set to -10 to 10 eV around the valence band maximum. We use `MADAS` (see Sec. 3.1 to compute the similarity matrix for these 144 fingerprints and display it in Fig. 4.6. The (symmetric) matrix is sorted such that lower indices, $i \leq 71$, correspond to LDA calculations and higher indices, $i > 71$, correspond to PBE results. The

LDA and PBE calculations are sorted separately by increasing numerical settings. The bottom panel of Fig. 4.6 shows the most crucial numerical parameters, *i.e.*, the total number of $\boldsymbol{k}$-points (N_{kpt}, blue) and the number of basis functions (N_b, orange). Furthermore, the calculations are sorted by a set of numerical settings,

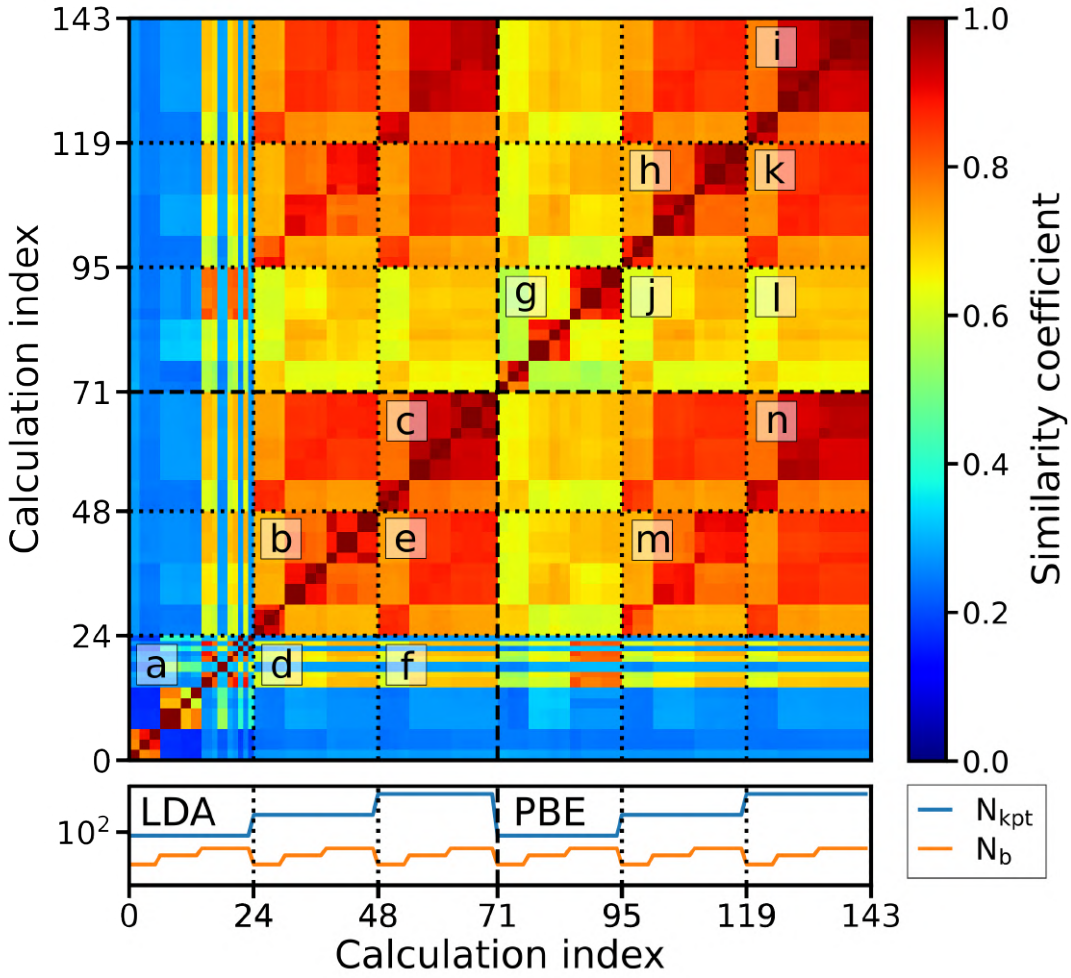


Figure 4.6: Similarity matrix of the DOS of h-BN, obtained with different basis-set sizes and $\boldsymbol{k}$-meshes and two different XC functionals. Low indices $i \leq 70$ correspond to LDA and high $i > 70$ to PBE data. The similarity scores are color coded. Dotted lines act as a guide to group sets of calculations with the same $\boldsymbol{k}$-point mesh. Letters mark blocks of the matrix that are discussed in the text. The bottom panel shows the number of $\boldsymbol{k}$-points (blue) and the number of basis functions (orange). Figure from.[33]

called "light", "tight", and "really tight", an `FHI-aims`-specific feature characterizing the basis set. Consecutive calculations with otherwise identical settings use different relativistic treatment for electrons, *i.e.*, "ZORA", or "atomic ZORA".[22] Using this specific way of sorting the matrix, which was done by repeatedly sorting and visualizing the matrix, distinct patterns emerge, which we discuss below. Dotted lines inside the matrix indicate sets of calculations with the same number of $\boldsymbol{k}$-points and act as a guide to the eye. Letters are used to label individual (unique) blocks of the matrix that will be further discussed.

First, we focus on the convergence of LDA results (indices $i \leq 70$) alone. Here, a block structure can be seen, where calculations having the lowest N_{kpt} (index $i \leq 23$) are notably dissimilar to all other blocks (d and f) as well as among themselves (block a). Nevertheless, they show pairwise similarity, suggesting that relativistic approximations have a minor impact on convergence of the DOS. Better agreement of the electronic structure is visible among calculations with medium ($24 \leq i \leq 47$) and high ($48 \leq i \leq 70$) numbers of $\boldsymbol{k}$-points (blocks b and c). When comparing them with each other, as seen in block e, calculations with low N_b show increased similarity among different N_{kpt}, but show lower similarity to calculations with higher N_b and *vice versa*.

Interestingly, parts of the convergence pattern of PBE calculations ($i \geq 71$) differ: Even calculations with low N_{kpt} exhibit moderate similarity to those employing high N_{kpt} (j and l). Calculations sharing the same N_b show high similarity to one another, even at low N_{kpt}, as seen in block g. Equivalent to LDA, calculations with medium N_{kpt} values attain high similarity scores with calculations using the maximum $\boldsymbol{k}$-point density (block k), if N_b is sufficiently high. The remaining diagonal blocks of the PBE data, *i.e.*, blocks h and i, reveal high similarity between calculations that employ the same N_b. When comparing LDA directly to PBE calculations, as it is done in the off-diagonal blocks of the matrix (indices $i \geq 71$ on the x-axis and $i \leq 72$ on the y-axis), again, N_b is the dominant parameter for the similarity of the DOS (see, for example, blocks m and n).

This example shows that, given that well prepared and annotated benchmark datasets are available, such as the data used here, calculations that yield similar results can easily be identified. Notably, the N_b used for the calculations has a

significant impact on the convergence of the results. This is especially interesting because the sets of numerical settings, *e.g.*, "light" or "tight", that are used by `FHI-aims`, also include suggestions for selecting basis functions as default inputs for DFT calculations. In the present dataset, both are varied independently. The analysis shown here suggests that the impact of these settings is clearly dominated by N_b, while other parameters, such as the parameters of the confinement potential[3], may play only a minor role.

The analysis shown here can be automatized. This allows to construct interoperable datasets from calculations with different computational parameters, as regions in parameter space can be identified which leave the DOS, and therefore the electronic structure, invariant. Then, data can be collected from within these regions in parameter space, rather than enforcing identical settings for all calculations in a dataset.

4.2.2 Grouping by mean similarity

The example above illustrates that –having a full description of the data– it is possible to find, and comprehensively visualize, sets of calculations with very similar results, *i.e.*, data that are interoperable. In the following, we approach the task from the opposite side, *i.e.*, we search for sets of calculations that give similar results without using the data provenance. This also means that the calculation that is considered most accurate is not known, *i.e.*, no reference exists to verify the precision for a given computational setup. In the following example,[87] we show that it may not be necessary to have such a highly converged reference, as the structure of the data itself can reveal the level of convergence.

Figure 4.7 shows a similarity matrix for $AlGaO_3$ which was computed using DOS fingerprints for the same dataset[113] as the example above. Likewise, all calculations were performed at the experimentally determined equilibrium volume, and the computational parameters were varied in a systematic way. In the bottom panel, we show again the most crucial convergence parameters, *i.e.*, the number

[3]For details we refer to the `FHI-aims` manual at `https://fhi-aims.org/uploads/documents/FHI-aims.240507.pdf`.

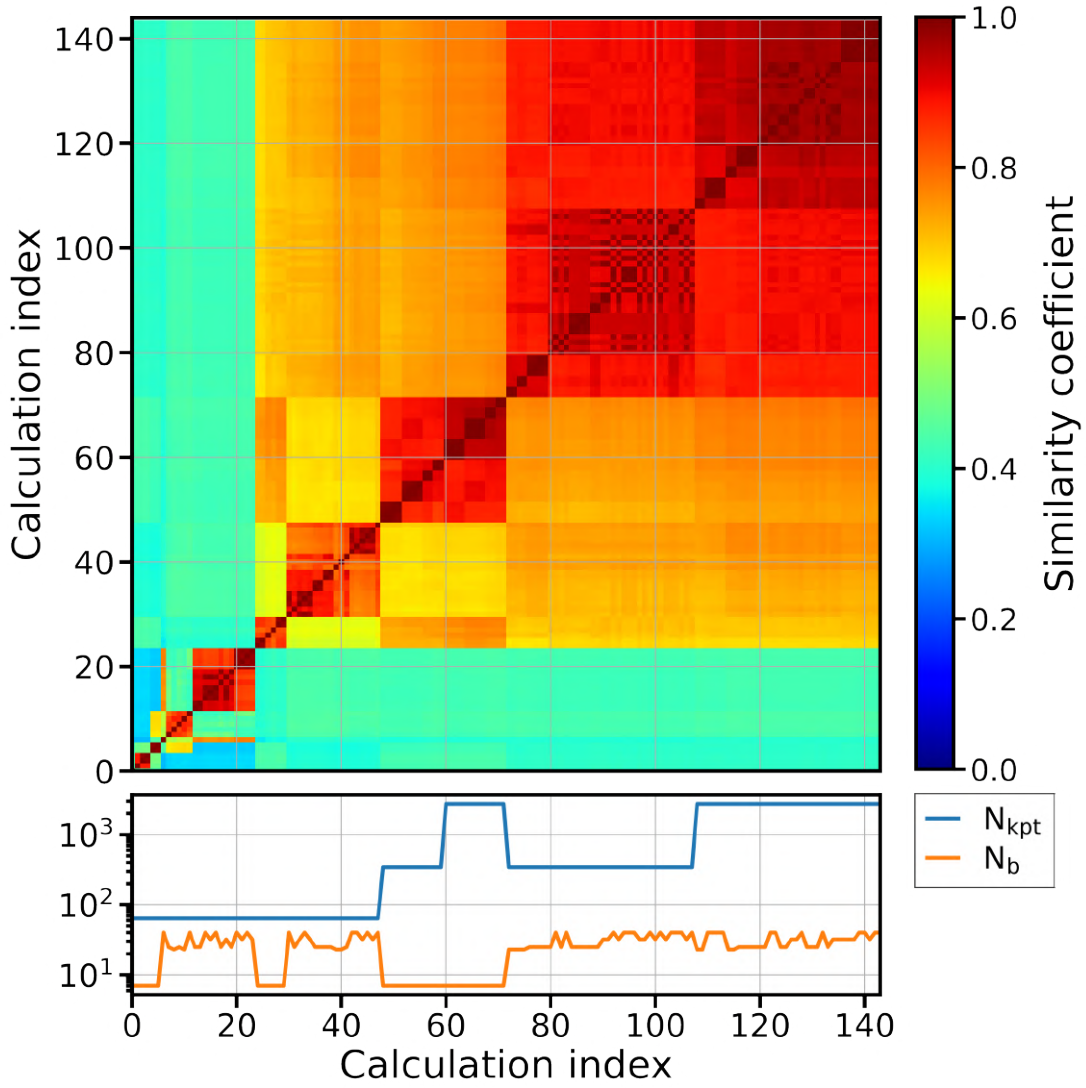


Figure 4.7: Similarity matrix of the DOS of $AlGaO_3$ from data obtained with different number of basis functions, N_b, and number of k-points, N_{kpt}. The matrix is sorted such that the calculation with the highest average similarity to the rest of the dataset has the highest calculation index. The bottom panel shows which N_{kpt} and N_b were used for the ground-state calculations. The DOS was obtained using 9 times more k-points. The color code indicates the similarity score, ranging from 0 (dark blue) to 1 (yellow). Figure from.[87]

of $\boldsymbol{k}$-points (N_{kpt}, blue) and the number of basis functions (N_b, orange). However, they are not used to sort the matrix, but only to illustrate the level of convergence

of the calculations. Instead, the matrix is sorted by the mean similarity $\bar{S}$ of each entry to the rest of the dataset. This means that after sorting, the calculation that is most similar to all others has the highest calculation index i.

From this sorting, structure emerges in the matrix: Subsets of calculations with similar DOS form clusters, as seen by the large red areas in the figure. Comparing these clusters with the most crucial convergence parameters in the bottom panel reveals a high level of correlation between these parameters and the DOS. The calculations that are most dissimilar to the rest of the data, *i.e.*, those with low indices ($i \leq 23$), all have the lowest number of $\boldsymbol{k}$-points. Furthermore, $\mathrm{N_b}$ appears to have no visible impact on the convergence of the DOS. Remarkable is that the next two groups of calculations, *i.e.*, $24 \geq i \geq 47$, contain calculations with the same $\mathrm{N_b}$ and $\mathrm{N_{kpt}}$, however, are significantly more similar to the rest of the data. We trace the discrepancies in the first group back to artifacts that appear in the DOS when the scalar ZORA approximation is used for the relativistic treatment in combination with too few $\mathrm{N_{kpt}}$. If a sufficient number of $\boldsymbol{k}$-points is used, these artifacts disappear. The next cluster, $48 \geq i \geq 71$, contains calculations with medium and high $\mathrm{N_{kpt}}$, but low $\mathrm{N_b}$. They show high similarity within the cluster, however, only moderate similarity with calculations with higher i. This can be interpreted such that the small basis cannot represent the correct ground-state wavefunction, irrespective of all other parameters. For higher indices, $i > 71$, the DOSs are highly similar to each other, since sufficient numbers of $\boldsymbol{k}$-points and basis functions are used. Here, $\mathrm{N_{kpt}}$ appears to have a larger impact on the electronic structure, as seen from the plateaus in the bottom panel.

In this example, we have shown that using a simple average of the similarity matrix row can be sufficient to obtain valuable information about the convergence behavior of the DOS, and also to find combinations of parameters that lead to numerical artifacts. This is possible because unconverged calculations, *i.e.*, those with too low numerical parameters, are dissimilar to both other unconverged and to converged calculations. Conversely, converged calculations are more similar to other converged calculations, thus their mean similarity is higher.

This analysis can be easily automated, *e.g.*, using `MADAS`, which was used for this example already. To obtain the same plot for another dataset, only the code

for downloading the data needs to be changed. However, the analysis still requires qualitative (human) interpretation. Despite its compelling simplicity, using the mean similarity of every row is not necessarily the best way of sorting the matrix for every material. Better results may be achieved using a clustering algorithm. However, the latter, *e.g.*, k-means[64] or DBSCAN,[114] often don't automatically sort the clusters in a meaningful way, but initialize the cluster labels randomly. This makes it harder to find correlations with the relevant convergence parameters.

4.3 Summary

In this chapter, we have shown how similarity measures can be used to quantify the effects of different approximations, settings, and methods employed by the scientific community. First, we have stressed the importance of understanding the data produced by HT approaches and their uncertainty. This was done on the example of the differences between the unit-cell volumes of NaCl, obtained from different computational databases. Shifting the focus to the electronic structure of materials, with the example of SiC, we quantified differences in DOS results calculated at the level of LDA and G_0W_0, respectively. The observed rigid shift of the conduction bands towards higher energies could be well captured using spectral fingerprints. For PbI_2, we could quantify the impact of SOC and exact exchange on the DOS. With these two examples, we illustrated how focusing the fingerprint on distinct energy ranges enables one to acquire additional insights into the effects of approximations employed within DFT. We then used spectral fingerprints to encode optical absorption spectra obtained using the BSE formalism, focusing on the convergence with increasing number of $\boldsymbol{k}$-points. For experimental spectra we could show the veracity of data obtained from different sources. These examples show that, even when only small amounts of data are available, quantifying their (dis)similarity is possible. Having more data at hand, this will allow us to understand the correlation of different methods, experimental conditions, and sample quality.

We then investigated how well our method of identifying converged results generalize when applied to systematically generated datasets. To do so, we used

`MADAS` to compute similarity matrices representing the DOS of different materials and sorted them by distinct criteria. First, we used the full information about the data provenance to show how the differences between DFT data computed with different approximations and computational settings can be understood and quantified. We showed for bulk h-BN that the semi-local XC functionals LDA and PBE show slightly different convergence behavior with the number of $\boldsymbol{k}$-points and basis functions, despite being generally expected to give very similar results in terms of the electronic structure. Going one step further, using only information about the similarity of the DOS, we have shown for $AlGaO_3$ that calculations with converged settings are on average more similar to the rest of the dataset. This may allow to identify converged calculations *without* having access to the full data provenance. Comparing the last two examples, which stem from the same dataset, we find that, as expected, the number of $\boldsymbol{k}$-points plays a larger role for the metal, $AlGaO_3$, than for the semiconductor, bulk h-BN.

Note that for the examples in this chapter, we use datasets that were created specifically for monitoring the convergence behavior. This is obviously not the case for heterogeneous data collections such as NOMAD.[4,9,28] However, by systematically advancing the here presented methodology and applying them to more materials, we expect to be able to generalize the results. A potential approach would be the usage of more complex similarity scores to automatically identify for which material, *e.g.*, the G_0W_0 approximation introduces a rigid shift, or alters the shape of the electronic band structure. Our method allows to automate convergence tests and perform systematic studies. Therefore, it enables the analysis of the convergence behavior of large amounts of data, which cannot be addressed by manual inspection. This will eventually allow for the identification of interoperable subsets of large datasets, *i.e.*, calculations (or measurements) that are *representative* for a given material. While the data obtained with the highest settings are generally expected to be most precise, these data are expensive to obtain and therefore rare. Using only these data is too restrictive for applications such as machine learning, which rely on the availability of large datasets. Thus, finding the data that are based on lower settings, but are "good enough" to be interoperable is an important goal.

5 Exploration of data spaces

For interoperable datasets, similarity measures can be used to explore the relationships between materials. In this chapter, we focus on the electronic structure of materials, making use of the spectral fingerprints introduced in Sec. 3.2. In several examples, we will use data from the C2DB.[8,53] An overview of the data contained in this HT database can be found in Sec. 2.2.2. Here, we mainly make use of the orbital projected DOS (PDOS, see Eq. 2.9), and compute the total DOS by summing over all atomic and orbital contributions. This quantity is given per unit cell in the C2DB. To be able to compare them regardless of the unit cell size of the materials, we divide it by the surface area. We furthermore use data from the NOMAD Repository and Encyclopedia.

First, we show how similarity searches can be applied to material properties, and how the feature region of spectral fingerprints affects the results of them. We then focus on clustering of materials data using similarity matrices, and explain how to interpret the clusters found by our analysis.

5.1 Similarity searches

As introduced in Sec. 2.5, similarity searches are used to scan a (large) dataset for entries that are most similar to a specific *reference*. Here, we present an adaptation of this approach to materials. To do so, we choose an appropriate fingerprint and a suitable dataset. The search is performed by selecting a reference material and then computing similarity scores of all dataset members to it. Then, these results are sorted to identify the most similar materials. To achieve this, we used the `MADAS` framework (see Sec. 3.1).

First, we present the results of a similarity search for GaAs in a large dataset of almost 1.9 million materials. We discuss the results of the search and highlight the exceptional nature of highly similar materials, even in large datasets. We show how treating spin channels individually can uncover unexpected similarities using Au and CoFe as examples. Finally, we study the influence of the feature region of the spectral fingerprint on the results of similarity searches.

5.1.1 Finding similar materials

To perform a similarity search, closely following Ref. 97, we compute spectral fingerprints of the electronic DOS for $\sim$ 1.9 million materials from the NOMAD Encyclopedia. These fingerprints employ a rather coarse, non-uniform grid with the following parameters: $\Delta\varepsilon_{\text{min}} = 0.05$, $\Delta\rho_{\text{min}} \sim 0.001$, $N = 21$, $N_H = 11$, $E_{\text{ref}} = -2$ eV, $W = W_H = 7$ eV, $N_\rho = 56$, and restrict the fingerprint to the energy region of -10 to 5 eV. The comparatively low value of N_ρ, leading to the coarse grid, is required to reduce the memory footprint of the fingerprints.

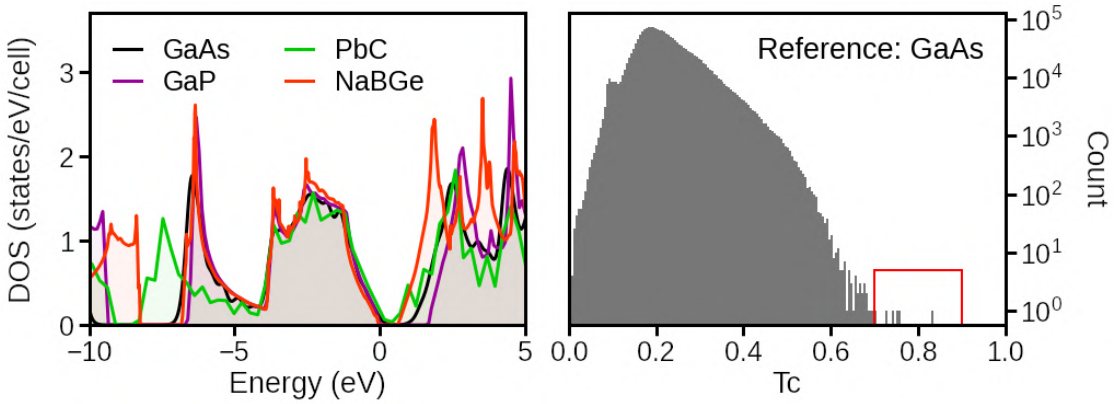


Figure 5.1: Results of a similarity search within approximately 1.9 million entries in the NOMAD Encyclopedia,[9,28] using the semiconductor GaAs as the reference material. The left panel displays the DOSs of the three materials most similar to GaA: GaP with Tc = 0.83, PbC with Tc = 0.75, and NaBGe with Tc = 0.74. The right panel shows the distribution of Tc values of the reference to all considered materials, where the red box indicates those most similar. Figure adapted from Ref. 97.

Figure 5.1 presents the results of this similarity search. The left panel shows

the materials that are most similar to the semiconductor GaAs.[115] The highest similarity score of Tc = 0.83 is found for GaP,[116] followed by PbC[117] (Tc = 0.75) and NaBGe[118] (Tc = 0.74). The four compounds share the higher-lying valence band structure and differ either in the lower valence region or the conduction bands. Comparing them individually, the high similarity between GaAs and GaP is to be expected. Both crystallize in the zinc-blende structure and are found to form stable alloys[119] containing all three elements, Ga, P, and As. The similarity of the electronic structure can be understood considering the electronic configurations of As and P: Both share the same valence configuration with half-filled $4p^3$, and $3p^3$ shells, respectively. This results in very similar DOSs, with deviations mainly stemming from the larger PBE band gap of GaP, of 1.56 eV compared to 0.51 eV of GaAs[1]. In the lower valence states (~ -10 eV to ~ -5 eV), the DOS of GaAs exhibits a prominent gap, which is smaller in GaP. The high similarity between GaAs and GaP was also found by previously, using $\sim$ 20.000 materials in the AFLOW database.[78] Less expected are PbC and NaBGe. The former is a purely hypothetical compound, also in zinc-blende structure. Attempts to synthesize it have been published, but they have not been reproduced. 2D materials made from Pb and C have been subject of recent theoretical studies.[120] Similarly, for the half-heusler compound NaBGe, little information is available. OQMD reports high formation enthalpies for both compounds[2], suggesting that they are likely to be unstable.

The right panel of Fig. 5.1 shows the distribution of similarity scores between the reference calculation, GaAs, and all 1.9 million compounds on a logarithmic scale. Notably, the mean of the distribution is at Tc $\approx$ 0.2. This means that the vast majority of materials have low similarity scores. We also emphasize that materials with high similarity scores, *i.e.*, Tc $>$ 0.7 for this material, are exceptional. For different reference materials, the shape of the distribution is similar, but its mean may be located at higher or lower similarity scores.

Effectively finding materials for novel applications requires a large pool of data, potentially larger than what is currently available in NOMAD. This can be seen

[1] Note that the band gap of GaAs is is not determined very accurately due to the comparatively large smearing applied in the calculation.

[2] $\Delta H_f = 1.468$ eV/atom for PbC,[121] $\Delta H_f = 0.479$ eV/atom for NaBGe.[122]

from the low number of materials that are highly similar. Additionally, novel, and more complex similarity measures should be used, considering not only one material property at a time, but combinations of them. In the present case, a measure of stability may be included in the search. Such additional metrics can drive the search results towards materials that are more likely to be synthesized.

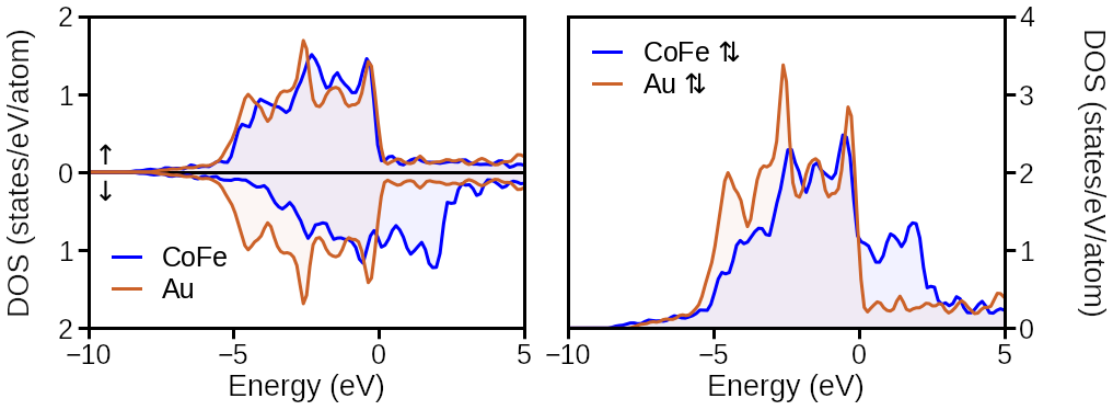


Figure 5.2: DOSs of Au and CoFe, for the spin components separately (left) and combined (right). The DOS of the majority spin channel shows high similarity (Tc = 0.80). For the minority spin, the similarity is significantly lower (Tc = 0.47). Thus, the similarity score for the total DOS is only 0.67. Figure adapted from[97]

Figure 5.2 shows on the left the impact of using spin-resolved DOSs for the two metals Au and CoFe. The DOSs of their majority spin channels show a high similarity of Tc = 0.80, reflecting the similar nature of the occupied 5*d* Au and 3*d* bands of Co, respectively. The minority spins (Tc = 0.47) differ mainly due to a rigid shift of the partially filled Co 3*d* minority band by about 2.5 eV towards higher energies. The right panel shows the total DOSs of both materials. Considering both spin channels combined results in a moderate similarity of Tc = 0.67. These results represent an application for complex similarity measures. In cases where spin-resolved DOSs are available, half-metals, *i.e.*, materials having a band gap in only one spin channel, may be identified, with potential applications for spintronics.

We also computed the four most similar materials for each of the ∼ 1.9 million fingerprints from the NOMAD Encyclopedia. This was made possible by using

the `BatchedSimilarityMatrix` implemented in `MADAS` (see Sec. 3.1.3). We used an HPC cluster with 50 MPI tasks, each utilizing 20 CPU cores, *i.e.*, 1000 cores in total. With this level of parallelization, it was possible to compute the matrix within 4 days. Using single-precision float values with a size of 32 bits, storing all unique entries of the similarity matrix requires slightly more than 7 TB of space.

Note that the data we used from the NOMAD Encyclopedia are not strictly homogeneous, *i.e.*, they contains calculations with different levels of precision. We expect more consistent and better interpretable results when working with more homogeneous datasets.

5.1.2 Impact of fingerprint parameters

The parameterization of fingerprints can crucially influence the results obtained in similarity searches. Following our previous work published in Ref. 87, we demonstrate this on the example of the feature region of spectral fingerprints using the DOSs of 2D materials stemming from the C2DB,[8,53] as introduced in the beginning of the chapter.

Figure 5.3 presents the materials most similar to $ZrTe_2$. The left side shows the DOSs of these materials, when the feature region is located in the conduction bands ($E_{ref} = 2$ eV, $w = 4$ eV, top panel), or in the valence bands ($E_{ref} = -2$ eV, $w = 4$ eV, bottom panel). The most similar material to $ZrTe_2$ is $HfZr_3Te_8$, irrespective of where we put the focus of the fingerprint, the similarity score results in $Tc = 0.83$ in both cases. The similarity of the electronic structures of these two materials can easily be easily verified by comparing the gray and blue curves in Fig. 5.3, which overlap in a large energy range. However, the second most similar material depends on the choice of the feature region: If the focus is on the valence bands, the second most similar material is $Hf_2Zr_2Te_8$ (turquoise), which matches the DOS of $ZrTe_2$ well over a large energy window between ~ -2 and ~ 3 eV, reaching a similarity of $Tc = 0.75$. Setting the focus on the valence bands, the second most similar material is $TiZr_3Te_8$ (yellow), which matches the DOS of $ZrTe_2$ especially well in the lower-energy region, reaching a similarity of $S = 0.76$.

Taking advantage of the flexibility of parameterizing spectral fingerprints, we

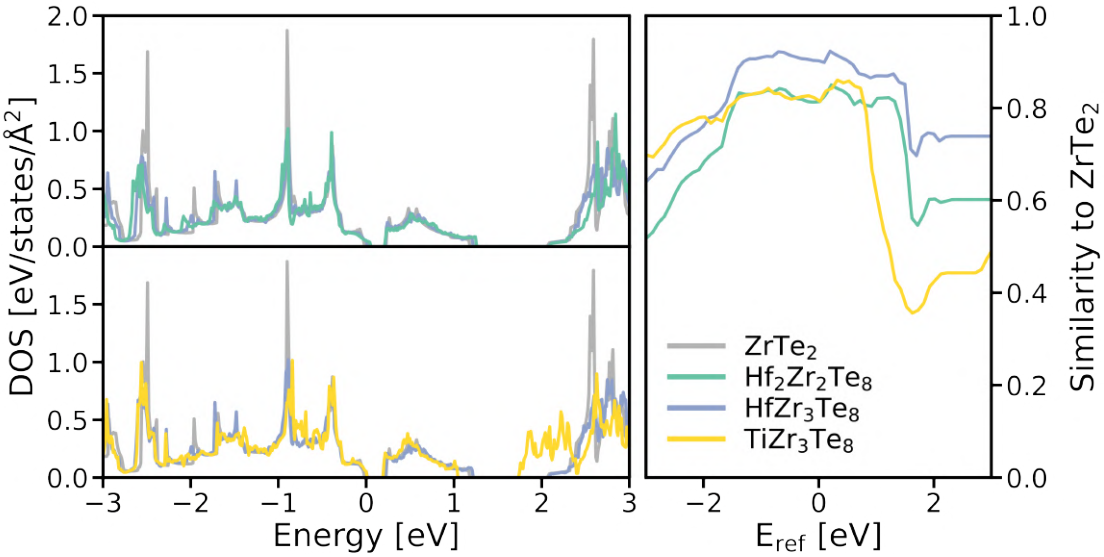


Figure 5.3: Results of a similarity search for $ZrTe_2$ for different choices of the feature region of the spectral fingerprint. The left side shows the DOSs of the most similar materials when focusing on the conduction (valance) bands in the top (bottom) panel. In the right panel, the similarity to $ZrTe_2$ in a small energy window around the reference energy E_{ref} is shown for the same materials. Figure from Ref. 87.

can quantify which energy regions have the largest impact on the similarity. To do so, we set the cutoff parameters (see Sec. 3.2.1) to $\varepsilon_{\text{min}} = -1$ and $\varepsilon_{\text{max}} = 1$ around the reference energy, and compute fingerprints of the four materials for a range of different reference energies between -3 and 3 eV. We use these fingerprints to study the similarity of the three most similar materials to $ZrTe_2$ discussed above. The right panel of Fig. 5.3 shows the results of this analysis. In the low energy range (-3 eV to ~ -2 eV) the material with the highest similarity is $TiZr_3Te_8$, shown in yellow. For the remaining energy range (~ -2 eV to 3 eV), $HfZr_3Te_8$ (blue) shows the highest similarity. In the highest energy range (~ 1 eV to 3 eV), $TiZr_3Te_8$ shows very low similarity. Comparing this to the bottom left panel, we see that this dissimilarity can be understood by the additional states above ~ 2 eV, which are only present for this material.

This kind of analysis is enabled by the modular architecture of `MADAS`, which allows for the seamless computation of fingerprints with different parameteriza-

tions. We furthermore implemented quantitative comparison between spectra within MADAS, which can be used in an automated manner, allowing to discuss spectra not only in a qualitative, but also in a quantitative way.

5.2 Clustering

Going beyond similarity searches, one may ask how materials relate to each other on a larger scale, *i.e.*, considering the similarity between all materials at the same time. This can be done with unsupervised machine learning, as introduced in Sec. 2.3.2. Specifically, we use clustering to find sets of materials that are similar to each other.

We first demonstrate that clusters of materials with similar electronic structures can be found in consistent datasets. We analyze these clusters based on the atomic and electronic structure of their members and showcase how a large number of compact clusters can be analyzed efficiently. In this analysis, we ask the question why the cluster members are similar, and which descriptors can be used to identify the underlying reason in an automated way. Then, we investigate the correlation between different similarity measures by comparing the clusters found using them. We visualize our results using similarity matrices and analyze them qualitatively.

5.2.1 Finding clusters of 2D materials

Here, we study the similarity of materials in terms of their electronic structure,[89] based on the DOSs of 3491 materials stemming from the C2DB,[8,53] as introduced in the beginning of this chapter. More information about the database can be found in Sec. 2.2.2. The DOSs are encoded in spectral fingerprints (see Sec. 3.2), using the parameters $\Delta\varepsilon_{min} = 0.05$ eV, $\Delta\varepsilon_{max} = 1.05$ eV, $N = \Delta\varepsilon_{max}/\Delta\varepsilon_{min} = 21$, $\varepsilon_{\mathrm{ref}} = 0$ eV, $W = W_H = 4$ eV, $N_\rho = 512$, $\rho_{min} = N_\rho\Delta\rho_{min} = 0.25$, $\rho_{max} = 2.75$, $N_H = \rho_{max}/\rho_{min} = 11$, $\varepsilon_{\mathrm{min}} = -3$ eV, and $\varepsilon_{\mathrm{max}} = 3$ eV. As a similarity metric, we use the Tanimoto coefficient Tc. An early version of MADAS (see Sec. 3.1) was used for all steps of this work.

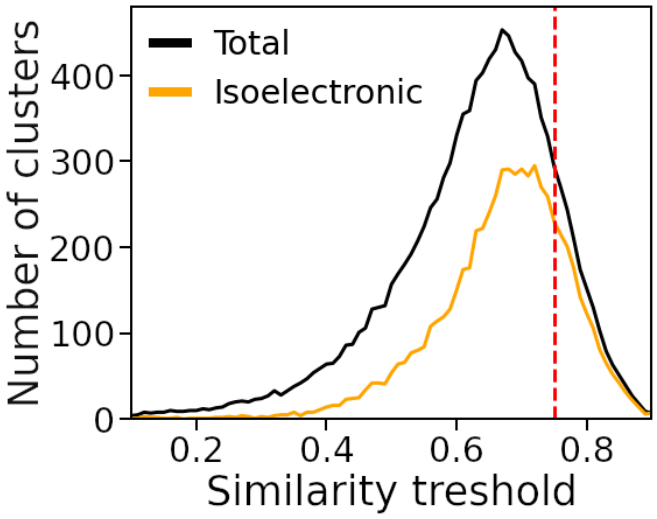


Figure 5.4: Relationship between the similarity threshold and the number of clusters. The total number of clusters is shown in black, isoelectronic clusters in orange. The clustering threshold that employed in the remainder of this section is indicated by the dashed red line. Figure adapted from Ref. 89.

Using the fingerprints as defined above, we computed the similarity matrix of all materials is the dataset. From this matrix, we obtain clusters with the algorithm introduced in Sec. 3.3.

Clustering process

We probe the impact of the only parameter of the clustering algorithm, *i.e.*, the similarity threshold, by repeating the clustering process for a wide range of values. Figure 5.4 shows the results of this analysis. For low threshold values, the total number of clusters is small, because all the materials are contained in few, very large clusters. As the threshold is increased, these clusters split, and the number of clusters increases monotonically. It reaches its maximum at $S_{\text{thres}} = 0.67$, which corresponds to an overlap of the areas below the DOS of around $\sim 50\%$[3]. We use a slightly larger threshold value of $S_{\text{thres}} = 0.75$, indicated by the dashed red

[3]The relationship between the Tanimoto coefficient and the area overlap is introduced in Sec. 3.2.2

line. This choice makes the clusters more compact: All cluster members have a similarity larger than S_{thres} to the cluster centroid, and all cluster members have a similarity of at least $S = 0.5$ to each other. Therefore, the overlap of the area under the DOSs between any two cluster members is at least 67% and $\sim$ 86% with the centroid. Our choice is scientifically meaningful, because the electronic structure around the Fermi level has a strong impact on the structural stability and dielectric response of a material and the reactivity of its surfaces. Thus, it plays an important role in, *e.g.*, catalysis processes and chemisorption.[123–125] Finding compact clusters with large overlap of the DOSs means that other properties of their members are likely to be similar.

Note that the threshold has a large impact on the results of the clustering process. Therefore, a threshold value has to be found that provides a balance between the desired characteristics of the clusters. Larger thresholds result in more compact clusters, but significantly reduce the number of clusters and their size. Therefore, interesting, chemically diverse clusters may not be found. Conversely, lowering the threshold leads to more clusters or larger clusters. At the same time, it allows cluster members to be less similar to each other, such that meaningful relations between cluster members may not be visible anymore.

Statistical analysis

Before focusing on the members of individual clusters, we investigate the distribution of clusters sizes. Additionally, we quantify the compactness of a cluster by its *radius*, as defined in Eq. 3.8. In total, we find 294 distinct clusters, containing $\sim$ 23% of the materials in the entire dataset.

Figure 5.5 shows a histogram of the sizes of these clusters, as well as the maximum and mean radii of all clusters of a given size. 68% of the clusters contain only two materials. For these clusters, due to the definition of the clustering algorithm, the cluster radius cannot exceed $r_c = 0.25$. However, the mean cluster radius is even smaller, which means that many clusters are even more compact. As the cluster size increases, the mean cluster radius increases to $r_c \sim 0.4$, which means that the overlap of the area under the DOSs is larger than 75%. The high-

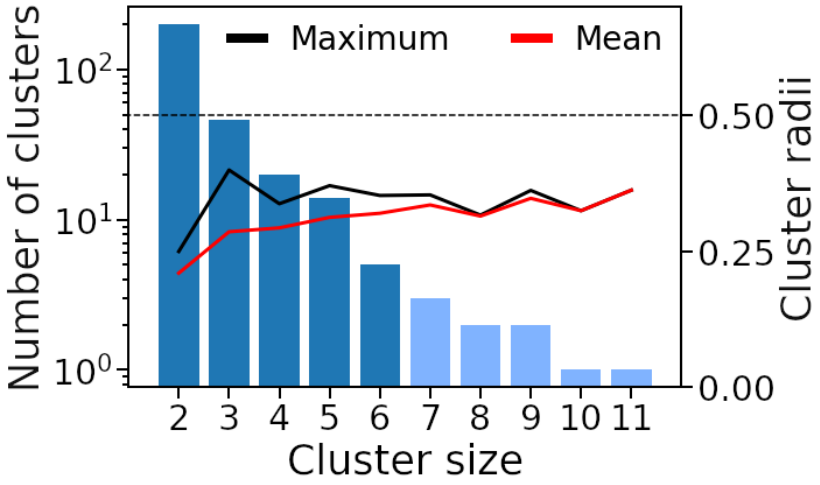


Figure 5.5: Distribution of cluster sizes (blue bars) and the respective maximum (black line) and mean (red line) cluster radii using a threshold of $S_{\text{thres}} = 0.75$. The dashed line represents the upper limit of the cluster radius for the given threshold. The light blue bars correspond to the clusters shown in Figure 5.6. Figure adapted from Ref. 89.

est cluster radius is found for clusters with three members. Still, its value does not surpass 0.4, despite its theoretical maximum being $r_c = 0.5$. Due to the low number of large clusters, the mean and maximum of the cluster radii approach the same value for larger cluster sizes.

Figure 5.6 shows a similarity matrix of the materials that are contained in the largest clusters. The clusters are indicated by red boxes. The matrix is sorted such that the largest cluster, *i.e.*, the cluster with 11 members (compare also Fig. 5.5), is located at low indices, shown in the bottom left of the matrix. Clusters shown at higher indices have a smaller number of members. Focusing on the matrix entries, which show the similarity color coded, we see that some clusters appear to be isolated, *i.e.*, the similarity of their members to all other materials in this matrix is low. This is especially pronounced for the cluster at the top right of the matrix. Others, like the largest and second largest cluster, show similarities that are almost as large as the similarity between their respective members. If a lower clustering threshold were used, these clusters would merge.

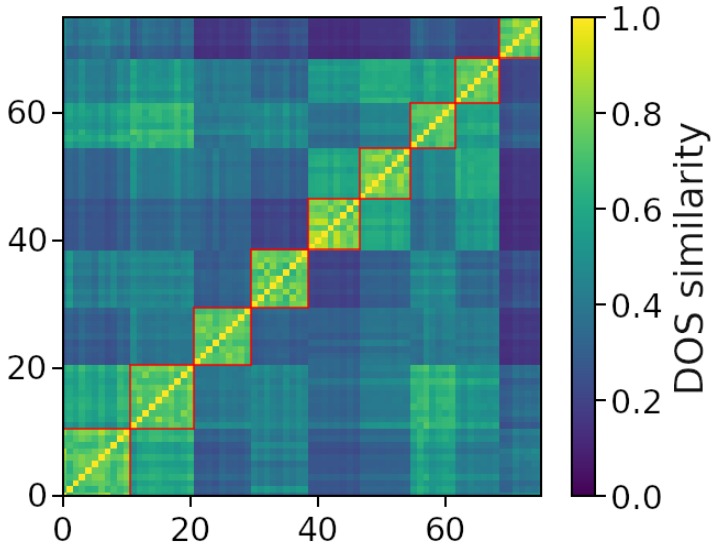


Figure 5.6: Similarity matrix of materials that are members of large clusters. The similarity between materials is color coded, according to the scale on the right. The clusters are highlighted by red boxes. All of them have more than six members; they are indicated by the light blue bars in Fig. 5.5. Figure adapted from.[89]

Orphans

About 77% of the materials are orphans, *i.e.*, they do not belong to any cluster. Out of these, there are 2643 materials with a similarity score lower than $S_{\text{thres}} = 0.75$ when compared to any other material in the dataset. We attribute the high number of orphans to the chemical diversity of the dataset. The remaining 54 orphans, approximately 2%, have at least one neighbor with a similarity score of $S \geq S_{\text{thres}}$, but this neighbor(s) belongs to another cluster that is either larger or more compact[4].

Among the orphans, there are six materials that are orphans even for very small similarity thresholds of $S_{\text{thres}} = 0.1$ (and naturally also higher values). In Fig. 5.7, we show their DOSs. Four of them, *i.e.*, $FeHfF_6$, $Li_2Cl_2O_4$, $FeZrCl_6$, and $Na_2F_2O_2$, are characterized by well-localized peaks in the DOS. These peaks most likely originate from the atomic structure of these materials. For example,

[4]This situation arises due to the definition of the clustering algorithm, see Sec.3.3 for details.

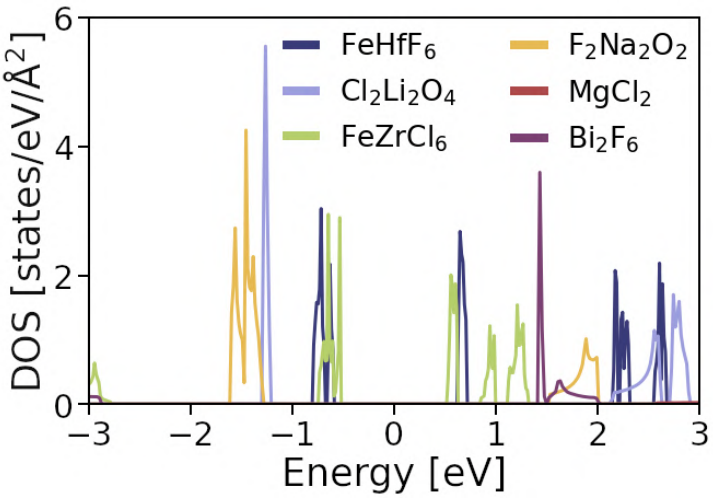


Figure 5.7: Electronic DOSs of materials that are orphans for all values of the clustering threshold in the range $0.1 \leq S_{\text{thres}} \leq 0.9$, as shown in Fig. 5.4. These materials show either narrow peaks in the DOS or very large band gaps. Figure adapted from.[89]

$Na_2F_2O_2$ is composed by a checkerboard pattern of FO molecules alternating with Na atoms[5]. The former give rise to sharp, almost molecular peaks of large height at -1.5 and 1.8, while the majority of Na states lies outside of the observed energy range. Due to these localized peaks, the overlap with the DOSs of other materials is low. The other two examples of orphans presented here, $MgCl_2$ and Bi_2F_6, have exceptionally large band gaps, such that no significant number of states is in the considered energy range. Thus, their similarity to other materials is low.

Isoelectronic compounds

Next, we focus on analyzing the members of the clusters that we found in the dataset. The first example is shown in Fig. 5.8, presenting the DOSs (a) and crystal structures (b) of five transition-metal dichalcogenides (TMDC), as well as their orbital-projected PDOSs (c). The radius of this cluster, with a value of $r_c = 0.28$, is close to the mean of clusters of this size (as presented in Fig. 5.5). Focusing first on the DOSs of the cluster members, it can be seen that their

[5]See also `https://cmrdb.fysik.dtu.dk/c2db/row/F2Na2O2-feda03610e19`.

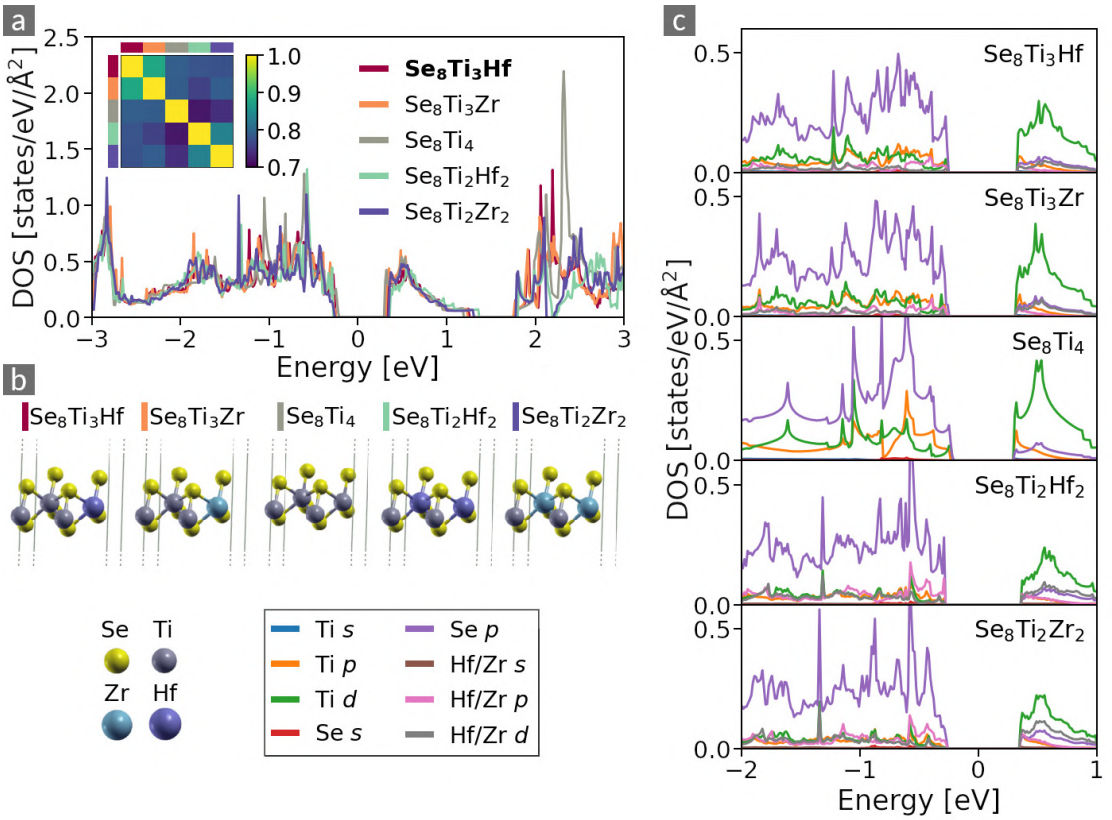


Figure 5.8: Overview of the electronic and atomic structures of materials in a selected cluster. The total DOSs (a) shows high overlap between all cluster members. The bold font in the legend indicates the cluster center, Se_8Ti_3Hf. The inset shows a similarity matrix of the cluster members, the color code is adjusted for better visibility. Their atomic structures (b) are similar up to substitutions with isoelectronic atoms. The corresponding PDOSs (c) indicate that despite slight variations depending on the composition, the DOSs of all materials sum up to a very similar total DOS. The Fermi level is located at $E = 0$ eV. Figure from Ref. 89.

PBE band gaps range from 0.52 eV ($TiSe_2$) to 0.65 eV ($Hf_2Ti_2Se_8$). Furthermore, the shape of the spectra is qualitatively very similar within the feature region $|E| \lesssim 2$eV. At higher energies, the DOSs become more dissimilar, which can be understood in terms of their coarser representation outside the feature region of the spectral fingerprint. At lower energies, *i.e.*, $E < -2$ eV, the DOSs of all cluster members show high overlap, although the fingerprint allows for larger deviations

in this region. The similarity matrix of the cluster members can be seen in the inset of panel (a), where the rows and columns are color coded according to the corresponding material. It can be seen that the highest similarity of $S \sim 0.9$ is found for $HfTi_3Se_8$ and $ZrTi_3Se_8$, and for $Hf_2Ti_2Se_8$ and $Hf_2Zr_2Se_8$, respectively.

The similarity of the cluster members is also rooted in their crystal lattice, shown in panel (b) of Fig. 5.8, which consists of a layer of transition metals (TM) between layers of Se in all cases. Thus, the cluster composition includes the binary phase $TiSe_2$ alongside ternary phases, characterized by the substitution of one or two Ti atoms with the respective number of either Hf or Zr atoms. The high similarity of the DOSs of all cluster members indicates the minor impact of these substitutions on the electronic structure. Considering the PTE, the reason for this observation becomes apparent, as these elements are all isoelectronic, *i.e.*, they are found in group 4 of the PTE.

This isoelectronic behavior can be further understood by considering the PDOSs of these materials, as shown in panel (c) of Fig. 5.8. It can be seen that the valence bands are dominated by fully occupied Se p states. For the conduction bands, the largest contribution comes from Ti d states. Additional contributions come from Zr and Hf d states when these elements are present.[126] Small numbers of Se-p states can be found in the conduction bands, as well as small amounts of TM d states in the valence region, indicating the hybridization of these orbitals. Overall, we can conclude that the substitution with isoelectronic elements does not alter the valence band in these materials, while the shape of the DOSs in the conduction bands, consisting of empty d states, is preserved due to the same amount of empty states for all elements. This conclusively explains the similarity of the electronic structure of the members of this cluster.

Given the combinatorial nature of the C2DB database, the dataset also contains other combinations of these elements with the same structural prototype, such as the binary compounds Se_8Hf_4 and Se_8Zr_4. These are contained in a separate cluster, which has members with band gaps ranging from 0.72 eV (Se_8TiHf_3) to 0.82 eV (Se_8HfZr_3). If a larger similarity threshold is chosen, these clusters merge.

We found that the vast majority of clusters in the dataset contain members

that differ only by such isoelectronic substitutions. To discover them efficiently, we make use of the PTE descriptor (see Sec. 3.4, Eq. 3.10) and filter the list of clusters accordingly: If the descriptor value $\overline{c}_m$ is identical for all cluster members, *i.e.*, the similarity according to the PTE fingerprint is 1, we conclude that the cluster is formed by isoelectronic materials, which we call *isoelectronic cluster* in the following. Our definition of isoelectronicity in this context focuses solely on electron count, without considering the specific electronic configurations. As such, an identical $\overline{c}_m$ can be constructed from either two Si atoms, two C atoms, or the combination of one Al and one P atom. Using this definition, we find that the dataset contains 230 clusters that have the same PTE descriptor for all of their members, corresponding to 78% of all clusters, making up 16.5% of all materials in the dataset. Furthermore, 88.8% of all clusters have at least two members with the same $\overline{c}_m$. As such, we identify isoelectronicity as the main reason for the formation of clusters in this dataset. Given the results in Sec. 5.1, where we have shown that the most similar material for GaAs is also isoelectronic, and in line with physical intuition, we expect these results to be more general.

We also study the influence of the similarity threshold on isoelectronic clusters. Their number is shown by the orange curve in Fig. 5.4 and can be compared with the total number of clusters, shown in black. It can be seen that isoelectronic clusters are observed over a wide range of similarity thresholds S_{thres}, and become dominant at $S_{\text{thres}} \approx 0.6$. As the threshold approaches 1, and the clusters become more compact, almost all clusters have exclusively isoelectronic members.

Isoelectronic surface groups

Filtering all isoelectronic clusters from the total list of clusters, we are able to identify the second most common origin of similarity of the electronic structure in the C2DB. These are exemplified in Fig. 5.9, which shows the DOSs (a) and atomic structures (b) of four metallic compounds. Their atomic structures consist of five alternating carbon and TM layers. The TMs are either Ta or Nb, which are isoelectronic and from group 5 of the PTE. The surface layers of material consist of either F atoms, or OH molecules.

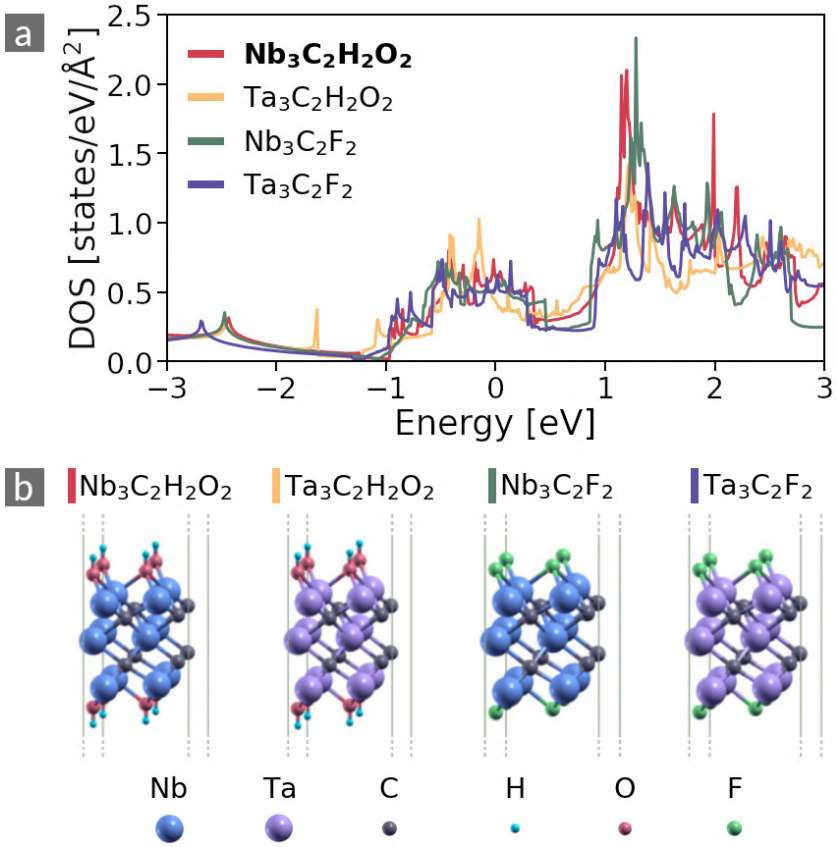


Figure 5.9: Materials with isoelectronic surface groups forming a cluster. Panel (a) shows the DOSs, the bold font in the legend indicates the cluster center. The Fermi level is at $E = 0$ eV. Panel (b) presents the corresponding atomic structures. The unit cells are repeated in the in-plane directions to increase visibility. Figure from Ref. 89.

Focusing on the electronic structure, we find that both compounds with F atoms at the surface have a characteristic valley in the DOS between 0.5 eV and 1 eV, which is not present in the compounds with OH surface groups. Other than that, the general shapes of the DOSs are qualitatively similar. The cluster radius, with a value of $r_c \sim 0.28$, is close to the mean of all clusters with four members.

We can further understand the similarity of the electronic structure by considering the PDOSs[6]. The DOS in the considered energy range is dominated by the TM d states, which hybridize with C and TM p states. In the conduction bands,

[6]See, *e.g.*, `https://cmrdb.fysik.dtu.dk/c2db/row/C2H2O2Nb3-137a187a149c`.

around $E \simeq 1.5$, there are also significant contributions from either O or F p states, depending on which surface atoms are present. That means, that the H-saturated O atom plays a similar role as the F atom, making the materials effectively isoelectronic. There are minor differences, such as the narrow peaks below -1 eV. These stem from multiple van Hove singularities, which are very sensitive to the exact location of band extrema and flat bands. Therefore, some discrepancies can be expected.

In the present dataset, we find 33 clusters containing materials where the surface is terminated by either F or OH groups. The majority of these cases also include materials that differ solely in their isoelectronic substitutions. It is important to recognize that while the isoelectronic behavior of fluorine atoms and hydroxyl groups is well known among domain experts of chemistry and electronic-structure theory, implementing this knowledge such that it can be used by automated systems, *e.g.*, search interfaces of databases, is not trivial. This is especially true given the increasing interdisciplinarity of scientific fields, and thus the broad spectrum of users of scientific data platforms with diverse scientific backgrounds. Therefore, domain-specific knowledge must be integrated into data models and represented in a comprehensive way. Currently, many search interfaces of materials databases, as well as ML descriptors, rely on structural features such as the chemical formula or the number of atoms in the unit cell. Features such as isoelectronicity, which was shown to be relevant above, are hardly represented, and are unlikely to be found as search criteria. This situation limits the *findability* (see Sec. 2.2.1) of materials data. Likewise, ML applications, not considering such aspects in descriptor design, may result in an increase of required model complexity to capture these similarities correctly.

Stacking patterns

The electronic properties of materials with the same chemical composition can be very different. As an example, we show the DOSs of three different phases of In_2S_2, which belongs to the material class of post-TMDCs, in Fig. 5.10. Panel (a) shows their DOSs, the corresponding crystal structures are depicted below in panel (b).

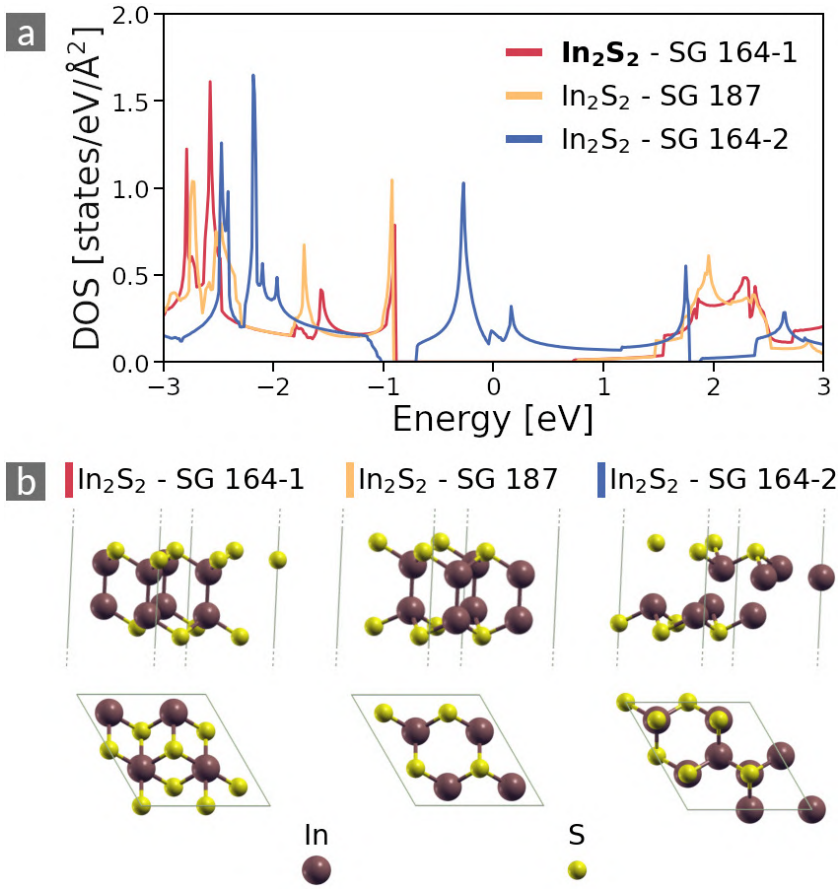


Figure 5.10: Electronic DOSs (a) and atomic structures (b) of different phases of In_2S_2. The first two structures, SG 164-1 and SG 187, form a DOS cluster. The third structure, SG 164-2, is not part of the cluster because of it's dissimilar DOS. The Fermi level is located at $E = 0$ eV, and to increase visibility, the unit cells of the atomic structures are repeated in both in-plane directions. Figure from Ref. 89.

For all three structural phases, the material consists of two layers of In between two layers of S. We categorize these materials additionally by the value of their SG descriptor (see Sec. 3.4). Two of these materials have SG 164, which we indicate as SG 164-1 and SG 164-2, respectively. The third material has SG 187. Despite the fact that two materials have the same SG, the stacking patterns of the In and S layers are different in all three materials. SG 164-1 can be classified as stacking pattern ABBC, SG 187 shows ABBA stacking, and for SG 164-2, it is ABDC.

Structures SG 164-1 and SG 187 have very similar DOSs, with a Tc of 0.76, thus they form a cluster that is found by our clustering algorithm. Both materials show a medium-sized band gap of 1.60 eV (SG 164-1) and 1.675 (SG 187)[7]. The shape and magnitude of their DOSs are qualitatively similar over the entire energy range considered. The third phase, SG 164-2, however, shows metallic behavior, and is therefore not part of the cluster. The corresponding Tc values w.r.t. the phases SG 164-1 and SG 187 are 0.32 and 0.34, respectively.

The similarity of the first two phases, and their dissimilarity to the third phase, can be understood in terms of the electronic configuration of the In and S atoms. In the (very similar) semiconducting phases, the In atoms are tetrahedrally coordinated, with three S atoms and one In atom as direct neighbors. They form covalent bonds, where the valence bands show the character of hybridized S and In p states.[127] The In-In bonds have a length of $d_{\mathrm{In-In}}$=2.82 Å. For the metallic phase, SG 164-2, the In atoms are coordinated with three other In atoms, with a bond length of $d_{\mathrm{In-In}}$=3.62 Å, similar to that of bulk metallic In (3.38 Å). It is found to be dynamically unstable, exhibiting metastability relative to the semiconducting phases[8]

In this example, the coordination of the In atoms plays a key role for the electronic properties of these materials. From the SG or the PTE descriptors, this cannot be understood, and only explicitly considering more advanced descriptors reveals their relationship. This means that the characterization of materials data based on simple descriptors is challenging. Note that the differences between these phases are also captured by the atomic structure fingerprint based on the SOAP descriptor, that we introduce in Sec. 3.4.

Outliers

The vast majority of the clusters found by our algorithm can be explained by the mechanisms described above. However, we found 25 clusters in the dataset with materials that have similar electronic structure, but are neither isoelectronic

[7] Values from `https://cmrdb.fysik.dtu.dk/c2db/row/In2S2-1b7899449ed6` and `https://cmrdb.fysik.dtu.dk/c2db/row/In2S2-172ef584c4a6`, respectively.

[8] See also `https://cmrdb.fysik.dtu.dk/c2db/row/In2S2-ef93efd2b5c0`

nor share the same crystal lattice. Figure 5.11 presents the members of one of

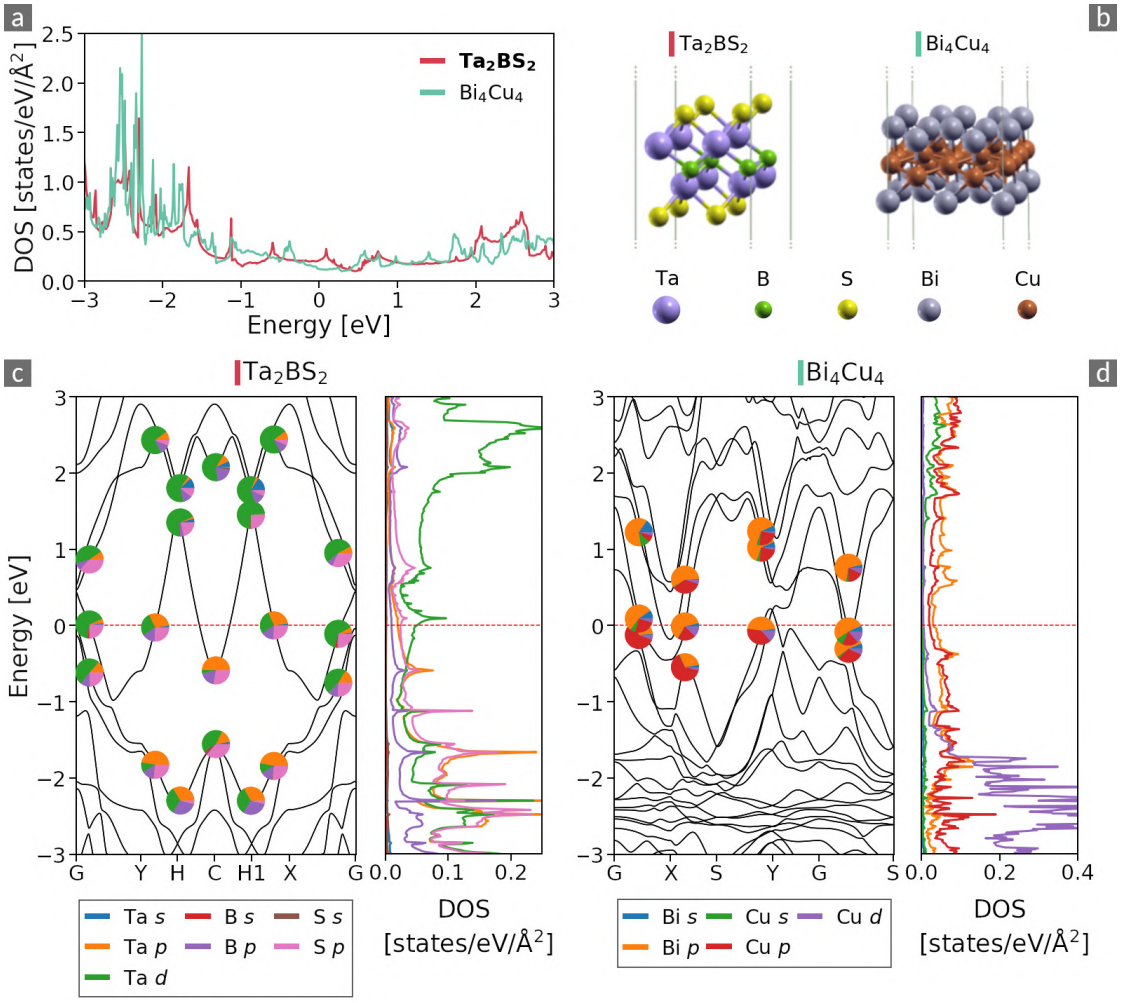


Figure 5.11: Electronic and atomic structures of Ta_2BS_2 and Bi_4Cu_4, materials with distinct atomic species and crystal structures, forming a cluster. In all panels, the Fermi level is located at $E = 0$ eV. Panel (a) shows the total DOSs, which are similar over the entire energy range considered. The atomic structures (b) of both materials are distinct, their unit cells are repeated in both in-plane directions to increase visibility. Panels (c) and (d) show the band structures with indicated atomic characters together with the PDOSs. Figure from Ref. 89.

these clusters. Focusing first on the total DOSs, shown in panel (a), we find that both materials are metallic and exhibit a nearly constant DOS over a wide energy range around the Fermi level in the interval -1.5 eV $\leq \varepsilon \leq 2$ eV. Below -1.5

eV, the DOS of both compounds increases in a similar way. Above 2 eV, the differences are larger. The similarity score of Tc = 0.76 is slightly higher than the clustering threshold. Structurally, the materials are very different. Their unit cells are shown in panel (b) of Fig. 5.11. Ta_2BS_2 has the trigonal SG 164, and Bi_4Cu_4 has an orthorhombic lattice with SG 51. Obviously, the similarity of their DOSs is not related to their atomic structure. In the lower panels of Fig. 5.11, we present the BS, with indicated atomic character of the bands, and the PDOS of both materials. In Ta_2BS_2, shown in panel (c), a single band crosses the Fermi level. The bands are predominately of p and d character, with contributions from all atomic species. The PDOS shows some hybridization of Ta d states with S p states at the Fermi level, while the conduction bands are dominated by Ta d states. In Bi_4Cu_4, shown in panel (d), several bands cross the Fermi level. Comparing the band character, reveals that the flat region in the DOS, -1.5 eV $\leq \varepsilon \leq 2$ eV, contains almost exclusively Bi and Cu p states. The unoccupied bands above 2 eV show additional Cu s character. In the lower valence bands ($\varepsilon < -1.5$), significant contributions of Cu d bands can be seen. None of these observations allow us to qualitatively understand the similarity of the DOSs, and we therefore conclude that it is in fact *accidental*. Thus, the present cluster can be understood as an outlier. We note that such kind of outliers are well known in molecular similarity.[85]

We want to highlight the analysis performed in this section as an efficient method for processing large numbers of clusters. This is achieved by iteratively filtering them through confirmatory analysis, *i.e.*, identifying the origin of similarity for a subset of clusters, constructing a descriptor to capture these cases, and removing clusters that can be fully explained by the descriptor. Using this process, we were able to efficiently identify outliers. These outliers are undoubtedly the scientifically most interesting cases, but require more in-depth analysis beyond the scope of this work. Automatizing the analysis presented here will allow to find such outliers in any dataset and may uncover unexpected connections between materials.

5.2.2 Comparing similarity measures

The examples shown above exclusively focus on the similarity of the electronic DOS using the spectral fingerprint. In the following, we use the same clustering algorithm, but three different fingerprint types (see Sec. 3.1.2), and compare the cluster assignments qualitatively. For this example, which follows our publication in Ref. 87, we have downloaded the crystal structures and electronic DOSs for a dataset of 3847 cubic perovskites, which stem from the AFLOW database and are also accessible through NOMAD. The dataset is downloaded using `MADAS`[9], and calculate PTE, DOS, and SOAP fingerprints (see Sec. 3.1.2 and 3.4) and the respective similarity matrices. We stress here again that the SOAP fingerprints do not distinguish between atomic species. From the similarity matrices, we obtain clusters using the algorithm presented in Sec. 3.3, using similarity thresholds of $S_{\text{thres}} = 0.75$ for both the DOS and SOAP matrices, and $S_{\text{thres}} = 1$ for the PTE matrix. In the latter case, all cluster members have identical PTE descriptors, *i.e.*, they are considered isoelectronic.

We then sort the similarity matrices based on the clusters that are found for each fingerprint type, and show the results on a 3×3 grid in Fig. 5.12. Each column in this figure shows similarity matrices based on the different fingerprint types, *i.e.*, from left to right, PTE, SOAP, and DOS fingerprints. The rows correspond to the sorting of the matrices, which is based on the clusters found in the PTE, SOAP, and DOS matrices, from top to bottom.

First, we focus on the diagonal panels in the figure, *i.e.*, the matrices that are sorted based on the clusters obtained from the same fingerprint type. The PTE matrix sorted by PTE clusters, shown in the top left panel, exhibits a large number of small clusters of similar size. This comes to no surprise, as the combintatorial HT approach used in AFLOW results in an almost even distribution of elements from the PTE. Note that, however, 1.814 of these materials contain oxygen, and 1920 contain fluorine.

Focusing on the similarity based on the atomic structure, *i.e.*, sorting the SOAP

[9]The respective code can be found at: `https://github.com/kubanmar/madas-examples/blob/master/notebooks/analyze_similarity_correlations.ipynb`.

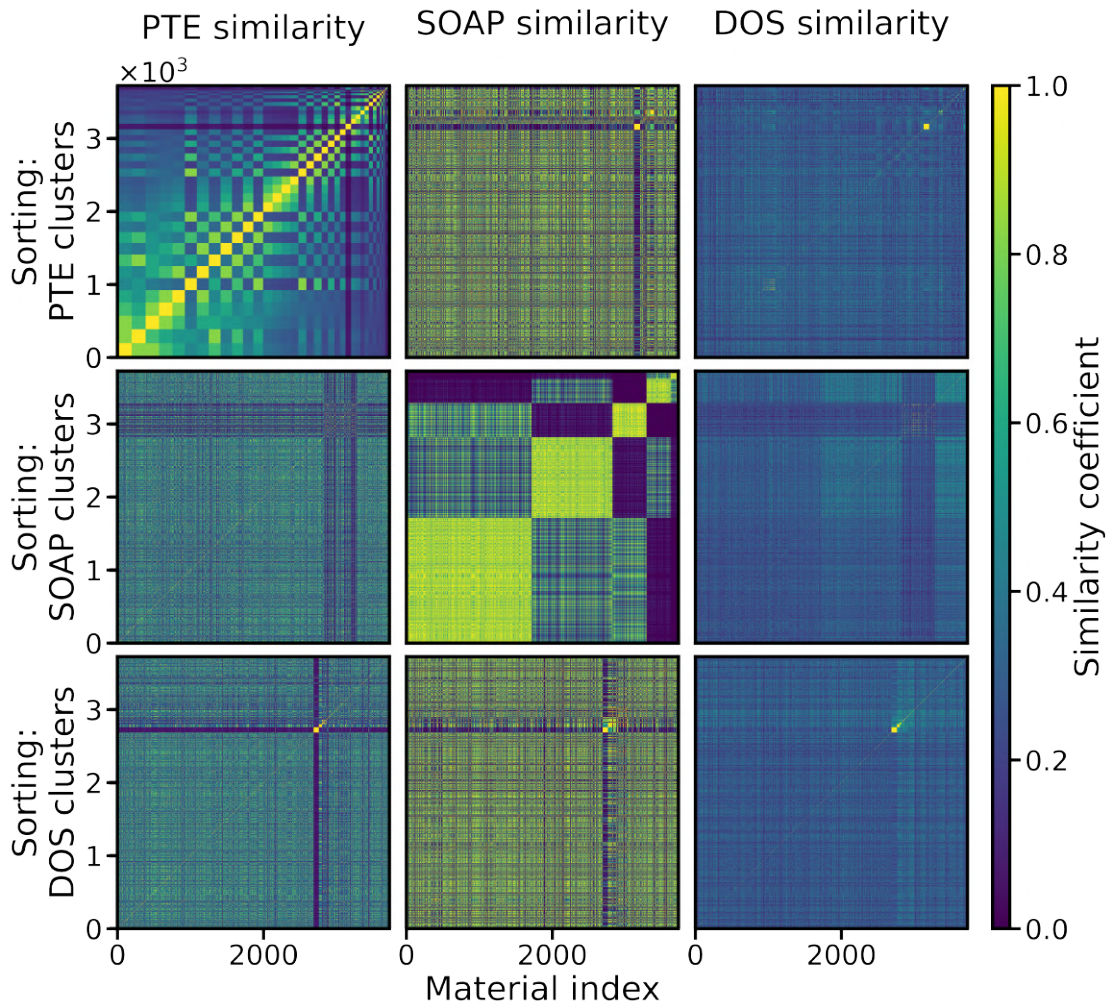


Figure 5.12: Similarity matrices of $\sim$ 3800 cubic perovskites, sorted by cluster assignments according to different similarity measures. The panels are arranged in a grid, where the columns correspond to PTE (left), SOAP (middle), and DOS (right) fingerprints. The rows of the grid indicate which similarity was used for clustering. The top, middle, and bottom rows correspond to sorting by PTE, SOAP, and DOS clusters, respectively. The similarity score is color coded, where dark blue represents lowest similarity ($S = 0$) and yellow indicates highest similarity ($S = 1$). Figure from Ref. 87.

matrix by SOAP clusters, as shown in the middle panel, reveals 10 clusters of different sizes. The largest one contains 1716 members and shows, on average, a rather high similarity to the members of the two next smaller clusters. The second

cluster is strictly dissimilar to the third, but shows higher similarity to the fourth cluster. The remaining, smaller clusters are rather isolated from the rest of the dataset. The descriptor we use for this fingerprint does not distinguish between atomic species, thus it is only sensitive to atomic positions and the unit cell. The current dataset consists exclusively of cubic perovskites, *i.e.*, all crystal lattice symmetries are identical and the atomic positions are similar. We therefore tend to attribute the cluster formation to the differences in cell volumes.

The last panel on the diagonal of the figure, the DOS matrix sorted by DOS clusters, reflects the distinct behavior of the electronic structure. The vast majority of materials, *i.e.*, those with indices < 3182, are less similar to any other material than $S_{\text{thres}} = 0.75$ and are therefore outliers. All clusters that are found are shown at higher matrix indices and are very compact, underlining the chemical diversity that can be found in this structurally rather homogeneous dataset.

Information about the relationship between different similarity measures can be detected in the off-diagonal panels of Fig. 5.12. Looking at the PTE similarity matrix sorted by DOS clusters in the bottom left panel, one feature is especially noticeable: The largest cluster, containing 75 materials (indices 3182 to 3257), can also be seen in the PTE matrix and the SOAP matrix (bottom center). Investigating the members of this cluster individually, we find that they are all calculations of $BPBa_3$[10]. This finding demonstrates the potential of a similarity-based approach for detecting duplicates and filtering databases. We note that structural similarity measures where already used by Valle and Oganov[68] to filter duplicate entries in a evolutionary algorithm for crystal structure generation. Our approach of combining different similarity measures to infer whether entries are duplicates extends this concept.

The DOS similarity matrix, when sorted by PTE clusters (top right panel), shows a slight correlation of the DOS descriptor with the PTE descriptor. This can be seen from the block-like pattern that emerges within the matrix. Even though the similarity of members of PTE clusters in the DOS matrix is small in most cases,

[10]These calculations are all contained individually in the AFLOW database. $BPBa_3$, with space group number 221, has 66 unique entries in the database at the time of writing of this manuscript, the remaining entries are likely discarded entries from prior iterations of AFLOW, but available in NOMAD.

statistically, the they are still more similar to each other than to other materials. This can be seen qualitatively by the lighter colors of the matrix entries close to the diagonal. This correlation is not surprising, because, as we demonstrated in the prior sections, materials with the same composition are more likely to have similar electronic structure. However, their similarity ultimately depends on many different factors, which are not captured by such a simple descriptor as the PTE.

When the DOS matrix is sorted by SOAP clusters, as shown in the middle panel on the right, no correlation between the largest SOAP cluster with the DOS can be found. However, the smaller clusters (index > 1717) qualitatively show a slightly increased DOS similarity among the members of SOAP clusters. This can be seen from the similar patterns in the center and middle right panel. Comparing the PTE and SOAP matrices sorted by SOAP and PTE clusters (top middle and center left panel, respectively), they do not appear to be correlated. Since they are agnostic of each other by construction, this was to be expected.

5.3 Summary

In this chapter, we have explored potential applications of similarity concepts in materials science, given the availability of interoperable data. The first application, similarity searches, allowed us to find materials that share desirable properties. We exemplified this on the prototypical semiconductor GaAs, finding the materials with the most similar electronic DOS among almost 1.9 million materials from the NOMAD Encyclopedia.[28] We found a material that was expected to be most similar, GaP, confirming previous experiments conducted by others,[78] but also finding unexpected compounds. When considering the distribution of similarity scores, we found that their mean is low and only few materials are highly similar to GaAs. This highlights the variety of different DOSs that are found among materials. Introducing a spin-resolved similarity measure, revealed an unexpected similarity between bulk Au and CoFe in their spin-majority channel.

We have also shown the impact of the feature region of spectral fingerprints on the results of similarity searches. In our example, the material that is most similar

to the reference did not change when focusing on either the valence or conduction bands, but the degree of similarity of other materials did depend on this choice. We expect larger differences for spectral properties spanning larger energy ranges. To confirm our analysis, we used spectral fingerprints with narrow feature regions and cutoffs to investigate in which energy regions the DOSs are most similar. The tool that we created for this purpose is available in `MADAS` and can be used to detect highly similar regions in the DOS in a quantitative, machine-readable way.

Similarity searches are, despite all optimizations, computationally expensive. One way to mitigate this problem is to compute the most similar materials for a reasonable set of fingerprint parameters *a priory*, such that they are available when researchers need them. However, this currently requires computing the full similarity matrix, which scales quadratically with the number of fingerprints. In the future, this may be optimized by using more efficient search technologies, such as k-d trees,[128] which partition the search space to allow finding the k most similar materials without comparing the reference to all other entries.

Alternatively, clustering can be used to find sets of similar materials. Here we used this approach for the DOSs of about 4000 2D materials from the C2DB.[8,53] We found that the majority of materials with similar electronic structure share the same lattice structure and differ only by isoelectronic substitutions. This finding is consistent with the results of the similarity search for GaAs mentioned above. This confirms that substitutions with isoelectronic elements are good candidates for obtaining materials with similar electronic structure. We also found that for this particular dataset, the most compact clusters contain exclusively isoelectronic materials, which shows that the differences in the electronic structure stemming from isoelectronic substitutions can be almost arbitrarily small.

In the future, this analysis can be extended and further automated. The current implementation uses `MADAS` and allows for efficient data handling, but many clusters still require quantitative analysis. To overcome this, implementing more (diverse) fingerprints may be beneficial: An example can be seen in the PTE fingerprint, which we have used to filter out isoelectronic clusters. Similarly, one may define fingerprints that can identify the similar behavior of F atoms and OH groups at surfaces. Such fingerprints can be used to filter data efficiently, which

allows to analyse even large datasets with many clusters in an efficient way.

Further potential for scientific discovery lies in the feature region of the spectral fingerprint. By setting the focus of the similarity analysis to a different energy range, for example the conduction or semi-core region, more similarities can be explored. Using `MADAS`, this analysis can be further automated, allowing to scan large datasets for outliers or unexpected phenomena.

By comparing the clustering results for different similarity measures we have shown how they relate to each other. We confirmed that clusters based on the PTE and the species-agnostic SOAP descriptors, *i.e.*, similarity measures that describe different material properties, show little correlation, except for those that are duplicate entries in the dataset. The similarity matrices that show correlations, *e.g.*, those comparing clustering results between DOS fingerprints and PTE or SOAP fingerprints, indicate that these fingerprints describe correlated properties. This type of analysis can be performed using `MADAS` and can assist in the development and performance benchmarking of novel descriptors and fingerprints.

6 Discussion

In this thesis, we addressed the challenge of performing similarity analysis by developing the computational framework `MADAS` that supports all steps of similarity analysis, including the collection and storage of data, the development and computation of fingerprints and similarity scores, and the seamless integration of data analysis and machine-learning methods. The software is written in a modern way, emphasizing modularity and usability, making it highly extensible. This, combined with the release of the source code under an open-source license, as well as the provision of extensive documentation, tutorials, and software tests, supports a long lifetime of our code.

Some vital parts of our software, however, cannot be tested, thus their functionality cannot be verified. This concerns, most drastically, the API interfaces (see Sec. 3.1.1) that connect data analysis pipelines to the APIs of public database providers. This means that data analysis pipelines written with `MADAS` are not guaranteed to run unless the data is provided with the analysis. The issue is mitigated by adding an abstraction layer into the software architecture, *i.e.*, the interaction of `API` classes with external databases is completely self-contained, and all data exchange with other components of `MADAS` is modeled through `Material` objects. Thereby, the impact of a potential mismatch of the data description between the external API and the API interface on the entire data pipeline is minimized. This design choice places the responsibility for ensuring compatibility with external APIs to the users of the framework. We believe that this choice is justified, because of the benefits mentioned above, and because the workload of maintaining the source code of API interfaces to multiple large databases is not feasible for a single developer.

We have also developed a spectral fingerprint that allows to encode any spectral

property, *e.g.*, the electronic DOS of materials, into a binary-valued raster-image. This was achieved by applying a non-uniform transformation to the spectrum, which allows to set the focus of the fingerprint on a user-specified range of values of the independent variable, *e.g.*, the energy. Using the Tanimoto coefficient Tc as the similarity score, this fingerprint can be used for a range of use-cases. It requires, however, some parameters to be chosen that affect, *e.g.*, the coarseness of the discretization of the spectrum. These can have a non-negligible effect on the results that can be obtained with subsequent analysis. To ensure the correctness of this analysis, manual testing and optimization of the fingerprint parameters is currently required.

Similarly, the quality of the spectra used as input for the fingerprint can affect the results. For example, using a non-smooth DOS can result in artificially reduced similarity between two spectra, because of the low overlap of numerical artifacts. This effect is minimized through the integration step in the computation of the fingerprint (see Sec. 3.2.1), but cannot be completely avoided. The extent of these effects, and how to reduce them, will be studied elsewhere. To this point, all results obtained with the spectral fingerprint were checked carefully for consistency and reproducibility.

Binary-valued fingerprints based on the atomic structure of molecules are well established in medicinal chemistry and drug design. Their success is based on the fact that, in many cases, the biological activity of a molecule correlates with specific structural features (see Sec. 2.5). The situation is different for materials. Here, the (local) atomic structure, while influencing it, does not solely dictate the electronic structure and therefore many derived properties. Instead, the electronic structure reflects the complex non-local many-body interactions of electrons in extended systems. Thus, measuring the similarity of materials based on the local atomic structure is not very promising. Therefore, novel, advanced fingerprints are required to capture these effects. The spectral fingerprint used here can serve as a guideline for future developments.

While the development of novel fingerprints can help to capture material characteristics, it is worth noting that, in many cases, the definition of similarity can vary significantly between different fingerprints, depending on the context and purpose.

For example, similarity may be well defined only for highly similar compounds, as observed in molecular similarity studies.[85] However, when dealing with less similar compounds, comparisons become challenging. In molecular similarity, so-called *activity cliffs*[79] are reported. They refer to the observation, that the correlation between the similarity score and a material property can drop drastically when the similarity score falls below a specific value. This highlights the subjective nature of similarity and the critical importance of defining different measures for different purposes.

Based on the spectral fingerprint, we performed a clustering of materials from the C2DB database. For this purpose, we developed a custom algorithm that fulfils our requirement of finding compact clusters. There are many alternatives published in the scientific literature, that have been applied to, *e.g.*, the selection of diverse subsets from molecular databases.[129] To this point, we have not systematically studied the impact of the clustering algorithm on the results, which we leave for future research. Another aspect that influences the results obtained with our method is a well-known size dependency of the Tanimoto coefficient.[85,130] In general, fingerprints with high filling factors[1] tend to systematically reach higher similarity scores. This results in large clusters with large cluster radii, containing many materials with overlapping spectral shapes, but smaller features *do not* overlap. Despite the high similarity scores, these materials are not very similar. The consequences of this effect may be mitigated, *e.g.*, by clustering the members of a large clusters with a higher clustering threshold. Using `MADAS` as a computational framework, these tasks can be addressed efficiently.

[1] We define the filling factor of a binary-valued fingerprint as the number of components of the fingerprint that are 1, divided by the total length of the fingerprint vector.

7 Conclusions

Understanding whether and under what conditions materials are similar to each other is a pressing issue in materials science. To find substitute materials, *e.g.*, to replace toxic or rare elements, similarity has been evaluated only qualitatively, based on the knowledge of domain experts. In this thesis, we showed that, by quantifying similarities using material fingerprints, the rules that these expert used can be written out explicitly, allowing to test large amounts of data in an automated way. We applied this methodology to address two major challenges in materials discovery, *i.e.*, assessing data quality and exploring large data spaces.

It is well known in the data-driven materials-science community that variety and veracity of data impose significant limits on the knowledge that can be extracted from large databases. The physical and numerical approximations that are required to compute material properties with a given accuracy, *i.e.*, their deviation from exact results, depend on the material itself. Thus, in order to obtain accurate results when extending the chemical diversity of the data, the diversity of the data in terms of employed methods has to increase as well. Using a consistent set of parameters and approximations for diverse data, conversely, reduces the accuracy and precision for those materials, that require higher parameter settings or levels of theory. Here, we showed how similarity measures can be used address this challenge.

Focusing on the electronic structure, we first demonstrated that differences between computed material properties that arise due to various approximations, can be captured using spectral fingerprints. This was achieved by first measuring the similarity of the electronic DOS over a large interval around the Fermi energy, and then restricting the fingerprint individually to the valence and conduction bands. By comparing the similarity scores in the different energy regions, we could iden-

tify where the different approaches that were used to obtain the DOS affected the electronic structure.

We then applied our method to optical absorption spectra. Using theoretical data obtained with the BSE formalism, we showed that the convergence of the absorption spectra of h-BN with respect to the number of $\boldsymbol{k}$-points can be well captured by spectral fingerprints. Specifically, as the number of $\boldsymbol{k}$-points is increased, the similarity of consecutive calculations increases monotonically. This shows how similarity measures can be used to describe the convergence of critical numerical parameters. We also compared absorption spectra obtained by different experiments. Using bulk silver as an example, we were able to quantify the significant differences between these measurements with our fingerprints. We emphasized that the differences between these data cannot be understood without more information about the samples and the conditions under which the experiment was performed.

Having shown that our methods are suitable for capturing the relevant effects to characterize data quality, we computed similarity matrices for larger datasets. By sorting the similarity matrix of 144 calculations of h-BN with respect to the numerical settings used, we could explicitly show the impact of the most relevant parameters on the DOS, notably the number of $\boldsymbol{k}$-points and the number of basis functions. This allows to directly select subsets of calculations which, despite employing different parameters, obtain the very similar results. By sorting of the similarity matrix, the relationship between calculations based on different numerical settings can be made visible.

By sorting the similarity matrix by the similarity of each entry to the rest of the dataset, we showed that this mean similarity can serve as a measure to identify sufficiently converged calculations. We showed this by relating clusters of similar calculations, which formed based on the sorting, to the relevant convergence parameters. The analysis we used for this example can be easily repeated for other materials, potentially allowing to automatically find and classify calculations with converged parameters. Since this approach only depends on the similarity scores, it can be applied even if the values of important convergence parameters are not known.

We then showed how the concept of similarity can be used to explore and learn from large datasets. Similar to what is known from molecular similarity searches in drug design, similarity searches can be performed for materials data. We demonstrated this by finding the most similar materials of GaAs from a large dataset of $\sim$ 1.9 million materials. Using this material as an example, we showed that, besides the large amount of data used in our search, highly similar materials are often exceptional. Finally, we computed the most similar materials for all members of the large dataset and made them available through the NOMAD Encyclopedia.

We then investigated the effect of the feature region of the spectral fingerprint on the results of similarity searches. We showed that setting the feature region to the valence or conduction bands individually changes which materials are found in a similarity search. To make these results more interpretable, we developed a tool to compute the similarity of spectra as a function of their independent variable. This tool can be used to support qualitative comparisons of spectra through quantitative analysis. Furthermore, it can be used to automatically discover the energy range in which two materials are highly similar.

Clustering, applied to datasets of interoperable calculations, can be used to analyze the similarities between many all materials at once. We demonstrated this for a set of 2D materials stemming from the C2DB. Starting from a detailed analysis of the clustering process, we selected a suitable threshold for our clustering algorithm. The latter is designed to obtain compact clusters. We analyzed the clusters that were found, focusing on understanding the physics behind the similarity of the electronic structure. We found that the vast majority of clusters contain materials that differ only by isoelectronic substitutions. After classifying these, we were able to discover the next most frequent mechanism, *i.e.*, the substitution of surface atoms with isoelectronic surface groups (or *vice versa*). We furthermore highlighted that the impact of structural patterns in materials on the electronic structure can be large, and that these differences are not well captured by typical search interfaces of materials databases such as the space group or the composition. Finally, we were able to discover outliers, which could not be explained by any of the previously discussed mechanisms. These cases can serve as a starting point for further investigations, allowing for the discovery of unexpected

connections between otherwise distinct materials. Furthermore, our iterative approach to identify the origin of the similarity in these clusters presents itself as a novel method to incorporate physical reasoning into material databases. This would allow to annotate existing data, transforming material databases from pure databases to knowledge bases.

Finally, we demonstrated the correlation between different descriptors by finding clusters in similarity matrices computed with different fingerprints. This type of analysis can be used for the development of novel fingerprints, or to study the correlation of existing ones with different material properties.

To conclude, in this thesis, we discussed the key methodologies that are necessary to perform similarity-based characterization of materials, and the analysis of materials data in general. Starting from existing concepts and a background in DFT, we established workflows for analyzing materials data. They can be applied to datasets ranging in size from two to millions of calculations. We would like to emphasize that all developed methods are available as open-source software, which greatly improves their reproducibility and reusability. They are complemented with extensive documentation and tutorials, allowing others to quickly adopt similarity-based methods for their research.

8 Outlook

Our work paves the way for future research, offering a comprehensive foundation for advancements and novel developments of similarity-based analysis in materials science. Despite the existence of large datasets computed with consistent settings, the total number of consistent datasets is rather small. Thus, data quality control and the compilation of interoperable datasets are pressing challenges. The development of `MADAS` will therefore focus on this aspect. An important step in this direction will be the design of autonomous workflows for data analytics. These will allow to handle large amounts of data, to find precise calculations that are representative of a specific material, and to discover outliers. A very first step can be done by applying the workflows presented in Sec. 4.2 to other, already available datasets. Comparing the results from these analyses of different materials can help to better understand the convergence of DFT calculations with respect to the relevant parameters.

Several aspects of the methodology presented in this thesis can be further automated. One aspect concerns the choice of the optimal parameterization of the material fingerprints, *i.e.*, the number of pixels or the minimum integration interval used. This would allow to, *e.g.*, select the smallest number of pixels that still represent the spectrum well, effectively reducing the memory footprint of the descriptor. Doing so, will also increase the applicability of the descriptor for non-expert users and therefore the reproducibility of scientific results. To find these parameters, their influence on the similarity of materials can be studied, *e.g.*, by computing and comparing similarity matrices for given benchmark datasets. To support this analysis, the spectral fingerprint already automatically computes several metrics, describing, *e.g.*, the amount of fingerprint components that are "1", or the number of states in the histogram bins that lie outside of the grid.

For the analysis of clusters, as presented in Sec. 5.2, several improvements can be made. One promising concept is the automated selection of clustering thresholds from the distribution of similarity scores within a dataset. In molecular similarity, the use of statistical tests for this purpose has been reported.[79] Using well-defined criteria could improve the comparability of clusters obtained from different fingerprints or with different similarity metrics.

Another interesting task is to find novel descriptors that allow for interpreting the results of clustering. In Sec. 5.2.1, we have described how the discovered clusters can be iteratively filtered to find all sets of clusters that can be explained with certain descriptors. Alternatively, such descriptors can be found using ML techniques. One option would be the usage of symbolic regression methods, such as SISSO.[72] Alternatively, techniques like inductive logic programming have been proposed[85] to generate rules that can be used to understand the formation of clusters. Given sufficiently large datasets, rules may be found that explain the unexpected similarity between cluster members.

To increase the range of applicability, our methodology can be extended in different directions. Clearly, the number of available fingerprints and similarity metrics can be increased. Promising candidates for additional similarity metrics are, *e.g.*, the Wasserstein distance[131] or cross-correlation functions. Furthermore, asymmetric similarity scores, such as the Tversky[88] coefficient, can be used for similarity searches. They allow to put a focus either on unique descriptor features of the reference or of the candidate compounds.[79] The performance of the code that computes spectral fingerprints and the similarity scores between them can be increased by implementing CPU intensive operations in a more efficient programming language, such as C, Rust, or Fortran.

Another interesting concept that can be borrowed from molecular similarity searching is *data fusion*.[83,84,86] Here, the similarity score between different molecules is computed using multiple similarity measures, either by combining the descriptor values, or by using, *e.g.*, weighted averages of multiple similarity scores as a new score. This is found to increase the effectiveness of similarity searches.[83] Within `MADAS`, these approaches can be tested by, *e.g.*, adding and normalizing similarity matrices, or by introducing new fingerprint types.

Finally, the work described here, also opens the door for completely new applications. One options is to directly infer material properties from similarity matrices and the properties of known compounds, *i.e.*, using supervised machine learning. This can be achieved by using existing implementations of kernel-based machine-learning algorithms. In preliminary tests, we found that these methods do not reach the state-of-the-art performance of, *e.g.*, deep learning methods. However, relevant applications can still be found, *e.g.*, for datasets that are too small for other methods, or when similarity matrices have already been computed for a different purpose and can be reused. To improve the accuracy of predictions from similarity-based supervised ML, the concept of data fusion, as described above, may be used.

The examples presented in this work illustrate that measuring the similarity of materials enables a wide variety of applications that help to address several pressing challenges of materials science. The methods that we have implemented can be used by others and adapted to fit their needs. Applied carefully, they allow researchers to better understand the data they are using and therefore improve the reliability of results obtained with statistical methods. Doing so, is especially challenging as ML methods are entering a wide range of scientific fields, requiring experts from both ML and domain sciences to understand the details of the respective other field. Similarity measures, bridging qualitative and quantitative analysis of scientific results, present an efficient way of approaching this challenge.

Acknowledgements

First and foremost, I would like to thank Prof. Claudia Draxl for supervising my PhD thesis. Her scientific vision laid the foundation for my work and her guidance throughout the whole project enabled me to complete it. I am very grateful for all the feedback, mentoring, and opportunities that I received, which allowed me to develop my skills as a scientist.

I thank my mentor Dr. Santiago Rigamonti. His expertise, rich knowledge, and critical thinking not only allowed me to understand the many details of theoretical materials science, but also helped me to discover and strengthen weak points of my own work. He was always available to discuss the even the smallest theoretical and computational details.

Special thanks goes to my co-authors for their work on the articles that we wrote together, and all past and current members of the SOL group for maintaining a welcoming and rich work environment.

Besonderer Dank geht an meine Familie, die mich in jeglichener Hinsicht unterstützt hat und mich immer ermutigt hat neugierig zu sein und meinen eigenen Weg zu gehen. Ohne euch wäre nichts von all dem möglich gewesen.

Furthermore I would like say *muito obrigado* to Matheus, my brother from another mother, *grazie mille* to Cecilia, who always was a great friend and colleague, and *merci beaucoup* to Ninon, who became my friend when I needed one the most.

非常感谢, 可迪.

Thank you for reading.

Bibliography

[1] `https://mgi.gov/`.

[2] Stefano Curtarolo, Gus L. W. Hart, Marco Buongiorno Nardelli, Natalio Mingo, Stefano Sanvito, and Ohad Levy. The high-throughput highway to computational materials design. *Nature Materials*, 12(3):191–201, Mar 2013.

[3] Isao Tanaka, Krishna Rajan, and Christopher Wolverton. Data-centric science for materials innovation. *MRS Bulletin*, 43(9):659–663, 2018.

[4] Claudia Draxl and Matthias Scheffler. Big data-driven materials science and its fair data infrastructure. *Handbook of Materials Modeling: Methods: Theory and Modeling*, pages 49–73, 2020.

[5] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C. Wolverton. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM*, 65(11):1501–1509, Nov 2013.

[6] Stefano Curtarolo, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kesong Yang, Richard H. Taylor, Lance J. Nelson, Gus L.W. Hart, Stefano Sanvito, Marco Buongiorno-Nardelli, Natalio Mingo, and Ohad Levy. Aflowlib.org: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science*, 58:227–235, 2012.

[7] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

[8] Sten Haastrup, Mikkel Strange, Mohnish Pandey, Thorsten Deilmann, Per S Schmidt, Nicki F Hinsche, Morten N Gjerding, Daniele Torelli, Peter M Larsen, Anders C Riis-Jensen, Jakob Gath, Karsten W Jacobsen, Jens Jørgen Mortensen, Thomas Olsen, and Kristian S Thygesen. The computational 2d materials database: high-throughput modeling and discovery of atomically thin crystals. *2D Materials*, 5(4):042002, sep 2018.

[9] Claudia Draxl and Matthias Scheffler. Nomad: The fair concept for big data-driven materials science. *MRS Bulletin*, 43(9):676–682, sep 2018.

[10] Christian Carbogno, Kristian Sommer Thygesen, Björn Bieniek, Claudia Draxl, Luca M. Ghiringhelli, Andris Gulans, Oliver T. Hofmann, Karsten W. Jacobsen, Sven Lubeck, Jens Jørgen Mortensen, Mikkel Strange, Elisabeth Wruss, and Matthias Scheffler. Numerical quality control for dft-based materials databases. *npj Computational Materials*, 8(1), apr 2022.

[11] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[12] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.

[13] John P. Perdew and Yue Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Physical Review B*, 45(23):13244–13249, June 1992.

[14] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868, Oct 1996.

[15] Cecilia Vona, Dmitrii Nabok, and Claudia Draxl. Electronic Structure of (Organic-)Inorganic Metal Halide Perovskites: The Dilemma of Choosing the Right Functional. *Advanced Theory and Simulations*, 5(1):2100496, 2022.

[16] Kurt Lejaeghere, Gustav Bihlmayer, Torbjörn Björkman, Peter Blaha, Stefan Blügel, Volker Blum, Damien Caliste, Ivano E. Castelli, Stewart J. Clark, Andrea Dal Corso, Stefano de Gironcoli, Thierry Deutsch, John Kay Dewhurst, Igor Di Marco, Claudia Draxl, Marcin Dułak, Olle Eriksson, José A. Flores-Livas, Kevin F. Garrity, Luigi Genovese, Paolo Giannozzi, Matteo Gi-

antomassi, Stefan Goedecker, Xavier Gonze, Oscar Grånäs, E. K. U. Gross, Andris Gulans, François Gygi, D. R. Hamann, Phil J. Hasnip, N. A. W. Holzwarth, Diana Iuşan, Dominik B. Jochym, François Jollet, Daniel Jones, Georg Kresse, Klaus Koepernik, Emine Küçükbenli, Yaroslav O. Kvashnin, Inka L. M. Locht, Sven Lubeck, Martijn Marsman, Nicola Marzari, Ulrike Nitzsche, Lars Nordström, Taisuke Ozaki, Lorenzo Paulatto, Chris J. Pickard, Ward Poelmans, Matt I. J. Probert, Keith Refson, Manuel Richter, Gian-Marco Rignanese, Santanu Saha, Matthias Scheffler, Martin Schlipf, Karlheinz Schwarz, Sangeeta Sharma, Francesca Tavazza, Patrik Thunström, Alexandre Tkatchenko, Marc Torrent, David Vanderbilt, Michiel J. van Setten, Veronique Van Speybroeck, John M. Wills, Jonathan R. Yates, Guo-Xu Zhang, and Stefaan Cottenier. Reproducibility in density functional theory calculations of solids. *Science*, 351(6280):aad3000, 2016.

[17] Charles Kittel. *Introduction to solid state physics*. John Wiley & Sons, 8th edition, 2005.

[18] Francis Birch. Finite elastic strain of cubic crystals. *Phys. Rev.*, 71:809–824, Jun 1947.

[19] G Kresse and J Hafner. Norm-conserving and ultrasoft pseudopotentials for first-row and transition elements. *Journal of Physics: Condensed Matter*, 6(40):8245–8257, October 1994.

[20] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54:11169–11186, Oct 1996.

[21] Gianluca Prandini, Antimo Marrazzo, Ivano E. Castelli, Nicolas Mounet, and Nicola Marzari. Precision and efficiency in solid-state pseudopotential calculations. *npj Computational Materials*, 4(1):72, Dec 2018.

[22] Volker Blum, Ralf Gehrke, Felix Hanke, Paula Havu, Ville Havu, Xinguo Ren, Karsten Reuter, and Matthias Scheffler. Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications*, 180(11):2175–2196, 2009.

[23] Sebastian Kokott, Florian Merz, Yi Yao, Christian Carbogno, Mariana Rossi, Ville Havu, Markus Rampp, Matthias Scheffler, and Volker Blum. Efficient all-electron hybrid density functionals for atomistic simulations beyond 10,000 atoms. *The Journal of Chemical Physics*, 161(2):024112, 07 2024.

[24] Andris Gulans, Stefan Kontur, Christian Meisenbichler, Dimitrii Nabok, Pasquale Pavone, Santiago Rigamonti, Stephan Sagmeister, Ute Werner, and Claudia Draxl. **exciting**: a full-potential all-electron package implementing density-functional theory and many-boyd perturbation theory. *J. Phys.: Condens. Matter*, 26(36):363202, 2014.

[25] Andris Gulans, Anton Kozhevnikov, and Claudia Draxl. Microhartree precision in density functional theory calculations. *Phys. Rev. B*, 97:161105, Apr 2018.

[26] Jonathan Schmidt, Mário R. G. Marques, Silvana Botti, and Miguel A. L. Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1), August 2019.

[27] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), mar 2016.

[28] Claudia Draxl and Matthias Scheffler. The NOMAD laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials*, 2(3):036001,

may 2019.

[29] Casper W. Andersen, Rickard Armiento, Evgeny Blokhin, Gareth J. Conduit, Shyam Dwaraknath, Matthew L. Evans, Ádám Fekete, Abhijith Gopakumar, Saulius Gražulis, Andrius Merkys, Fawzi Mohamed, Corey Oses, Giovanni Pizzi, Gian-Marco Rignanese, Markus Scheidgen, Leopold Talirz, Cormac Toher, Donald Winston, Rossella Aversa, Kamal Choudhary, Pauline Colinet, Stefano Curtarolo, Davide Di Stefano, Claudia Draxl, Suleyman Er, Marco Esters, Marco Fornari, Matteo Giantomassi, Marco Govoni, Geoffroy Hautier, Vinay Hegde, Matthew K. Horton, Patrick Huck, Georg Huhs, Jens Hummelshøj, Ankit Kariryaa, Boris Kozinsky, Snehal Kumbhar, Mohan Liu, Nicola Marzari, Andrew J. Morris, Arash A. Mostofi, Kristin A. Persson, Guido Petretto, Thomas Purcell, Francesco Ricci, Frisco Rose, Matthias Scheffler, Daniel Speckhard, Martin Uhrin, Antanas Vaitkus, Pierre Villars, David Waroquiers, Chris Wolverton, Michael Wu, and Xiaoyu Yang. Optimade, an api for exchanging materials data. *Scientific Data*, 8(1):217, Aug 2021.

[30] Matthew L. Evans, Johan Bergsma, Andrius Merkys, Casper W. Andersen, Oskar B. Andersson, Daniel Beltrán, Evgeny Blokhin, Tara M. Boland, Rubén Castañeda Balderas, Kamal Choudhary, Alberto Díaz Díaz, Rodrigo Domínguez García, Hagen Eckert, Kristjan Eimre, María Elena Fuentes Montero, Adam M. Krajewski, Jens Jørgen Mortensen, José Manuel Nápoles Duarte, Jacob Pietryga, Ji Qi, Felipe de Jesús Trejo Carrillo, Antanas Vaitkus, Jusong Yu, Adam Zettel, Pedro Baptista de Castro, Johan Carlsson, Tiago F. T. Cerqueira, Simon Divilov, Hamidreza Hajiyani, Felix Hanke, Kevin Jose, Corey Oses, Janosh Riebesell, Jonathan Schmidt, Donald Winston, Christen Xie, Xiaoyu Yang, Sara Bonella, Silvana Botti, Stefano Curtarolo, Claudia Draxl, Luis Edmundo Fuentes Cobas, Adam Hospital, Zi-Kui Liu, Miguel A. L. Marques, Nicola Marzari, Andrew J. Morris, Shyue Ping Ong, Modesto Orozco, Kristin A. Persson, Kristian S. Thygesen, Chris Wolverton, Markus Scheidgen, Cormac Toher, Gareth J. Conduit, Giovanni Pizzi, Saulius Gražulis, Gian-Marco Rignanese, and Rickard Armiento. Developments and applications of the optimade api for materials discovery,

design, and data exchange. *Digital Discovery*, 3:1509–1533, 2024.

[31] Luca M Ghiringhelli, Carsten Baldauf, Tristan Bereau, Sandor Brockhauser, Christian Carbogno, Javad Chamanara, Stefano Cozzini, Stefano Curtarolo, Claudia Draxl, Shyam Dwaraknath, et al. Shared metadata for data-centric materials science. *Scientific Data*, 10(1):626, 2023.

[32] `https://doi.org/10.17487/RFC8259`.

[33] Daniel Speckhard, Tim Bechtel, Luca M. Ghiringhelli, Martin Kuban, Santiago Rigamonti, and Claudia Draxl. How big is big data? *Faraday Discuss.*, pages –, 2024.

[34] Vinay I. Hegde, Christopher K. H. Borg, Zachary del Rosario, Yoolhee Kim, Maxwell Hutchinson, Erin Antono, Julia Ling, Paul Saxe, James E. Saal, and Bryce Meredig. Quantifying uncertainty in high-throughput density functional theory: A comparison of AFLOW, Materials Project, and OQMD. *Phys. Rev. Mater.*, 7:053805, May 2023.

[35] Stefano Curtarolo, Wahyu Setyawan, Gus L.W. Hart, Michal Jahnatek, Roman V. Chepulskii, Richard H. Taylor, Shidong Wang, Junkai Xue, Kesong Yang, Ohad Levy, Michael J. Mehl, Harold T. Stokes, Denis O. Demchenko, and Dane Morgan. Aflow: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226, 2012.

[36] Sebastiaan P. Huber, Spyros Zoupanos, Martin Uhrin, Leopold Talirz, Leonid Kahle, Rico Häuselmann, Dominik Gresch, Tiziano Müller, Aliaksandr V. Yakutovich, Casper W. Andersen, Francisco F. Ramirez, Carl S. Adorf, Fernando Gargiulo, Snehal Kumbhar, Elsa Passaro, Conrad Johnston, Andrius Merkys, Andrea Cepellotti, Nicolas Mounet, Nicola Marzari, Boris Kozinsky, and Giovanni Pizzi. Aiida 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Scientific Data*, 7(1), sep 2020.

[37] Anubhav Jain, Shyue Ping Ong, Wei Chen, Bharat Medasani, Xiaohui Qu, Michael Kocher, Miriam Brafman, Guido Petretto, Gian-Marco Rignanese, Geoffroy Hautier, Daniel Gunter, and Kristin A. Persson. Fireworks: a dy-

namic workflow system designed for high-throughput applications. *Concurrency and Computation: Practice and Experience*, 27(17):5037–5059, 2015.

[38] Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Computational Materials*, 1(1):15010, Dec 2015.

[39] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, jun 2017.

[40] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

[41] Alec Belsky, Mariette Hellenbrandt, Vicky Lynn Karen, and Peter Luksch. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallographica Section B*, 58(3 Part 1):364–369, Jun 2002.

[42] Marco Esters, Corey Oses, Simon Divilov, Hagen Eckert, Rico Friedrich, David Hicks, Michael J. Mehl, Frisco Rose, Andriy Smolyanyuk, Arrigo Calzolari, Xiomara Campilongo, Cormac Toher, and Stefano Curtarolo. aflow.org: A web ecosystem of databases, software and tools. *Computational Materials Science*, 216:111808, 2023.

[43] Corey Oses, Marco Esters, David Hicks, Simon Divilov, Hagen Eckert, Rico Friedrich, Michael J. Mehl, Andriy Smolyanyuk, Xiomara Campilongo, Axel van de Walle, Jan Schroers, A. Gilad Kusne, Ichiro Takeuchi, Eva Zurek, Marco Buongiorno Nardelli, Marco Fornari, Yoav Lederer, Ohad Levy, Cormac Toher, and Stefano Curtarolo. aflow++: A c++ framework for autonomous materials design. *Computational Materials Science*, 217:111889, January 2023.

[44] Michael J. Mehl, David Hicks, Cormac Toher, Ohad Levy, Robert M. Hanson, Gus Hart, and Stefano Curtarolo. The aflow library of crystallographic prototypes: Part 1. *Computational Materials Science*, 136:S1–S828, August 2017.

[45] David Hicks, Michael J. Mehl, Eric Gossett, Cormac Toher, Ohad Levy, Robert M. Hanson, Gus Hart, and Stefano Curtarolo. The aflow library of crystallographic prototypes: Part 2. *Computational Materials Science*, 161:S1–S1011, April 2019.

[46] David Hicks, Michael J. Mehl, Marco Esters, Corey Oses, Ohad Levy, Gus L.W. Hart, Cormac Toher, and Stefano Curtarolo. The aflow library of crystallographic prototypes: Part 3. *Computational Materials Science*, 199:110450, November 2021.

[47] Camilo E Calderon, Jose J Plata, Cormac Toher, Corey Oses, Ohad Levy, Marco Fornari, Amir Natan, Michael J Mehl, Gus Hart, Marco Buongiorno Nardelli, et al. The aflow standard for high-throughput materials science calculations. *Computational Materials Science*, 108:233–238, 2015.

[48] Vinay I. Hegde, Muratahan Aykol, Scott Kirklin, and Chris Wolverton. The phase stability network of all inorganic materials. *Science Advances*, 6(9), February 2020.

[49] A. R. Akbarzadeh, V. Ozoliņš, and C. Wolverton. First-principles determination of multicomponent hydride phase diagrams: Application to the li-mg-n-h system. *Advanced Materials*, 19(20):3233–3239, September 2007.

[50] Scott Kirklin, Bryce Meredig, and Chris Wolverton. High-throughput com-

putational screening of new li-ion battery anode materials. *Advanced Energy Materials*, 3(2):252–262, September 2012.

[51] Anubhav Jain, Joseph Montoya, Shyam Dwaraknath, Nils E. R. Zimmermann, John Dagdelen, Matthew Horton, Patrick Huck, Donny Winston, Shreyas Cholia, Shyue Ping Ong, and Kristin Persson. The materials project: Accelerating materials design through theory-driven data and tools. pages 1751–1784, 2020.

[52] Kiran Mathew, Joseph H. Montoya, Alireza Faghaninia, Shyam Dwarakanath, Muratahan Aykol, Hanmei Tang, Iek-heng Chu, Tess Smidt, Brandon Bocklund, Matthew Horton, John Dagdelen, Brandon Wood, Zi-Kui Liu, Jeffrey Neaton, Shyue Ping Ong, Kristin Persson, and Anubhav Jain. Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. *Computational Materials Science*, 139:140–152, November 2017.

[53] Morten Niklas Gjerding, Alireza Taghizadeh, Asbjørn Rasmussen, Sajid Ali, Fabian Bertoldo, Thorsten Deilmann, Nikolaj Rørbæk Knøsgaard, Mads Kruse, Ask Hjorth Larsen, Simone Manti, Thomas Garm Pedersen, Urko Petralanda, Thorbjørn Skovhus, Mark Kamper Svendsen, Jens Jørgen Mortensen, Thomas Olsen, and Kristian Sommer Thygesen. Recent progress of the computational 2D materials database (C2DB). *2D Materials*, 8(4):044002, jul 2021.

[54] 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, Jan 2005.

[55] Peder Lyngby and Kristian Sommer Thygesen. Data-driven discovery of 2d materials by deep generative models. *npj Computational Materials*, 8(1), November 2022.

[56] Markus Scheidgen, Lauri Himanen, Alvin Noe Ladines, David Sikter, Mohammad Nakhaee, Ádám Fekete, Theodore Chang, Amir Golparvar, José A. Márquez, Sandor Brockhauser, Sebastian Brückner, Luca M. Ghiringhelli, Felix Dietrich, Daniel Lehmberg, Thea Denell, Andrea Albino, Hampus

Näsström, Sherjeel Shabih, Florian Dobener, Markus Kühbach, Rubel Mozumder, Joseph F. Rudzinski, Nathan Daelman, José M. Pizarro, Martin Kuban, Cuauhtemoc Salazar, Pavel Ondračka, Hans-Joachim Bungartz, and Claudia Draxl. Nomad: A distributed web-based platform for managing materials science research data. *Journal of Open Source Software*, 8(90):5388, 2023.

[57] Luigi Sbailò, Ádám Fekete, Luca M Ghiringhelli, and Matthias Scheffler. The nomad artificial-intelligence toolkit: turning materials-science data into knowledge and understanding. *npj Computational Materials*, 8(1):250, 2022.

[58] K. Lejaeghere, V. Van Speybroeck, G. Van Oost, and S. Cottenier. Error estimates for solid-state density-functional theory predictions: An overview by means of the ground-state elemental crystals. *Critical Reviews in Solid State and Materials Sciences*, 39(1):1–24, October 2013.

[59] Emanuele Bosoni, Louis Beal, Marnik Bercx, Peter Blaha, Stefan Blügel, Jens Bröder, Martin Callsen, Stefaan Cottenier, Augustin Degomme, Vladimir Dikan, Kristjan Eimre, Espen Flage-Larsen, Marco Fornari, Alberto Garcia, Luigi Genovese, Matteo Giantomassi, Sebastiaan P. Huber, Henning Janssen, Georg Kastlunger, Matthias Krack, Georg Kresse, Thomas D. Kühne, Kurt Lejaeghere, Georg K. H. Madsen, Martijn Marsman, Nicola Marzari, Gregor Michalicek, Hossein Mirhosseini, Tiziano M. A. Müller, Guido Petretto, Chris J. Pickard, Samuel Poncé, Gian-Marco Rignanese, Oleg Rubel, Thomas Ruh, Michael Sluydts, Danny E. P. Vanpoucke, Sudarshan Vijay, Michael Wolloch, Daniel Wortmann, Aliaksandr V. Yakutovich, Jusong Yu, Austin Zadoks, Bonan Zhu, and Giovanni Pizzi. How to verify the precision of density-functional-theory implementations via reproducible and universal workflows. *Nature Reviews Physics*, 6(1):45–58, nov 2023.

[60] Daniel T. Speckhard, Christian Carbogno, Luca Ghiringhelli, Sven Lubeck, Matthias Scheffler, and Claudia Draxl. Extrapolation to complete basis-set limit in density-functional theory by quantile random-forest models, 2023.

[61] Ghanshyam Pilania, Chenchen Wang, Xun Jiang, Sanguthevar Rajasekaran,

and Ramamurthy Ramprasad. Accelerating materials property predictions using machine learning. *Scientific Reports*, 3(1), September 2013.

[62] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, 3(1), dec 2017.

[63] Sue Sin Chong, Yi Sheng Ng, Hui-Qiong Wang, and Jin-Cheng Zheng. Advances of machine learning in materials science: Ideas and techniques. *Frontiers of Physics*, 19(1):13501, November 2023.

[64] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning. *Springer Series in Statistics*, 2009.

[65] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, June 2015.

[66] Lauri Himanen, Marc O.J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, feb 2020.

[67] Artem R. Oganov and Mario Valle. How to quantify energy landscapes of solids. *The Journal of Chemical Physics*, 130(10), mar 2009.

[68] Mario Valle and Artem R. Oganov. Crystal fingerprint space – a novel paradigm for studying crystal-structure sets. *Acta Crystallographica Section A Foundations of Crystallography*, 66(5):507–517, aug 2010.

[69] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18), may 2013.

[70] Luca M. Ghiringhelli, Jan Vybiral, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler. Big data of materials science: Critical role of the descriptor. *Physical Review Letters*, 114(10), mar 2015.

[71] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):1–7, 2016.

[72] Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, and Luca M Ghiringhelli. Sisso: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials*, 2(8):083802, 2018.

[73] Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769, 2016.

[74] Haoyan Huo and Matthias Rupp. Unified representation of molecules and crystals for machine learning. *Machine Learning: Science and Technology*, 3(4):045017, nov 2022.

[75] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, 104:136403, Apr 2010.

[76] Michael J. Willatt, Félix Musil, and Michele Ceriotti. Atom-density representations for machine learning. *The Journal of Chemical Physics*, 150(15), apr 2019.

[77] James P. Darby, James R. Kermode, and Gábor Csányi. Compressing local atomic neighbourhood descriptors. *npj Computational Materials*, 8(1), August 2022.

[78] Olexandr Isayev, Denis Fourches, Eugene N. Muratov, Corey Oses, Kevin Rasch, Alexander Tropsha, and Stefano Curtarolo. Materials cartography: Representing and mining materials space using structural and electronic fingerprints. *Chemistry of Materials*, 27(3):735–743, jan 2015.

[79] Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jürgen Bajorath. Molecular similarity in medicinal chemistry: Miniperspective. *Journal of Medicinal Chemistry*, 57(8):3186–3204, nov 2013.

[80] Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature Communications*, 8(1), June 2017.

[81] Chiheb Ben Mahmoud, Andrea Anelli, Gábor Csányi, and Michele Ceriotti.

Learning the electronic density of states in condensed matter. *Phys. Rev. B*, 102:235130, Dec 2020.

[82] Nikolaj Rørbæk Knøsgaard and Kristian Sommer Thygesen. Representing individual electronic states for machine learning gw band structures of 2d materials. *Nature Communications*, 13(1), feb 2022.

[83] Peter Willett. Similarity-based data mining in files of two-dimensional chemical structures using fingerprint measures of molecular resemblance. *WIREs Data Mining and Knowledge Discovery*, 1(3):241–251, mar 2011.

[84] Dagmar Stumpfe and Jürgen Bajorath. Similarity searching. *WIREs Computational Molecular Science*, 1(2):260–282, feb 2011.

[85] Andreas Bender and Robert C. Glen. Molecular similarity: a key technique in molecular informatics. *Organic & Biomolecular Chemistry*, 2(22):3204, 2004.

[86] Peter Willett, John M. Barnard, and Geoffrey M. Downs. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38(6):983–996, jul 1998.

[87] Martin Kuban, Santiago Rigamonti, and Claudia Draxl. Madas: A python framework for assessing similarity in materials-science data, 2024.

[88] Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, July 1977.

[89] Martin Kuban, Santiago Rigamonti, Markus Scheidgen, and Claudia Draxl. Density-of-states similarity descriptor for unsupervised learning from materials data. *Scientific Data*, 9(1):646, Oct 2022.

[90] David J. Rogers and Taffee T. Tanimoto. A computer program for classifying plants. *Science*, 132(3434):1115–1118, 1960.

[91] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert. A survey of binary similarity and distance measures. *Journal of systemics, cybernetics and informatics*, 8(1):43–48, 2010.

[92] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an

appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1), may 2015.

[93] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[94] Atsushi Togo and Isao Tanaka. `Spglib`: a software library for crystal symmetry search, 2018.

[95] David C. Lonie and Eva Zurek. Identifying duplicate crystal structures: Xtalcomp, an open-source solution. *Computer Physics Communications*, 183(3):690–697, 2012.

[96] Sebastiaan P. Huber, Emanuele Bosoni, Marnik Bercx, Jens Bröder, Augustin Degomme, Vladimir Dikan, Kristjan Eimre, Espen Flage-Larsen, Alberto Garcia, Luigi Genovese, Dominik Gresch, Conrad Johnston, Guido Petretto, Samuel Poncé, Gian-Marco Rignanese, Christopher J. Sewell, Berend Smit, Vasily Tseplyaev, Martin Uhrin, Daniel Wortmann, Aliaksandr V. Yakutovich, Austin Zadoks, Pezhman Zarabadi-Poor, Bonan Zhu, Nicola Marzari, and Giovanni Pizzi. Common workflows for computing material properties using different quantum engines. *npj Computational Materials*, 7(1), August 2021.

[97] Martin Kuban, Šimon Gabaj, Wahib Aggoune, Cecilia Vona, Santiago Rigamonti, and Claudia Draxl. Similarity of materials and data-quality assessment by fingerprinting. *MRS Bulletin*, 47(10):991–999, sep 2022.

[98] `https://nomad-lab.eu/entry/id/kUqd_BDKGmTWhJcbvy_puqwA9vbi`.

[99] Lars Hedin. New method for calculating the one-particle green's function with application to the electron-gas problem. *Phys. Rev.*, 139:A796–A823, Aug 1965.

[100] R.G. Humphreys, D. Bimberg, and W.J. Choyke. Wavelength modulated absorption in sic. *Solid State Communications*, 39(1):163–167, 1981.

[101] https://dx.doi.org/10.17172/NOMAD/2021.10.26-1.

[102] Ch. Gähwiller and G. Harbeke. Excitonic Effects in the Electroreflectance of Lead Iodide. *Phys. Rev.*, 185(1141), Sep 1969.

[103] R. Ahuja, H. Arwin, A. Ferreira Da Silva, C. Persson, J. M. Osorio-Guillén, J. Souza De Almeida, C. Moyses Araujo, E. Veje, N. Veissid, C. Y. An, I. Pepe, and B. Johansson. Electronic and optical properties of lead iodide. *J. Appl. Phys.*, 92(7219), 12 2002.

[104] Chenhai Shen and Guangtao Wang. Electronic and optical properties of bilayer PbI2: A first-principles study. *J. Phys. D: Appl. Phys.*, 51(035301), 1 2018.

[105] Stephan Sagmeister and Claudia Ambrosch-Draxl. Time-Dependent Density Functional Theory versus Bethe–Salpeter equation: An All-Electron Study. *PCCP*, 11(22):4451–4457, 2009.

[106] https://dx.doi.org/10.17172/NOMAD/2022.01.23-1.

[107] Wahib Aggoune, Caterina Cocchi, Dmitrii Nabok, Karim Rezouali, Mohamed Akli Belkhir, and Claudia Draxl. Dimensionality of excitons in stacked van der Waals materials: The example of hexagonal boron nitride. *Phys. Rev. B*, 97:241114(R), 2018.

[108] H. Ehrenreich and H. R. Philipp. *Phys. Rev.*, 128:1622, 1962.

[109] S. Robin. Propriétés optiques de l'argent et du palladium dans l'ultraviolet lointain, 1966.

[110] H. J. Hagemann, W. Gudat, and C. Kunz. *Phys. Rev. B*, 65:742, 1975.

[111] G. Leveque, C. G. Olson, and D. W. Lynch. *Phys. Rev. B*, 27:4654, 1983.

[112] W. S. M. Werner, K. Glantschnig, and C. Ambrosch-Draxl. *J. Phys. Chem. Ref. Data*, 38:1013, 2009.

[113] https://dx.doi.org/10.17172/NOMAD/2020.07.27-1.

[114] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[115] https://nomad-lab.eu/entry/id/zkkMIAPyn4OCbdEdW21DZTeretQ3.

[116] https://nomad-lab.eu/entry/id/yMEbPhw-ttwsWZEsEVycez647OKH.

[117] https://nomad-lab.eu/entry/id/qLoIniorAfKleyQnlAudrk_GuKFF.

[118] https://nomad-lab.eu/entry/id/GtPHkCo0qq8VDHmZyjOBP7mAR1jl.

[119] E. D. Pierron, D. L. Parker, and J. B. McNeely. Coefficient of expansion of gaas, gap, and ga(as, p) compounds from -62° to 200°c. *Journal of Applied Physics*, 38(12):4669–4671, November 1967.

[120] H Muñoz, R O Escamilla, J M Cervantes, J León-Flores, M Romero, E P Arévalo-López, E Carvajal, and R Escamilla. Theoretical study on the structural, electronic, mechanical, vibrational, thermodynamical, and optical properties of the two-dimensional pbc nanomaterials. *Physica Scripta*, 99(1):015921, December 2023.

[121] Pbc sg 216 at oqdm. https://dx.doi.org/20.500.12856/oqmd.v1-en.1218947.v1.

[122] Pbc sg 216 at oqdm. https://dx.doi.org/20.500.12856/oqmd.v1-en.908542.v1.

[123] M. H. Cohen, M. V. Ganduglia-Pirovano, and J. Kudrnovský. Orbital symmetry, reactivity, and transition metal surface chemistry. *Phys. Rev. Lett.*, 72:3222–3225, May 1994.

[124] M. H. Cohen, M. V. Ganduglia-Pirovano, and J. Kudrnovský. Electronic and nuclear chemical reactivity. *The Journal of Chemical Physics*, 101(10):8988–8997, nov 1994.

[125] W Yang and R G Parr. Hardness, softness, and the fukui function in the electronic theory of metals and catalysis. *Proceedings of the National Academy of Sciences*, 82(20):6723–6726, 1985.

[126] Banabir Pal, Yanwei Cao, Xiaoran Liu, Fangdi Wen, M. Kareev, A. T. N'Diaye, P. Shafer, E. Arenholz, and J. Chakhalian. Anomalous orbital structure in two-dimensional titanium dichalcogenides. *Scientific Reports*, 9(1):1896, 2019.

[127] Houlong L. Zhuang and Richard G. Hennig. Single-layer group-III monochalcogenide photocatalysts for water splitting. *Chemistry of Materials*, 25(15):3232–3238, 08 2013.

[128] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, September 1975.

[129] Mark Ashton, John Barnard, Florence Casset, Michael Charlton, Geoffrey Downs, Dominique Gorse, John Holliday, Roger Lahana, and Peter Willett. Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quantitative Structure-Activity Relationships*, 21(6):598–604, dec 2002.

[130] John D. Holliday, Naomie Salim, Martin Whittle, and Peter Willett. Analysis and display of the size dependence of chemical similarity coefficients. *Journal of Chemical Information and Computer Sciences*, 43(3):819–828, apr 2003.

[131] Szymon Majewski, Michal Aleksander Ciach, Michal Startek, Wanda Niemyska, Blazej Miasojedow, and Anna Gambin. The Wasserstein Distance as a Dissimilarity Measure for Mass Spectra with Application to Spectral Deconvolution. In Laxmi Parida and Esko Ukkonen, editors, *18th International Workshop on Algorithms in Bioinformatics (WABI 2018)*, volume 113 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 25:1–25:21, Dagstuhl, Germany, 2018. Schloss Dagstuhl – Leibniz-Zentrum f"ur Informatik.