Towards Efficient Novel Materials Discovery
Acceleration of High-throughput Calculations and Semantic Management of Big Data using Ontologies

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. Maja-Olivia Lenz-Himmer

Präsident (komm.) der Humboldt-Universität zu Berlin:
Prof. Dr. Peter Frensch

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter*innen:	Prof. Dr. Matthias Scheffler Prof. Dr. Claudia Draxl Prof. Dr. Sören Auer
Tag der mündlichen Prüfung:	25. Januar 2022

Acknowledgement

I would like to express my deepest appreciation to my supervisor Matthias Scheffler who gave me the opportunity to pursue my PhD studies at the Fritz Haber Institute. The enormous amount of resources available at the institute allowed me to focus on my research without any struggles to finance computation resources or traveling to conferences and workshops. He was never short of good ideas and suggestions to support my research. His focus on applicability of semantic technology shaped my view on the field. I would like to extend my deepest gratitude to my direct supervisors. In the first two years and the resulting Part I of this thesis, this was Christian Carbogno who always listened to my doubts and problems and was never short of ideas for solutions or next projects. For the second part of this thesis, I am truly grateful for the support and supervision by Carsten Baldauf and Luca Ghiringhelli. Both were always available to discuss new ideas or findings and to find a balance on which paths to follow and which not. Furthermore, I am deeply indebted to Thomas A. R. Purcell, who helped refine and continue parts of the work presented in Part I. Apart from fixing bugs he implemented the transformation of the Hessian, extended the benchmark study to a representable number of materials and converged the polaron relaxations in the large MgO supercell. When I was on parental leave, he took over responsibility for this project and finalized our publication. Thank you – without your help, this thesis would look a lot different. Within the STREAM project I often consulted Markus Scheidgen whom I want to thank for being continuously helpful for everything related to NOMAD. For example, the technical realization of the DCAT interface was implemented by him and he was always quick to respond to questions in detail. I also want to thank the STREAM project partners, especially Javad Chamanara, Tatyana Sheveleva and Adham Hashibon, for fruitful discussions, meetings and workshops. Thank you, Matthias Weber for the organization and moderation. My very first contact with ontologies was facilitated by a two-day visit of Jesper Friis, which gave me a great start into ontology development. I spent 2 weeks in Durham, NC in the group of Stefano Curtarolo to effectively collaborate with the AFLOW developers for the parametric relaxation project which resulted in my first publication. I wish to thank Stefano for this opportunity and all the organization around it. The whole group was very welcoming, friendly and helpful, and made my stay a real pleasure. Here, I want to mention in particular Corey Oses and David Hicks. The last chapter of this work is the applications of semantic technologies to the field of heterogeneous catalysis which was enabled by the work of the group of Annette Trunschke at the Inorganic Chemistry Department of the FHI. I want to thank Annette for her work and Lucas Foppa for the introduction to this field in which I had to learn a lot as it so different from what I have been doing before.

I would like to acknowledge also the administrative staff at the FHI: Julia Pach, Hanna Krauter, Annika Scior and Steffen Kangoswki. Finally, thank you Hagen-Henrik Kowalski and Sebastian Kokott for sharing office T1.23 with me – it has been enlightening. ;) I am glad we met. Thanks also to Florian Knoop for asking so many questions, indeed questioning everything and thereby providing a real scientific atmosphere. There are a lot more people I want to thank: Marcel Langer, Christopher Sutton, Yuanyuan Zhou, Xiaojuan Hu and many more. Every colleague whose name can not be found in this acknowledgement should not worry, I still remember you and found it wonderful to meet you all and spend the last years (at least before the pandemic) with you.

Finally, I am thanking my family for any type of support in all parts of my life.

Abstract

The discovery of novel materials with specific functional properties is one of the highest goals in materials science. Screening the structural and chemical space for potential new material candidates is often facilitated by high-throughput methods. Fast and still precise computations are a main tool for such screenings and often start with a geometry relaxation to find the nearest low-energy configuration relative to the input structure. In part I of this work, a new constrained geometry relaxation is presented which maintains the perfect symmetry of a crystal. How this also saves time and resources is shown in a benchmark study with hundreds of materials throughout all symmetry groups. As another example, two materials with meta-stable phases are relaxed with these new constraints that would otherwise loose their symmetry in a fully unconstrained relaxation at zero Kelvin. The proposed constraints also allows for local symmetry preservation or breaking enabling maximum flexibility when systems with distortions or defects are investigated. This is demonstrated at the example of a small polaron distortion.

Apart from improving such computations for a quicker screening of the materials space, better usage of existing data is another pillar that can accelerate novel materials discovery. While many different databases exists that make computational results accessible, their usability depends largely on how the data is presented. We here investigate how semantic technologies and graph representations can improve data annotation. A number of different ontologies are developed enabling the semantic representation of crystal structures, materials properties as well experimental results in the field of heterogeneous catalysis. In a first use-case a computational dataset of hybrid organic-inorganic perovskites is expressed as a knowledge graph using these ontologies. Results from subsequent studies, such as similarity relations, are added to showcase the flexibility of graph-based storage approaches. As a second application, an experimental dataset for nine vanadium-containing catalyst materials is transformed to a knowledge graph. We discuss the breakdown of the knowledge-graph approach when knowledge is created using artificial intelligence and propose an intermediate information layer. Gaining new insights was not possible simply by using semantic techniques for storage and annotation. The underlying ontologies can provide background knowledge for possible autonomous intelligent agents in the future. However, mathematical support is needed in the natural sciences to infer new logical consequences from semantic descriptions. We conclude that making materials science data understandable to machines is still a long way to go and the usefulness of semantic technologies in the domain of materials science is at the moment very limited.

Zusammenfassung

Die Entdeckung von neuen Materialien mit speziellen funktionalen Eigenschaften ist eins der wichtigsten Ziele in den Materialwissenschaften. Das Screening des strukturellen und chemischen Phasenraums nach potentiellen neuen Materialkandidaten wird häufig durch den Einsatz von Hochdurchsatzmethoden erleichtert. Schnelle und dennoch genaue Berechnungen sind eins der Hauptwerkzeuge für dieses Phasenraum-Screening und der erste Schritt sind oft Geometrierelaxationen um die nächstgelegene Konfiguration mit einem lokalen Energieminimum zu finden. In Teil I dieser Arbeit wird eine neue Methode der eingeschränkten Geometrierelaxation vorgestellt, in welcher die perfekte Symmetrie des Kristalls erhalten bleibt. Anhand von hunderten Materialien quer durch alle Symmetriegruppen wird in einer Benchmark-Studie gezeigt, wie dieser Ansatz Zeit und Resourcen spart. Als zweites Beispiel werden zwei Materialien mit metastabilen Phasen eingeschränkt relaxiert, die in einer vollen unbeschränkten Geometrieoptimierung bei null Kelvin ihre Symmetrie verlieren würden. Die neue Methode erlaubt ebenfalls das lokale Erhalten oder Brechen von Symmetrien was wiederum ein Maximum an Flexibilität bedeutet, insbesondere für Systeme mit Verzerrungen und Defekten. Am Beispiel eines kleinen Polarons wird dies demonstriert.

Neben der Verbesserung solcher Berechnungen um den Materialraum schneller durchleuchten zu können ist auch eine bessere Nutzung vorhandener Daten ein wichtiger Pfeiler, der die Entdeckung neuer Materialien beschleunigen kann. Obwohl schon viele verschiedene Datenbanken für computerbasierte Materialdaten existieren ist die Nutzbarkeit stark abhängig von der Darstellung dieser Daten. Hier untersuchen wir inwiefern semantische Technologien und Graphdarstellungen die Annotation von Daten verbessern können. Verschiedene Ontologien werden entwickelt anhand derer die semantische Darstellung von Kristallstrukturen, Materialeigenschaften sowie experimentellen Ergebenissen im Gebiet der heterogenen Katalyse ermöglicht werden. Zunächst wird als Beispiel ein DFT-Datensatz von hybriden organisch-anorganischen Perowskiten als Wissensgraph mithilfe dieser Ontologien repräsentiert. Ergebnisse späterer Studien, wie beispielweise Ähnlichkeitsrelationen, werden dem Graphen hinzugefügt um die Flexibilität solch graph-basierten Speichermethoden aufzuzeigen. In einer weiteren Anwendung wird ein Wissensgraph aus einem experimentellen Datensatz für neun vanadiumbasierte Katalysematerialien erstellt. Wir diskutieren, wie der Ansatz Ontologien und Wissensgraphen zu separieren, zusammenbricht wenn neues Wissen mit künstlicher Intelligenz involviert ist. Eine Zwischenebene wird als Lösung vorgeschlagen. Semantische Technologien zur Speicherung und Annotation führen nicht automatisch zu einem Erkenntnisgewinn, sondern bilden eher die Grundlage für zukünftige autonome Agenten, die Ontologien als Hintergrundwissen verwenden. Es ist in den Naturwissenschaften allerdings notwendig mathematische Konzepte zu unterstützen um neue logische Schlüsse aus semantischen Beschreibungen ziehen zu können. Solche Ansätze existieren bisher nicht. Zusammenfassend ist es noch ein langer Weg bis Materialdaten für Maschinen verständlich gemacht werden können, so das der direkte Nutzen semantischer Technologien nach aktuellem Stand in den Materialwissenschaften sehr limitiert ist.

Contents

List of Abbreviations

AI	Artificial intelligence
API	Application programming interface
DFT	Density functional theory
EMMO	European Materials and Modeling Ontology
FAIR	Findable, accessible, interoperable, re-usable / re-purposable
KOS	Knowledge Organization System
ML	Machine learning
NOMAD	Novel Materials Discovery
OWL	Web Ontology Language
PES	Potential energy surface
RDF	Resource Description Framework
SCF	Self-consistent field (method/cycle)
SPARQL	SPARQL Protocol And RDF Query Language
STREAM	BMBF project: Semantic representation, linking and curating of quality-ensured materials data

List of Figures

List of Tables

General Introduction

Technological progress is today closely connected with advanced materials. From the touch screen display of our smartphones over the coatings of gas turbines to the right catalyst material for carbon dioxide elimination in our atmosphere: Every area of modern life depends on the choice of the appropriate material. Finding and designing new materials that meet the demands of specific applications is one of the most challenging tasks in materials science. The amount of possible materials is defined by the immensity of combinations in the chemical and structural space and is practically infinite. Even with today's fast supercomputers only a tiny portion of this space has been explored so far. Experimentally, even less materials are known. Properties of interest are often the result of complicated computational or experimental workflows requiring expensive calculations or setups, respectively, and are thus available for an even smaller portion of materials. Not so long ago, physics and the natural sciences were divided into experimental and theoretical studies. With the development and improvement of computers, a third pillar has evolved: the computational physics and chemistry. Today, the amount of data created every day has become tremendously large and a so called fourth paradigm of materials science research (or research in general) has risen. This is the age of big-data driven materials science which aims to find patterns and anomalies in big data using for example artificial intelligence (AI). [1] In order to find the needle in the haystack[1], not only a sufficient amount of data is needed but also the quality of these data must be reliable enough to not mistake false input data for interesting anomalies. For future data production therefore quality and quantity are similarly important factors. High-throughput approaches are very popular nowadays as means to scan the chemical and structural space for possibly interesting materials. Here, one method is reducing accuracy just enough to maintain a balance between efficiency and quality so that qualitative trends can still be identified reliably. Exploiting symmetry in crystal structure is a widely used method to both accelerate calculations and improve their accuracy at the same time. Traditionally, crystal symmetries are incorporated already at the electronic-structure level. This saves memory and workload especially for highly symmetric systems. Such global symmetry constraints, e.g. space group conservation, are not sufficient to describe systems with distortions or defects. Selectively breaking the symmetry locally is necessary to address these effects that may crucially alter a material's properties. In Part I of this thesis, we present an effective way to treat global and local symmetries equally during geometry relaxations. Utilizing parametric constraints, both global and local symmetry can be preserved or symmetry can selectively be broken locally when needed. We explore how this approach enables relaxations of dynamically stabilized structures that would otherwise not be addressable easily. In a benchmark study, we show that it furthermore accelerates relaxations for stable systems and is therefore a high-throughput-ready method. Finally its power to handle local distortions in relaxations of supercells is demonstrated at the example of a small polaron in MgO. Besides accelerating and facilitating calculations, these constraints also classify materials. Global symmetries do

[1] a perfect material for a specific use-case in the huge amount of possibilities

in fact determine the space group of materials, whereas local symmetry breaking classifies different types of defect formations and distortions. Such clear and unique classifications are essential to harvest the wealth of data produced nowadays.

Facilitating high-throughput methods naturally leads to a rapid increase in the amount of produced data. Storage and annotation of these data has therefore become an important pillar of materials research with its own challenges and strategies. Several databases and repositories are nowadays established in the field, among which the NOMAD Repository and Archive [2] is the largest due to synergetic relationships with other major databases like AFLOW [3], Materials Project [4], and OQMD [5]. Each data or dataset is fully characterized by a set of attributes called meta-data that allows to clearly describe the data and their provenance. The aforementioned databases all adopt their own meta-data schemas of which the NOMAD Metainfo is the most sophisticated one that not only structures, categorizes and describes the meta-data terms but also provides basic relations between them. A plethora of different meta-data efforts have been taken since the 1980s: the Chemical Markup Language (CML [6]) for chemical meta-data; the European Theoretical Spectroscopy Facility (ETSF) File Format Specifications [7], and the Electronic Structure Common Data Format (ESCDF) [2] are only a few of them. More recently, the OPTIMADE consortium [8] has built a first version of an API providing unified access to a subset of common meta-data terms of different data sources. According to the FAIR principles [9], data as well as meta-data should be findable, accessible, interoperable and re-usable or, as termed in NOMAD, re-purposable. The NOMAD Metainfo fulfills these data principles in many aspects and therefore serves as a good starting point to add another layer of information: semantics. Semantics is the branch of linguistics considering the *meaning* of words and their relations. In computer science, this refers to the formal expression of these meanings within standardized frameworks. One such framework is an ontology that enables annotating linked data and provides means to describe knowledge in a machine-readable formal manner. Ontologies have gained increasing interest in the last decades especially because reasoning software can be used on top of them. One desired goal of such reasoners is to infer logical consequences automatically that have not been put into the ontology directly. The European Materials and Modelling Counsil (EMMC) has focused on the development of an upper ontology for the physical sciences (EMMO) [10] that is based on descriptions from physics, analytical philosophy, and information and communication technologies. Focusing on small chemical compounds, the ChEBI [11] ontology (Chemical Entities of Biological Interest) is a dictionary for general knowledge about molecular entities. Other efforts like the Materials Design Ontology [12], the Materials Ontology [13] and MatOnto [14] reflect that ontology development is currently a hot topic in materials science. Most ontologies are designed to either improve data exchange among heterogeneous databases, therefore covering a wide range of the most common properties and concepts, or for specific narrow sub-domains like nanoparticle domain [15, 16]. In contrast, part II of this thesis attempts to not only design ontologies that correctly represent a material and its properties but also to find applications that showcase the unique value ontologies can provide in the materials sciences. The NOMAD Metainfo is converted to an ontology and can be used directly to represent data in the NOMAD Archive in a linked data format. Further, the Metainfo ontology is enhanced and semantified using a number of ontologies developed within the NOMAD ecosystem. For an example dataset, a knowledge graph is created and enhanced using new and old semantic technologies. Using heterogeneous catalysis as a contrary experimental use-case, an ontology for catalytic characterization and testing experiments is developed. A small set of real experimental results in combination with the outcomes of a machine-learning study on this data is represented utilizing this ontology. Finally, limitations of semantic technologies as they are available and usable today are identified and

discussed.

Part I

Accelerated Materials Discovery in High-throughput Computations: Parametrically Constrained Geometry Relaxations

Chapter 1

The Basics: From the Many-Body Problem to Polarons

This chapter introduces the fundamental concepts that are used in electronic-structure theory. After a short description of the many-body problem and its Hamiltonian, we will explain the widely used Born-Oppenheimer approximation as well as density-functional theory as the main method for electronic-structure calculations. Based on that, the atomic positions that make up the crystal structure are discussed together with symmetry considerations and *ab initio* methods of structure relaxations. Finally, temperature effects like the vibrational motion of the nuclei and quasiparticles like phonons and polarons are introduced.

1.1 The Many-Body Problem

Fully describing a solid material like a crystal requires solving the many-body time-independent Schrödinger equation given by

$$H\Psi = \mathcal{E}\Psi \tag{1.1}$$

with the many-body wave function Ψ, which depends on the coordinates of both, all electrons and all nuclei. The complexity of this becomes obvious when looking at the full non-relativistic Hamiltonian for a system with N electrons and N_{nuc} nuclei:

$$H = T_{\boldsymbol{R}} + T_{\boldsymbol{r}} + V_{\text{nn}} + V_{\text{ee}} + V_{\text{en}} \tag{1.2}$$

$$= -\sum_{I}^{N_{\text{nuc}}} \frac{\hbar^2}{2M_I}\nabla_I^2 - \sum_{i}^{N} \frac{\hbar^2}{2m}\nabla_i^2 \tag{1.3}$$

$$+ \frac{1}{2}\sum_{I\neq J}^{N_{\text{nuc}}} \frac{Z_I Z_J e^2}{|\boldsymbol{R}_I - \boldsymbol{R}_J|} + \frac{1}{2}\sum_{i\neq j}^{N} \frac{e^2}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} - \sum_{i}^{N}\sum_{I}^{N_{\text{nuc}}} \frac{Z_I e^2}{|\boldsymbol{r}_i - \boldsymbol{R}_I|},$$

where the indices i, j run over the electrons with mass m and I, J over the nuclei with masses M_I and nuclear charges Z_I. $T_{\boldsymbol{R}}$ and $T_{\boldsymbol{r}}$ are the nuclear and electronic kinetic energy operators, V_{ee} and V_{nn} the electron-electron and nucleus-nucleus Coulomb repulsion terms excluding self-interaction and V_{en} is the attractive Coulomb potential between the nuclei and electrons. For a representative piece of matter, N_{nuc} would be in the order of 10^{23} leading to the necessity of a number of approximations to solve the many-body problem.

1.1.1 Adiabatic Approximation

One of the most important approximations being made is the adiabatic or Born-Oppenheimer (BO) approximation. It uses the fact that the motion of the nuclei happens on a much larger timescale than the motion of the electrons due to the M_I/m times heavier mass. Hence, the electrons follow the nuclear movement almost simultaneously and their dynamics can be decoupled. The Hamiltonian can be split into an electronic (H_e) and a nuclear (H_n) part

$$H_\mathrm{n} = T_{\boldsymbol{R}} + V_\mathrm{nn}\,, \qquad H_\mathrm{e} = T_{\boldsymbol{r}} + V_\mathrm{ee} + V_\mathrm{en}\,. \tag{1.4}$$

that can be solved consecutively. First, the electronic Schrödinger equation can be solved regarding the nuclear positions $\{\boldsymbol{R}_I\}$ only as parameters. The electronic energy eigenvalues $E_i(\{\boldsymbol{R}\})$ depending on these parameters then serve as a potential for the nuclear eigenvalue problem.

Formally, the total wavefunction can be expanded in the basis of the orthonormal electronic eigenfunctions that depend parametrically on the nuclear positions. The expansion coefficients are the nuclear wave functions. When the nuclear kinetic energy operator $T_{\boldsymbol{R}}$ acts on this product wave function, non-adiabatic couplings between different electronic states appear that are due to the nuclear motion in the off-diagonal matrix elements. Neglecting these as well as their diagonal elements lead to the Born-Oppenheimer (BO) approximation.

Whenever non-adiabatic effects become important, the Born-Oppenheimer approximation is not sufficient anymore to describe the systems appropriately. An example are electron-phonon interactions as well as Jahn-Teller like distortions. Furthermore the nuclear coupling terms become large when two Born-Oppenheimer surfaces lie close together. Consequently another assumption for the BO approximation is that the energy surfaces are well separated.

Once the electronic problem is solved, all information needed to describe the motion of the nuclei is given. The quantum nuclear dynamics are determined by the nuclear Schrödinger equation. However, it is in many cases sufficient to solve the classical Newton equations to obtain reasonable results for the relatively slow movement of the nuclei. Therefore, a combination of *ab initio* quantum electronic-structure calculations with classical equations for the nuclei equilibration is considered often. [17] In the Born-Oppenheimer approximation, the nuclei move in the effective potential created by the electrons as calculated for example using density-functional theory.

1.1.2 Density Functional Theory

Despite the Born-Oppenheimer approximation, which regards the nuclear coordinates only as parameters, solving the electronic problem still remains impossible for systems with many electrons.

Based on the idea of Thomas [18] and Fermi [19], in **density-functional theory** (DFT) the N-electron wavefunction is replaced by the 3-dimensional electron density, hence reducing the degrees of freedom by a factor N. In 1964, Hohenberg and Kohn recognized that the many-electron wavefunction is too complex and could prove that the electron density can be used instead from which all other ground-state properties follow. [20] While it is not surprising that the ground state wavefunction and therefore the density can be uniquely determined

from the external potential $v_{\text{ext}}(\boldsymbol{r})$ [1], the **Hohenberg-Kohn** theorem implies also the opposite, namely that from the density the external potential can be concluded and thus also all other properties of the system. One can say that the wavefunction as well as properties like the ground state energy E are functionals of the external potential and, because of the Hohenberg-Kohn theorem, also of the density. The HK energy functional $E^{\text{HK}}[\rho(\boldsymbol{r}); v_{\text{ext}}(\boldsymbol{r})]$ exists and is unique and it is "minimal at the exact ground-state density, and its minimum gives the exact ground-state energy of the many-body problem" [20]. According to the variational principle this energy functional needs to be minimized with respect to the density

$$E^{\text{HK}}[\rho(\boldsymbol{r}); v_{\text{ext}}(\boldsymbol{r})] = \underbrace{T[\rho(\boldsymbol{r})] + V_{\text{ee}}[\rho(\boldsymbol{r})]}_{\substack{\text{universal functional} \\ F[\rho(\boldsymbol{r})]}} + \int v_{\text{ext}}(\boldsymbol{r})\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} \ . \tag{1.5}$$

The density-functional theory as presented so far is exact up to disregarded phenomena like spin, magnetism or degeneracy. It has to be noted that the universal functional $F[\rho(\boldsymbol{r})]$ is unknown and has to be appropriately approximated.

The most often used framework to practically apply DFT is the **Kohn-Sham** (KS) scheme, which assumes that "for each non-uniform ground-state density $\rho(\boldsymbol{r})$ of an interacting electron system there exists a non-interacting electron system (the Kohn-Sham system) with the same non-uniform ground-state density"[20]. This leads to the decomposition of the ground-state density $\rho(\boldsymbol{r})$ of an *interacting* system into the sum of N *independent* one-electron orbital contributions . When the kinetic energy of the Kohn-Sham system is denoted by T_0 and the Hartree potential V_{Hartree} is defined as the Coulomb repulsion between the ith electron and the electron density produced by all electrons, the universal functional in Equation 1.5 can be further split into

$$F[\rho] = T_0[\rho] + V_{\text{Hartree}}[\rho] + E_{\text{xc}}[\rho] \tag{1.6}$$

with the exchange-correlation energy functional E_{xc} containing all many-body quantum effects. A variational calculation of the functional 1.5 leads to the Kohn-Sham equations that are Hartree-Fock-like single particle equations with a local effective potential and need to be solved iteratively usually in a self-consistent field (SCF) approach.

As the exact exchange-correlation energy functional $E_{\text{xc}}[\rho]$ is unknown, many different **density-functional approximations** have been developed with different levels of accuracy [21]. *Local density approximations* (LDA) regard the exchange correlation energy as a function of only the value of the electron density at each point in space neglecting any derivatives of the density. A step further goes the class of *generalized-gradient approximations* (GGA) taking the first derivative, i.e. the gradient, of the density into account to correct for the non-locality of real systems. Whereas the LDA is local, GGAs fall under the name of semi-local approximations. The proposed parametrization by Perdew, Burke and Ernzerhof [22] (PBE) is one of the widest used GGA and also used in this work alongside its later modification for solids PBEsol [23]. *Meta-generalized gradient approximations* (MGGA) additionally consider the kinetic energy densities τ, which are the Laplacians of the orbitals. Even more accurate but also computationally costly are *hybrid* functionals mixing a fraction of exchange energy from the Hartree-Fock calculation into a GGA. Especially popular is the Heyd–Scuseria–Ernzerhof screened hybrid functional (HSE2006) [24], which can be necessary to obtain more accurate electronic Kohn-Sham levels or for charge-transfer calculations.

[1] The external potential is in our case given by the electron-nucleus attraction.

1.2 Crystal Structure

A crystal structure is characterized by its unit cell that is spanned by three lattice vectors ($\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$) and contains one or more atoms. Each atom's position can then be written either in Cartesian coordinates or in fractional coordinates referring to the lattice vectors as basis. The discrete lattice that is generated by a set of discrete translations defined by the lattice vectors is called Bravais lattice. In three dimensions only 14 different Bravais lattices exist.

1.2.1 Symmetry Groups

Due to the periodic nature of a crystal, several operations are possible that leave at least parts of the crystal invariant. A point-symmetry operation is an operation where exactly one point in the structure is left unchanged. The corresponding mathematical groups where these operations are classified are the so called **point groups**. They include groups for plane reflection, rotation about an axis and combinations of them leading to 32 unique three-dimensional crystallographic point groups. Another type of symmetry operation is translation, which itself does not leave any point of the crystal invariant. However, certain combinations of the point symmetry operations with translations yield another type of symmetry operations, which similarly bring the crystal structure into self-coincidence. Together with the screw axis and glide plane operations, 230 different **space groups** can be defined in three dimensional space. A space group fully determines the symmetry of a periodic system. It also restricts the possible positions for the atoms within the unit cell to the symmetrically equivalent sites exhibiting the same local site point group symmetries (symmetry sites). These sites are called **Wyckoff positions** and are tabulated for each space group in the International Tables for Crystallography (ITC) [25, 26] and the Bilbao Crystallographic Server [27, 28]. Still, how these Wyckoff positions are decorated with atoms leaves us with infinite possibilities and explains the tremendous amount of crystal structures.

1.2.2 Structure Prototypes

Apart from the symmetry groups, another classification can be made into structural prototypes, which are undecorated structures with given space group and Wyckoff positions. This is useful because often the same structure is found among different chemical compositions. It is common practice in the search for novel materials to decorate known structure prototypes systematically with different elements to obtain potential new materials. [29, 30] An example is the crystal structure of Wurtzite, which is found in many different chemical compositions, e.g. ZnO or GaN. The probably largest library collecting structure prototypes today is the AFLOW Library of Crystallographic Prototypes [31, 32, 33]. As of February 2021 it contains 1100 unique prototypes across all 230 space groups. They are sorted by space group, stoichiometry and occupied Wyckoff sites, as calculated with AFLOW-SYM [34].

1.2.3 Structure Relaxation

Regardless whether a prototype has just newly been decorated with elements or a known material is investigated with different computational settings (e.g. number of k-points or another XC functional), a structure relaxation is most often the first step. Even if the experimental

values for the real lattice parameters and atomic positions are known, the actual computational equilibrium configuration associated with the electronic ground state depends on the physical and numerical approximations being made (e.g. XC functional, basis set, k-grid). Relaxing the structure is an iterative process in which the BO potential-energy surface is explored to find a local minimum. In each step the electronic structure is calculated from first principles with DFT, the forces on the atoms as well as the stress on the lattice is calculated and the geometry is optimized until the forces are above a chosen threshold. A necessary condition for a local minimum is that all forces on the atoms vanish. The inter-atomic forces are the derivatives of the Born-Oppenheimer potential-energy surface with respect to the nuclear coordinates. Because the latter are only included as parameters (see Section 1.1.1), the **Hellmann-Feynman theorem** [35, 36] can be applied stating

$$\frac{\mathrm{d}E_\lambda}{\mathrm{d}\lambda} = \left\langle \Phi_\lambda \middle| \frac{\mathrm{d}H_\lambda}{\mathrm{d}\lambda} \middle| \Phi_\lambda \right\rangle \tag{1.7}$$

for a continuous parameter λ. With that, the force acting on the kth atom becomes

$$\boldsymbol{F}_k = -\left\langle \Phi_\mathrm{g}(\boldsymbol{r};\boldsymbol{R}) \,|\, \nabla_{\boldsymbol{R}_k} H_\mathrm{e} \,|\, \Phi_\mathrm{g}(\boldsymbol{r};\boldsymbol{R}) \right\rangle \ . \tag{1.8}$$

For electronic basis functions that depend on the nuclear coordinates (like atom-centered orbitals) additional correction terms appear. These are the Pulay forces, which arise from the derivative of the orbitals with respect to the nuclear positions. Besides, also a lack of self-consistency and therefore inaccuracies in the energy lead to such correction terms. [37] Furthermore, the lattice degrees of freedom need to be relaxed along with the atomic coordinates. Equivalent to the forces on the atoms, a deformed lattice feels a stress, which also needs to vanish at a local minimum. It is in general described by the stress tensor σ and defined as the first order change of the total energy with respect to a strain deformation ε relative to a reference system

$$\sigma_{ij} = \frac{1}{V} \frac{\partial E}{\partial \varepsilon_{ij}}\bigg|_{\varepsilon=0} \tag{1.9}$$

for a unit cell with volume V.The indices i and j stand for the three Cartesian coordinates, i.e. the stress acts on a plane normal to the i-axis in the direction j. At a local maximum or saddle point of the PES, forces and stress vanish as well, so that an additional constraint is therefore a positive second derivative of the energy, which is commonly known as the Hessian $\mathcal{H}$ or force constant matrix

$$\mathcal{H}_{ij} = -\frac{\partial \boldsymbol{F}_i}{\partial \boldsymbol{R}_j} \ . \tag{1.10}$$

A plethora of different optimization algorithms exist (see e.g. [38]), which can be categorized into *direct* methods using only the function value, *gradient* methods using the first derivative, and *Newton* methods using also the second derivative. The best technique depends as so often on the system size and the available computational resources. Gradient algorithms like the steepest descent or conjugate gradient first choose a step direction and then a step size, classifying them as line-search methods. Newton methods assume a quadratic approximation around the current point. The validity of this approximation can be specified with a trust radius making this technique a trust-region method where the step size (the trust radius) is chosen before the step direction. [39] Quasi-Newton methods fall in between steepest descent and Newton methods by updating an approximated Hessian in each step instead of re-calculating it. This makes quasi-Newton approaches the most efficient and widely used optimization strategy for local minimum detection on a PES. Many

different quasi-Newton methods exist, all implementing their own updating functions for the approximate Hessian. All have in common that the Hessian approximation must satisfy the quasi-Newton condition, the secant equation, which can be viewed as a finite difference approximation or Taylor expansion of the gradient itself. Because in more than one dimension this equation is underdetermined, additional constraints are needed to obtain a solution. The most popular quasi-Newton algorithm is the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. [40, 41, 42, 43] In BFGS, the Hessian does not only need to be symmetric and positive definite but also must be sufficiently close to the Hessian in the previous iteration. To ensure this closeness the matrix norm of the difference of two subsequent Hessians has to be minimized, which turns out to be equivalent to adding a symmetric, rank-two matrix to the Hessian. The update is therefore called a rank-two update. In practice, not the Hessian but its inverse is updated directly, so that no additional matrix inversion is necessary reducing the computational cost from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2)$. This method is very robust and shows a self-correcting behavior if at some iteration the Hessian approximation is far from the true Hessian. One limitation is the system size. Approximated Hessian matrices are usually dense even if the true Hessian is sparse. Thus, the cost for storage of the approximated Hessian and working with it can become crucial for large n.

1.3 Electronic-structure Calculations and Relaxations in FHI-aims

Electronic-structure calculations in this work were performed with the all-electron, full-potential electronic-structure code FHI-aims (Fritz Haber Institute *ab initio* molecular simulations package). [17] To solve the Kohn-Sham equations, it uses numeric atom-centered orbitals (NAOs) of the form

$$\phi_i(\boldsymbol{r}) = \frac{u_i(r)}{r} Y_{lm}(\Omega) \tag{1.11}$$

as basis functions for the Kohn-Sham orbitals. The shape of the radial part $u_i(r)$ is element-dependent and numerically tabulated. $Y_{lm}(\Omega)$ are the real and imaginary parts of the complex spherical harmonics. This choice ensures the ability to gradually increase accuracy (and with it computational cost) by adding more basis functions. The correct behavior of the radial functions near the nuclei, i.e. for $r \to 0$, is given by including occupied free-atom orbitals in the basis. [17] Each radial function is a solution to a Schrödinger-like radial equation with a potential $v(r) = v_i(r) + v_{\text{cut}}(r)$ where the second potential cuts the tails of the radial functions ensuring they are zero outside a confining radius r_{cut}. The exponential form of the confining potential in a region $r_{\text{onset}} < r < r_{\text{cut}}$ is critical to avoid discontinuities in the basis functions and their derivatives. Such a strict spatial separation ensures the good efficiency of the necessary numerical integrations for larger systems with $\mathcal{O}(N)$ instead of $\mathcal{O}(N^3)$ as usual. [44] When periodic systems are treated, the Kohn-Sham Hamiltonian and its solutions become $\boldsymbol{k}$-dependent. The basis functions $\phi_i(\boldsymbol{r})$ can then be generalized to Bloch-like basis functions of the form

$$\chi_{i,\boldsymbol{k}}(\boldsymbol{r}) = \sum_N \exp\left[i\boldsymbol{k} \cdot \boldsymbol{T}(\boldsymbol{N})\right] \cdot \phi_i\left[\boldsymbol{r} - \boldsymbol{R}_{\text{at}} + \boldsymbol{T}(\boldsymbol{N})\right] \tag{1.12}$$

where $\boldsymbol{T}(\boldsymbol{N})$ is the translation vector for the unit cell with the unique three-dimensional index $\boldsymbol{N}$. The resulting $\boldsymbol{k}$-dependent matrix elements are numerically integrated not in one sweep but as partial integrals that each extend only over the volume of one unit cell. In practice, the basis functions for each element are grouped into different *tiers* that build on each

other hierarchically. The minimal tier consists of core and valence functions for spherically symmetric free atoms.

The choice of these NAOs result in additional terms for the atomic forces. A first correction term appears because the electrostatic potential is truncated and the missing terms "move" along with the nuclear positions. Also the so called Pulay terms arise that appear due to the incompleteness of the basis set and because the basis function depend on the nuclear positions. [45] The Hellmann-Feynmann theorem is based on the assumption that the derivations of the wavefunctions vanish, which does not hold in this case. The stress is calculated analytically in FHI-aims and includes corrections due to multiple approximations [46].

Because forces and stress are readily available after an electronic-structure calculation, FHI-aims is also able to perform structure relaxations. The preferred algorithm is a trust-radius enhanced BFGS method, which covers all use cases. As quickly mentioned in Section 1.2.3, trust-radius refers to the step size that is chosen before the step direction in the type of optimization techniques called trust-region methods [39].

1.4 Nuclear Dynamics in the Harmonic Approximation

In Section 1.2.3 we have placed the nuclei in the local minima of the potential-energy surface, which represent their equilibrium positions at zero Kelvin. However, in a quantum description localized ions have a non-vanishing kinetic energy even at zero Kelvin due to the uncertainty principle. This is known as the zero-point motion or zero-point energy (ZPE). It influences the cohesive energy as well as the lattice constant and other equilibrium properties at absolute zero [47]. Even more crucial are the effects of nuclear motion at finite temperatures.

A good starting point to investigate the nuclear dynamics is Taylor expanding the PES around the static equilibrium $\boldsymbol{R}_0$

$$E(\{\boldsymbol{R}_0 + \Delta\boldsymbol{R}\}) = E(\{\boldsymbol{R}_0\}) + \sum_i \frac{\partial E}{\partial \boldsymbol{R}_i}\bigg|_{\boldsymbol{R}_0} \Delta\boldsymbol{R}_i + \frac{1}{2}\sum_{i,j} \frac{\partial^2 E}{\partial \boldsymbol{R}_i \partial \boldsymbol{R}_j}\bigg|_{\boldsymbol{R}_0} \Delta\boldsymbol{R}_i \Delta\boldsymbol{R}_j + \mathcal{O}(\Delta\boldsymbol{R}^3) \tag{1.13}$$

where the first term is the static equilibrium energy from DFT, the second summand vanishes in the local minimum and the third term contains the Hessian already defined in Equation 1.10. Neglecting the terms of third and higher orders is called the harmonic approximation and deemed valid for small displacements $\Delta\boldsymbol{R}$. The Hessian can either be calculated using density functional pertubation theory (DFPT) within linear-response theory [48, 49] or using the finite-differences approach [50]. For a real solid the number of atoms becomes practically infinite. Applying periodic boundary conditions to the unit cell, the Hessian can however be represented equivalently in the reciprocal space as the so called Dynamical Matrix

$$D_{kl}(\boldsymbol{q}) = \frac{1}{\sqrt{M_k M_l}} \sum_i \exp\left(\mathrm{i}\,\boldsymbol{q}(\boldsymbol{R}_{0,i} - \boldsymbol{R}_{0,l})\right) \mathcal{H}_{ki}\,, \tag{1.14}$$

which is the mass-weighted Fourier transform of the Hessian. The equation of motion for the nuclei becomes now an eigenvalue problem

$$\boldsymbol{D}(\boldsymbol{q})\,\nu_s(\boldsymbol{q}) = \omega_s^2\,\nu_s(\boldsymbol{q}) \tag{1.15}$$

with the eigenvectors ν_s and the eigenvalues ω_s^2. In real space, the analytical solutions for the nuclear movement are superpositions of harmonic oscillators with eigenfrequencies ω_s. It can

be shown that these vibrational modes are quantized leading to the concept of a phonon – a quantized lattice vibration. In this work, the python package phonopy [51] was used to calculate phonon properties. It uses a finite-difference approach and creates a symmetry-reduced set of atomic displacements for which the forces are calculated. The force constants are then obtained calculating a Moore-Penrose pseudoinverse[2] by fitting all symmetry-reduced elements of the force constants to the symmetry-expanded atomic forces of the atoms in the supercells [52]. This is called a modification of the Parlinski-Li-Kawazoe method [50].

One of the properties that result from a phonon calculation is the Helmholtz free energy $F = U - TS$ with the internal energy U, and T and S being temperature and entropy respectively. The Helmholtz free energy is minimal at equilibrium when a system is kept at constant temperature. Because volume (V) and particle number N do not change in such calculations, it describes a canonical ensemble (constant NVT). It takes the place of the cohesive energy for elevated temperatures and determines whether a phase is stable or not (or more stable than another one) at a given temperature. In contrast, when the pressure p is kept constant and the volume is allowed to change, a closed thermodynamic system is called isothermal-isobaric or Gibbs ensemble (constant NpT) and is described by the Gibbs free energy.

Finally, it should be noted that the harmonic approximation is a very crude approximation and its ability is limited. Anharmonic effects that can not be described within the harmonic approximation include temperature dependence of equilibrium properties like the thermal lattice expansion as well as phase transitions and heat transport. In the quasi-harmonic approximation the volumetric dependence of the force constants is included, but higher-order terms in the Taylor expansion are often necessary to accurately describe a material at finite temperatures [53].

1.5 Electron-phonon Interaction

So far, the electronic and nuclear motions were investigated separately. Many interesting phenomena in solid-state physics, however, rely on the interaction of these systems. The temperature dependence of charge-carrier mobility in semiconductors or of the electronic energy bands are just two simple examples where electron-phonon interactions (EPI) play a crucial role [54]. A particularly interesting scenario is the reduced conductivity when a charge carrier (e.g. in a photovoltaic material) interacts with polar phonon modes. This leads to the formation of a quasiparticle called *Polaron* consisting of a charge carrier (electron or hole), which is dressed by a lattice distortion. The size of this distortion depends on the strength of the electron-phonon coupling. Small polarons usually inhibit strong EPI and can be regarded as a type of point defect [55, 56]. Such point defects often show a Jahn-Teller like behavior [57, 58, 59]. The Jahn-Teller theorem states that "any non-linear molecular system in a degenerate electronic state will be unstable and will undergo distortion to form a system of lower symmetry and lower energy, thereby removing the degeneracy." [60]. Mostly, this occurs in octahedral and tetrahedral complexes in form of elongated or compressed axial and equatorial bonds. Such a system will be discussed in Section 3.3 with MgO. Because polarons can significantly reduce the charge-carrier mobility, is it crucial to understand how they form and move through a material. Many applications like catalysis [61, 62, 63] or thermo-electricity [56] rely on the mobility of the charge carriers.

[2]The Moore-Penrose inverse, also called pseudoinverse, is a widely known generalization of an inverse matrix. It is obtained from singular value decomposition.

Instead of formulating and solving the complicated interaction Hamiltonian, the supercell approach is often chosen in which the distortion is placed subsequently in supercells of increasing sizes to study how the defect affects the system's total energy. Extrapolating to the dilute limit is then possible if a scaling law is known.

Systems with local distortions as discussed here are a good example for why it is useful to constrain structure relaxations to a reduced parameter space as presented in the next chapter.

Chapter 2

Constraining a Relaxation

2.1 The Role of Symmetry

Symmetry is one of the most fundamental concepts in materials science. It determines for example selections rules of electronic transitions, which is why we talk about symmetry-allowed and symmetry-forbidden transitions. Many properties and applications therefore require certain symmetries to be maintained. Not only global crystallographic symmetries like space group or point group matter, but also local symmetry breaking. For example, polaronic distortions reduce the charge carrier mobility, which can be either favorable or disadvantageous, e.g. for photo-voltaic applications.

Many first-principle codes exploit crystallographic symmetries already at the electronic-structure level because it leads to significant savings in cost and computational resources for highly symmetric crystals. In VASP [64], ABINIT [65] and exciting [66] the $\boldsymbol{k}$-space is sampled in the irreducible Brillouin zone and in PARSEC [67] the sampling happens in symmetry-defined "irreducible wedges" in real space. This ensures that also the atomic forces as well as the stress on the lattice reflect these symmetries. As discussed in Section 1.2.3, geometry relaxations rely on the forces and stresses to calculate the new geometric configuration in each relaxation step. Therefore, the global symmetry such as the space group is inherently preserved in this approach even during structure relaxations. Local symmetry breaking is, however, not possible and typically requires lifting all constraints to involve all atomic and lattice degrees of freedom in the electronic structure calculations and relaxations. A three-dimensional crystal has a $(9 + 3N)$-dimensional potential-energy surface resulting from the three components of the three lattice vectors and the $3N$ components of N atoms in the unit cell. Trajectories can become long and inefficient for unconstrained relaxations. More degrees of freedom and more first-principles calculations because of more relaxation steps hence increases the computational cost. Sometimes, fixing the lattice, atomic or internal degrees of freedom [68, 69] (as done in Quantum Espresso [70], VASP or ABINIT) can help to ensure that the chosen crystal structure or at least the space group is retained during a relaxation. However, this requires manual inspection of the structures and the mentioned examples of systems with local distortions may not always be expressible within such simple constraints.

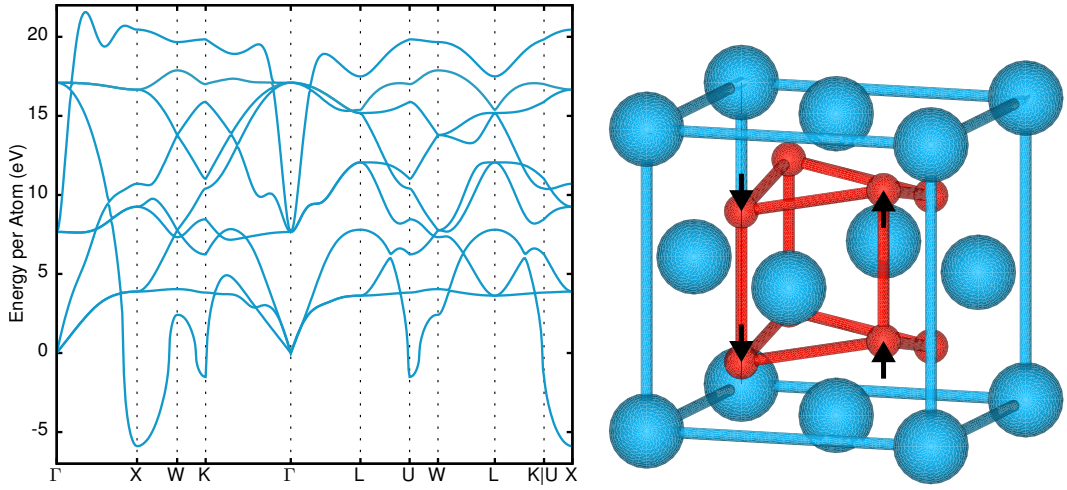


Figure 2.1: Left: Phonon band structure for cubic ZrO_2 with an imaginary soft mode at X indicating instability along this eigenvector. Right: Ball-and-stick model for the 12-atom supercell of tetragonal ZrO_2. Blue balls represent Zr and red balls O atoms. Black arrows indicate the eigenvector of the imaginary phonon mode that pushes the oxygen atoms out of the cubic symmetry resulting in a tetragonal structure.

2.2 New Parametric Constraints

2.2.1 Motivating Example of Zirconia

The three lattice vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ are the edges of the unit cell and are usually expressed in Cartesian coordinates and it is common to use them as basis vectors for the atomic positions, which serves not only human readability but is also practical for structure relaxations where the shape of the unit cell changes. The more symmetric a crystal structures is, the fewer parameters are necessary to describe the shape of the unit cell. A cubic unit cell for example consists of only one lattice constant, referred to as a. If one of the lattice vectors has a different length, c, the unit cell is tetragonal. To illustrate why relaxing within these reduced parameters can be advantageous let us look at the example of zirconia, ZrO_2. In its pure form it exists in three different crystal phases: a high-temperature ($T > 2370\,^\circ$C) cubic phase, a tetragonal phase at intermediate temperatures ($1170\,^\circ\text{C} \leq T \leq 2370\,^\circ$C), and a low-temperature ($T < 1170\,^\circ$C) monoclinic phase [71]. The cubic phase however is only dynamically stabilized, i.e. it is actually a thermodynamical average of lower-symmetry structures [72]. In fact the cubic structure is a special case of the tetragonal structure with a ratio $c/a = 1$ and no displacements of the oxygen atoms. As it corresponds to a saddle point on the PES, it lies in between the six symmetry-equivalent tetragonal structures (two for each cartesian direction). In other words, the cubic structure can be regarded as a transition state between each two tetragonal phases. Figure 2.1 depicts the phonon band structure of the hypothetical, pure cubic ZrO_2 structure. This exhibits an imaginary phonon mode at the X point [50], shown here as negative value. An antiparallel distortion of oxygen atom pairs along the associated eigenvector leads to a stretched unit cell and thus to the tetragonal structure [71]. As shown in figure 2.2, the cubic crystal phase constitutes a saddle point on the potential-energy surface, which lies inbetween two minima corresponding to equivalent tetragonal phases. Here, the PES is shown for a reduced parameter set describing the lattice constant a and the atomic motion along the imaginary mode z_2. Freely relaxing cubic zir-

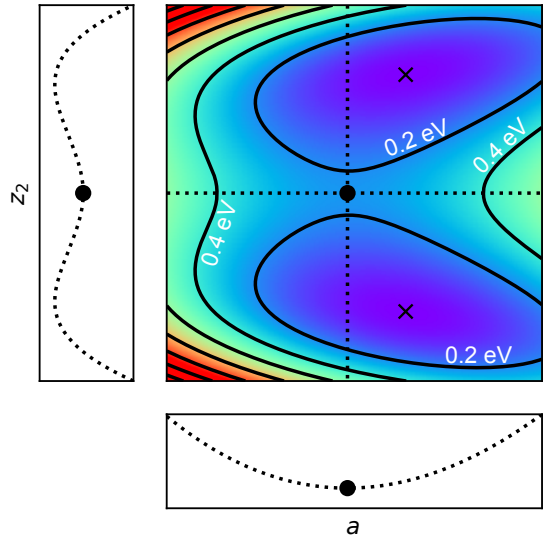


Figure 2.2: Two-dimensional potential-energy surface of ZrO_2. The black crosses are the local energy minima (set to 0.0 eV) representing the tetragonal phase and the cubic phase is the transition state between them indicated by the black dot. The outsets show how the two parameters, lattice constant a and oxygen distortion z_2, change if the other is kept fix.

Table 2.1: Parametric expressions for each cartesian component of the lattice vectors in the 12-atom cubic and tetragonal ZrO_2 supercells

	Cubic			Tetragonal		
	x	y	z	x	y	z
$\boldsymbol{a}$	a	0	0	a	0	0
$\boldsymbol{b}$	0	a	0	0	a	0
$\boldsymbol{c}$	0	0	a	0	0	c

conia in its 12-atom conventional unit cell on this PES would push the material out of the cubic phase into a local minimum. To maintain the cubic structure, the relaxation needs to be constrained to optimize only the lattice parameter a. The respective lattice and atomic constraints for the cubic and tetragonal structures are given in tables 2.1 and 2.2. Within a space group the possible atomic positions are defined by the Wyckoff positions as introduced in Section 1.2.2. In the case of cubic zirconia there are no free parameters in the fractional coordinates so that indeed the whole crystal structure is fixed up to the lattice constant. The parametrically constrained relaxation needs to act only on a while all other degrees of freedom remain untouched. To explore the imaginary phonon modes at X in Figure 2.1, i.e. to allow relaxation to the tetragonal polymorph, the constraints can be lifted stepwise by adding more parameters to the analytic expression of the (originally cubic) structure. It can be seen in Tables 2.1 and 2.2 that the pairwise distortion of the oxygen atoms is included as additional parameter z_2 in the constraints along with an additional lattice parameter c accounting for the stretching of the lattice, which leads to the tetragonality. Using this method ensures that the system relaxes truly to the tetragonal structure and not any other possibly existing near-by local minima on the PES. Cubic and tetragonal zirconia could as well have been relaxed in their primitive cells with six and three atoms respectively to maintain symmetry. However it is not always desirable or feasible to work with primitive cells. When

Table 2.2: Parametric expressions for each fractional component of the atomic positions in the twelve atom cubic and tetragonal ZrO_2 supercell

	Cubic			Tetragonal		
Atom	$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{c}$	$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{c}$
Zr	0.00	0.00	0.00	0.00	0.00	0.00
Zr	0.50	0.50	0.00	0.50	0.50	0.00
Zr	0.00	0.50	0.50	0.00	0.50	0.50
Zr	0.50	0.00	0.50	0.50	0.00	0.50
O	0.25	0.25	0.25	0.25	0.25	$0.25 - z_2$
O	0.25	0.75	0.25	0.25	0.75	$0.25 - z_2$
O	0.75	0.75	0.75	0.75	0.75	$0.75 + z_2$
O	0.25	0.25	0.75	0.25	0.25	$0.75 + z_2$
O	0.75	0.25	0.25	0.75	0.25	$0.25 - z_2$
O	0.25	0.75	0.75	0.25	0.75	$0.75 - z_2$
O	0.75	0.75	0.75	0.75	0.75	$0.75 + z_2$
O	0.25	0.25	0.25	0.25	0.25	$0.25 + z_2$

investigating thermal properties often the phonon supercell approach is used whose name already suggests the need for larger unit cells. Furthermore, there exist other materials where different polymorphs have the same number of atoms in their primitive cells, e.g. the later discussed bismuth oxide [73, 74].

2.2.2 Transformation to Reduced Space

In practice, the three lattice vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ can be conveniently written as a (3×3)-dimensional lattice vector matrix, $\boldsymbol{\mathcal{L}}$, where the rows contain the lattice vectors. Equivalently the fractional atomic positions are combined in an $(N \times 3)$-dimensional matrix $\boldsymbol{\mathcal{R}_\mathcal{F}}$. This matrix notation is also chosen for the forces on the Cartesian atomic positions, $\boldsymbol{\mathcal{F}_\mathcal{R}}$, which are usually calculated directly in first-principles codes, as done in FHI-aims. Similarity, when the stress on the lattice is known, we can obtain a generalized force on the lattice as derived in the following [75]. From Equation 1.9 we know that the stress is defined as

$$\sigma_{ij} = \frac{1}{V} \frac{\partial E}{\partial \varepsilon_{ij}} \bigg|_{\varepsilon=0} \tag{2.1}$$

where the strain tensor ε describes the distortion of the lattice vectors compared to their relaxed states

$$\boldsymbol{\mathcal{L}'}^T = (\mathbb{1} + \varepsilon)\boldsymbol{\mathcal{L}}^T \ . \tag{2.2}$$

Here, the lattice vector matrix is transposed because the strain tensor acts on vectors and the Cartesian coordinates need to be indicated by the rows. Performing the matrix multiplication element-wise, this reads

$$\mathcal{L}'_{ij} = \sum_{k=1}^{3} (\delta_{jk} + \varepsilon_{jk})\, \mathcal{L}_{ik} \ , \tag{2.3}$$

now referring to the un-transposed lattice matrix again after index-swapping. With the chain rule the energy derivative in the stress definition becomes

$$\frac{\partial E}{\partial \varepsilon_{ij}} = \sum_{l,m=1}^{3} \frac{\partial E}{\partial \mathcal{L'}_{lm}} \frac{\partial \mathcal{L'}_{lm}}{\partial \varepsilon_{ij}} \tag{2.4}$$

and the second factor can be rewritten using Eq. 2.3 as

$$\frac{\partial \mathcal{L'}_{lm}}{\partial \varepsilon_{ij}} = \sum_{k=1}^{3} \delta_{mi}\delta_{kj}\mathcal{L}_{lk} \ . \tag{2.5}$$

After making use of the δ functions, the energy derivative with respect to the lattice vector matrix can be formulated as

$$\frac{\mathrm{dE}}{\mathrm{d}\boldsymbol{\mathcal{L}}} = \boldsymbol{\mathcal{L}}^{T^{-1}} V \cdot \sigma \ . \tag{2.6}$$

Finally, a generalized force on the lattice, $\boldsymbol{\mathcal{F}_{\mathcal{L}}}$ is obtained after removing atomic force contributions

$$\boldsymbol{\mathcal{F}_{\mathcal{L}}} = -\frac{\mathrm{dE}}{\mathrm{d}\boldsymbol{\mathcal{L}}} - \boldsymbol{\mathcal{R}_{\mathcal{F}}}^{T} \boldsymbol{\mathcal{F}_{\mathcal{R}}} \ . \tag{2.7}$$

To transform these matrix quantities denoted by calligraphic letters to the parameter-reduced space it is useful to flatten them to one-dimensional vectors that we will name $\boldsymbol{R_F}$, $\boldsymbol{F_R}$, $\boldsymbol{L}$, and $\boldsymbol{F_L}$. Their reduced parametric counterparts are given in small letters: the M_R-dimensional vectors $\boldsymbol{r}$, $\boldsymbol{F_r}$ and the M_L-dimensional $\boldsymbol{l}$ and $\boldsymbol{F_l}$. M_R and M_L are exactly the number of free parameters in the atomic and lattice degrees of freedom. Assuming a linear relationship between the full coordinates and the reduced parameters, transforming back and forth becomes possible by defining the Jacobian matrices $\boldsymbol{\mathcal{J}_R}$, $\boldsymbol{\mathcal{J}_{Rf}}$ and $\boldsymbol{\mathcal{J}_L}$ as well as translation vectors $\boldsymbol{t_{Rf}}$ and $\boldsymbol{t_L}$ that account for additional constant shifts missing in the Jacobians. The reduced lattice and atomic parameters are obtained from the lattice vectors and fractional positions via

$$\boldsymbol{r} = \boldsymbol{\mathcal{J}}_{\boldsymbol{Rf}}^{-1} \left(\boldsymbol{R_F} - \boldsymbol{t_{Rf}}\right) \tag{2.8a}$$

$$\boldsymbol{l} = \boldsymbol{\mathcal{J}}_{\boldsymbol{L}}^{-1} \left(\boldsymbol{L} - \boldsymbol{t_L}\right) \ . \tag{2.8b}$$

Because the Jacobians are not square matrices, they are not regularly invertible. Instead we use the generalized left inverse [76] defined for a matrix $\boldsymbol{\mathcal{A}}$ as

$$\boldsymbol{\mathcal{A}}^{-1,L} = \left(\boldsymbol{\mathcal{A}}^{T}\boldsymbol{\mathcal{A}}\right)^{-1} \boldsymbol{\mathcal{A}}^{T}, \tag{2.9}$$

provided $\boldsymbol{\mathcal{A}}$ has full column rank. Despite working with fractional coordinates, the atomic forces are usually given in a Cartesian coordinate system. To transform the atomic coordinates from Cartesian to reduced space, we define the third Jacobian, $\boldsymbol{\mathcal{J}_R}$ as

$$\boldsymbol{\mathcal{J}_R} = \begin{pmatrix} \boldsymbol{\mathcal{L}}^T & 0 & \dots & 0 \\ 0 & \boldsymbol{\mathcal{L}}^T & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\mathcal{L}}^T \end{pmatrix} \boldsymbol{\mathcal{J}_{Rf}} \ . \tag{2.10}$$

With this, the atomic forces as well as the generalized forces on the lattice can be transformed to their reduced counterparts via

$$\boldsymbol{F_r} = \boldsymbol{\mathcal{J}}_{\boldsymbol{R}}^T \boldsymbol{F_R} \tag{2.11a}$$

$$\boldsymbol{F_l} = \boldsymbol{\mathcal{J}}_{\boldsymbol{L}}^T \boldsymbol{F_L} \ . \tag{2.11b}$$

A back-transformation of the coordinates to real space is done by inverting Equations 2.8 in each relaxation step after the reduced parameters were updated. Back-transforming the forces is not necessary but can still be done in the same manner by inverting Equations 2.11. Thereby, symmetrized Cartesian or fractional forces can be investivated to check for convergence of the relaxation or to explicitly monitor forces that would drive the system out of its constraints.

The implied linear relationship is automatically fulfilled for the atomic positions when fractional coordinates are used. For the unit cell, non-linear expressions appear when angles are introduced as is the case for the monoclinic and triclinic lattice systems. In these cases, these expressions can be substituted with independent parameters. As an example the components $c \cdot \cos\beta$ and $c \cdot \sin\beta$ in the $\boldsymbol{c}$-vector of the monoclinic primitive cell can be replaced by individual independent parameters c and d respectively.

2.2.3 Implementation

In FHI-aims, a trust-radius enhanced BFGS relaxation algorithm is used, where the Hessian matrix is not calculated directly but initialized with an approximate guess and updated in each step. Therefore the Hessian, $\boldsymbol{\mathcal{H}}$, which is initialized in the full coordinate space, needs to be transformed to the reduced space as well. First, it is divided into atomic and lattice blocks, $\boldsymbol{\mathcal{H}_\mathcal{R}}$ and $\boldsymbol{\mathcal{H}_\mathcal{L}}$ respectively. These are then transformed individually into the reduced parameter space via

$$\boldsymbol{\mathcal{H}_r} = \boldsymbol{\mathcal{J}}_{\boldsymbol{R}}^T \boldsymbol{\mathcal{H}_R} \boldsymbol{\mathcal{J}_R} \tag{2.12a}$$

$$\boldsymbol{\mathcal{H}_l} = \boldsymbol{\mathcal{J}}_{\boldsymbol{L}}^T \boldsymbol{\mathcal{H}_L} \boldsymbol{\mathcal{J}_L}. \tag{2.12b}$$

The full reduced Hessian is recombined as

$$\boldsymbol{\mathcal{H}} = \begin{pmatrix} \boldsymbol{\mathcal{H}_r} & 0 \\ 0 & \boldsymbol{\mathcal{H}_l} \end{pmatrix} \ . \tag{2.13}$$

Atomic and lattice degrees of freedom are handled by the optimizer as if they were at the same scale. The atomic parameters coming from the fractional coordinates are however multiple orders of magnitude smaller than the lattice parameters. Before passing them to the optimizer, the atomic coordinates are therefore scaled by the average unit vector length, $V^{1/3}$. Similarly, the reduced atomic forces as well as the atomic block of the Hessian are scaled by $V^{-1/3}$.

Although it would principally be possible to construct the Jacobians at each relaxation step by describing the real space coordinates as a function of the reduced parameters, it is much simpler to create them once at the start of the calculation and re-use them in every step. As mentioned this is however only possible by assuming a linear relationship.

The workflow of relaxing structures within the parametric constraints is shown in Figure 2.3. In each step of the relaxation, forces and stress for the current geometry of the system are

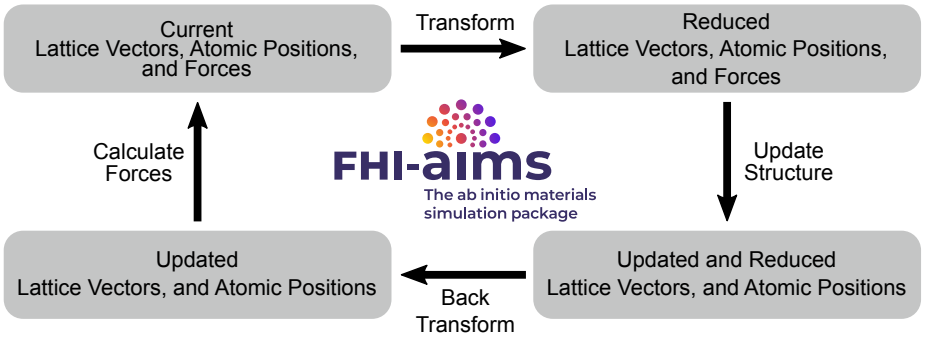


Figure 2.3: Workflow of the relaxation constrained to the parameter reduced space

obtained from a full SCF cycle. Convergence is achieved when the forces are below a given threshold. In this case the current geometry is returned and the relaxation stops. Otherwise the transformations in Equations 2.8, 2.11 and 2.12 are applied to map all degrees of freedom onto the reduced space. After the optimization, the real space representation of the structure is obtained via back-transformation and is passed to the next relaxation step.

To initialize such a constrained relaxation, the analytic parameter representation of the structure needs to be given. In FHI-aims this relaxation scheme is triggered when the keywords `symmetry_n_params`, `symmetry_params`, `symmetry_lv`, and `symmetry_frac` are found in the geometry.in file and a relaxation calculation is started. The two latter keywords specify exactly the parametric expressions as listed in Tables 2.1 and 2.2 respectively. Additionally, the parametric constraints were implemented in the Python package ASE, Atomic Simulation Environment [77], and can be used therein using the classes

- `ase.constraints.FixParametricRelations`,
- `ase.constraints.FixScaledParametricRelations`, and
- `ase.constraints.FixCartesianParametricRelations` .

2.3 Automated Parameter Representations

The proposed formalism requires that the reduced parameter representation of the system and how it is related to its full geometry is known analytically before generating the input files. For crystals, they could in principle be manually constructed from the space group and its Wyckoff positions. Fortunately, the AFLOW Library of Crystallographic Prototypes collects crystal structures and provides their analytical expressions for more than 1100 different prototypes across all 230 space groups. The concept of structure prototypes was already introduced in Section 1.2.2. As there are already three editions of this library, the number of prototypes is likely to increase even more. Materials with the same stoichiometry, space group and occupied Wyckoff positions as calculated with AFLOW-SYM [34] are grouped together in the same crystal prototype. Such a prototype is then uniquely identified by an ID that is created from concatenating the undecorated chemical composition, the Pearson symbol, the space group and the occupied Wyckoff positions. For the example of cubic Zirconia, this reads `AB2_cF12_225_a_c`. Because of the large amount and variety of structure prototypes in the library, it is well suited as a starting point for high-throughput studies. AFLOW is able to create input geometry files automatically for VASP [64], FHI-aims [17], Quantum Espresso [70], Abinit [65] and more codes allowing usage of the prototypes for further investigations. The option `--add_equations` is available in AFLOW since version 3.1.204 to add

the parameter blocks needed for the constrained relaxation to FHI-aims' `geometry.in` file automatically. Adding additional constraints is straight-forward as will become clear when this method is applied to relax distortions in the next chapter. In the same manner, parameters can be removed selectively to constrain certain parts of the structure even further.

Chapter 3

Applications and Results

This chapter demonstrates the necessity and usefulness of constraints in a geometry relaxation for meta-stable and unstable systems. In a benchmark study, the ability of the approach to accelerate high-throughput calculations is presented. Finally, using the example of a small polaron in magnesium oxide, it is shown that even systems with local distortions can be treated more efficiently with constraints. Throughout this chapter, the full-potential, all-electron electronic-structure code FHI-aims is used that was already introduced in Section 1.3.

3.1 Relaxing Metastable and Unstable Systems

The work of this section has largely been done by the author. Thomas Purcell contributed by refining calculations on Bi_2O_3, especially after a reviewer from [78] pointed out different refinements for the β phase.

Some materials have very complex potential-energy surfaces with several local minima representing different meta-stable polymorphs. These stable or meta-stable phases can be relaxed in an unconstrained fashion by choosing an initial geometry close enough to the respective global or local minimum on the PES. As we have seen for zirconia, there can as well be phases that are unstable at zero-point conditions but may be stabilized at higher temperatures or pressures. For the cubic zirconia structure, we have seen that this transition structure corresponds to a saddle point on the energy landscape in between two stable tetragonal structures. A free relaxation would always drive the system towards such a nearby local minimum with lower symmetry. To showcase how parametric constraints introduced in the previous chapter can be used to maintain the structure of these phases, we relax cubic zirconia in its twelve-atom conventional unit cell as it is given in the AFLOW Library of Crystallographic Prototypes. While usually relaxations are performed on the primitive cell of a structure, the choice of the conventional cell for zirconia allows the free relaxation to converge to another polymorph which has a different number of atoms in its primitive cell. It is therefore used as a simple demonstrative example. As we will see later for bismuth oxide, there are indeed systems featuring different polymorphs with the same number of sites in the primitive cell. We expect that the relaxation scheme would not be affected by a further increase of accuracy due to larger basis sets or hybrid functionals.[1] Figure 3.1 shows the convergence behavior of the

[1] The calculations were performed using the 'tier 1' basis sets with 'light' settings that have been shown to yield good results for the lattice parameters and cohesive energies for face-centered cubic gold, i.e. an accuracy

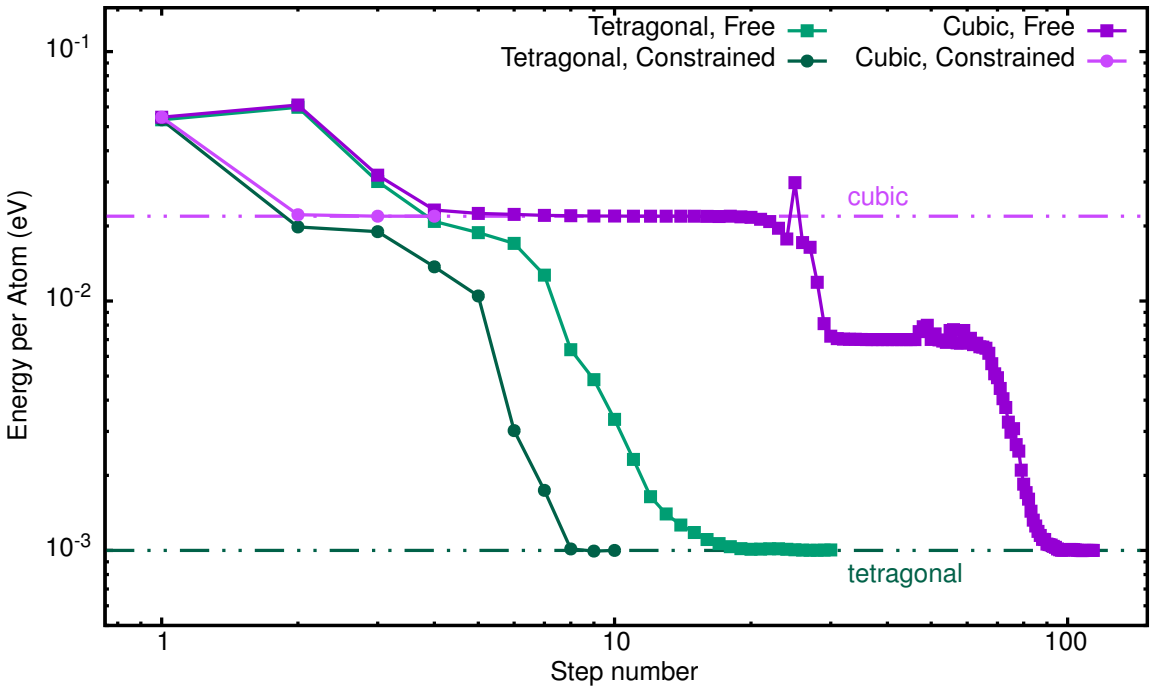


Figure 3.1: Relaxation convergence for ZrO_2. Unconstrained, i.e. free, (squares) and parametrically constrained (circles) relaxations are compared for the tetragonal (green) and cubic (purple) phases. The cubic phase can only be maintained in a constrained relaxation. The energy zero is set to 1 meV below the energy minimum.

free and constrained relaxations in comparison. The initial tetragonal structure is the same as the cubic but with the aforementioned pairwise distortion of the oxygen atoms as seen in Figure 2.1. Initially, the lattice constant is $a \approx 4.968$ Å for both systems. The magnitude of the initial pairwise distortion is chosen as 0.01 units in fractional atomic positions which corresponds to roughly 0.05 Å for the chosen lattice constant. This matches the distortions given in [72]. Of course, different initial configurations will likely yield different results. The focus here lies on showcasing the ability of the parametric constraints to converge to meta-stable or unstable systems. Both, free and constrained relaxations maintain the correct tetragonal symmetry, but we see an increase in efficiency by a factor of 3 when using constraints. Due to only one parameter being optimized, the cubic phase is relaxed even quicker with constraints in only four steps. In contrast, the free relaxation does not manage to converge within the symmetry until the maximum forces are below the requested threshold. The full potential-energy surface is high-dimensional and therefore much more complex than the simple sketch along the a- and z-axis in Figure 2.2. It can comprise multiple other local minima or local energy barriers that complicate the trajectory in the unconstrained relaxation. In fact, it seems as if a small energy barrier needs to be overcome in step 25 to break the symmetry and converge after 37 steps to an intermediate plateau which corresponds to a simple-cubic structure. Only when the convergence criterion, the threshold for the maximum force component, is reduced further to 0.001 eV/Å, the tetragonal structure is finally reached after 114 steps. This not only demonstrates that constraints are crucial to investigate this polymorph but also that the results of a free relaxation depend strongly on the numeric settings. For both struc-

of 0.001 Å and 20 meV respectively. In the SCF cycles, the electron density was converged up to 10^{-6} eV/Å and the atomic forces up to 5×10^{-4} eV/Å. The exchange-correlation functional was chosen to be PBEsol and the relaxation is deemed converged when the maximum forces are below 0.005 eV/Å.

tures we see an initial increase in energy in the unconstrained relaxation. Because the initial lattice constant is chosen too small, the volume of the unit cell is increased significantly in the first step. Because this works well, the next Quasi-Newton step is taken too large leading to a counter-productive iteration that is neglected for further geometry updates.

As a second example for a material with different meta-stable phases, we investigate bismuth oxide. Similarly to zirconia, its different polymorphs [73] exist only in certain temperature ranges. Moreover, some of them can only be obtained in one direction: either heating the material up or cooling it down. The low-temperature monoclinic α-phase transforms into the high-temperature, face-centered cubic phase, denoted δ, upon heating at around 730°C. Further heating does not bring up new structures but leads to melting at approximately 825°C [74]. Both, the low- and high-temperature phases are considered stable within their temperature ranges. Only during the cooling process, one of the two meta-stable phases, the tetragonal β- or the body-centered cubic γ-phase, can be obtained at roughly 650 °C or approximately 640 °C, respectively [74]. Which of these structures is formed depends on the cooling procedure. Cooling down to even lower temperatures, the β-phase eventually returns to the stable α-Bi_2O_3 at around 300°C. At which temperature the γ-phase transforms back to α depends on the cooling rate [74]. Two more meta-stable polymorphs of Bi_2O_3 are known, the orthorhombic ε- and triclinic ω-phase, which require very specific synthesis conditions and are not considered here. To accurately describe monoclinic structures containing oxygen it has proved helpful to add an f-orbital to the 'tier 1' basis sets for oxygen. This has been found in benchmark calculations from [72] because bindings in such structures are often not oriented along the main axes so that p_x-, p_y and p_z-orbitals do not allow for the necessary flexibility. Therefore, for Bi_2O_3 the 'light' settings were replaced by so called 'intermediate' settings and basis sets in FHI-aims which also bring along better accuracies.

As initial geometries for the meta-stable polymorphs we took the relaxed structures from the Materials Project [79] with initial parameters taken from ICSD because this is a very common starting point for high-throughput studies. For zirconia we chose the twelve-atom conventional unit cell for the cubic structure to make the different unit cells commensurate – to allow them being transformed into each other without adding or removing atoms. In bismuth oxide the cubic γ-phase belongs to space group 197 and has 30 atoms in its primitive cell. It it thus not commensurate with the 20-atom primitive cells of the β- and α-phases and consequently could not relax to one of these structures in a free relaxation. This does however not prevent the cubic phase from breaking its symmetry by seeking lower-energy and lower-symmetry structures. Constrained and free relaxation behaviors are shown in Figure 3.2 for the α-, β- and γ-phases of bismuth oxide. For the monoclinic α- and tetragonal β-phases both, the free and the constrained relaxations converge in the correct structure and the constraints accelerate the relaxation needing fewer steps until convergence. In the low-temperature monoclinic polymorph 20% of all relaxation steps are saved (5 steps out of 25) when constraints are used compared to the unconstrained case. As expected, the step savings are larger for the higher symmetric tetragonal structure: instead of 66 steps when all degrees of freedom relax freely, only 17 iterations are needed when the structure is relaxed within the parameter space. A more crucial effect of the constraints can be observed during relaxation of the γ-phase where a free relaxation yields to a previously unknown, non-symmetric structure (space group 1) after 160 steps. Reaching one of the lower-symmetric structures was unsuccessful due to the incompatible unit cells discussed above. Weather the found structure is experimentally realizable potentially at higher pressures or in hetero-epitaxy needs further investigations that are not the topic of this work.

Including the constraints instead, the body-centered cubic structure is conserved and con-

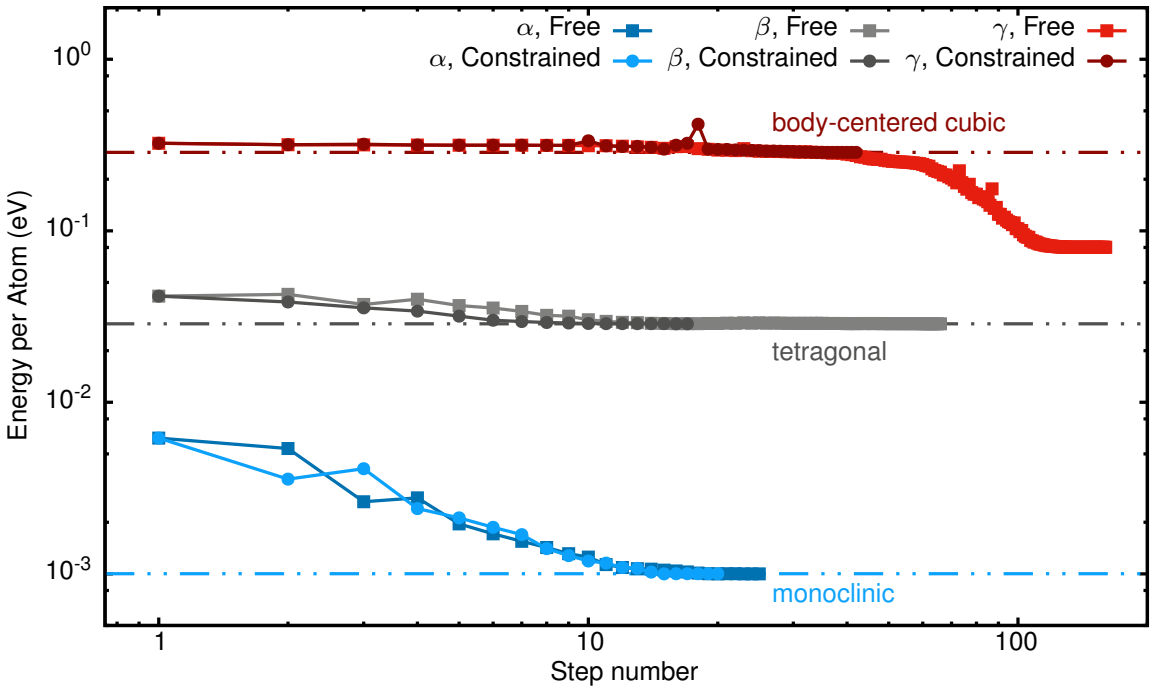


Figure 3.2: Relaxation convergence for Bi_2O_3. The α- (blue) and β- (grey) phases converge correctly in both, unconstrained (free) and constrained, relaxations whereas the γ-phase relaxes to a hitherto unknown lower-symmetry structure without constraints.

verges in 42 steps. Unlike cubic zirconia, cubic Bi_2O_3 does not get close to convergence until the symmetry is completely broken. There are no plateaus in the relaxation curve in Figure 3.2 where the maximum force components drop significantly.

This example illustrates the usefulness for constraining relaxations to their structural prototypes since a structure that relaxes to another symmetry does not represent the same material anymore. While in ZrO_2 one of the known phases is reached eventually, the final structure in Bi_2O_3 might represent not only an unknown material but also potentially a structure that is either not experimentally realizable or under very different conditions and with very different properties. Any further investigations on such a system are therefore either impossible or yield results that are not representative for the initially chosen structural composition. It may well be interesting to study this relaxed system but this way no new information about the original prototype can be gained. Hence, blindly relaxing without constraints, when the material of interest is in a particular known structure, can lead to wrong results that easily stay undetected in high-throughput or machine learning studies especially when long and complicated workflows are involved. A systematic way of detecting such errors would be necessary which is extremely difficult due to the inconsistency of the symmetry breaking. Likewise, correcting the departed geometries requires manual inspection and intervention which is practical only for a small number of systems. Crystal structure prediction is one of fields that could benefit from these constrained relaxations since often the newly discovered structures need to be refined within exact known lattice types.

Table 3.1: Summary of the materials used in the test dataset

AFLOW Prototype	Space Group	# of Materials	Atoms per Unit cell	Free Parameters	Full d.o.f. / # Free Parameters
AB_oP8_62_c_c	62	8	8	7	4.71
A2B_oP12_62_2c_c	62	35	12	9	5.00
A2BC4_tI14_82_bc_a_g	82	35	7	5	6.00
A2BC4D_tI16_121_d_a_i_b	121	29	8	4	8.25
AB2_hP3_164_a_d	164	25	3	3	6.00
AB_hP4_186_b_b	186	37	4	2	5.25
AB_cF8_216_c_a	216	37	2	1	15.00
ABC_cF12_216_b_c_a	216	54	3	1	18.00
AB2_cF12_225_a_c	225	13	3	1	18.00
AB2C_cF16_225_a_c_b	225	14	4	1	21.00
AB_cF8_225_a_b	225	19	2	1	15.00
A_cF8_227_a	227	3	2	1	15.00
A2BC4_cF56_227_d_a_e	227	50	14	2	25.50

3.2 Bench-marking the Algorithm

We have seen that using the parametric constraints not only preserves symmetry in metastable and unstable systems but also accelerates relaxations due to fewer degrees of freedom being optimized. In a benchmark study, this is now quantified using 359 materials across 13 different structural prototypes covering eight space groups in four lattice systems. Most calculations and analyses in this section have been performed by Thomas Purcell who extended the benchmark study to a representable amount of materials starting from a handful of calculations by the author of this thesis. The structure prototypes were chosen to include common materials in well-known structures, for example the Wurtzite, Zincblende, Rocksalt, and Diamond structures as well as Half-Heuslers, Heuslers, and Fluorites. This ensures also that the number of available materials within a prototype is sufficient to understand how the method performs for different materials in the same structure. Additionally, a varied ratio of all degrees of freedom over the number of reduced parameters was important for the choice of the included prototypes. Table 3.1 summarizes the dataset grouped by AFLOW Prototype and ordered by space group. Different composition types from elementary to quaternary and varying number of atoms per unit cell are represented. In this high-throughput study the Atomic Simulation Environment (ASE) [77] was used to transform the initial structures to a consistent format. The original geometries are taken from AFLOW [3] or the Materials Project [79] database. For high-throughput studies, these are very typical entry points to obtain a large amount of different structures throughout all space groups. Of course, it is also possible to manually construct the structure by choosing a space group and Wyckoff positions therein. This defines the parameters needed for the parametric relaxation. All materials are relaxed with the same numerical settings as zirconia from the previous section using both, the PBEsol and PBE functionals.

Before quantitatively investigating the relaxation performance, the similarities between the fully and constrained relaxed structures are studied. Structures can be compared using the AFLOW-XtalFinder [80] which not only identifies the correct prototype of a material but also allows to calculate the similarity between two structures or materials. It uses a misfit value

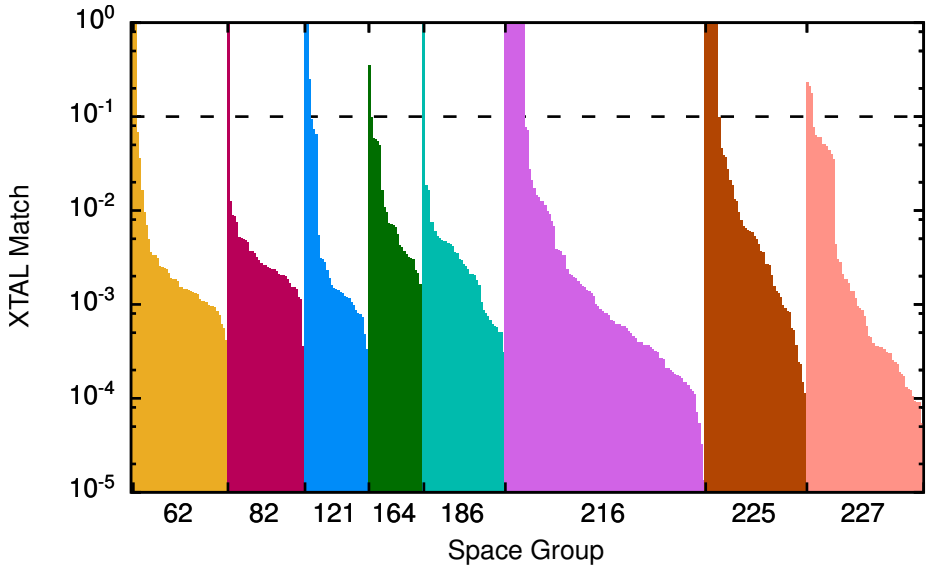


Figure 3.3: Histogram of all misfit values between the fully and constrained relaxed structures using AFLOW-XtalFinder. The horizontal line indicates the threshold of 0.1 below which structures are matching. The bars represent different materials and are ordered by m for each space group.

m proposed by Burzlaff and Malinovsky [81]

$$m = 1 - (1 - m_{\text{latt}})(1 - m_{\text{atom}})(1 - m_{\text{fail}}) \tag{3.1}$$

considering deviations between two sets of lattice vectors m_{latt}, displacements of atomic positions m_{atom} as well as a figure of failure m_{fail} indicating that the distance between two mapped atoms is larger then half the atom's nearest neighbor. With m lying in the range between zero and one, the mapping

$$\begin{aligned} m \leq 0.1 &: \text{structures are similar} \\ 0.1 < m \leq 0.2 &: \text{structures are within same family} \\ m > 0.2 &: \text{structures are not compatible} \end{aligned}$$

determines whether two structures are considered a match or are at least within the same family which means that they have common symmetry subgroups [81]. Figure 3.3 shows a histogram of all misfits between the fully and constrained relaxed structures grouped by space group and sorted by the misfit value m.[2] Roughly 7% of all freely relaxed materials are not similar to their initial structure anymore according to the AFLOW-XtalFinder when PBEsol is used as exchange-correlation functional. In absolute numbers this corresponds to 26 of the 359 investigated materials. With the PBE functional this number slightly drops to 6%, so only 22 materials loose their initial structure (see Table 3.2). In fact, most of them do not even stay in the same family, i.e. the misfit value lies much higher than 0.2. All these cases have similar relaxation behaviors to what was seen for cubic zirconia and γ-bismuth oxide. Consequently, constraining their relaxations is crucial to maintain the physical relevance of any further computations or investigations. The average misfit value m is 0.07654 for PBE and 0.1322 for PBEsol. Apparently, the probability of breaking the symmetry in a free relaxation is larger

[2] Here, AFLOW version 3.1.223 was used for the misfit calculations which ignores space group symmetry and Wyckoff positions by default. Newer AFLOW versions give consistent results when the `--ignore_symmetry` flag is used, because otherwise misfits are detected where symmetries do not match.

when using PBEsol. Even when the extreme cases where $m = 1$ are excluded, the average misfits compare 0.01630 for PBE and 0.02038 for PBEsol. It is likely that this tendency of the PBEsol functional comes from general larger differences between the starting and relaxed structures which would be supported by a larger number of steps until convergence.

Even when the structures are similar, the perfect symmetry is always broken to some degree in a free relaxation because numerically the forces never completely vanish. The extent of this symmetry breaking can be measured by calculating the space groups of the freely relaxed structures and compare them with the initial space groups. For this, the popular python package *spglib* is used which calculates the space group by first finding the primitive cell and then its symmetry operations [82]. A symmetry operation maps all atoms in the primitive cell to sites that are occupied by he same atom type within a tolerance distance ε. Using the default tolerance factor of 10^{-5} Å, not a single structure preserves its space group during a free relaxation. Increasing ε to 10^{-2} Å is required to obtain the initial space groups for approximately 70% and 63% of all freely relaxed materials for PBE and PBEsol respectively. Such an increase in accuracy is consistent with the results of a larger study on symmetry calculations performed by Hicks et al. [34]. Table 3.2 lists the number of materials within a prototype for which the free relaxation preserves the space group. In a parametrically constrained relaxation, the symmetry is by definition always perfectly preserved. This in turn can have a tremendous effect on the computational resources required for further analyses and calculations. For example, phonon calculations are often performed using the finite difference method where atoms are displaced in supercells according to the symmetry of the system. Lower-symmetric systems require many more atomic displacements and corresponding also more force evaluations compared to highly symmetric systems. In particular *ab initio* calculations of the forces in supercells can be very time-consuming, so that every non-necessary computation effort is to be avoided. The widely used python package *phonopy* calculates phonon spectra using the finite difference approach [51] and internally calculates the symmetry of a structure with *spglib*'s default settings. Blindly relying on these out-of-the-box solutions would increase the time to compute phonon properties in a way that can be a bottle-neck in high-throughput studies. Feeding relaxed systems from parametrically constrained relaxations with their perfect symmetries to *phonopy* circumvents this problem.

Apart from the advantages perfectly symmetric systems have on further calculations, there is also an immediate effect of using the constraints for relaxations even when both, the free and constrained relaxations end up in the same final structure. The number of steps taken until the relaxation trajectory converges is significantly reduced on average by 33.11% and 52.43% for calculations with PBE and PBEsol respectively. Across all materials including the relaxations with incorrect final structures the saving are 34.68% and 53.80% as noted in Table 3.2. These step savings S are calculated as

$$S = \frac{N_{\text{free}} - N_{\text{constrained}}}{N_{\text{constrained}}} \times 100\% \tag{3.2}$$

with the numbers of steps needed to converge the free and constrained relaxations, N_{free} and $N_{\text{constrained}}$ respectively. As expected from the larger misfit values for the PBEsol functional, the relaxed structures are further away from the starting structures compared to when using PBE. Thus, the savings are higher for this functional because the free relaxation trajectories are longer. Unfortunately, the savings are inconsistent across the different structural prototypes. The total numbers of steps needed until the structures are relaxed are compared in Figure 3.4 for the constrained and unconstrained relaxations with PBEsol. Each bar represents one material and for each space group these are sorted by descending number of steps

Table 3.2: Summary of the free and constrained relaxation performance by AFLOW prototype.

AFLOW Prototype	Space Group	# of Materials	PBE Average Savings	PBE # Preserved Space Group Free	PBE # XTAL Match	PBEsol Average Savings	PBEsol # Preserved Space Group Free	PBEsol # XTAL Match
AB_oP8_62_c_c	62	8	10.23	3	8	24.61	4	8
A2B_oP12_62_2c_c	62	35	10.84	19	29	18.32	20	33
A2BC4_tI14_82_bc_a_g	82	35	40.32	29	32	59.98	28	34
A2BC4D_tI16_121_d_a_i_b	121	29	44.01	23	26	58.13	21	26
AB2_hP3_164_a_d	164	25	7.03	10	24	19.68	5	24
AB_hP4_186_b_b	186	37	30.34	23	36	41.68	19	36
AB_cF8_216_c_a	216	37	29.71	28	36	54.83	31	36
ABC_cF12_216_b_c_a	216	54	33.35	44	54	74.03	36	46
AB2_cF12_225_a_c	225	13	57.63	8	9	54.20	7	9
AB2C_cF16_225_a_c_b	225	14	19.77	12	14	80.02	7	12
AB_cF8_225_a_b	225	19	32.50	11	19	51.72	7	19
A_cF8_227_a	227	3	35.12	3	3	47.62	3	3
A2BC4_cF56_227_d_a_e	227	50	67.10	37	47	73.58	37	47
Full Dataset		359	34.68	69.64%	93.87%	53.80	62.67%	92.76%

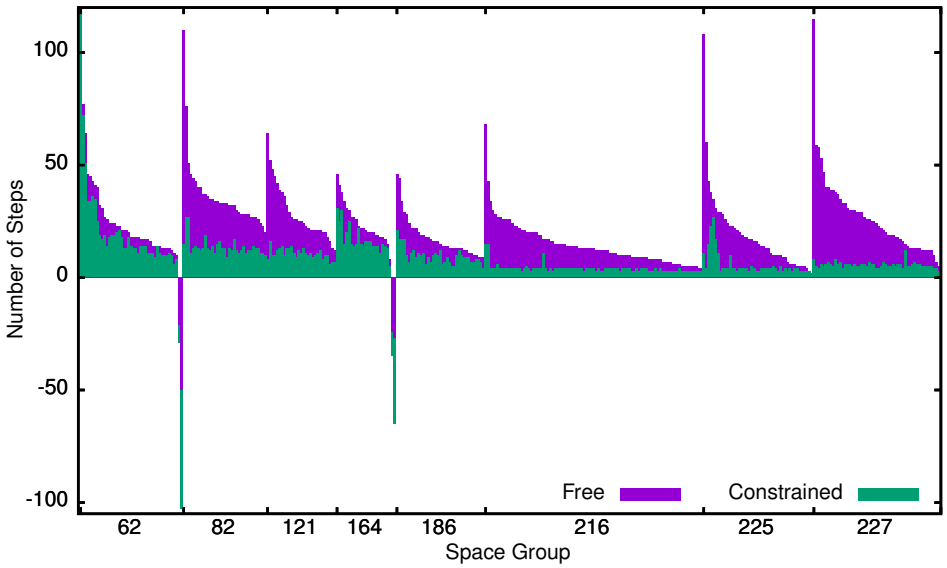


Figure 3.4: Comparison of the number of steps taken until convergence for the constrained and free relaxations. Negative step numbers indicate that the constrained relaxation needs more steps than the free relaxation. The left most bar for N_2O is cut for visualization because it takes 310 and 296 steps to converge in the free and constrained relaxations respectively.

in the free relaxation. Generally the trend of fewer constrained steps for higher-symmetric systems can be observed as expected. Vice versa, as the number of reduced parameters gets closer to the number of all degrees of freedom in a material, the savings decrease. Exceptions are however present as in space group 225 for the lead chalcogenides PbS, PbSe, and PbTe. Seemingly, their potential energy surfaces exhibit some special features that are not captured by the constraints leading to smaller savings compared to the free relaxations. Very conspicuous are also the low average savings for the hexagonal prototype in space group 164. From Table 3.2 we see that only 5 of the 25 materials in this prototype remain in their space group in a free relaxation with PBEsol although all except $CoCl_2$ are considered matching according to the AFLOW-XtalFinder. By driving the system out of its symmetry, the free relaxation converges faster and the performance advantage of the constraints decreases. For the two hexagonal materials platinum sulfide, PtS_2, and vanadium chloride, VCl_2, the constrained relaxation takes more steps to converge than the free relaxation. I found that this is also observed for orthogonal OF_2 and $ThSe_2$ in space group 62 and is indicated in Figure 3.4 by negative step numbers. Different behaviors can lead to this inversion as exemplified in Figure 3.5 but in all cases the optimizer takes unproductive steps. The additional degrees of freedom in the free relaxation for PtS_2 allow the optimizer to overcome problematic regions in the PES faster and to converge in about half of the steps that the constrained relaxation needs. In $ThSe_2$, the parameter space even introduces a large energy barrier between the initial and final structures that is not present in full space and prolongs the relaxation.

Not shown in Figure 3.4 are the calculations with the PBE functional where in some cases the trajectories contain a few extra steps at the end where convergence is almost reached.

Despite the discussed outliers, the general trend of increased performance especially for higher-symmetric structures is beneficial and provides, together with the symmetry preservation, a crucial advantage for many high-throughput studies.

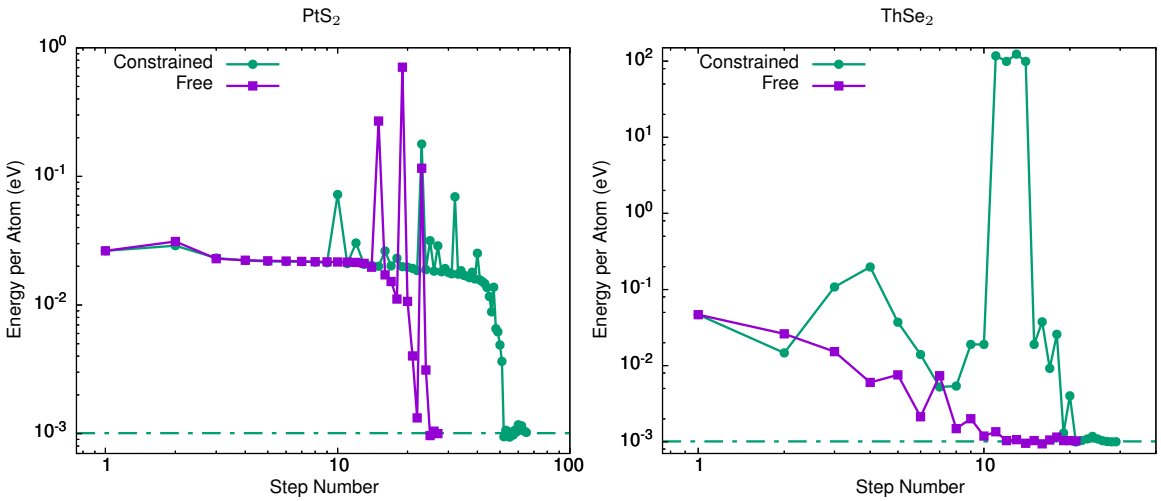


Figure 3.5: Relaxation trajectories for PtS_2 (left) and $ThSe_2$ (right). For both materials, the constrained relaxation needs more steps to converge than the free relaxation.

3.3 Systems with Local Symmetries or Distortions

Thomas Purcell and the author contributed approximately equally to the work presented in this section.

Previously, we have seen the advantages that relaxing within a parameter-reduced space can offer for stable, meta-stable and unstable structures. Maintaining symmetry and accelerating relaxations would however also be possible using symmetrized forces. The real benefit and key advantage of this approach lies in the ability to locally break the symmetry for example to include point defects. This can help to tremendously reduce relaxation times of large unit cells containing defects when the supercell approach is used. Such an approach studies how defects behave in real solid systems by sequentially increasing the unit cell sizes and thereby the number of atoms, which in turn significantly increases the computational resources needed in each step to calculate *ab initio* energies and forces, e.g. from DFT. Many defects act on a short range order, so that their effect on the surrounding atoms rapidly diminishes with increasing distance. This allows including the defect in forms of additional degrees of freedom in the parameter presentation discussed above. Here, as an example, a polaronic distortion in rock-salt magnesium oxide, MgO, will be discussed [55]. The lattice distortions for an electron hole polaron in MgO are Jahn-Teller like, which is typical for octahedral structures like rock-salt. Figure 3.6 shows the conventional unit cell of MgO. When an electron hole is placed in the unit cell, the surrounding oxygen and magnesium atoms are attracted and repelled from the hole, respectively, as indicated by the arrows. Here, the hole was located on a fixed oxygen atom in the center of the cell. In parameter space, this movement along a line through the center of the cell is realized by calculating the unit distance vector of each atom with position $\boldsymbol{r}_i$ to the fixed atom in the center at $\boldsymbol{r}_0$ and distort its initial "perfect" position in this direction. The distorted position is then

$$\boldsymbol{r}_i' = \boldsymbol{r}_i + \lambda_i \frac{\boldsymbol{r}_i - \boldsymbol{r}_0}{|\boldsymbol{r}_i - \boldsymbol{r}_0|} \,. \tag{3.3}$$

The sign of this small perturbation, λ, is negative for oxygen and positive for magnesium. Using such constraints, the main interactions are of electrostatic nature assuming that all other perturbations have only a minor effect on the final geometry and can be mapped onto

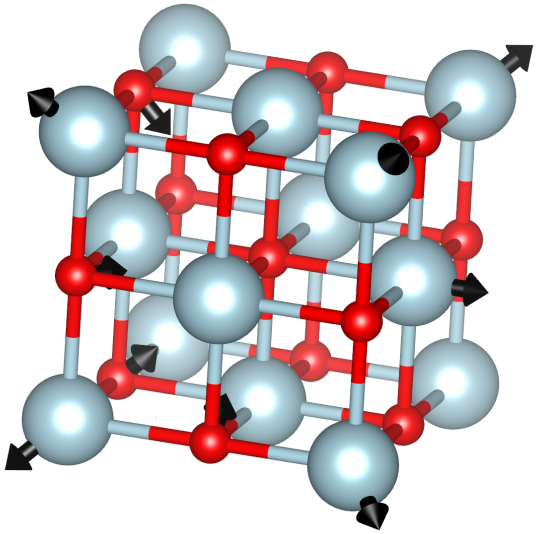


Figure 3.6: Ball-and-stick model of rock-salt MgO with oxygen as red balls and magnesium as grey balls. Arrows indicate the additional degrees of freedom for a constrained relaxation of the electron hole polaron when the hole sits at the center oxygen. Oxygen is attracted to the hole while magnesium atoms are repelled from it.

these radial Coulomb distortions. For a supercell with N atoms, this creates $N-1$ additional parameters $\{\lambda_i\}$ that enable relaxing the polaronic distortion using our parametric constraints.

An important property of a polaron is its binding energy

$$E^{\mp}_{\text{binding}} = E_{\text{polaron}}(N \pm 1) - E_{\text{perfect}}(N \pm 1) \tag{3.4}$$

where E^{+}_{binding} refers to electron removal (hole polaron) and E^{-}_{binding} to an electron addition (electron polaron). This formula does not contain corrections for the finite-size effect.

For the DFT calculations the HSE 06 functional for the exchange-correlation was used with a screening parameter of $\omega = 0.11$ Bohr^{-1} and a fraction of exact exchange $\alpha = 1$ as suggested in [55]. Because of the localized charge of the electron hole, it is necessary to use this more accurate and costly hybrid functional. As convergence criteria for the self-consistent field cycles 10^{-4} eV/Å was chosen for the density and forces while the total energy and the eigenvalues were converged to 10^{-5} eV and 10^{-2} eV, respectively. Following the settings from Kokott *et al.*, five Kohn-Sham energy states above the occupied levels are computed and the relaxation is deemed converged when the forces are below 10^{-4} eV/Å. Using the constraints from Equation (3.3) for MgO, the relaxation trajectory for the hole polaron can be compared to the free relaxation as shown in Figure 3.7 in terms of the uncorrected polaron binding energies. The notation of the supercells is relative to the conventional cubic unit cell, so that the $2 \times 2 \times 2$ super cell doubles the length of each unit cell vector resulting in a 64-atom supercell, while the $3 \times 3 \times 3$ supercell contains 216 atoms. Only 11 steps are needed by the constrained relaxation to converge the distorted 64-atom supercell, which is only one-eighth of the steps needed in the unconstrained case. Even more significant are the savings in the 216-atom supercell: 10 steps instead of 234 steps are taken to converge the distorted geometry, which is equivalent to 96% saved relaxation steps when using the constraints. The fact that the step number with constraints does not increase with the number of atoms in the cell reflects that less fine-tuning is needed to converge the distortions due to other weaker

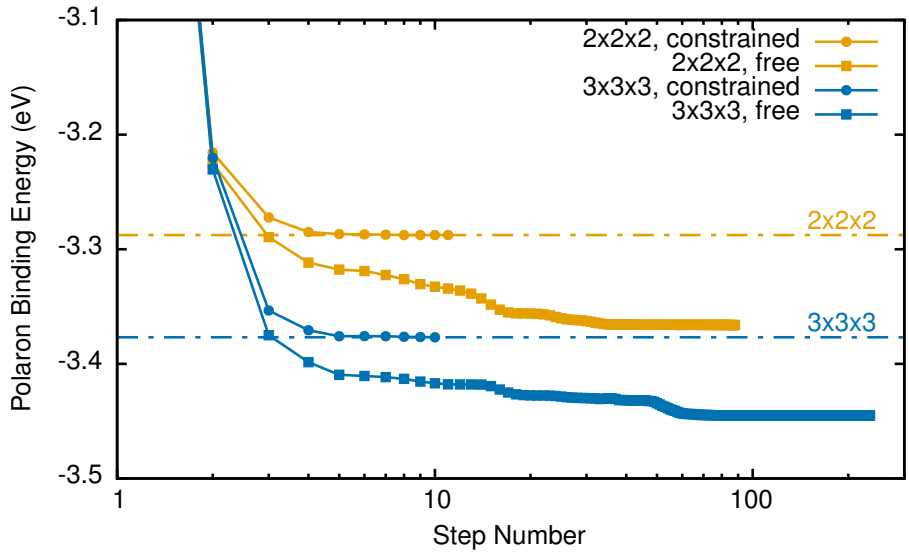


Figure 3.7: Uncorrected polaron binding energy and its relaxation behavior in the $2 \times 2 \times 2$ and $3 \times 3 \times 3$ charged supercells for free and constrained relaxations. Dashed lines represent the final binding energies in the constrained relaxation.

non-Coulombic interactions. Another effect of the chosen constraints is the slightly higher energy. Although the final geometries are almost identical (with an AFLOW mismatch value of $m \approx 0.007$ and $m \approx 0.003$ for the 64- and 216-atom cells respectively), tiny changes in the freely relaxed atomic positions can sum up to a significant energy contribution especially in very polar materials. Here, this energy difference in the polaron binding energy is only 78.4 meV for the $2 \times 2 \times 2$ supercell and 69.1 meV for the $3 \times 3 \times 3$ supercell. In both cases, the free and constrained relaxation, the finite supercell size leads to artificial interactions of the polaron with its periodic images. Hence, increasing the size of the supercell diminishes these artifacts and leads to a 89.2 meV and 79.1 meV lower binding energy of the polaron in the constrained and free relaxation, respectively.

If more accurate results are necessary or the constraints are deemed too restrictive, it is possible to use the constrained relaxation as a starting point and continue from the relaxed constrained geometry with a free relaxation or additional parameters to account for more degrees of freedom. This way, a large amount of steps might still be saved while maintaining accurate results.

Chapter 4

Summary: Advantages and Risks

In this part of the thesis, a new relaxation scheme has been proposed, which uses a parameter-reduced representation of a material's structure to constrain a geometry relaxation. Transformation of both, the structure as well as the atomic and generalized lattice forces, creates a parameter space that is much smaller than the full $(9+3N)$-dimensional space in which an unconstrained relaxation typically is performed. This naturally reduces the number of available directions in which a structure can change and fixes the system to the chosen shape. If a structure prototype is known, this is a useful way to ensure that the symmetry and the exact prototype are kept during a relaxation. In the last chapter, we presented how such constraints are useful to maintain the perfect structure of meta-stable or unstable systems. Also stable systems were shown to benefit from the approach: The fewer parameters are optimized, the faster a relaxation converges, which is particularly useful for highly-symmetric systems. Further, adding or removing individual parameters gives the user a high degree of flexibility to selectively break or preserve local symmetries, for example around defects or distortions as shown at the example of polaronic distortions in MgO.

It has to be noted however that in many cases, a free and unconstrained relaxation is favorable. Predicting new crystal phases is not possible when a structure is strongly constrained. The local minima on a reduced potential-energy surface projected along the parametric representation may not always coincide with the real local minima on the full PES. Statements about the stability of such a parametrically converged system at zero Kelvin are therefore only possible after a subsequent unconstrained relaxation.

Part II

Semantic Data Management in Computational Materials Science: Meta-data and Ontologies

Chapter 1

Towards a semantic world in materials science

The huge amount of data that have been and are being produced in materials science has led to the fourth paradigm of research: The era of big data-driven science [83]. Consequently, new ways to store and annotate data are necessary to ensure findability, accessibility, interoperability and re-usability of data and their meta-data, in short to fulfil the FAIR principles [9], see Section 1.2. Therefore a machine-readable and even machine-actionable representation of knowledge is highly desirable.

1.1 Metadata

Any data produced needs some form of annotation if it is meant to be stored and to be found. This annotation data is the simplest form of meta-data, which is often defined as data about data. However, depending on the context certain types of information may or may not be recognized as meta-data but rather as data (or vice versa).[1] A definition of meta-data is needed that circumvents these cases and clearly distinguishes between data and meta-data. In [84] meta-data of a given data object is defined as the "set of attributes that is necessary to locate, fully characterize, and – ultimately – reproduce other attributes that are identified as data. The meta-data include a clear and unambiguous description of the data, and their full provenance.". This definition is well suited for science because it emphasizes the role of data reproducibility. In practical terms, the meta-data structure to annotate data in a database is defined and organized in a meta-data schema. In accordance with the given definition, the *Metainfo* [2] was designed and implemented as part of the Novel Materials Discovery (NOMAD) Laboratory and consists of a unique name, a human-readable description, expected format (e.g. scalar, string, array), allowed values (e.g. array shape, explicit lists, intervals) and provenance in form of a hierarchy. It annotates and structures data in the NOMAD Archive, the largest database for atomistic calculations containing billions of normalized data objects calculated by more than 40 different codes.

[1] For example, consider a DFT total-energy calculation of a crystal with certain atomic positions and a crystal unit cell. The total energy is the desired result and therefore seen as data, whereas the atomic positions and the unit cell have served as input to the calculation and may therefore be called meta-data of this calculation. Further analysis and computations can use this total energy as input, thereby reclassifying it as meta-data for another calculation.

1.2 The FAIR Principles

The so called FAIR Principles [9] were compiled by the FORCE11 group[2] in 2016 and are nowadays widely accepted guiding principles for data storage. As already stated above, they aim at improving findability, accessibility, interoperability and re-usability. These four terms can be seen as subsequent steps to achieve FAIR compliance. Even before these guidelines were published, the NOMAD Repository and Archive had already implemented most of the concepts.

Findable: First of all, data and their respective meta-data should be easy to find. Human-readable descriptions as well as machine-readable meta-data ensure automatic discovery of stored data. This requires assignment of globally unique and persistent identifiers often realized in form of uniform resource identifiers (URIs). Rich meta-data need to be defined to describe data and identify it explicitly. On the technical side the data and meta-data should be registered in searchable resources.

Accessible: Once found, accessibility needs to be ensured using the data identifier and standardized, open and free protocols allowing for authentication and authorization if needed.[3] Moreover, meta-data should be accessible even when data are no longer available.

Interoperable: Interoperability concerns the ability to integrate with other data as well as different applications or workflows. This can be achieved by using formal, broadly applicable languages or formats for knowledge representation. Furhter referencing other (external) meta-data or data enhances interoperability. One way to improve interoperability is the use of ontologies (see Section 1.3).

Re-usable: Finally, enabling and optimizing re-use of data would be the largest benefit of data storage. This is particularly important for data that is hard to reproduce, e.g. due to computational complexity. Re-use can mean the use in a different application, setting or workflow. Accurate and relevant description attributes, data usage licenses, detailed provenance, community standards help achieving re-usability. In NOMAD, we prefer the "R" to refer to re-purposing or recycling of data rather than reusing to emphasize that data can be used for other purposes than the initial one it has been produced for.

1.3 Ontologies and Knowledge Graphs

Originally, ontology is the philosophical study of being, and it concerns everything related to existence. Nowadays, computer science has borrowed the word from philosophy to describe a new form of semantic knowledge organization system. Within the last twenty years, ontologies have gained increasing interest in computer science and data intensive research fields as can be seen by evaluating the number of publications concerning this topic on the

[2] `https://www.force11.org/group/fairgroup`

[3] Note, that not the data themselves are required to be open for everyone, but only the protocols. FAIR data is not necessarily open data.

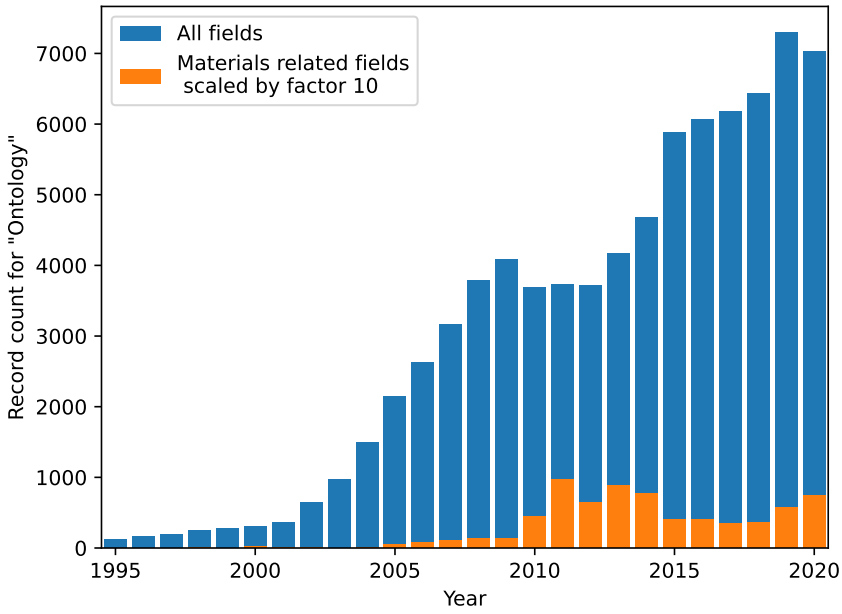


Figure 1.1: Record counts for Web of Science entries on the topic “Ontology”. Blue bars represent the entries for all fields, whereas the orange bars (plotted on top of blue bars) stand for materials science related fields (excluding engineering) and are scaled by a factor of 10 for better visibility.

Web Of Science (WOS)[4] in Figure 1.1. One milestone was the development and publication of the Gene Ontology in 2000 [85], which is the most cited paper for this topic on WOS. Since then, ontologies have established quite well in the area of biology and medicine. The authors’ analysis of the Web of Science entries shows that about 20% of all publications with the topic “Ontology” are in a biological or medical related field. Physics, chemistry and materials science together only hold 3% of the publications reflecting that ontological methods are being explored only recently in these fields. Interestingly, 15% is already published in the engineering areas. The leader is of course computer science, which makes up almost 69% of all publications.

Although the idea of semantic technologies is already quite old, recent developments led to recognition of ontologies also in industry. This is reflected in the appearance of ontologies (and graphs) in Gartner’s most recent hype cycle for emerging technologies from July 2020. Companies worldwide use this famous diagram to evaluate modern technologies for adoption and investment even though the graph has been criticized for its lacking justification. Interestingly, ontologies are placed in the “trough of disillusionment” illustrating that the public’s high expectations from the past could not be met.

1.3.1 What is an Ontology?

Defining “ontology” is not such an easy task as even until today, multiple different definitions co-exist and are being constantly discussed. Originally a computational ontology was defined

[4] `www.webofknowledge.com`

by Gruber in 1993 as an "explicit specification of a conceptualization" [86]. This leaves open many questions, so that since then several modifications and extensions to this definition emerged. One widely accepted extension is by Studer et *al.* in 1998:
Definition 1 (Ontology [87]). *"An ontology is a formal, explicit specification of a shared conceptualization." where 'formal' means the ontology should be machine readable; 'explicit' requires all concepts, properties, relations, functions, constraints and axioms to be explicitly defined; 'shared' emphasizes that the ontology represents consensual knowledge, e.g. that it is accepted by a group; and 'conceptualization' is a abstract model of some phenomenon in the world.*

Still, this definition is somewhat cryptic and abstract, so that another definition that is more helpful for beginners will be given:
Definition 2 (Ontology [88]). *"An ontology is a formal description of knowledge as a set of concepts within a domain and the relationships that hold between them. To enable such a description, we need to formally specify components such as individuals, classes, attributes, relations, restrictions, rules, and axioms. As a result, ontologies do not only introduce a sharable and re-usable knowledge representation but can also add new knowledge about the domain."*

In short, an ontology is a knowledge organization system (KOS) with strong semantics. Traditional KOS exhibit only weak or no semantics, popular types are classifications or controlled vocabularies. They are often designed along typical search paths and not as actual domain representations. An example of a traditional KOS is a directory structure on a computer – files are put into folders where the user would typically search for them and not where they objectively and context-independently belong to. There is no *meaning* for a file to be in a particular folder; it is therefore not an instance or a subclass [89]. In contrast, ontologies are designed to represent objects in a semantic way regardless of whether this corresponds to human interaction. Instantiation and subclassing are essential in ontology development. Figure 1.2 depicts the *semantic ladder* that shows different knowledge organization systems

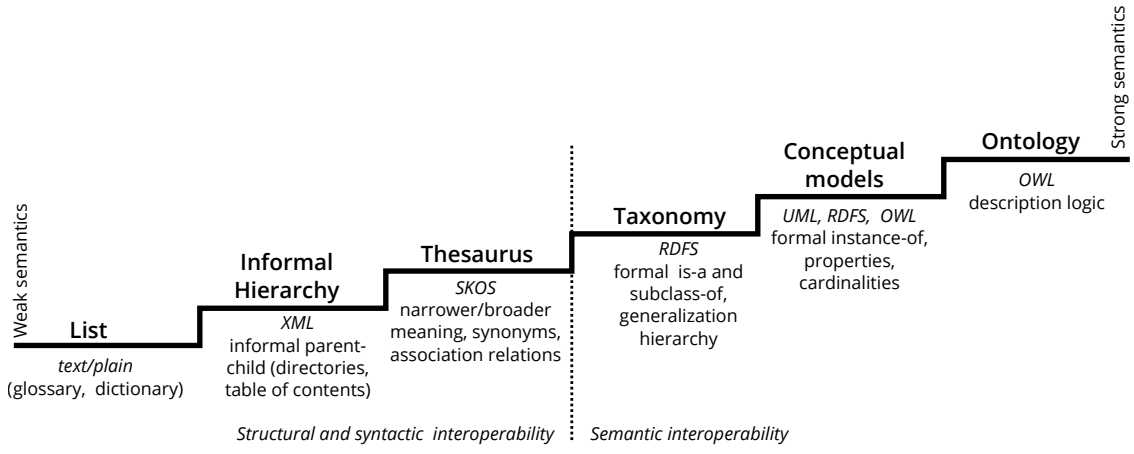


Figure 1.2: Semantic ladder of knowledge organization systems. Redrawn after [10].

ordered from weak to strong semantics with the ontology on the upper-most step. The ladder also lists the respective languages in which these systems are represented.

1.3.2 The Web Ontology Language and its Profiles

Ontologies and other graph-based representations can be written as a number of triple statements. Such a triple consists of a subject, a predicate and an object, which are visualized typically as two nodes connected by an edge (the predicate), i.e. as a directed graph.

The standard data model for directed graphs is the *Resource Description Framework* (RDF). All data in RDF are statements about resources and their relationships, and are modeled as triples. While subject and predicate always are resources, the object can also be a literal. A resource is identified uniquely by a uniform resource identifier (URI) or its extension, the internationalized resource identifier (IRI). A subset of such URIs are the well known URLs. Multiple syntaxes exist in which RDF statements can be formalized, such as XML or more human-readable syntaxes like Turtle. Storing RDF triples in a traditional relational database is not very efficient. Instead the preferred database type is a *triplestore*, also called *RDF store*.

RDF Schema [90] semantically extends RDF by classes and properties and therefore serves as a data schema to describe RDF resources. It provides a basic vocabulary and allows to formalize taxonomies – simple forms of ontologies. Similarly to object-oriented programming languages classes can be instantiated. In contrast however, a class is not defined through its properties, but a property is described by defining its domain (subject class) and range (object class), i.e. the classes that are connected by this property.

The *Web Ontology Language* (OWL) [91] extends RDF Schema to express relations between classes, cardinality, equality, characteristics of properties and much more. Every OWL document is a RDF document. OWL exists in three variants of different complexity and expressivity. These so called OWL profiles are OWL Lite, OWL DL and OWL Full. *OWL Lite* supports classification hierarchy and simple constraints but is too simple for most use cases. *OWL DL* provides maximum expressivity while staying computationally complete and decidable. One limitation is that a class cannot be at the same time an instance of another class. It is based on description logic, hence the name DL. This is the standard variant of OWL that is mainly used in ontologies nowadays. The main reason OWL DL evolved was to enable reasoning software. Such reasoners can infer logical consequences within the description logic framework. *OWL Full* has no computational guarantees but allows maximum freedom with RDF. For example, a class can at the same time be a collection of individuals and an individual itself. Due to its complexity, full reasoning with all its features will likely never be possible with OWL Full.

Intrinsic to OWL ontologies is the *open-world assumption* [92]: Every statement that is allowed can be true irrespective of our knowledge about its truth. In practice this means for example that distinct mutually exclusive concepts should be defined to be distinct, otherwise a reasoner could assume that an individual can perfectly be an instance of both classes.

1.3.3 Ontology vs. Knowledge Graph vs. Property Graph

In computer science, two types of statements are distinguished (following [89, 93]). The *terminological component*, consisting of so called TBox statements, defines classes and properties. An example are classes and properties in an object-oriented programming language. The TBox builds the framework to express actual facts or data. In the *assertion component*, made of ABox statements, such facts or data are expressed using the vocabulary defined by the TBox. Staying with the example of object-oriented programming, instances of a class

belong to the ABox. Properties defined in the TBox are used to relate two or more instances or to connect literal values (numbers, strings) to the instance. It is often stated that TBox and ABox together form a knowledge base consisting of the vocabulary and the data. In graph-based knowledge representations, the TBox is typically an ontology and data expressed using it (the ABox) is stored in a knowledge graph together with the ontology. Ontologies are however not restricted to classes and properties. They can also contain instances of classes, in particular when these instances are part of the specification for a domain. Weather an instance belongs to an ontology depends on the scope and purpose of the respective ontology, hence there is no clear right or wrong. An example of ABox statements (instances) that are part of an ontology comes later in this thesis when a Structure Ontology is developed by the author (see Section 2.2). There, the class `CrystalSystem` is created (TBox) and also its seven instances (ABox) representing the crystal systems in three dimensions are added to the ontology. Whenever the distinction between ABox as part of an ontology and ABox as actual data representation becomes important this will be clarified using the terms ontological individuals (or instances) and data individuals (or instances). Strictly, this separation is important only for the description logic profile of the Web Ontology Language (OWL DL), because it breaks down in OWL Full where a class can also be an individual.

Another type of graph databases are labeled property graphs. Both, knowledge graphs and property graphs have nodes and edges. Property graphs differ from knowledge graphs in that values can be assigned not only to the nodes but also to the edges. Moreover, subject and object in a triple do not necessarily need to be resources. However, property graphs are lacking semantics and do not have a standardized data model. The Resource Description Framework (RDF) used for ontologies and knowledge graphs does not support these *edge properties*. One recent development in this direction is the extension RDF* [94] proposed in 2019. More and more applications support the usage of RDF* but it is not yet a recommended standard of the World Wide Web Consortium (W3C).

1.3.4 Ontology Development

Ontologies can be developed either bottom-up or top-down. Both approaches have their advantages. Bottom-up starts with what is already there and builds on that by converting and connecting. It is the practical approach to quickly be able to query and use the ontology, and populate it with existing data to build a knowledge graph. The top-down approach on the other side starts from the concepts and defines the domain semantically correct without caring for usability with existing databases. Deciding for one or the other strategy strongly influences whether and how the ontology will be used.

Before an ontology can be developed, the domain and scope of the ontology need to be set [95]. One way to define the scope is to formulate competency questions that the ontology should be able to answer. These are usually very different from the kind of questions that the respective knowledge graph – the knowledge base using the ontology – can answer. For example the ontology about crystal structures can answer the question “What is the crystal system for space group 12?” because the crystal systems and space groups are part of the ABox specification in this ontology. However the question “Which space group has cubic Silicon?” cannot be answered because it requires data about cubic Silicon, which is not part of the ontology. Some people use the term competency question for both types of questions. As this can lead to confusion, this work strictly separates them and refers to the second type as applications or use cases.

Once domain and scope are defined, one should consider reusing existing ontologies. These

can be either upper ontologies like BFO[5] (Basic Formal Ontology), DOLCE[6] (Descriptive Ontology for Linguistic and Cognitive Engineering) or ontologies covering closely related domains with an overlap of concepts. In materials science, there is also EMMO [7], the European Materials Modeling Ontology, that can serve as an entry point and upper ontology.

Now, important terms in this domain need to be identified and can then be defined as classes in a class hierarchy. The next step is defining properties (or slots) of classes and their facets. Facets include the domain and range of the property as well as value type and cardinality. In general, they determine which type of subjects and objects are allowed in a statement with this property. To be more precise, if a property $P(A, B)$ with A as domain and B as range is used on an individual a to relate it to individual b, a reasoner automatically infers that a is an instance of A and b is an instance of B. These properties can then be used to add restrictions onto classes. It is important to note that there are no relations between classes but only between individuals of these classes that are specified with class restrictions. Restrictions can be either existential (indicated often by $\exists$) or universal (indicated by $\forall$). An existential restriction of the form $A \xrightarrow{p} B$ ensures that an instance a of class A always is connected to *some* instance b of class B via p. In contrast, a universal restriction of the same form would indicate that an instance a of class A can *only* be related via p to another instance b if b is an instance of class B. Generally, there are three types of properties p: Object properties relate individuals to other individuals whereas datatype properties link individuals to literal values like strings or integers. The third type is annotation properties used for human readable labels, descriptions or similiar non-semantic annotations. The latter are ignored by reasoners.

Finally, ontological instances can be created and possibly linked using the defined properties.

A reasoner can and should be used to validate the ontology by inferring logical consequences. Reasoning helps finding inconsistencies in the ontology, so it should assist the development at all stages.

To test whether all competency questions can be answered with the ontology, they have to be formulated as SPARQL queries, as introduced in Section 1.3.6.

1.3.5 Building a Knowledge Graph

Building an ontology is only the first step in creating a semantic linked data network. The real benefit lies in *using* the ontology to express data that would have previously only been stored in traditional databases – in the best case. The vision of an ultimate knowledge graph for materials science is very similar to the vision of the Semantic Web [96]. It is to create a "giant global graph" [96] that makes the most use out of all the information spread across different databases or websites respectively.

Two distinct strategies [97] can be applied to express data as RDF. *Graph materialization* is the process of converting data from existing formats to RDF using the ontology as annotation and loading it into a triplestore. As all data are made available in graph format, SPARQL (see Section 1.3.6) can extract triples directly making this approach favorable for further processing or data analysis. In principle even a reasoner can be used to infer new triples,

[5] `https://basic-formal-ontology.org/`

[6] `http://www.loa.istc.cnr.it/dolce/overview.html`

[7] `https://emmo.info/`

however reasoning on large triplestores of ontological individuals is still very inefficient and a topic of current research [98]. Depending on the pace of database updates, a triplestore may rapidly become outdated. Frequently rerunning the transformation and reloading the triples into the triplestore may be necessary in this case as a form of synchronization. This is however very costly in terms of time, memory and CPU. A lot of storage space is occupied by a triplestore due to the heavy RDF format.

On the other hand lies the *query rewriting* strategy utilizing rule based mappings to fetch data from the database at run-time. Run-time here refers for example to a user query. The graph-based representation of only query-relevant data is then realized. Heterogeneous data sources are queried as virtual RDF graphs by a rewritten query using a declarative rule language such as RML – the RDF Mapping Language or R2RML for relational databases. This approach ensures that always the most recent data are fetched and it scales better with large data sets. When complex data analysis is required, performance decreases quickly due to the created overhead.

The use and application of such knowledge graphs include semantic search, context-related recommendations, data validation and transparency due to data provenance. Promising is in particular the idea to discover new interesting information by using network analysis and visualization tools.

1.3.6 Accessing Ontologies and Linked Data with SPARQL

As semantic query language for databases, SPARQL is the means of choice. It resembles SQL in large parts with the difference that is designed for graphs. Four different request types are summarized in Table 1.1. Apart from those, also update requests in form of INSERT statements are possible for triplestores. Typically, a SPARQL query is sent, e.g. via HTTP, to a SPARQL endpoint that is able to process the query. Federated queries over multiple SPARQL endpoints are possible as an extension using the `SERVICE` keyword. Parts of this query are then sent to a remote endpoint. This is an extension that is standardized and widely used Multiple other extensions exist whose availability depends on the SPARQL implementation. For example, SPARQL supports only very simple mathematical operations. Filtering for more complex quantities whose computation requires some post-processing is therefore not possible. A solution to this can be the use of custom functions. Most SPARQL implementations allow for the definition of such custom functions but the language and implementation varies a lot between providers. For really complex post-processing a mixture of SPARQL with other frameworks might be the better choice.

Let us quickly review the most important aspects and keywords in a SPARQL request to better understand the examples snippets in the following chapters. Every SPARQL request has a `WHERE` clause, which contains the graph pattern (or triple pattern) to match. Such

Keyword	Query Purpose
`SELECT`	Extract data like raw values (Table, CSV, JSON)
`CONSTRUCT`	Construct a new graph (RDF) from information retrieved
`ASK`	Answer simple true or false questions
`DESCRIBE`	Extract RDF graph about information related to one resource

Table 1.1: Request types in SPARQL.

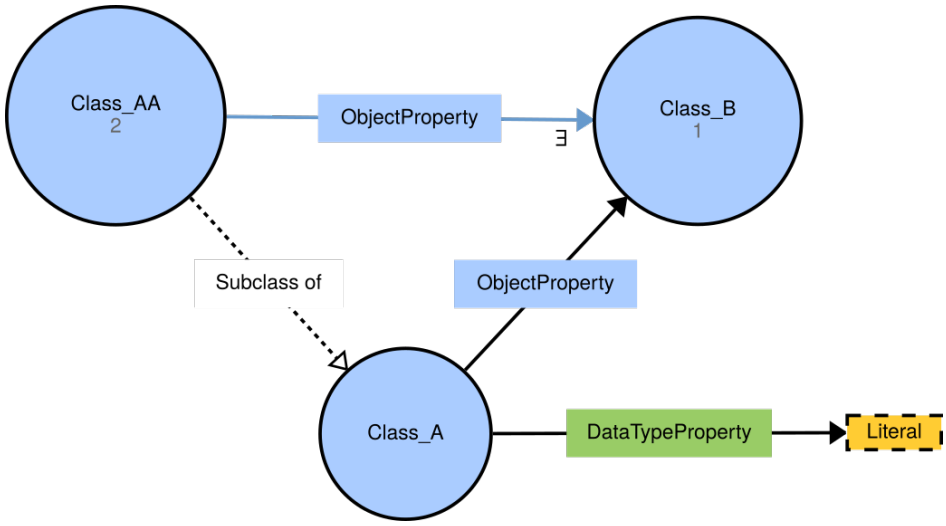


Figure 1.3: Visualization of an example ontology.

patterns look like one or more normal RDF triples where subjects, predicates and objects can be replaced by variables (starting with "?"). Matching triples in the queried triplestore or graph database are then returned. Before the `WHERE` clause stands the query clause, i.e. one of the keywords in Table 1.1. In a `SELECT` query, the variables that make up the headers of the requested table are given. For a `CONSTRUCT` query, another graph pattern is required that is then created as output. Prefixes for names spaces can be defined at the very top of a SPARQL request and used to shorten full URIS. Additionally, results can be ordered or grouped by specific quantities using the `ORDER BY` and `GROUP BY` keywords below the `WHERE` clause. Queries can include sub-queries that are handled first and whose results can be used in an outer query. The `BIND` keyword binds values such as strings, URIs or numbers to a variable.

1.4 Visualization and Interactive Exploration

1.4.1 Ontology Visualization

In the graph-based visualization of ontologies classes are usually depicted as circles or ellipses. Instances of these classes that might exist within the ontology are typically not shown at all. Sometimes only the number of instances is indicated by a small figure and the circle size is then scaled by this number as in Figure 1.3. In this work, we use VOWL [99], in particular WebVOWL[8], for ontology visualization. Hierarchical structures are presented using the `rdfs:subClassOf` property with dashed lines. This ontology example graph also shows the definition of an object property as black arrow: Class A is the domain and `Class B` is the range of this property. In the definition of the subclass `Class AA` this pre-defined property is then used in a restriction stating that for each instance `aa` there exists (∃) some instance `b` that is related to `aa` via `ObjectProperty` indicated by a blue arrow. Finally, a datatype property with a literal in its range is visualized. Such literal values are shown as rectangles. This convention will be used throughout the remaining thesis.

Unfortunately, conceptual modeling is not restricted to such easy examples. Nested statement

[8] `http://www.visualdataweb.de/webvowl/`

and nested restrictions are often needed to correctly describe what a concept means. If in the above example `b` would only be related to `a` if it is at the same time an instance of `Class C` (not shown), this could not be depicted anymore. In fact, VOWL does not even show half of this statement but ignores it completely. Visualizing these combined restrictions does not seem to be the focus in the community as there are no attempts to solve this issue yet.

For large ontologies like the DBPedia [100] or the Gene Ontology [85] this graph-based visualization is not practical anymore and only implemented for local user-driven exploration around a concept of interest. Web-based interfaces with search engines and tabular views of definitions are used instead.

A recent comprehensive overview of ontology visualization methods and tools is given in [101].

1.4.2 Knowledge Graph Visualization

Visualizing ontologies is already a challenging task when complicated statements are involved. Because an ontology is in its structure more complicated than a knowledge graph, it should be easier to visualize the latter. Considering the data character of knowledge graphs however it becomes clear that often an even larger amount of individuals and therefore nodes are involved. Not only can this mean that the graphical rendering of knowledge graphs becomes performance-critical but such visualization would also be impossible to interpret for a human. Extracting valuable information is the most important goal of visualization and a large, densely connected network with multiple different node- and edge types is not always the best choice for this.

In general, it is a good idea to follow Shneiderman's visual information seeking mantra [102]: "Overview first, zoom and filter, then details-on-demand". As discussed, a graph-based overview is not always practical or applicable to knowledge graphs. Zooming and filtering can be translated to searching for keywords of interest, e.g. via SPARQL. Details-on-demand means further exploring related objects and properties with more complex SPARQL queries including filters and possibly visualizing the query results as simplified networks (see Section 1.4.3). A network-based navigation tool for material databases and similarities was proposed only very recently in [103].

Often users' questions are directly expressible within SPARQL so that no visualization is necessary and a simple text output is sufficient to answer the question. A comprehensive but not up-to-date overview of Linked Data visualization tools is given by Dadzie and Rowe [104]. General guidance how to develop visualization tools for linked data is given by the Linked Data Visualization Model [105].

1.4.3 Complex Networks and Network Analysis

Networks, sometimes also called graphs, are a collection of nodes and edges[9]. In a complex network the collective behavior differs substantially from the one of individual components. The field of complex network analysis studies the properties, behaviors and structures of such networks. Social networks are most often used as real-world examples but also neurons in our brain forming neural networks or communication infrastructures are just a few examples showing that network structures exist in almost every imaginable field.

[9] Nodes are sometimes called vertices and edges may be called links.

Probably the first attempt to utilize the methods of complex networks and network analysis in the Materials Sciences was called "Materials Cartography" where fingerprints based on their bandstructures and densities of states were calculated [106]. Investigating network dynamics over time is another interesting idea recently used to predict materials discovery and synthesizability [107]. A third example how networks have recently gained interest in Materials Science is the representation of a generalized n-dimensional convex hull as a phase stability network of all inorganic materials [108].

The main difference between a network and a knowledge graph is that in networks usually only one type of nodes and one type of edges is present. A knowledge graph can be interpreted as a directed multi-dimensional multi-mode graph. When two types of nodes are present, a network is called bipartite or two-mode network. Similarly, k types of nodes the network is k-partite or a multi-mode network. Multi-dimensional refers to the presence of more than one edge type. For convenience however such a network is often projected onto a one-mode network to be able to use existing network analysis algorithms. A more in-depth introduction to networks can be found for example in [109, 110, 111].

For an individual node, its *degree*, namely the number of edges connected to it, is an interesting measure. A high degree can indicate particularly influential behavior. The number of incoming edges, the so called in-degree and vice-versa the out-degree can also be considered separately. A network can form clusters of nodes: *communities* have more nodes within its group then to other groups and *cliques* are defined as a collections of nodes where each pair of nodes is connected. An important measure to detect communities is the *modularity* [112, 113]

$$Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}(\boldsymbol{e}) - ||\boldsymbol{e}^2|| \,, \tag{1.1}$$

where the $k \times k$-symmetric matrix $\boldsymbol{e}$ has elements e_{ij} that represent the fraction of edges in the network that connect two nodes in the communities i and j, so that an appropriate community division has a high value for its trace. Further, it is $a_i = \sum_j e_{ij}$ and $||\boldsymbol{x}||$ stands for the sum of the elements in $\boldsymbol{x}$. Global or local *clustering coefficients* can give insight about the clustering behavior of the network. Real-world networks are often *scale-free*, i.e. their degree distribution follows a power law

$$P(k) \propto k^{-\gamma} \,, \tag{1.2}$$

where γ is positive and $P(k)$ stands for the fraction of nodes with degree k. Another type of network is the *small-world* network where the neighbors of two random nodes are likely to be neighbors too and therefore the number of edges between any two nodes is relatively small.

These are just a few examples of network characteristics. Analyzing a network requires choosing the most appropriate features to ensure interpretability, which is a challenging task given the huge amount of algorithms available today.

1.5 Current Status in the Materials Sciences

In the last years, several databases and repositories have been created and maintained with different focuses [2, 114], like the Materials Project [4], the Materials Cloud [115], the Open Quantum Materials Database (OQMD) [29], the Theoretical Crystallography Open

Database (T-COD) [116], the electronic structure project (ESP) [117], the Open Materials Database [118] and AFLOW [3], to name some important ones. Meta-data schemes exist for some of them to annotate data and make it accessible through an application programming interface (API). However, most of these repositories do not store the full input and output data, thereby not ensuring data provenance. Additionally most databases are designed for only one specific computer code. The approach of the OPTIMADE consortium is providing an API for a subset of common meta-data items that is independent of the specific models of each database [119]. Another initiative is CECAM's electronic structure library (ESL) with plans to develop the electronic structure common data format (ESCDF) [120]. As a flexible hierarchical data structure for materials storage the physical information file (PIF) has been suggested [121]. OpenKIM is a knowledge base for interatomic models, simulation codes and necessary reference data [122].

There is also a number of attempts to create an ontology for materials sciences. Some of them are very specific domain ontologies, like PLINIUS [123] for ceramics or an ontology engineering materials [124]. More general ontologies are MatOnto [125], PREMΛP [126], Materials Ontology [13], MatOWL [127], Materials Design Ontology [12] or the European Materials and Modeling Ontology (EMMO)[128]. A more detailed overview of some related ontologies can be found in Appendix B. This thesis focuses on EMMO as an upper ontology for the materials sciences (see Section 1.5.1) and uses it as framework for the newly developed ontologies.

In the following two sections, the EMMO as well as the different aspects of NOMAD are explained in more detail because they build the foundation for any work presented in the next chapters.

1.5.1 European Materials and Modelling Ontology

The European Materials and Modelling Council aims to develop a standard representational framework for the applied sciences, the EMMO [128]. It combines theories from physics, philosophy, and information and communication technologies. Starting from the concept of space-time, it allows in principle to define everything. Any real world object in EMMO extends in space and time. All relations are classified according to the following four primitive families:

1) **Taxonomy** defines the classification. An example is the `subClassOf` relation (subclassing in the sense of "is a").

2) **Mereotopology** includes Parthood and Slicing. For example the `hasPart` relation that defines the components by which an object is built from.

3) **Semiotic** is the branch of representational relations like `standsFor` or `hasProperty`.

4) **Set theory** is the theory of membership. A collection is made of unrelated items and links to them using the `hasMember` relation.

Most important for real world descriptions is the parthood relation, which can be subdivided into different kinds: A component can be a *spatial*, *temporal* or *spatio-temporal* part of an object. It is called a *direct* part if there is no additional parthood layer in between. A part is *proper* if it is smaller than the object itself. This differentiation makes it possible to provide a full description of the construction of complex objects. In many cases the use of the simple `hasPart` property may be sufficient.

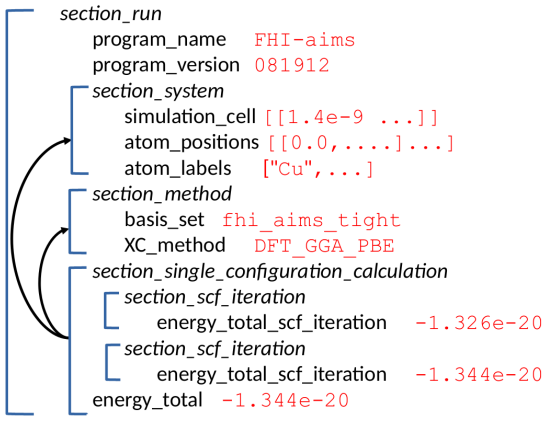


Figure 1.4: Extract from the NOMAD Metainfo. Black terms are meta-data and red values represent data. Sections (italic) structure quantities (roman). Black arrows indicate a reference relationship. Based on a draft courteously provided by Fawzi Mohamed.

The philosophical nature of EMMO can be illustrated at an example: Processes are defined as temporal parts of anything that stands for real world objects. A subclass is semiosis, the process that has the participant `Intepreter` that produces a `Sign` to represent another participant, the `Object`. For example, density functional theory is an `ElectronicModel`, which subclasses `Icon` and is therefore a `Sign`. As such it is a participant of some semiotic process. Although this might be the correct philosophical classification using Peirce's sign theory, it seems too complicated and unnecessary for any physical applications. EMMO is a pure TBox ontology, hence there are no ontological instances. Even the mathematical concepts of `Number`, `Integer` and `Real` are defined as classes although for numerical data other ontologies typically use datatype properties in connection with built-in datatypes for the values. The benefit of having all these classes is unclear and no applications exist in which this has proven useful. Thus, EMMO will be only used as an orientation and its principles will be followed where it is deemed reasonable.

1.5.2 NOMAD: Repository, Archive and Metainfo

With the Novel Materials Discovery (NOMAD) Repository a common place for full input and output storage was created supporting over 40 different codes. To date over 107 million total energy calculations are stored in more than 11 million entries. All data is normalized and accessible in a unified, code-independent form in the NOMAD Archive. It also includes original data from AFLOW, OQMD and Materials Project databases.

An API is available to explore both, data as well as meta-data that are used to annotate and structure data. This particular meta-data schema is called the NOMAD Metainfo [2] and annotates and structures data from electronic-structure theory or force-field calculations. As of July 2020 a new version of the Metainfo is available where the definitions are not stored in JSON files anymore but as python objects and are directly accessible through the API. A few concepts were also renamed, which is why the following explanation differs terminologically slightly from [2].

There are four different types of meta-data in the NOMAD Metainfo:

1) **Quantities** are the labels to the values (strings, scalars, vectors, ...) that are parsed by the parsers. In a relational database, these are the headers to the columns in a table.

2) **Sections** represent the different parts of a computer simulation (e.g. a section for the simulated system, a section for the simulation method). Sections can *contain* other sections (a) and quantities (b), so the sections build an hierarchical structure. Sections can also *refer* to other sections (c). Sections would be the tables in a relational database.

3) **Categories** are meta-data for the meta-data. They *describe* the type of data labeled by a "Quantity" or contained in a "Section" (d). For example, different energy values may all be related to the same abstract type "energy". Categories are also nested and build a hierarchy (a).

4) **Dimensions** are themselves "Quantities" but with the additional property of being integers that define the length of one dimension of another non-scalar "Quantity" (e).

This means, the NOMAD Metainfo already contains 5 types of relations between the meta-data: (a) *is subclass of*, (b) *is part of*, (c) *has reference*, (d) *has category* and (e) *has dimension*. Following the discussion in Section 1.1 meta-data (in particular quantities) and data are comparable to keys and values in a key-value list respectively. Metainfo's quantities are however rich objects with attributes like description, shape, units, datatype, category and an affiliation to a section. Figure 1.4 illustrates the difference between meta-data and data and shows how sections structure the quantities. The same figure also names the most important sections. The upper `section_run` contains everything related to a single program run. Direct children are for example the quantities `program_name` or `program_version`. It also includes subsections like `section_system` that stores information concerning the simulated system (molecule, crystal, ...) and data to construct the system like atomic positions and lattice vectors of a crystal unit cell. The `section_method` groups meta-data related to the calculation method, e.g. the basis set or exchange correlation functional in DFT calculations. Finally, `section_single_configuration_calculation` is the output of a calculation for a specific configuration of the system. It can be viewed as the result section containing for example total energies, densities of states, band structures, forces and more.

The NOMAD universe fulfills the FAIR principles from Section 1.2 and is constantly improving on it even further. For example, in this thesis especially improving findability and interoperability is addressed using semantic technologies. Proving a rich human- and machine-readable meta-data schema – the Metainfo – and unique paths (similar to URIs) for each term and data item ensures findability (F). An API and search resources connected with the NOMAD Archive make data and meta-data accessible (A). Using widely adopted formats and languages like JSON and python improves interoperability (I) with external services. Storing full input and output of the calculations as well as annotating *all* data and not only the important parts for a specific application makes data better re-usable/re-purposable (R). The ontologies developed in the following Chapter 2 aim to further refine interoperability.

Chapter 2

The Ontological Baseline for Materials Representation

This chapter introduces three ontologies that have been developed for Materials Science. They build the semantic baseline for representing materials. Their layered structure is visualized in Figure 2.1 as a stack with the most fundamental ontology "Core" at the bottom providing general basic concepts.

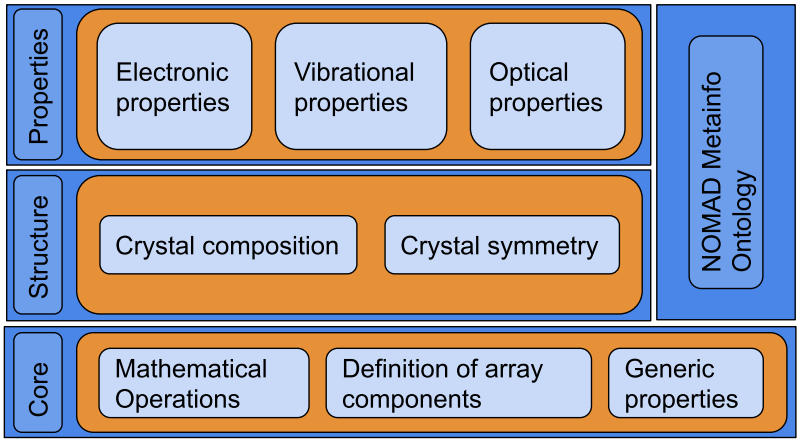


Figure 2.1: The NOMAD Ontology stack shows the hierarchy of the developed ontologies. At the bottom is the core ontology providing all basic concepts that are missing in EMMO. Building on this the top-down Structure and Properties ontologies are developed. Parallely, the NOMAD Metainfo as bottom-up ontology accompanies the two (Chapter 3)

2.1 Core Ontology

Before being able to describe materials structures and properties we need to identify several basic concepts that are important for the definition of more specific materials-related concepts. The concept of arrays and their representations is particularly important. Another example are basic mathematical operations and relations allowing us to assign values or describe functional dependence between physical quantities are needed for a correct semantic descriptions of the physics.

2.1.1 Representing Arrays

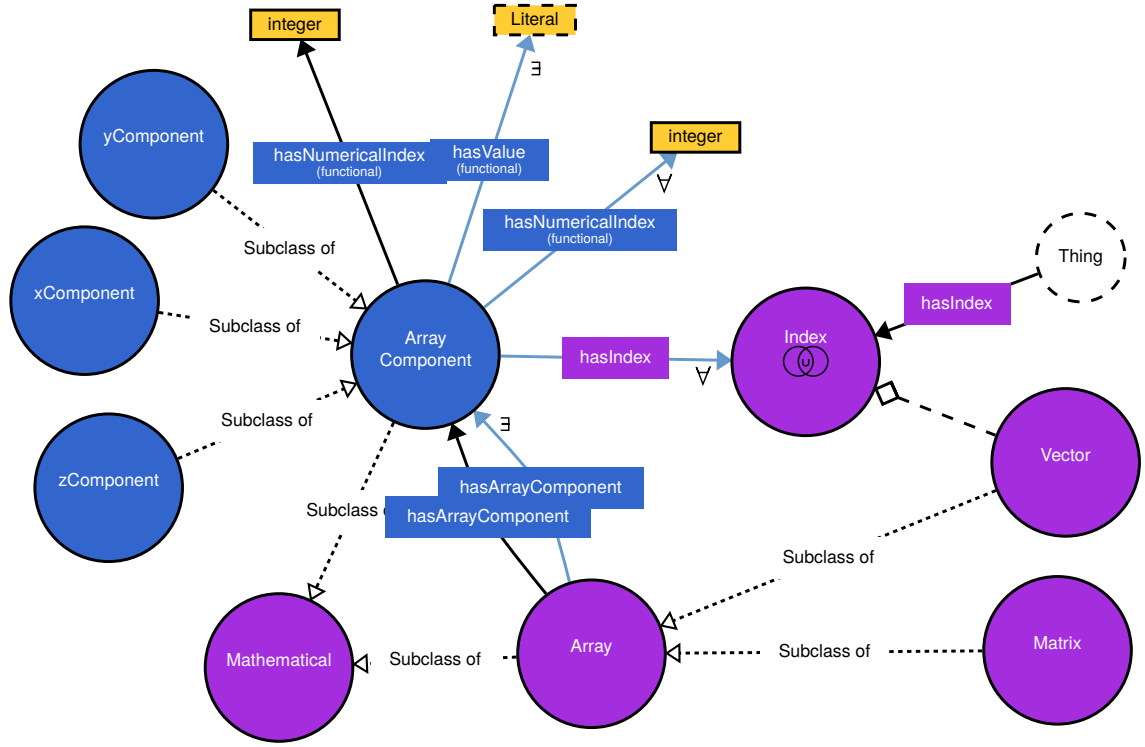
How to represent arrays in an ontology is a still unsolved issue in ontology research with several approaches. The easiest class able to represent a numerically ordered collection of items is the RDF Sequence container (`rdf:Seq`) which can be used together with container membership properties `rdf:_nnn` where `nnn` is a positive integer with no leading zeros. An RDF List (`rdf:List`) follows a different approach. It is a linked nested list where the first item is pointed to by the `rdf:first` relation and the rest of the list by the `rdf:rest` relation. This way multiple nested lists need to be accessed to read an item in the middle of the list. Only when the whole list is read in sequential order, this concept does perform well. For a general array definition where regular access to single items is required, large RDF Lists can quickly become a performance bottleneck because many lookups will be required. The *Ordered List Ontology* [1] was developed to represent music playlists or track lists on records. A `olo:OrderedList` has a `olo:Slot` which in turn has a positive integer as index and an item associated with it. Items are any resources and can also be related to their previous and next items. An interesting approach is the *RDF Data Cube Vocabulary* [2] which is very suitable for tables where each column and row have a specific meaning that is also an ontologically defined concept. Practically, this is not always feasible. All of the above methods have in common that the representation of higher-dimensional arrays is very complicated. Deeply nested structures would be created which makes querying for specific items costly. An ontological representation of arrays in computational materials science requires highest flexibility in these terms because user's requests can not be foreseen and depend on the use-case.

Therefore a simple but effective way to express arrays in an ontology was developed in the Core ontology as shown in Figure 2.2a. Each array has at least one array component indicated by the existential ($\exists$) class restriction `hasArrayComponent`. An array component cannot itself be an array, it represents an elementary part of an array. It must have a value that is not an instance of another class but a *literal* value like an integer or a string. To identify its position within the array, each array component additionally has either a numerical index or a multi-dimensional index which is a vector and thereby itself an array. The often used x-, y- and z-components of vectors in three-dimensional space are special subclasses with predefined numerical indices. In this model it is easy to access a single item anywhere in a multi-dimensional array. A draw-back of any ontological array representation discussed so far is the required storage space. Because every single item, its value and index are annotated and written out explicitly in RDF, the file size can be more than an order of magnitude greater than saving it for example in JSON format. Having a native way of storing numeric arrays via a dedicated data-type in the same manner as integers, strings, decimals are defined in XML Schema[3] would be preferable.

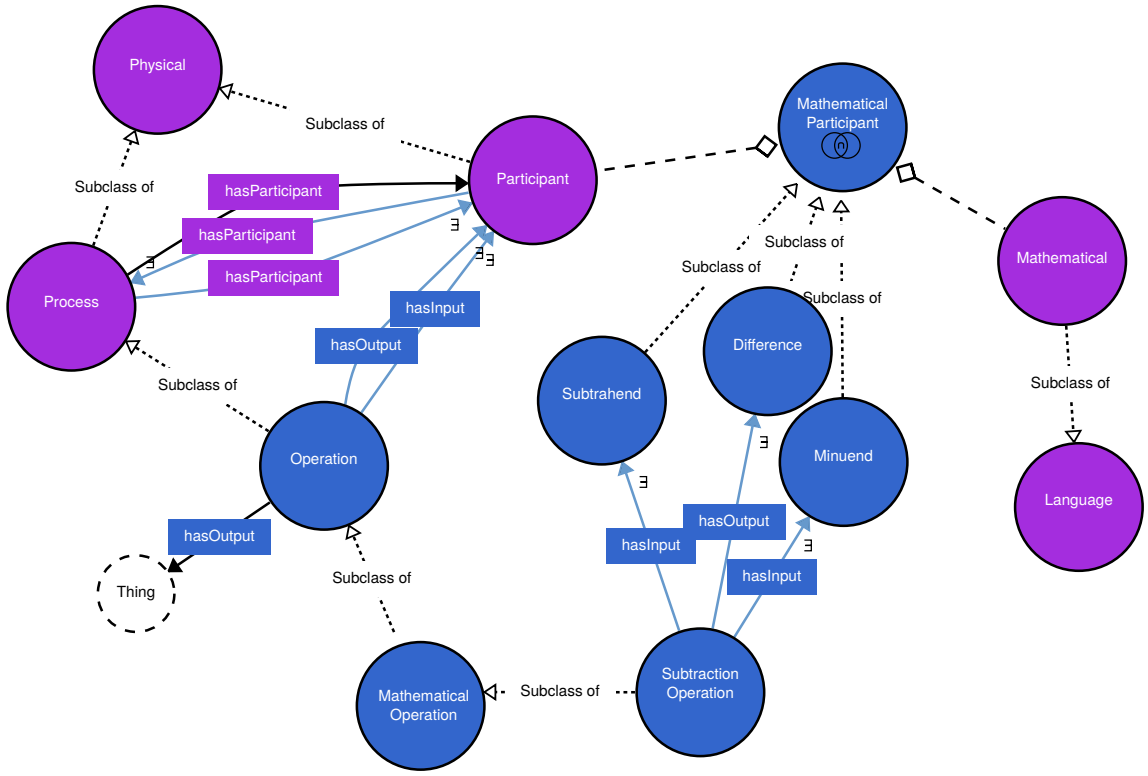
2.1.2 Mathematical Operations

Semantically expressing physical quantities is only possible when mathematical relations are taken into account. One prominent example is the band gap of semiconductors and insulators which is an energy difference. Defining the mathematical concept of subtraction is therefore needed to express how the band gap is related to the energy bands. Existing ontologies on mathematics have focused on the classification of math concepts like in $OntoMath^{\text{PRO}}$ [129].

[1] `http://purl.org/ontology/olo/core#`
[2] `https://www.w3.org/TR/vocab-data-cube/`
[3] `https://www.w3.org/TR/xmlschema-0/`



(a) Array definition



(b) Algebraic operations at example of subtraction

Figure 2.2: Extract from the Core ontology inheriting from EMMOs `Mathematical` branch. Purple concepts are from EMMO, blue are new in the Core ontology. Black solid lines represent property definitions while blue solid lines denote class restrictions using ∃ and ∀ for existential and universal restrictions respectively.

Using these concepts to connect real-world physical quantities is however entirely new and no appropriate maths ontologies or similar attempts could be found in literature. Therefore, the Core ontology includes concepts for basic algebraic operations like addition, subtraction, multiplication and division as shown in Figure 2.2b at the example of subtraction. Making these classes equivalent to the respective classes in the mathematics ontology $OntoMath^{\text{PRO}}$ would be a strong statement. Any future changes in either our or this external ontology like additional relations to other ontologies can possibly lead to reasoning issues. Therefore the mathematical concepts in our ontology are defined as subclasses of their existent counterparts in the external math ontology so that only the Core ontology is affected if something changes. The four operations are defined in a very general way as mathematical operations with an input and an output. More complicated operations like derivatives or Fourier transforms can be defined in the same way. Even though this approach is machine-readable, it is not machine-actionable, that means the computer does not know how to calculate the difference or Fourier transform. OWL ontologies do not allow for this kind of operational definitions. Theoretically, it would be possible to develop a software similar to a reasoner that interprets the individuals of these mathematical classes and calculates the results. In practice however, these are still dreams of the future.

2.1.3 Generic Properties

Two philosophies can be followed in ontological modeling:

1) It is possible to define each property with a very specific domain and range, so that it can only ever be used on the same type of instances. It has the advantage that by simply looking at the property an instance's type is obvious. However, this might require a very large amount of properties that are principally very similar. For example the subtraction operation from Figure 2.2b could be equally defined using sub-properties `hasMinuend`, `hasSubtrahend` and `hasDifference`.

2) Another approach is to keep the number of properties to a minimum. This way a user querying a knowledge graph does not need to know the specific sub-property and can still request a certain type of instance by adding this restriction to the instance itself. Such models tend to have a higher re-usability.

Here, simplicity and re-usability are considered particularly important, so that the second method is chosen throughout the whole ontology modeling process. The list of required properties has grown during the iterative ontology development. Whenever possible without introducing too much reasoning complications, the EMMO properties hierarchy was used to classify new properties into one of the four property types from Section 1.5.1. The most important property is the data-type property `hasValue` which can be used whenever an object's value is specified regardless of its data-type. In the ontology and knowledge graph community such generic property is rarely used because each value usually has its own data-type property. In materials science, a single value can represent a complex concept that is better defined as an instance of a class (e.g. a total-energy value). Table 2.1 lists newly created object properties that have been identified as necessary in an alphabetical order. Sometimes the semantic difference between two properties is small: `characterizes` and `determines` seem similar but while the first refers to a characterization process that characterizes a material in a particular state, the second is applied to properties of a material that are determined by a specific technique. When a physical quantity `isFunctionOf` another quantity, this refers to a mathematical relation whereas if it `dependsOn` another object this can mean something more general.

Table 2.1: Alphabetical list of newly defined generic object properties in the Core ontology.

characterizes	hasClassification	hasOutput
dependsOn	hasImaginaryPart	hasProjection
determines	hasInput	hasRealPart
isDerivedFrom	hasInverse	hasRepresentation
isFunctionOf	hasLabel	hasStatistics
isInverseProportionalTo	hasMeanValue	resultsFrom
isProportionalTo	hasMeasure	refersTo
isSpannedBy	hasMethod	usesModel
hasArrayComponent	hasNumericalIndex	

Of course this list is not complete. There are many more generic property concepts that could be included. The listed ones are however sufficient for all applications within this work.

2.2 Structure Ontology

The second layer in the ontology stack (Fig. 2.1) is the structure ontology that was developed mainly for crystals – the workhorse of solid-state materials science. Aiming to represent crystals is therefore the logical first step for any materials science related ontology development. While a number of different dictionaries with standardized nomenclature for most crystallographic concepts is provided in the International Tables of Crystallography [130] (Crystallographic Information Framework, short CIF, dictionaries), no Crystal ontology has established yet. Only recently, the EMMO efforts are being extended to create an ontology based on CIF which is still work-in-progress. As described in Section 1.3.4, competency questions can be formulated to restrict domain and scope of an ontology. The following questions are examples for what should be answerable with the structure ontology. The list is by no means complete but gives a good overview of the capability of the ontology.

- *What is a crystal made of?*
- *How can a crystal be represented?*
- *What is the crystal system for space group 12?*
- *How many different crystal systems exist?*
- *Which concepts lead to the lattice system classification?*

The first two questions describe what a crystal is and how scientists talk about it. For the last three questions the definition of symmetry concepts is required. To further clarify which concepts should be defined in the ontology, a word list was collected with most common terms used in this domain (Table 2.2). Sometimes there are several words that describe the same underlying concept. An example is the “basis” which is just another word for the atoms and their positions in the crystal unit cell. Once the words and the concepts behind those are set, relations between them can be defined. Not all concepts from the word list are included yet in this ontology, in particular more in-depth discussion is needed to formalize concepts like surface or defect.

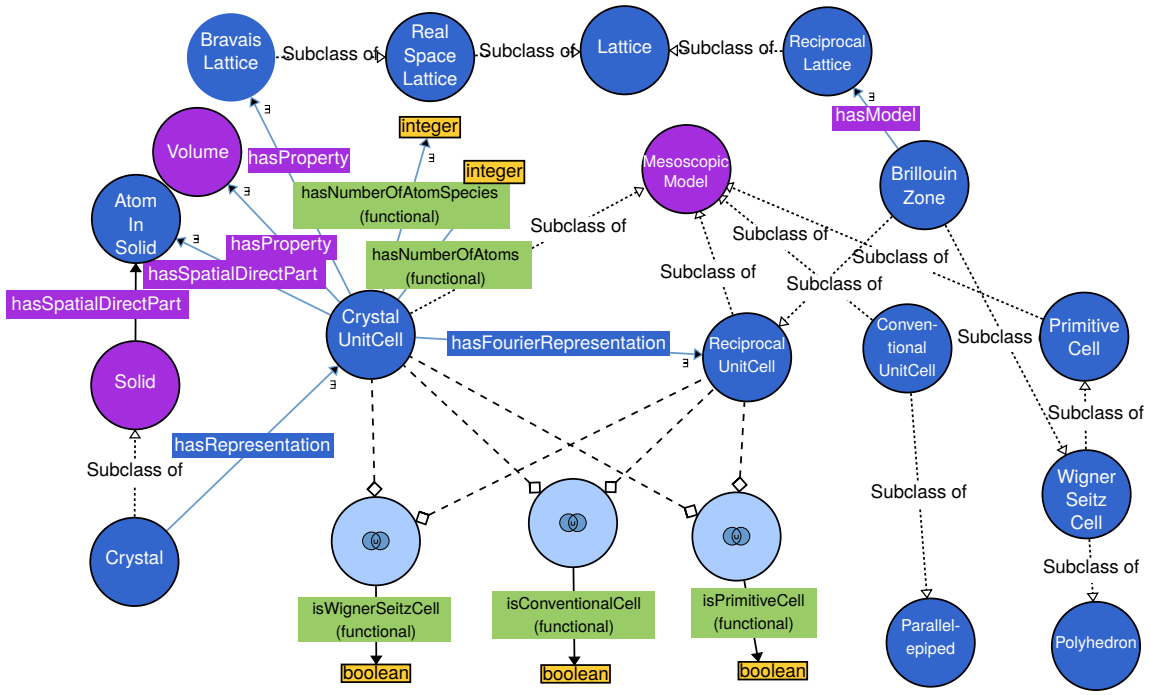
Table 2.2: Word list for structure ontology (incomplete), not alphabetically ordered, but neighborhood might suggest semantically connected concepts.

Crystal	Atom species	Crystal symmetry	Wyckoff multiplicity
Crystal unit cell	(Chemical) element	Crystal point group	Wyckoff letter
Conventional unit cell	Bravais lattice	Space group	Wyckoff site symmetry
Primitive unit cell	Brillouin zone	Crystal system	Hermann Mauguin
Super cell	Periodic boundary conditions	Lattice system	Surface
Lattice	Crystal structure	Crystal family	Boundary
Lattice vector	Structure prototype	Lattice centering	Interface
Basis	Reciprocal space	Wyckoff position	(Point) Defect
Atom postions	k point	Schönflies symbol	Distortion
Fractional positions	High-symmetry point	Pearson symbol	Vacancy

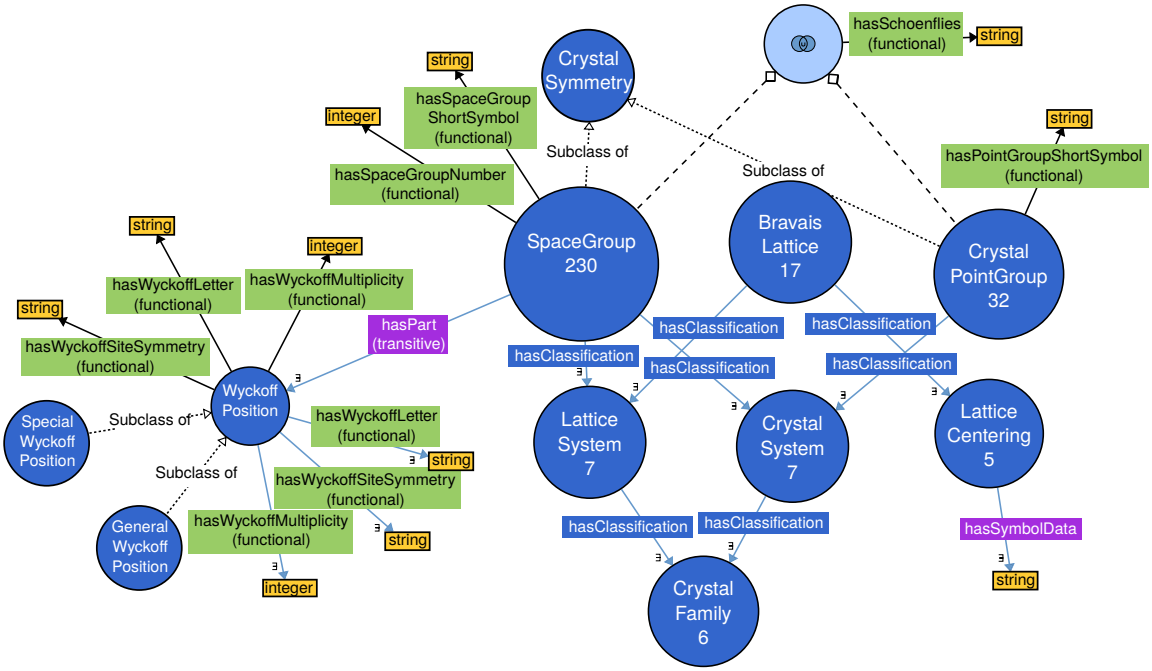
2.2.1 Crystal Unit Cell as Representation of a Crystal

In computational solid state physics, crystals are almost always represented via their crystal unit cells. One could say that a crystal unit cell is part of a crystal. However this parthood relation brings multiple incorrect implications: If a crystal unit cell is a spatial direct part of a crystal, then the atoms are *not* direct parts of the crystal, but only of the unit cell. The unit cell is not unique so that multiple different unit cells can be chosen to represent a crystal. Also, as a human-made model, the unit cell should not be made existential part of a crystal. A better alternative is to introduce the object property `hasRepresentation` to connect any unit cell to a crystal (as shown in Figure 2.3a). Different classifications for unit cells and their Fourier transforms, reciprocal unit cells, can be defined. Primitive cells are defined to contain exactly one lattice point. Note, that a lattice point is not to be confused with the atomic sites because one lattice point is occupied by an atomic basis which can consist of more than one atom. If the primitive cell is constructed in a special way using the so called Voronoi or Dirichlet construction, it is called a Wigner-Seitz cell. This construction leads to a polyhedron shaped cell which has the symmetry of the crystal. The Wigner-Seitz cell in reciprocal space is the Brillouin zone. In contrast, most other unit cells form parallelepipeds and are spanned by three lattice vectors. Conventional cells also have the crystal's symmetry but are not primitive. The characterization of a unit cell into these types can happen via boolean datatype properties (`isWignerSeitzCell`, `isConventionalCell`, `isPrimitiveCell`) or subclassing (of the classes `WignerSeitzCell`, `ConventionalCell`, `PrimitiveCell`). Any instance of these classes is equivalent to having the respective datatype property value "true" (although this is not depicted in Figure 2.3a). The domain of these datatype properties is a union of the crystal unit cell class and the reciprocal unit cell class indicated in the Figure by the union symbol $\cup$. In other words, they can be applied to any instance that instantiates any of these two classes.

Just like a crystal, a unit cell contains atoms, i.e. it connects to the atom concept with the spatial direct parthood relation. In contrast to a crystal, a unit cell is characterized by a well-defined number of atoms and atom species which is included using integer datatype properties. This representation is visualized in Figure 2.3a. The volume of the crystal unit cell, a measure often used to compare different structures, can be modelled using the predefined `Volume` class in EMMO.



(a) Representation of the crystal unit cell.



(b) Crystal symmetry related concepts.

Figure 2.3: Excerpts from the Structure ontology developed in this thesis. Purple classes and properties are inherited from the EMMO. The small white numbers in the blue nodes indicate the number of ontological instances that exist for this class.

Query 2.1: WHERE part of a SPARQL Query for structural prototypes, their spacegroups and crystalsystems as defined in the structure ontology (CSO).

```
WHERE{
    ?proto a cso:StructurePrototype .
    ?proto cso:hasSymmetry ?spacegroup .
    ?spacegroup core:hasClassification ?crystalsystem .
    ?crystalsystem a cso:CrystalSystem .
}
```

2.2.2 Crystal Symmetry

Another very important aspect to consider when representing materials is symmetry. Apart from the computational advantages in exploiting symmetry, these concepts are often used to search and find specific groups of materials – in particular because a material's structure strongly influences its properties. Figure 2.3b shows how the different classifications like space group and crystal system connect to each other. Besides the class definitions forming the TBox, this ontology also contains ontological instances which build the ABox. Here such ontological instances make sense from a semantic perspective: Each space group can exist only once, it is the same every time it is being used to describe a crystal structure. An example for such an individual is the space group with the short symbol Pn-3n and the international number 222 that is classified to belong to the cubic lattice system and crystal system. In an ideally linked data world, any crystal in any database would refer to the respective individuals in this ontology and not only saving the values as strings locally. This would create a world wide net of linked crystal structures and enable searches across domains and resources.

2.2.3 AFLOW Prototypes Knowledge Graph

The AFLOW Library of Crystallographic Protoypes [31, 32, 33] has already been used in Part I of this thesis to perform a high-throughput study across many different spacegroups in Section 3.2. Until 2021, three editions of this library have been published featuring 1100 different structural prototypes. Within this thesis, the author creates a knowledge graph representation of these prototypes using the freshly developed Structure Ontology. First, RML-based mappings are written to transform the tabular overview of the library from JSON format[4] to an RDF format using the Core and Structure ontologies for annotation. For symmetry information, it is sufficient to map the space group to the ontological space group instances. This already allows access to other information like crystal system and lattice system as stored in the ontology. How access to the ontological crystal system instance is gained with a SPARQL query is demonstrated in Query 2.1 showing the WHERE part of this request. For an overview of all prototypes and how they are distributed across the different crystal systems a network-based visualization in Figure 2.4 is created using Gephi [131]. Including information about the number of atoms and atom species per unit cell is possible for example through node sizes and colors as it is done here. For smaller networks an additional layer of information can be encoded in the node shapes. From the network plot it is obvious that only few prototypes belong to the triclinic crystal system covering only the two lowest symmetry space groups. Another fact that can be extracted from this figure is that prototypes in

[4] `http://aflowlib.org/CrystalDatabase/js/table_sort.js`

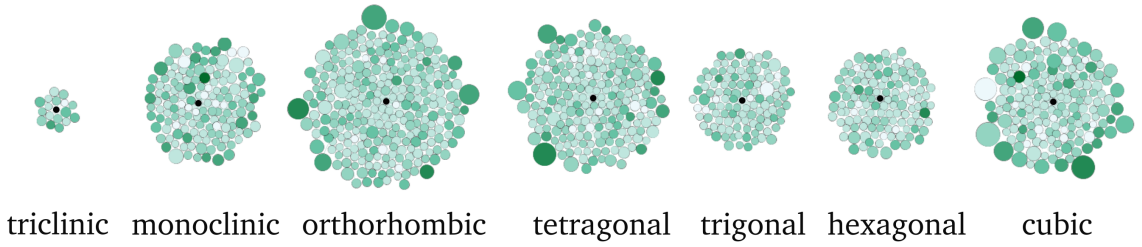


Figure 2.4: Network representation of AFLOW Prototypes (green nodes) in all seven crystal systems (small black center nodes) visualized from the knowledge graph created in this work. Prototype nodes are size-scaled by the number of atoms in the unit cell and color scaled by the number of atom species where darker shades indicate more different atom species per unit cell.

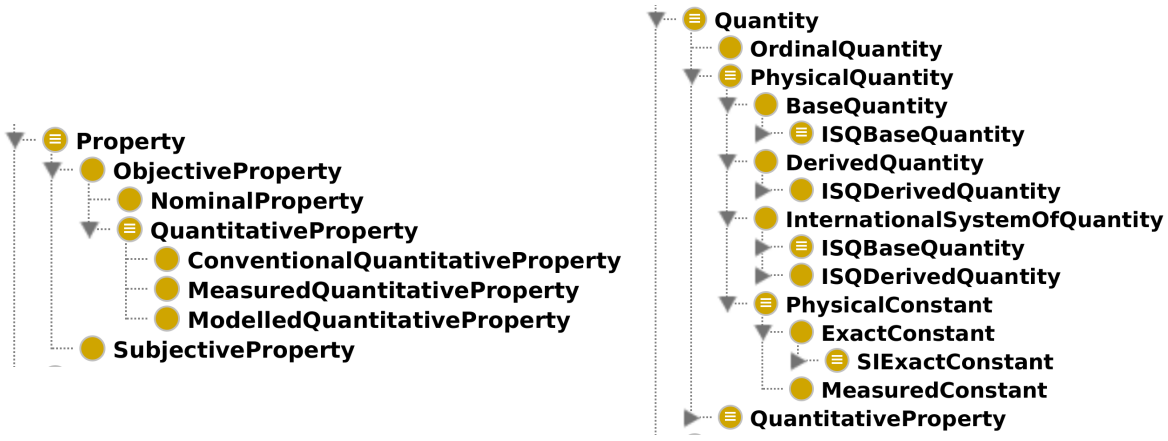
the trigonal and hexagonal crystal systems (hexagonal crystal family) have fewer prototypes with many atoms of many different species in their unit cells which would correspond to large dark green nodes. In detail, two extreme cases are the 105-atom unit cell of trigonal prototype A_hR105_166_bc9h4i is made of only one atom type, and six different atom species are used in hexagonal prototype A3BCD3E15F3_hP52_173_c_b_b_c_5c_c with 52 atoms per unit cell.

An interesting use-case would be to observe the allowed and occupied Wyckoff positions within each space group. So far the structure ontology does only include the Wyckoff positions as concepts. It requires more effort to add the allowed general and special Wyckoff positions explicitly to each space group instance. Creating the AFLOW knowledge graph, the occupied Wyckoff sites can then be added to each prototype. This allows for easy investigations which Wyckoff positions are rarely or even never occupied and which ones are predominantly used. Because the AFLOW prototypes are based on well known structures, this can indicate how to construct novel so far unknown structures that are possibly interesting for crystal structure prediction. Due to the close relationship between a material's structure and its properties, such unexplored geometric configurations might be potential candidates to discover novel phenomena or identify new materials classes.

2.3 Properties Ontology

Up to this section, this work considered only structural information of materials. Although one could think of such information as being *properties*, the structure forms a more fundamental part in defining a material compared to electrical or optical properties. Furthermore, it is possible to talk about structures and their features without having a particular chemical composition of this structure in mind. Our decision to separate structure and properties is also supported by the existence of structure libraries like the AFLOW Library of Crystallographic Prototypes [31, 32, 33].

EMMO distinguishes between properties and quantities. A quantity like $10\,\mathrm{kg}$ is only a property when it is assigned to an object because properties are always the result of an observation process. In EMMO, a property is either *subjective* or *objective* where only the latter results from a well-defined observation procedure. Furthermore objective properties can be either *nominal* if their value can not be quantified or *quantitative* in which case it can be *measured*, *modelled* or *conventional* (e.g. specified by a vendor), see Fig. 2.5a. EMMO already defines



(a) Properties (b) Quantities

Figure 2.5: Hierarchy of quantities and properties in EMMO depicted as a screenshot from the classes and object properties overviews in the Protégé ontology editor.

many basic quantities like force, energy or electric charge and classifies them according to the international system of quantities (ISQ) [132].

Now, we develop the materials-properties ontology which aims at specifying more concrete quantities as being properties of crystals, crystal unit cells or their constituents. In an ontology designed to represent materials properties there is no need for quantities that are not properties. A quantity related to an object via EMMO's `hasProperty` relation will automatically be inferred to be a quantitative property when a reasoner is run due to the defined range of the relation.

2.3.1 What is a Material?

To be able to discuss properties, the definition of a material has to be clarified first. Within the last year, we have been involved in several discussion with the OPTIMADE community with the goal of building an ontology for materials databases. The very open view of what a material is, is also reflected in the formal definition of a material that has come up during these meetings: "Something that can be expressed by a materials model". Such a model is described by a specific set of parameters specifically tied to that model. Instead of modeling materials, the focus has been shifted to conceptualize and relate materials models.

A more intuitive way is to regard a material as a collection of configurations of this material. Such a configuration can be a state that exists only for a limited time (due to vibrations, fluctuations or phase transitions etc.). We will call this a *material snapshot* and it can be represented by a single geometric configuration. Modifying this configuration even slightly alters also the electronic properties. In the NOMAD Metainfo such a snapshot is called system. For each system the total energy or even the full energetic spectrum can be calculated. Some more complex properties require multiple snapshots to be measured or computed. For example, the thermal conductivity can be calculated from molecular dynamics simulations on thousand of slightly different geometric configurations all snapshots of one single material.

This is a typical use case for a materials knowledge graph. A material node (an instance of a material) in this graph would be connected to multiple individual snapshots represented by

section_system in NOMAD each of which can have simple properties like energies, atomic forces etc. calculated with specific methods. This can be used for the identification of the relaxed state of the material, i.e. the configuration with the lowest energy and vanishing forces which can be marked as representative system for this material. Complex properties can directly be connected with the materials node and the configurations it has been obtained from thereby also tracking provenance. In practice however, it is very complicated to identify entries in NOMAD that belong together after they have been uploaded unless there is a clear description and links provided by the uploader.

2.3.2 Properties Classification

Multiple different classifications for materials properties exist: Often, physical properties are categorized as being either intensive or extensive according to how the property changes when the system size or amount of material changes. Extensive properties are additive and increase or decrease with the extent of the system like mass, volume or entropy. Intensive properties do not change with the system size as for example density or temperature. Volume and mass of the crystal unit cell are structural properties which belong to the structure ontology. Division of two extensive properties usually gives an intensive value: mass divided by volume yields the density. Moreover, there exist properties that fall in neither of these two categories. Therefore, this distinction will not be made in our ontology because it would only complicate the class hierarchy. Another way to categorize properties is to assign them to domains: One distinguishes mechanical, electrical, optical and thermal properties. More domains like magnetics or acoustics as well as manufacturing-related concepts can be defined. A clear distinction is not straight-forward for all properties, so that often one property belongs to multiple classes. Especially when a material is used in the engineering industry, the microscopic and macroscopic levels are separated. Our ontology incorporates this view via materials models and classifies for example the crystal unit cell as emmo:MesoscopicModel and a crystal structure as an emmo:AtomisticModel.

2.3.3 Modeling the Band Structure

One of the most important concepts used by materials scientists to describe the electronic structure of a periodic solid is the energy band structure. Even though it is a simplified model of the real energetic structure in a material, it is used extensively and will therefore be included as a property of a crystal snapshot. The band structure is made up of energy bands that are formed by energies with the same band index and continuously varying wave vector $\mathbf{k}$. Following the picture a material scientist has in mind when thinking of a band structure , the bands are modeled as spatial direct parts (using the EMMO relation hasSpatialDirectPart) of the full band structure. Each band also has spatial direct parts which are the individual measured or computed energy values which make up a band.

In principle every solid material has energy bands. However only for periodic solids (crystals) the band structure can be visualized along the paths between high symmetry points in the Brillouin zone. Amorphous materials do not have a Brillouin zone. Therefore a more general concept is needed to describe the energetic spectrum of any material, such as a three-dimensional random sampling of the crystal combined with statistical methods. We will focus on crystals here because they are one of the most important solid materials types. Figure 2.6 depicts the ontological concept of the valence-conduction band gap using the isDerivedFrom relation. Whether a particular band gap, represented by a band gap instance, is a direct band

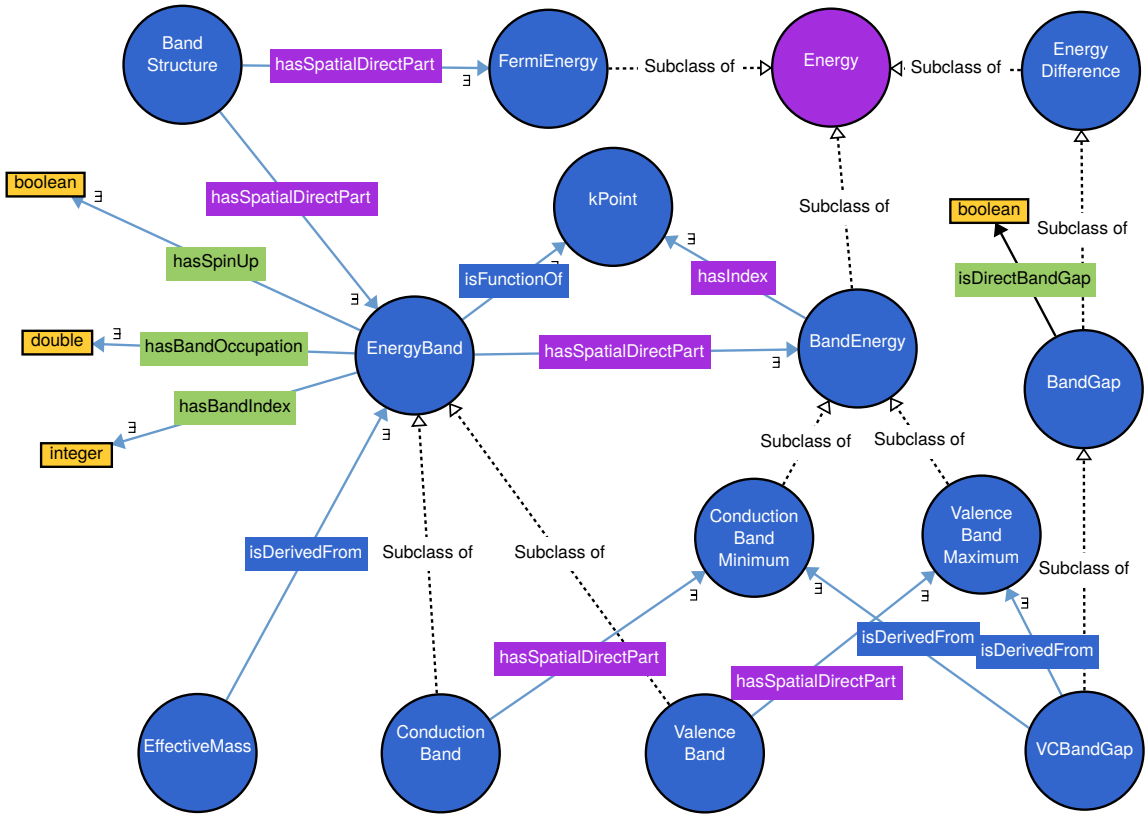


Figure 2.6: Bandstructure concept in Materials Properties Ontology developed in this work by the author. Purple classes and properties are inherited from EMMO.

Code 2.2: Nested definition statement for the valence conduction band gap as defined in the `subClassOf` field in the ontology editor Protégé.

```
inverse(hasOutput) some
  (SubtractionOperation and
    ( (hasInput some (Minuend and ConductionBandMinimum)) and
      (hasInput some (Subtrahend and ValenceBandMaximum))
    )
  )
```

gap (or not) is indicated at data level using the boolean datatype property `isDirectBandGap`. The ontology contains a more semantic description of the relation using the subtraction operation in a nested statement given in Listing 2.2. None of the known ontology visualization tools are however able to visualize these nested triples. Additional blank classes would need to be introduced making the graph too complicated. Band index and occupation as well as spin channels (up/down) have been modeled as datatype properties. Of course the spin in quantum physics is a more complex concept earning an own class. As an electronic model is developed here, it is clear that the spin is $\frac{1}{2}$ and can be oriented either up or down lifting spin degeneracy of the energy bands. In fact, a representation of the symmetry group of the wavefunction is also assigned to a band, however this is often not discussed.

The concept schema shown is equivalently valid for vibrational band structures in which the allowed energetic states of phonons are described. Instead of the k-points in momentum space, the phonon wave vector, q, is then plotted on the x-axis.

As the number of physical properties that can be defined is incredibly large, it is important to focus on a particular application or sub-field and the relevant properties therein. Data from the NOMAD Repository will be used in this thesis to build knowledge graphs, so that focusing on electronic properties is a good starting point.

Chapter 3

NOMAD Metainfo Ontology

The previous chapter introduced a collection of ontologies that was created top-down to avoid bias towards a particular database or meta-data schema. On the other hand, following the practical bottom-up approach (see Section 1.3.4) it seems natural to transform the well established NOMAD Metainfo from Section 1.5.2 to an ontology. In fact, it does already fulfill a number of requirements of an ontology: Each meta-data item has a unique identifier given by a path, a name as well as a rich human-readable description, several attributes, and its format as python object or JSON file makes it machine-readable too. Five different relations including two parallel hierarchies for the sections and the categories provide basic semantics to the meta-data structure making it superior to traditional meta-data schemas.

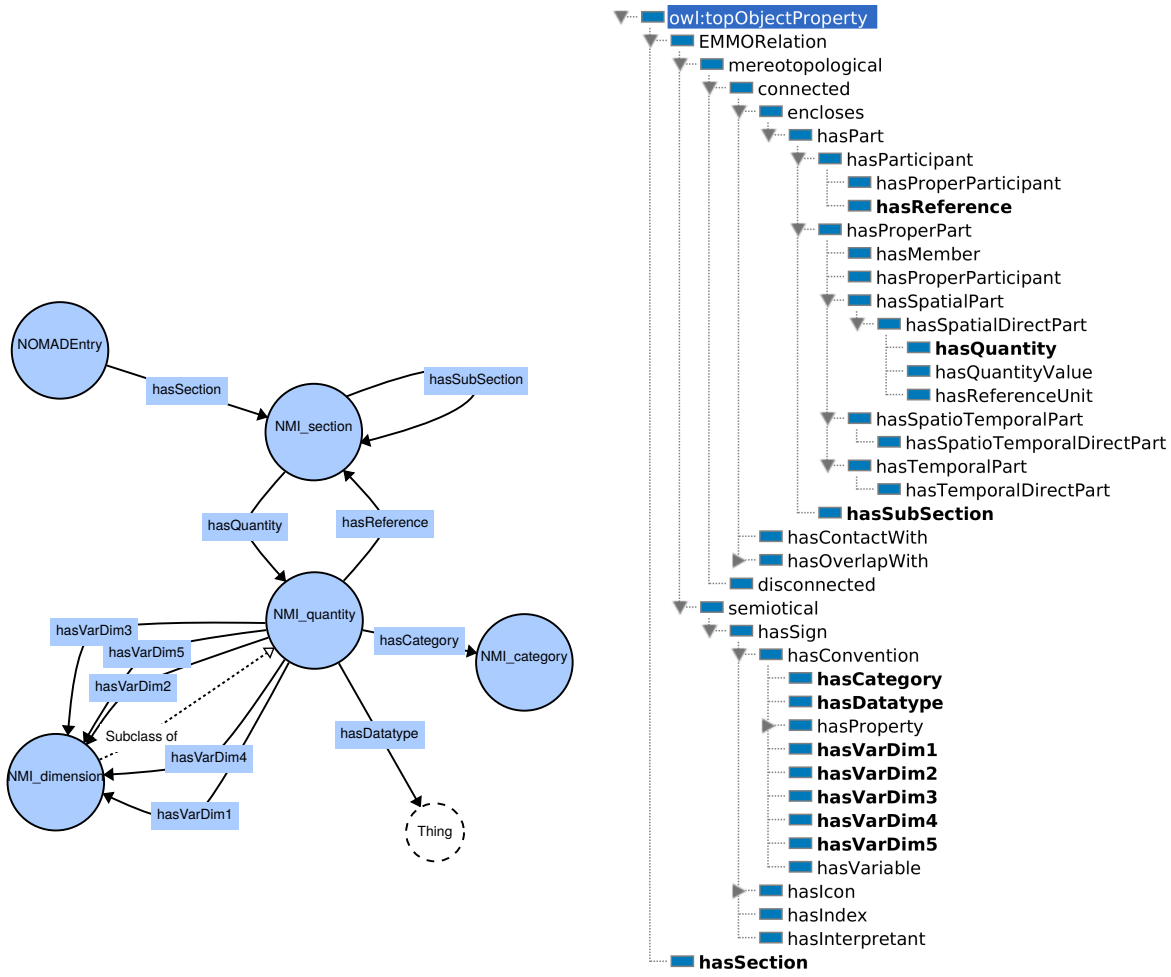
To ensure extensibility of the Metainfo ontology and synchronization with the current Metainfo implementation, two ontology layers were created. The *pure* Metainfo ontology fetches meta-data definitions via the NOMAD API, creates an ontology from them and stores it in the OWL format. Anything else like rules, additional classification or relations to other ontologies is stored in the *extended* Metainfo ontology that imports the pure one.

3.1 The *Pure* Metainfo Ontology

Each Metainfo item is represented as on ontological class and inherits from one of the four meta-data types from Section 1.5.2, i.e. Section, Quantity, Dimension or Category. Real data from calculations in the Archive can then instantiate these classes. The object properties by which they are connected are embedded in the EMMO properties hierarchy as shown in Figure 3.1b. Three relations are sub-properties in the mereotopological branch: `hasReference` as a participant of some referencing process, and `hasQuantity` and `hasSubSection` as parts of a section. All others are classified as sub-properties of `hasConvention` in the semiotic branch of EMMO. The meta-data types and their relations build the upper structure for the Metainfo ontology and are displayed in Figure 3.1a.

As can be seen, `NMI_dimension` is defined as a subclass of `NMI_quantity`. In fact, the current Metainfo implementation (as of July 2020) does not strictly distinguish between quantities and dimensions anymore. However, because this distinction is useful, the classification is re-introduced in the next chapter within the extended ontology to increase findability and semantics.

When the Metainfo is used to annotate real data, it quickly becomes obvious that a concept for



(a) Schematic view of the upper classes showing the four different types of meta-data and their relations.

(b) Hierarchy of the new object properties (bold) in the NOMAD Metainfo. They inherit from EMMO properties (roman). This is a screenshot from Protégé's view of the object properties.

Figure 3.1: Overview of the upper structure of the NOMAD Metainfo Ontology.

a single NOMAD entry is necessary to completely ensure unique identification of data. Such an entry in the NOMAD Archive is represented by the ontology concept `NOMADEntry`.

Multiple conceptional decisions need to be made during ontology development one of which was how to include multi-dimensionality of quantities. While array representation was already discussed in Section 2.1.1, now the shape of a quantity is to be described. First, variable and fix dimensions need to be treated differently. The meta-data `atom_positions` has the shape [`number_of_atoms`, 3], so the first dimension is a variable depending on the system and the second dimension is fix and represents the 3 components x, y, z. Whereas the variable component can be class, it does not make sense to create a class for each possible fix dimension. Therefore, an object property for variable dimensions is appropriate but fix variables are better described using a datatype property with range integer. This is also the reason why the technique of dimension chaining using `hasFirstDimension` and `hasNextDimension` relations does not work. Figure 3.1a only shows object properties but in principle multiple `hasFixDim` properties would point away from `NMI_quantity`. A datatype property `hasNumberOfDimensions` for the quantities additionally ensures user-friendly accessibility. Five variable and fix dimensions are sufficient to describe all quantities defined in the

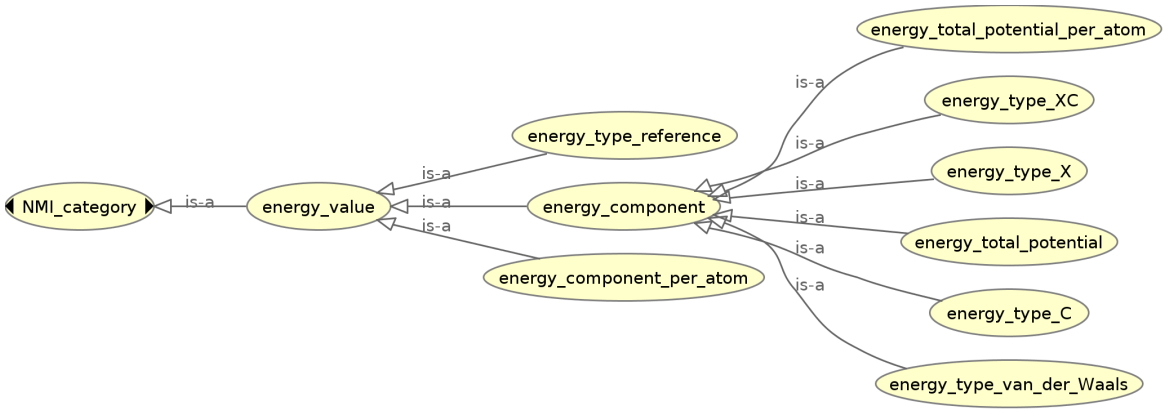


Figure 3.2: Subclass hierarchy for energy categories in the NOMAD Metainfo Ontology.

Metainfo to date.[1]

Quantities usually are expected to be provided in a specific datatype like integer, string or real. Here, external datatype classes from EMMO are used which at the same time avoids redundancy and improves interoperability and re-usability. The Metainfo also gives specifications for the units of a quantity using the python module Pint. For the ontology, these units are replaced by EMMO units and related using EMMO's `hasReferenceUnit` property. EMMO relates units to their corresponding concepts in IUPAC [133] and QUDT [134].

In ontologies, the *subclass-of* relation is equal to an *is-a* statement. If `A is subclass of B`, then `A is a B` and therefore inherits all properties of `B`. For the NOMAD Metainfo, the only meta-data type that can be expressed hierarchically using *is-a* statements are the categories. An example of different energy categories in the Metainfo ontology and their hierarchy is presented in Figure 3.2. This categorization provides a semantic layer to the remaining Metainfo terms.

Although the word "subsection" suggests some kind of subclass relation between sections, in fact a subsection does not contain the same things as its parent section. For this reason, the `hasSubSection` relation was created. All sections and their relations are shown in Figure 3.3 colored by communities that were identified using the modularity measure introduced in Section 1.4.3.

A particularly cumbersome part is transferring references from the current Metainfo implementation to the ontology. If a section like `section_system` refers to another section, it contains a quantity (e.g. `system_to_system_ref`) which has a Proxy object as value pointing to the actual system. However, such Proxy links are not representable within the Metainfo. The value is only given in the Archive when real data is considered. In the same manner, the ontology only defines the object property `hasReference` (Fig. 3.1a) but contains no further restrictions on its usage. It is only used to link individuals once they have been created (a process referred to as instantiation or ontology population). This work uses SPARQL to find referenced sections in the materialized knowledge graph and relates such instances afterwards.

The pure NOMAD Metainfo ontology counts 4272 axioms, 629 classes, 16 object properties and 6 datatype properties including directly referenced or subclassed external classes. It is a

[1] The use of edge properties was avoided here to eliminate any incompatibility issues for such a fundamental relation. However, it would be a perfect use-case for an edge property to specify which dimension a `hasVarDimension` or `hasFixDimension` relation refers to.

Figure 3.3: Sections in the NOMAD Metainfo represented as network with arrows visualizing the `hasSubSection` relations. Sections are grouped into communities using the modularity measure. Labels are size-scaled by out-degree. Note that the highest level section is `section_run` here. Any meta-data concerning the NOMAD entry are not displayed for simplicity.

Query 3.1: SPARQL CONSTRUCT Statement to classify quantities that are used as variable dimensions for other quantities. Refer to Section 1.3.6 for details of the SPARQL syntax.

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX nmi: <https://nomad-coe.eu/ontology/metainfo/pure#>

CONSTRUCT {
    ?dimension owl:subClassOf nmi:NMI_dimension .
}
WHERE {
    ?quantity rdfs:subClassOf [
        a owl:Restriction;
        owl:onProperty ?vardim;
        owl:someValuesFrom ?dimension
        ] .
    FILTER (?vardim IN (nmi:hasVarDim1, nmi:hasVarDim2,
        nmi:hasVarDim3, nmi:hasVarDim4, nmi:hasVarDim5) )
}
```

pure T-Box ontology, so it does not declare any individuals.

3.2 The *Extended* Metainfo Ontology

In the second layer of the NOMAD Metainfo Ontology further classification and relations are introduced that enhance semantics.

First, some Metainfo quantities are used as dimensions for other quantities. This classification is however not directly accessible in the Metainfo schema. The respective triples are therefore created within the extended ontology using the SPARQL Query 3.1 to identify these dimension quantities and create the respective triples.

Second, several Metainfo terms have information in their human-readable descriptions that are not captured by the schema. One of these is the method with which a quantity has been calculated. The object property `hasMethod` was therefore created which is equivalent to stating "was calculated using the method". New links are created in the Metainfo Ontology using this relation. Table 3.1 shows all method-value pairs that could be extracted from the Metainfo descriptions.

Categories already structure the Metainfo, albeit lacking semantic definitions and relations themselves. Referencing more semantic ontologies is therefore a means to enhance interoperability. Basic physical quantities like energy or force are present already in EMMO and therefore also represented in the Properties Ontology from the last chapter. All symmetry concepts that have been defined in the Structure ontology are made equivalent using `owl:EquivalentClass` statements.

In so called *General Concept Inclusion (GCI) Axioms* [135] it is possible to express general statements that can be interpreted as simple rules. Above, triples were constructed using SPARQL to include a classification based on a query. GCIs are an alternative approach that is

Table 3.1: Methods meta-data (right) describing value meta-data (left) as given in Metainfo human descriptions.

Metainfo term	is related via `hasMethod` to
`energy_C`	`XC_functional`
`energy_X`	`XC_method`
`energy_XC`	`XC_method`
`energy_van_der_Waals`	`van_der_Waals_method`
`energy_van_der_Waals_value`	`energy_van_der_Waals_kind`
`stress_tensor`	`stress_tensor_method`
`stress_tensor_value`	`stress_tensor_kind`

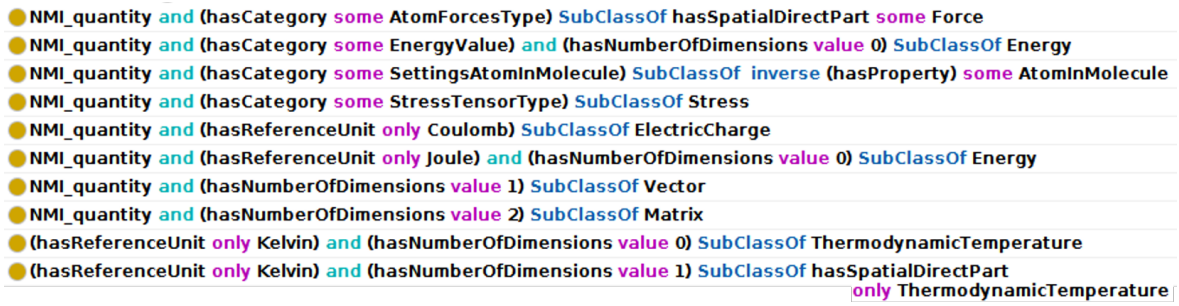


Figure 3.4: Screenshot of General Class Axioms in Protégé to semantify the NOMAD Metainfo. After reasoning all classes satisfying the LHS of `SubClassOf` are subclasses of the RHS.

less flexible but includes the rules themselves leaving it to the reasoner to infer the results. This is especially useful when an ontology is still growing or changing. Such general class statements have a left hand side (LHS) and a right hand side (RHS) combined by either of `rdfs:subClassOf`, `owl:EquivalentClass` or `owl:disjointWith`. Each side can consist of anonymous classes that do not have names themselves. In Figure 3.4 general class axioms are listed in the simplified human-readable Protégé syntax. Concepts in the NOMAD Metainfo Ontology are related to concepts in EMMO and the NOMAD ontologies. The structure of each statement is the same: the LHS is stated to be a subclass of the RHS. A reasoner can then infer for all classes that fulfill the LHS that they are subclasses of whatever is stated on the RHS. Because the pure Metainfo Ontology contains relations to categories, dimensions or physical units they may be used to infer that a certain Metainfo quantity is a subclass of a specific physical quantity. For example anything that has a unit of Joule is an energy or an array containing energies in its components depending on its number of dimensions.

3.3 Enhancing the NOMAD Metainfo with DCAT

The Data Catalog Vocabulary [136], usually only referred to as DCAT, is an RDF vocabulary for data catalogs and datasets published on the Web. It aims to enhance interoperability between different data repositories as well as discoverability of datasets within these. High level abstract meta-data about things like the title, creator and creation time as well as access rights and landing pages are annotated using the DCAT vocabulary and made available in a linked data format such as RDF. External applications can use this DCAT description and

enable dataset searches across multiple otherwise heterogeneous data catalogs. The second version, DCAT v2, is a recommendation of the World Wide Web Consortium (W3C) as of 2004 and widely used across the Web. The European Union public sector uses DCAT as a foundation for open dataset descriptions. An example for a search application is the Google Dataset Search [2] fetching datasets that are structured and annotated using either schema.org or DCAT. Datasets from the Materials Project or OQMD are already findable via the Google Dataset Search. It requires a structured markup of each dataset's landing page to find and identify it as dataset but does not crawl the databases directly. Within the STREAM project[3], we develop a mapping of meta-data to the DCAT vocabulary which is presented in this section. The conceptual mapping was done by this thesis' author in collaboration with the TIB. We further created a DCAT interface for NOMAD that is accessible through a dedicated DCAT API [4]. On a technical side, this was mainly realized by Markus Scheidgen from the NOMAD Laboratory. To make NOMAD data findable in the Google Dataset Search such standardized meta-data descriptions need to be added as markup to each dataset's landing page. Google crawls the HTML source code of these pages and extracts meta-data markup to index the dataset. DCAT borrows properties from other vocabularies like the DCMI Metadata Terms by the Dublin Core Metadata Initiative (DCMI) [137] and the PROV Ontology (PROV-O) for provenance information [138].

While it is possible to simply describe the Metainfo schema within the DCAT specification using the `dcat:Catalog` class, it is more useful to represent the information stored for each NOMAD entry in DCAT format. This way, external dataset search engines can access this information and find relevant NOMAD entries. An individual entry in the NOMAD Archive corresponds to a `dcat:Dataset`, because it represents a collection of data. This is a crucial distinction to the NOMAD usage of the term "dataset" where multiple Archive entries are combined in a single dataset. There is no equivalent in DCAT for a NOMAD dataset.

NOMAD's DCAT interface provides descriptive data about NOMAD entries in DCAT format. The technical realization refrains from using RML to define mappings from NOMAD to DCAT, since an RML processor would be necessary to generate the actual RDF files. As there are no RML processing implementations for Python, the mappings where implemented directly in Python using rdflib [139] to avoid the technical complexity of adding non-Python components in the NOMAD architecture.

Only very few DCAT dataset properties have exact matches in the NOMAD Metainfo, many are only similar concepts. It is possible to relate them as narrower or broader concepts, however this decreases findability. Thus, they are still mapped directly even though these slight differences should be kept in mind when further concepts are added to the Metainfo. Table 3.2 summarizes relevant properties of a `dcat:Dataset` for which related NOMAD terms exist or their values are known. Datatype properties marked with "DT" are easy to map as they require only literal values such as strings as given in NOMAD. More complicated are the object type properties ("O") where the value needs to be a resource. Creator and publisher are properties with the recommended range `foaf:Agent` [140]. Such an agent can be for example an organization or a natural person. An instance of this class is created first for each author and uploader (or respective publishing organization) to be able to map these persons (or organizations) stored in NOMADs user management system to the DCAT vocabulary. On the other hand, the contact information pointed to using `dcat:contactPoint` is recommended to be provided using the vCard Ontology [141]. Because of the open-world

[2] `https://datasetsearch.research.google.com/`

[3] The Fritz Haber Institute is one of five partners in the BMBF-funded project STREAM: semantic representation, linking and curating of quality-ensured materials data. More info at `https://stream-projekt.net/`.

[4] `https://nomad-lab.eu/prod/rae/dcat/`

Property of Dataset	DT/O	NOMAD Metainfo term or value	Comment
dct:description	DT	comment	comment does not always contain a description
dct:title	DT	formula	chemical formula was chosen as title
dct:issued	DT	upload_time	issued means the official publication date
dct:modified	DT	last_processing	
dct:isReferencedBy	O	references	
dct:creator	O	authors	
dct:publisher	O	uploader	publisher and uploader does not need to be the same
dct:identifier	DT	calc_id	historically there is also entry_id, upload_id
dcat:landingPage	O	**[1]**	can be constructed but not a Metainfo term
dct:license	O	**CC BY 4.0**	known, but not a Metainfo term, might change in future
dct:language	O	**English**	known, but not a Metainfo term
dcat:contactPoint	O	**section User**	mapping through use of vCard class
prov:wasGeneratedBy	O	**section program_info**	needs definition of programs as software agents in PROV-O
dcat:theme	O		needs definition of SKOS concepts like “Material”, “Molecule”
dct:type	O		needs definition of concepts like “Computer simulation of physical object”
dcat:distribution	O	**API, JSON**	Distribution objects need to be created

Table 3.2: Datatype (DT) properties and object (O) properties for a `dcat:Dataset` and the corresponding terms in NOMAD as well as comments about concerns. If not a Metainfo term but known directly it is marked blue.
[1] `http://nomad-lab.eu/prod/rae/gui/entry/id/[upload_id]/[calc_id]`

assumption that is made in OWL, an instance of `vCard:Individual` can at the same time be an instance of `foaf:Agent`. In this way, the `dct:creator` and `dcat:contactPoint` properties point to the same individual. This choice was taken to avoid duplication and enhance semantics because it is immediately clear that the creator is also the contact point.

The concepts mapped so far are very general and not specific for materials science. A user querying a particular material can therefore only get hints about which materials repositories contain relevant data by checking the chemical formula. Whether a dataset describes a computer simulation or an experiment as described in NOMADs `domain` keyword is not part of the DCAT representation. Mapping it via the `dct:type` property requires such concepts to be formalized in a controlled vocabulary such as the DCMI Type Vocabulary. Similar is true for the `dcat:theme` property with the range `skos:Concept` – formalizing materials classes or other features of a material in SKOS concepts [142] would enable the usage of this property.

Additionally, NOMADs DCAT API provides information about the distribution types: each NOMAD entry is distributed in multiple ways – via the API, in JSON format and as raw data. The respective access (or endpoint) URLs, media types and other related information are specified using DCAT.

3.4 A Knowledge Graph of Hybrid Organic-Inorganic Perovskites

3.4.1 Graph Materialization

As a first subset of NOMAD entries, a dataset on hybrid organic-inorganic perovskites (HOIP)[5] with different structures [143] is chosen to populate the knowledge graph. This materials class is promising in solar cell technology [144, 145]. Since their first use as photo-voltaic cells in 2009, efficiencies have rapidly increased. In 2021, an efficiency of 25.6% has been reported for perovskite solar cells [146]. The chosen dataset contains 8076 NOMAD entries on 1346 different HOIPs combining 16 organic cations, 3 group-IV cations and 4 halide anions.

As explained in Section 1.3.5, two different strategies are commonly followed to create a knowledge graph: graph materialization or query rewriting in the form of mappings. Because the data model of the Metainfo ontology was directly inherited from its original, ontological concepts can be instantiated and a knowledge graph is created without the need for complicated rules and mappings. This direct applicability is the immediate benefit of the bottom-up approach. The NOMAD Python API can be used to query and retrieve the dataset from the NOMAD Archive. Automatically looping through the section structure and creating resources from each item is then easily possible using the python packages owlready2 [147] or RDFlib [139]. Items are linked using the ontological relations from Section 3.1. Some of the items in the Metainfo are reference objects. They are used for example to indicate for which system and with which method a particular calculation output was obtained. These items are encoded as Proxy objects when retrieving them via the API. During the KG creation, they are replaced by actual links between the respective sections using the `hasReference` relation. Such references are resolved with simple SPARQL queries enabled by RDFlib. Here, the queries are performed within an individual knowledge graph representing one NOMAD entry and finish very fast. However, it has to be kept in mind that performing a large number of queries on a large knowledge graph can quickly become performance-critical.

[5] Dataset DOI: 10.17172/NOMAD/2017.03.15-1

The size of the resulting RDF files can similarly increase rapidly when many triples are created, for example if large arrays are transformed. Each single element in an array is an individual array component node with a value and an index that can itself be a vector (see Core Ontology in Section 2.1). For smaller arrays this representation works well, but band structures as stored in NOMAD can easily have tens of thousands of elements. As an example consider a VASP calculation on $SrZrO_3$ [6] whose legacy JSON representation has a file size of 1.1 MB. Execution of the conversion to RDF/Turtle format takes almost 5 minutes on a standard laptop and results in a 24 MB large Turtle file, which is a factor of 22 times larger than the original JSON file. Depending on the use-case for these newly created knowledge graphs, it is not always necessary to store numeric arrays in this storage-consuming manner. A much smaller file can be obtained by storing all arrays above a certain size as strings. This threshold was here chosen to be 5: any array with more than 5 elements is saved as string value. Of course, this way the semantics can not be leveraged but if needed such strings can easily be parsed as NumPy arrays. The final size of the RDF file using this compromise is 256 KB which is even smaller than the legacy JSON representation. Almost half of the conversion time is spent converting between different RDF syntaxes to maintain compatibility between owlready2 and RDFlib. Designing a better framework can therefore significantly reduce this time. Naturally, this is a problem that is straight-forward to parallelize because each entry forms its own knowledge graph. Using NOMADs Python API however, pagination and Proxy objects have to be dealt with that are incompatible with standard parallelization techniques.

To provide a single point of access to a given dataset or data repository, all RDF files can be loaded to a triplestore together with the used ontologies. Dozens of different triplestore implementations are available, both open-source and proprietary ones many of which have native SPARQL support and provide a SPARQL endpoint. Here, the proprietary triplestore and graph platform *Stardog* is used which natively supports edge properties using RDF*/SPARQL* as well as provides basic visualization.

3.4.2 Semantification on Instance Level

The extended Metainfo Ontology includes statements about the equivalence of some of its classes to classes in the Structure Ontology. For example the `nmi:crystal_system` is equivalent to `cso:CrystalSystem`. This means that all instances of one class are also instances of the other class. However, when creating the knowledge graph for the HOIP dataset, only string values are attached to the `nmi:crystal_system` instances. Neither the ontology nor a reasoner can know which string values correspond to which instances in the Structure Ontology. When this connection is not directly created during the graph materialization, two such instances can be connected using the `owl:sameAs` relation. A reasoner can then infer the equality of these instances. Similarly to the crystal system, also space group and bravais lattice have equivalent classes in both ontologies. These sameAs-relations are added to the knowledge graphs using SPARQL INSERT or CONSTRUCT statements as given in Query 3.2. While the querying is performed on a triplestore and INSERT statements would add the new triples directly to the graph, all knowledge graphs will finally be published in RDF format, so that the CONSTRUCT request is chosen whose results can be stored in RDF format together with the data.

[6] upload_id: `3xtCpsST9Wb7NEAV6ADGQ`, calc_id: `wDjLmh7A5kgSTtKqYzfgBqaJt8OI`

Query 3.2: SPARQL Insert statement for adding instance equality between Metainfo instances and Structure Ontology instances. Refer to Section 1.3.6 for details of the SPARQL syntax.

```
prefix nmi: <https://nomad-coe.eu/ontology/metainfo/pure#>
prefix core: <https://nomad-coe.eu/ontology/core#>
prefix cso: <https://nomad-coe.eu/ontology/structure#>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>

CONSTRUCT{
    ?NMI_CS owl:sameAs ?CSO_CS .
}
WHERE{
    # Crystal System
    # --------------
    ?NMI_CS a nmi:crystal_system ;
        core:hasValue ?CS_string .
    BIND(IRI(CONCAT(str(cso:CS),?CS_string)) as ?CSO_CS)
}
```

3.5 Applications to Real Life Problems

3.5.1 The Search for a Better Solar Cell Material

One often noted example of an application for an ontology is the *semantic search*. It differs from traditional keyword-based search in that it not only offers webpages or links but truly answers the user's question [148]. Examples for such search engines are Wolfram Alpha and the Google Knowledge Graph [149]. In contrast, a traditional *faceted search* gives the user options to constrain the number of search results based on so called facets [150]. For materials science, consider the example of the search for good or better solar cell materials. A semantic search engine would need to answer the question

(i) "What material is a good solar cell material?" or

(ii) "Which basic materials properties are needed for making a material a candidate for a good solar cell material?".

Clearly, to answer such questions the underlying search engine needs detailed knowledge of solar cells or what makes a good solar cell. Such information can be encoded in an ontology. Ideally, the ontology would include only the basic physics describing the physical phenomenon and an autonomous intelligent agent could determine the resulting required properties (such as the band gap or the effective mass). Unfortunately, no such agents exist yet and ontologies themselves cannot think, so the physical, economical and biological requirements for a solar cell material need to be explicitly stated in an ontology. Question (ii) is already answerable with such an ontology. Finally, to answer question (i), the ontology would help to formulate a faceted search against a number of knowledge graphs from different domains. The search then becomes semantic as the facets that are being queried are automatically identified by the ontology. Finding candidate materials in this way is today still unrealistic because too much information is still missing or unavailable to give valuable answers to such a complex multi-domain question.

Faceted searches in the domain of computational materials science are already realized within many of the available DFT simulation databases, e.g. NOMAD, AFLOW, OQMD or Materials Project. Nevertheless, in the following we will investigate the different aspects of a potential new solar cell material and discuss a simple example how different domains can be connected with the help of ontologies. This example was realized and implemented by the author of this work utilizing the NOMAD repository as primary data source.

The maximal theoretical efficiency for single p-n junction solar cells is given by the Shockley-Queisser limit and was recently calculated to be 33.7% for a band gap of 1.34 eV [151]. This optimal band gap value is already one of the most fundamental criteria in the search for solar cell materials. How the conduction and valence bands behave around this band gap is important for the efficiency of solar cells and is captured by the effective masses of the electrons and holes. A strongly curved conduction band yields a small effective mass and therefore a large charge carrier velocity leading to fast circuit reaction times. Additionally the material should of course be stable as well as obtain a strong optical absorption coefficient, be low in cost, not be toxic or hazardous and in the optimal case be compatible with existing technology.

Stability can refer to different concepts: Nuclear stable materials are isotopes that do not decay spontaneously and are therefore not radioactive. Thermodynamical stability occurs when a material is in chemical equilibrium with the environment, i.e. it is in its lowest energy state. Typically, the free energy is investigated to include finite temperature effects like the vibrational motion of the nuclei. Thermodynamically, a material is stable if its free energy is lower than the free energy of all possible decompositions into elementary, binary, ternary (and so on) compounds. For this, often the convex hull of stability is constructed depicting the formation energy as a function of the chemical composition [152]. Because free energy calculations are costly and often not available for enough materials to create a convex hull function, the total energy at 0 K can be used instead for the convex hull construction to obtain a first stability estimation if zero point motion is neglected. Total energy calculations are available for a huge amount of compositions and, if missing, not too expensive to run. NOMAD does not relate individual calculations to each other. This makes estimation still tedious to extract from existing data. The NOMAD AI Toolkit [7], was specifically designed to utilize artificial intelligence to fill the gap of missing data by predicting materials properties and also to gain insights about relationships. Another aspect of stability are meta-stable states as seen in Part I of this thesis. An example is diamond which is stable only at very high pressures and meta-stable at normal temperatures and pressures. Its conversion to the stable graphite phase is however hindered by a large activation barrier. Also meta-stable heterostructure alloys are a topic of current research to discover materials with specific functionalities [153].

Optical absorption spectra can be obtained using for example time-dependent DFT. Even though a few codes supported by NOMAD provide possibilities to calculate light-matter interaction, no standardized meta-data keys are defined in the Metainfo due to the lack of available data. The code-specific keyword `dmol3_optical_absorption` occurs only twice as test entries in the whole NOMAD Archive.

The *cost* is an economic factor that is probably impossible to predict for novel compounds that have never experimentally been realized. Simple estimates based on the abundance (and cost) of the constituent chemical elements neglect the cost of synthesis and fabrication and are therefore too rough to be valuable measures. In fact, prices of chemical elements are listed on Wikipedia [8] but converting the table to a linked data format is tedious. It re-

[7] `https://nomad-lab.eu/services/AIToolkit`

[8] `https://en.wikipedia.org/wiki/Prices_of_chemical_elements`

Query 3.3: SPARQL Query for substances that have an effect (wdt:P1542) that is either intoxication (wd:Q18621601) or a subclass of it (wdt:279).

```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX bd: <http://www.bigdata.com/rdf#>

SELECT ?substanceLabel ?effectLabel ?formula
WHERE{
    ?substance wdt:P1542 ?effect .
    ?effect wdt:P279* wd:Q18621601 .
    ?substance wdt:P274 ?formula .

    SERVICE wikibase:label {
        bd:serviceParam wikibase:language "en" .
    }
}
```

quires a mapping to ontologies and vocabularies that are designed for this domain. Intensive literature research is needed to ensure proper usage of existing ontologies. If none exist, new ontologies or vocabularies have to be developed first to be able to express the tabulated prices for chemical elements in a knowledge graph format. The benefit was regarded as too low to follow this path.

Biological aspects like *toxicity* and hazardousness are similarly difficult to find for compound materials and in general unavailable for any novel materials. Most publicly available data repositories for chemical compounds include only biomedically or pharmacologically relevant substances, such as PubChem [154], ChEBI [155], National Cancer Institute Thesaurus [156] or the National Drug File Reference Terminology [157]. Chemicals are often part of a complex classification that does not allow for an easy true-or-false statement about its toxicity. Furthermore, the chosen HOIP dataset has no overlap with these databases and the author's search for appropriate databases containing some of our HOIP materials was not successful. DBPedia's properties `dbp:health` and `dbp:flammability` are unfortunately rarely used. In Wikidata, a substance can have an effect which is a subclass of intoxication (Query 3.3). This includes however ethanol and other industrially harmless compounds. In total, 42 results are returned none of which is a perovskite. Alternatively, a substance can itself be a subclass of some noxa (contaminants) (Query 3.4). This results in a much longer list of 819 distinct substances. The chemical formulas follow the Hill system [9] with few exceptions. Comparing these formulas with the respective entries annotated by the NOMAD Metainfo key `chemical_formula_hill` can filter out any of the harmful substances from the HOIP dataset. Unfortunately, there is however again no overlap, so that this filter has no effect. In fact, none of the materials studied in our HOIP dataset have a corresponding entry in one of the mentioned databases where toxicity information is stored. Filtering by the contained elements that are potentially harmful is too hard as a constraint, especially because many useful materials contain such elements, for example lead in methylammonium lead iodide.

[9] In the Hill system, carbon atoms come first in the chemical formula, then hydrogen atoms and the remaining elements follow in alphabetical order.

Query 3.4: SPARQL Query for substances that are instances of a subclass of noxa (wd:Q50379880).

```
SELECT ?xLabel ?substanceLabel ?formula
WHERE{
    ?substance wdt:P31 ?x .
    ?x wdt:P279* wd:Q50379880 .
    ?substance wdt:P274 ?formula .

    SERVICE wikibase:label {
        bd:serviceParam wikibase:language "en".
    }
}
```

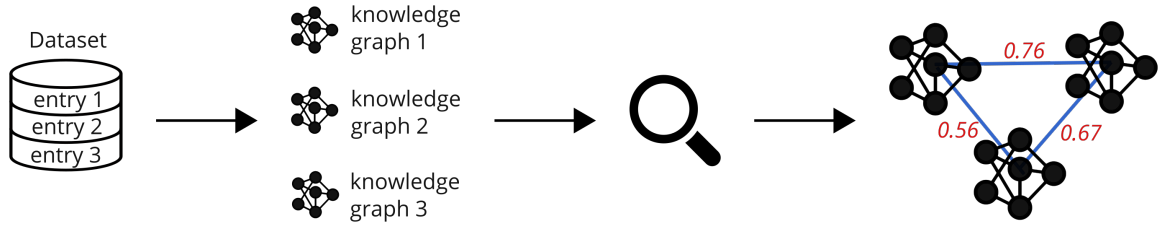


Figure 3.5: Simplified schematic view of the dataset enhancement workflow. A dataset from NOMAD containing multiple entries is converted to individual knowledge graphs. Further investigation, symbolized by the magnifier, can add additional links that connect the knowledge graphs finally building one large KG. These new relations (blue lines) can have edge properties (red).

3.5.2 Dataset Enhancement and Connections

Traditionally, when a research study builds on an existing dataset, this dataset is often cited by its DOI or by the publication in which it has been presented. Newly created knowledge can include for example connections between original data points or additional data values that enhance the original dataset. Such information are however hidden in text form and are unaccessible for a computer if no direct links are created. In Figure 3.5, a schema is shown how the knowledge graph approach is used in this work to overcome this limitation. One realistic and useful application of representing data as knowledge graphs is the ability to add new research results directly to the original graphs or at least directly using the IRIs of the original data. Due to the unique IRI representation of each data point, it is straight-forward to re-use such data. Sometimes access to the original triplestore is not given, so that new data can not directly be added to the existing graphs. In these cases simply referring to the respective IRIs is sufficient for a software agent (like a SPARQL endpoint) to understand the equality of two data points.

Due to the photo-voltaic interest in hybrid organic-inorganic perovskites [158], let us consider in more detail the electronic spectra of the materials in our dataset. As explained in Section 3.5.1, a material's band gap and the behavior of the energy bands around it play the most important role for its photo-electric functionality. The region of the band structure around the band gap of a potential new solar cell material is therefore likely to resemble the band structure of other well-known photo-voltaic materials. Because the search for solar cell

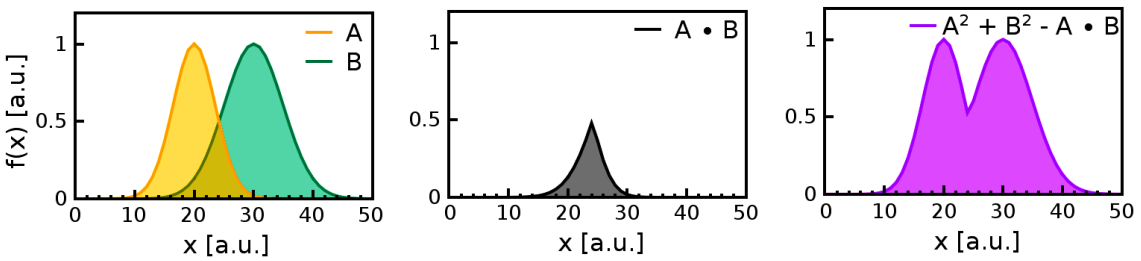


Figure 3.6: The components of the Tanimoto similarity coefficient visualized. Left: two individual functions, middle: intersection of the two functions, right: union of the two functions. Based on a plot courteously provided by Martin Kuban.

materials is focused on finding *better* materials, it is necessary to provide a reference point. Two very promising HOIP materials for the light-harvesting active layers in solar cells are methylammonium lead iodide ($CH_3NH_3PbI_3$, also $MAPbI_3$) [159, 160] and formamidinium lead iodide ($HC(NH_2)_2PbI_3$ or $FAPbI_3$) [161, 146]. They can therefore serve as reference materials and be compared to other perovskites in the dataset.

Density Of States Similarity Measure

The density of states (DOS) is a derived property that integrates values of the energy dispersion relation (band structure) over all states and k-points. It measures the number of available states at a particular energy. Very flat bands result in a high DOS whereas steep energy bands show up as low values in the DOS.

As introduced by Isayev *et al.* [106] DOS fingerprints can be defined representing the DOS as a binary descriptor. Sampling the DOS diagram into equally sized 256 bins and discretizing their values in 32 bits results in a 1024 byte fingerprint. The NOMAD encyclopedia implements a modification of this descriptor developed by Martin Kuban available through the python package *nomad_dos_fingerprints*. It improves the fingerprints by using a non-uniform grid where the bins are distributed around a reference energy by a Gaussian function. [162] This leads to a denser binning around this reference value which usually lies in the band gap region. More bins with smaller widths are desired in this region especially when photoelectric properties are investigated. The fingerprints of two materials are then used to calculate the Tanimoto coefficient [163] as a quantitative similarity measure. The Tanimoto similarity coefficient T_c as used here is also called Jaccard index and defined as the ratio between the intersection and the union of two sample sets A and B

$$T_c(A,B) = \frac{|A \cap B|}{|A \cup B|}\,, \tag{3.1}$$

where

$$0 \leq T_c(A,B) \leq 1\,. \tag{3.2}$$

A Tanimoto coefficient of 0 indicates no similarity at all whereas a value of 1 means that the two samples are identical. Figure 3.6 shows two distributions A and B (left) as well as their intersection (middle) and their union (right). For bit arrays like the DOS fingerprints, A and B are bit vectors and T_c can be calculated as

$$T_c(A,B) = \frac{A \cdot B}{A^2 + B^2 - A \cdot B}\,. \tag{3.3}$$

In the NOMAD encyclopedia, for each material that has a calculated DOS the top five materials with most similar densities of states are shown. Standard settings are an energy intervall of $(-10, 5)$ eV with the Gaussian centered around $\mu = -2$ eV and a width of $\sigma = 7$ eV. The python package *nomad_dos_fingerprints*[10] can be used to manipulate the Gaussian parameters, control the region in which the fingerprint is calculated or simply calculate T_c for materials that are not in the NOMAD Encyclopedia.

Similarity Relations as Edge Properties

Within this thesis, fingerprints and similarity measures are calculated for the chosen dataset of hybrid organic-inorganic perovskites. With an eye on the application as photo-voltaic materials, the DOS fingerprints are constrained to the region around the band gap. Because in this region there will naturally be only few states, many of the bins will be empty leading to a very small filling factor f (the number of '1' bits divided by the number of all bits in the fingerprint). When this filling factor approaches zero, this indicates that the region is too small to capture the surrounding bands. The fingerprint becomes therefore meaningless. [11] For the fingerprint calculation, the normalized density of states is used in which the energies are given relative to the highest occupied energy state. An asymmetric interval capturing the top 1 eV region of the valence band and the lower region of the conduction band is therefore reasonable. Assuming appropriate band gaps are smaller than 2 eV, the upper limit of the interval can thus be chosen to be 3 eV. With the Gaussian center at $\mu = -0.5$ eV and a width of $\sigma = 1$ eV, this yields an average filling factor of $f = 0.042 \pm 0.016$. This is appropriate when compared with standard settings (the full energy range) where $f = 0.016 \pm 0.003$. For each two fingerprints, the similarity T_c is calculated.

The similarities are added to the existing knowledge graphs enhancing the original dataset. Further studies on the same data can then make use of these relations more easily. First, we define a symmetric object property that expresses the quantitative similarity between two instances of the same class (`isQuantitativeSimilarTo`). To provide full provenance this can and should be used together with another specifically defined datatype property for the similarity value (`hasSimilarityValue`) and an annotation property describing the method with which this was calculated (`hasSimilarityMethod`). Because such relations are quite generic, they are added to the Core ontology from Section 2.1. For each pair of densities of states, a triple following the syntax in Code snippet 3.5 is then created. Here, so called edge properties as introduced in Section 1.3.3 are used with the novel RDF* specification. To avoid blowing up the graph, it is useful to only add this relation when the similarity value is greater than a threshold value, e.g. $T_c > 0.5$. Finally, these triples can either be added to the triplestore using SPARQL INSERT or simply be stored separately.

Similarity Network

Visualization of RDF*-based knowledge graphs is still rarely possible due to its novelty. For a network of similar DOSs with only one node type, one edge type and a similarity value that can be included as an edge weight, it is sufficient to utilize existing network tools. In

[10] `https://gitlab.mpcdf.mpg.de/nomad-lab/nomad-dos-fingerprints`

[11] An alternative approach would be to calculate two fingerprints: one for the occupied bands right below and one for the unoccupied bands right above the band gap. In combination with the band gap value itself, this would be a better measure for the bands' behavior. Here, we stick however with only one fingerprint to simply showcase the procedure.

Code 3.5: Schematic triple expressing the similarity between two DOSs and the similarity value and method using edge properties.

```
<< DOS_1 core:isQuantitativeSimilarTo DOS_2 >>
    core:hasSimilarityValue Tc ;
    core:hasSimilarityMethod "Tanimoto coefficent calculated
       using NOMAD DOS Fingerprint with parameters: ..." .
```

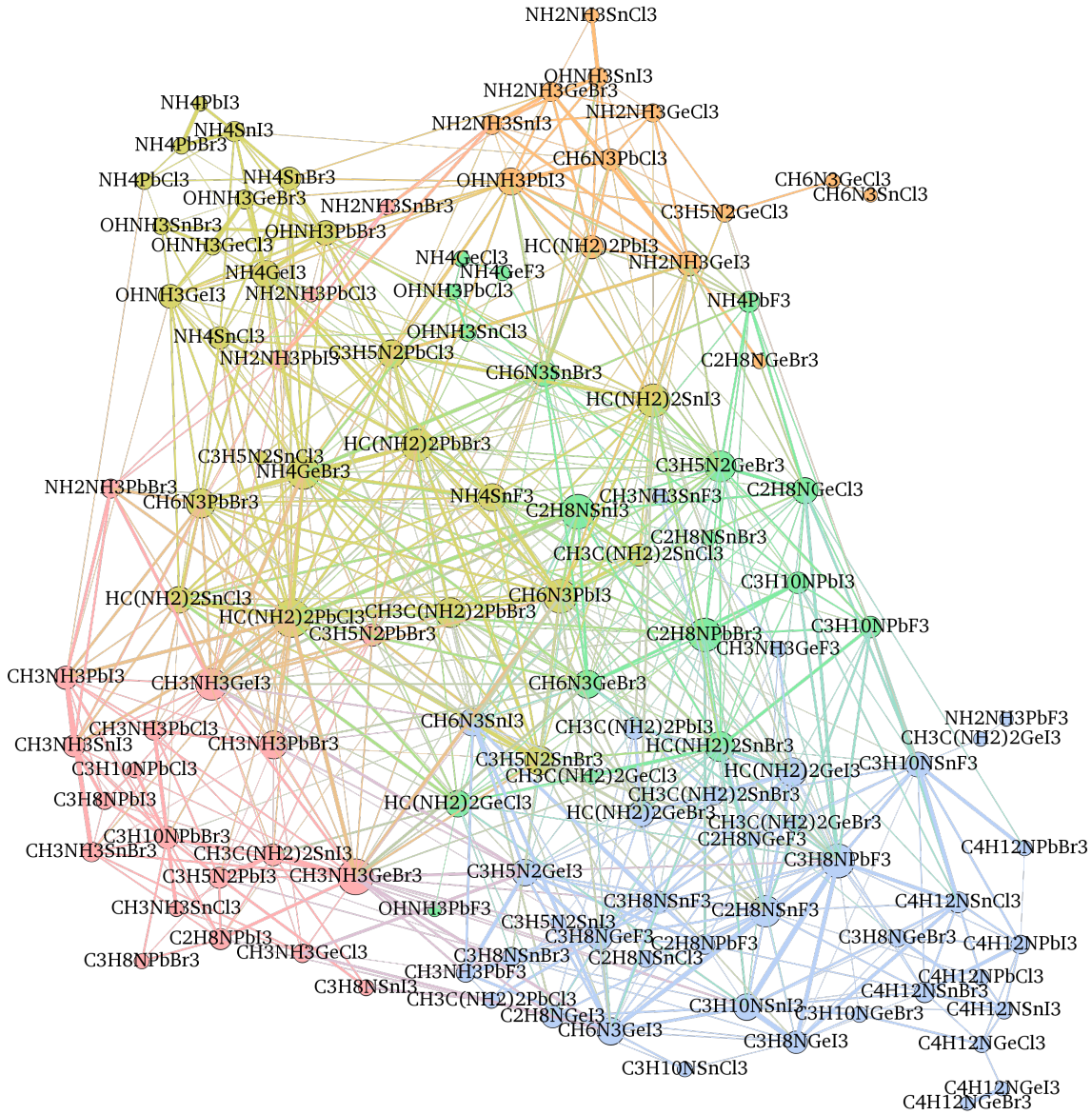


Figure 3.7: Network of densities of states (nodes) that are connected when their similarity values are $T_c > 0.7$ (weighted edges). The network is layouted using a force-directed graph algorithm. The nodes are colored by their modularity class and size-scaled by their degree.

Figure 3.7 the densities of states of our perovskite materials and their similarities are shown in a typical network layout using the force-directed graph drawing algorithm called Force Atlas as implemented in Gephi. Nodes represent densities of states and are connected by edges when their Tanimoto similarity is $T_c > 0.7$. Compared to a complete graph, in which each node is linked to every other node, this network shows 7% of all possible edges. As can be seen, it is still densely connected. Note that a different similarity threshold leads to a different network structure. The Tanimoto coefficient is reflected in the edge weights which determine the edge thickness.

We now run a modularities calculation which allows coloring the nodes according to different communities they belong to [164]. As introduced in Section 1.4.3, a modularity calculation divides the graph into communities according to the fraction of edges in a given group relative to that fraction in a random graph. In short, same colored nodes form a community which has more connections within the community than to nodes outside of it. This measure gives a rough overview of the network structure. Coloring the nodes by their contained halogen atom or group-IV atom does not yield such a clear cluster structure. It would be interesting to also look at the graph when all nodes are colored by their contained organic cation. Unfortunately, it is impossible to extract all the 16 different organic cations from the chemical composition in Hill form as stored on NOMAD. Seven of the cations share common Hill formulas:

- Dimethylammonium and Ethylammonium (C_2H_8N),
- Trimethylammonium, Propylammonium and Isopropylammonium ($C_3H_{10}N$), and
- Tetramethylammonium, Butylammonium ($C_4H_{12}N$) .

Because the materials also have the same symmetries, their differences are encoded in the atomic positions and not easily accessible. We therefore only distinguish the reduced chemical formulas of the cations. As found in [165], for methylammonium lead iodide, the organic cation contributes only to the core states of the DOS and should thus not be reflected in the similarity measure around the band gap. However, in our network in Figure 3.7 non-carbon-containing compositions cluster on the top having clearly less similarity to carbon-containing compositions occupying the bottom half of the graph. A further separation of the occupied and unoccupied parts of the DOS likely leads to a clearer distinction as the iodide p-state largely dominates the occupied part of $MAPbI_3$ whereas the lead p-state dominates the unoccupied DOS part but with generally fewer states [165]. In such a separated picture, materials with the same halogen or group-IV atom would likely have similar DOSs in the negative or positive regions relative to the Fermi level respectively.

Another interesting layout is the Circle Pack Layout in which the position of a node is determined by a number of node attributes. In Figure 3.8, we created a network featuring three hierarchies: The outer hierarchy is given by the halogen atom, the middle layer is structured according to the contained group-IV atom and within these each node position refers to one of the 12 different chemical compositions of the organic cations. The node colors reflect the same modularity classes as in Figure 3.7. Additionally, all nodes are size-scaled by their degree, i.e. the number of edges connected to them. Grey nodes are isolated, i.e. the DOS is not similar to any other DOS with $T_c > 0.7$. The number of grey isolated nodes are significantly higher in the fluorine square. Fluorine is the lighest halogen in the periodic table and the element with the highest electronegativity and therefore very reactive.

The knowledge graph allows us to identify materials with similar DOSs relative to the well known perovskites $MAPbI_3$ and $FAPbI_3$. With respect to the reference $MAPbI_3$, the perovskites with the most similar DOSs are $MASnBr_3$ with $T_c \approx 0.848$ and $MASnI_3$ with $T_c \approx 0.838$. Both have already successfully been synthesized [166, 167] and its photovoltaic

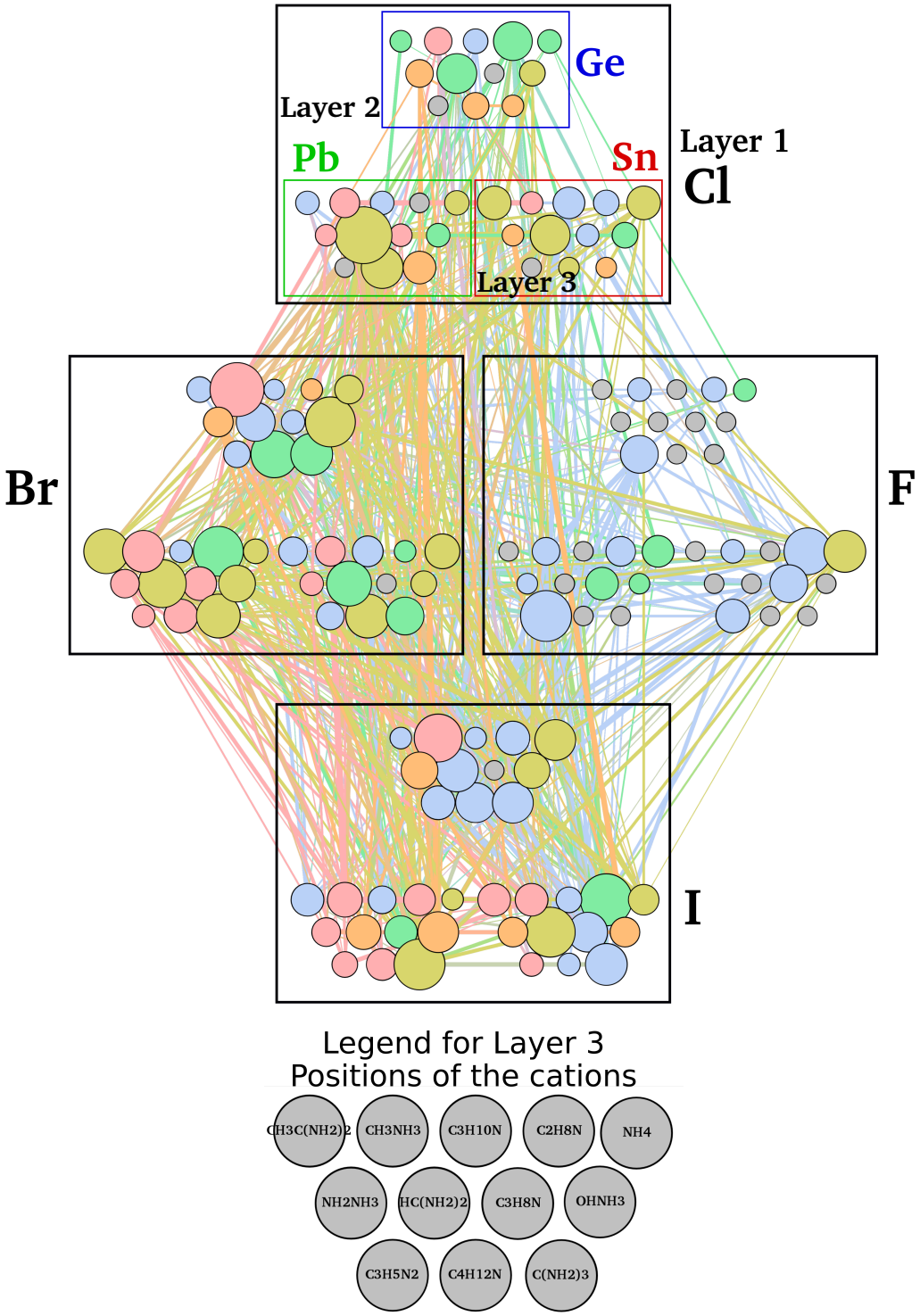


Figure 3.8: Circle pack layout of DOS similarity network structured in three hierarchies: halogen atoms, group-IV atoms and chemical composition of the organic cations. The order of the cations is shown in the outset with grey circles. Nodes are sized scaled by their degree and colored by modularity.

Table 3.3: Materials with a DOS that is (in the region around the band gap) similar to the DOSs of methylammonium lead iodide ($MAPbI_3$ or $CH_3NH_3PbI_3$) or formamidinium lead iodide ($FAPbI_3$ or $HC(NH_2)_2PbI_3$) with a $T_c > 0.7$.

Similar to $MAPbI_3$	T_c	Similar to $FAPbI_3$	T_c
$CH_3NH_3SnBr_3$	0.848	$C(NH_2)_3SnBr_3$	0.792
$CH_3NH_3SnI_3$	0.838	$OHNH_3PbI_3$	0.776
$HC(NH_2)_2PbCl_3$	0.800	$NH_2NH_3GeBr_3$	0.770
$NH_2NH_3PbBr_3$	0.782	$NH_2NH_3GeI_3$	0.767
$C(NH_2)_3PbBr_3$	0.774	$C(NH_2)_3PbCl_3$	0.756
$C_2H_8NPbI_3$	0.744	$OHNH_3SnI_3$	0.744
$CH_3NH_3GeBr_3$	0.740	NH_4PbF_3	0.726
$CH_3NH_3PbCl_3$	0.736	$C_3H_5N_2GeBr_3$	0.722
$CH_3NH_3GeI_3$	0.732	$C_3H_5N_2PbCl_3$	0.716
$HC(NH_2)_2SnCl_3$	0.728	NH_4SnF_3	0.714
$C_3H_{10}NPbBr_3$	0.724	$NH_2NH_3SnI_3$	0.714
$C_3H_8NPbI_3$	0.722	$HC(NH_2)_2SnI_3$	0.714
NH_4PbCl_3	0.714	$C_3H_5N_2PbBr_3$	0.709
$CH_3NH_3PbBr_3$	0.713	$CH_3C(NH_2)_2PbBr_3$	0.707
$OHNH_3GeI_3$	0.708	$C_2H_8NGeCl_3$	0.706

performance has been studied [168, 169]. For $FAPbI_3$ the two most similar materials are $C(NH_2)_3SnBr_3$ with $T_c \approx 0.792$ and $OHNH_3PbI_3$ with $T_c \approx 0.776$. While the latter, hydroxylammonium lead iodide, has been synthesized and physico-chemically characterized [170], the former, guanidinium tin bromide, was only been synthesized with the chemical composition G_2SnBr_4 [171], with guanidium abbreviated with G (equal to $C(NH_2)_3$).

A list of all similar materials (with $T_c > 0.7$) identified for these two compounds is given in Table 3.3.

For the purposes of easier visualization we left out the information whether the band gaps are direct or indirect. It would be another option to color the nodes according to this or to choose different node shapes. The actual size of the band gap could also be encoded in the node size. In the sense of network statistics we chose to include the node degree instead. A lot of different network visualizations are realizable each of one has their own insights and advantages. Picking only two is enough to showcase how such a procedure of information processing works.

It has to be noted that for a solar cell material candidate even more information are interesting. By capturing the region around the band gap of the DOS we included a valuable part of the electronic structure. However, electron and hole effective masses, that are crucial for the efficiency of a solar cell, require knowledge about the band curvature which is lost in the DOS when more states are close to the valence and conduction band edges. Automating effective mass calculations and including them in the approach as additional features is therefore more likely to enhance the quality of the approach. Furthermore, selection rules for the optical transitions have been neglected completely in our methodology so far. It is thus obvious that we have presented an oversimplified graph-based version of the search for solar cell materials which without graph technologies is already much more mature.

In March 2021, a paper by Veremyev *et al.* was published [103] that presents a very similar approach. Using different types of similarity and distance measures for the density of states,

they create a network, where each material node is linked to another one when the similarity is above a certain threshold. In contrast to this work, they deliberately refrain from using edge weights. 27000 ISCD materials are included in this network so that materials with very different structures and compositions are connected. They perform a handful of different network analyses and propose an interactive network-based visualization tool for these kind of investigations, as mentioned in Section 1.4.2. The negative and positive energy regions are handled separately as was suggested above. Hence, the band gap itself is explicitly left out of the similarity measure and included only as classification for the nodes into metals, semiconductors and insulators. In turn, this leads to the network showing similarities between materials with very different band gaps. With the photo-voltaic application in mind, this was intentionally avoided here.

Chapter 4

Clean Data in Heterogeneous Catalysis

The previous example of conducting studies on existing datasets to create new (derived) information is a typical situation not only in computational materials science. Often, experimental data are handed over to the theoreticians and computational scientists to extract further value from it. Experimental data can enter either as reference values for computer simulations, as parameters in functions or functionals or even as input data for machine learning studies. The latter case can in the best case create knowledge that is valid beyond the initially used experimental dataset and is therefore especially interesting. This chapter handles such a scenario in the field of heterogeneous catalysis.

4.1 The Importance of Clean Data in Catalysis

Heterogeneous catalysis is a highly complex function of a material which is affected by multiple interacting processes. A full computational modelling of such high-dimensional problems using first-principles methods is unfeasible. Data science and AI techniques therefore provide new means to investigate and understand the underlying chemistry. A pre-requisite to utilize such approaches is a sufficiently diverse high-quality dataset. Although over the last decades large amounts of high-quality data on catalysts and catalytic processes have been produced, datasets are often incomparable because different inconsistent conditions were applied or published datasets are incomplete. Only recently, the idea of handbooks was proposed [172] in which minimum requirements for catalyst investigations shall be defined for each class of reactions. Of course, this requires also commitment of the communities to comply with such standardized procedures for catalytic synthesis, testing and characterization. This ensures the availability of sufficiently good data for future AI studies or modern data analytics approaches. As an example a “Clean Data Handbook” for the selective oxidation of short-chain alkanes over mixed metal oxide catalysts was published [172]. Thirteen such metal-oxide catalysts were synthesized and studied within the development of this handbook and can be used for further analysis forming a consistent and complete dataset.

Providing guidelines for standardized measurements and experiments is only the first step towards a digitalized and FAIR data infrastructure in catalysis research. The optimal goal is storing and annotating data in a consistent way that fulfills the FAIR principles from Section 1.2. In this chapter, an ontology for catalyst characterization and testing was developed

with the specific purpose of representing existing experimental tabular data using this ontology.

4.2 A Path towards Ontological Representations in Heterogeneous Catalysis

Just like most field concerned with specific materials functions, the field of catalysis is too wide and complex to unify all concepts and knowledge in one ontology. In this work, the purpose is to show how an ontology can be used to

a) create a unified representation of data,

b) link different parts of a complex experiment,

c) combine multiple layers of data and knowledge.

The focus lies on applicability of the ontology and the ability to use it for representations of existing tabular datasets on catalytic characterization and testing. Therefore, we apply here a bottom-up ontology development approach to gain maximum usability. All work in this chapter was done by the author. All data was provided by the group of Annette Trunschke at the Inorganic Chemistry Department at the FHI.

4.2.1 Conceptual Modeling

In the beginning of the modeling process stands the question whether there are existing ontologies that can be re-used. While general crystallographic concepts have been modeled in multiple different approaches as could be seen in the last chapter, chemistry is largely lacking ontological representations so far. The IUPAC International Chemical Identifier (InChI) [173] offers a standardized way to identify chemical substances by text labels. More recently, this idea was extended and an international chemical identifier for reactions was suggested (RInChI) [174]. Furthermore, an open repository for chemical reactions on catalytic surfaces, the Catalysis-Hub.org [175], is available containing more than 110 000 chemisorption and reaction energies from electronic structure calculations on surfaces using DFT. Such databases improve the findability and accessibility of relevant data and thereby progress the FAIR compliance of the field of catalysis. The aforementioned identifiers additionally address findability and interoperability issues ensuring more important aspects of the FAIR principles. Ultimate re-usability of knowledge and data including their semantics requires semantic technologies such as ontologies. Ready-to-use ontologies are however still missing in heterogeneous catalysis. The ChEBI Ontology for chemical entities of biological interest [155] includes the concept of a catalyst as a "role" of a material. CheBI's database and ontology also includes a number of important chemical substances that will be re-used in this work whenever possible.

Catalyst Characterization

Catalyst characterization refers to the general material properties of the catalyst during all stages of the catalytic reaction. According to the "Clean Data Handbook", there are three such stages at which the catalyst material should be characterized:

1) The “fresh” catalyst is obtained after synthesizing, calcining, pressing and sieving the raw catalytic material.

2) The “activated” catalyst is obtained when the catalyst material is put under catalytic reaction conditions which can include for example a thermal treatment. Typically, this happens during an “induction” period before the reaction takes place. This process is called Activation.

3) During the catalytic reaction the catalyst may undergo dynamic re-structuring and is classified as “spent” after the reaction.

Although most fundamental properties of the catalytic material will not change during the whole process, some will be altered and relevant properties may even change significantly. This makes it necessary to fully characterize the catalyst at all three stages to ensure completeness of the dataset. Conceptually, we will model these stages as temporal parts of the catalyst material because they exist only for a limited time. The `hasTemporalPart` relation is already defined in EMMO. Fresh, activated and spent catalyst are therefore neither subclasses of the catalyst material nor instances of it. The term we chose to represent the overall class for such temporal parts of a (catalyst) material is “snapshot” (see also Section 2.3.1). Whereas the overall catalytic performance is still a property of the catalyst material, each of its snapshots has its own set of characterizing properties such as electronic structure, structural composition etc. We then add all the different suggested characterization experiments (or short characterizations) for each type of catalyst snapshot as classes to the ontology. The different characterizations are further made subclasses of `BasicCharacterization` and `AdvancedCharacterization` where the subclass relations are equivalent to “is-a” statements in natural language, see Section 1.5.1. In other words, each of the characterization has multiple superclasses (the opposites of subclasses), i.e. for example one characterization is a subclass of `FreshCatalystCharacterization` as well as `BasicCharacterization`. Such a characterization experiment uses one or more preferred experimental techniques which determine for example structural or electronic properties of the bulk or surface. The example characterization above is further defined by its use of the technique `XRD` which implies that the bulk structure is determined via measurements of the electron density. Including the concept of “characterization experiment” in the ontology creates a natural connection between the catalyst snapshot, its property and the used experimental technique. This allows flexibility if the same property was measured more than once with different techniques. Another approach to add information on experimental techniques would be to use edge properties in the form of nested statements similarly to the similarity relations in Section 3.5.2. Because such edge properties are a very new thing in knowledge graphs and rely on the realization in RDF*, there might be compatibility issues when graphs containing such features are loaded into older triplestores. It is thus a good idea to avoid complicated descriptions and unestablished new methods whenever possible. The choice to include the “characterization experiment” in the ontology further enables the classification into basic and advanced (or mandatory and obligatory) characterizations at ontology level.

Catalyst Testing

Catalytic performance refers to how well a catalyst is suited to aid a chemical reaction towards desired reaction products. It is measured and compared in terms of multiple properties of which the three most important are the catalyst *activity*, *stability*, and *selectivity*. The selectivity always refers to a desired reaction product, so that one usually speaks about the “selectivity towards reaction product p”. Because different ways to measure and calculate

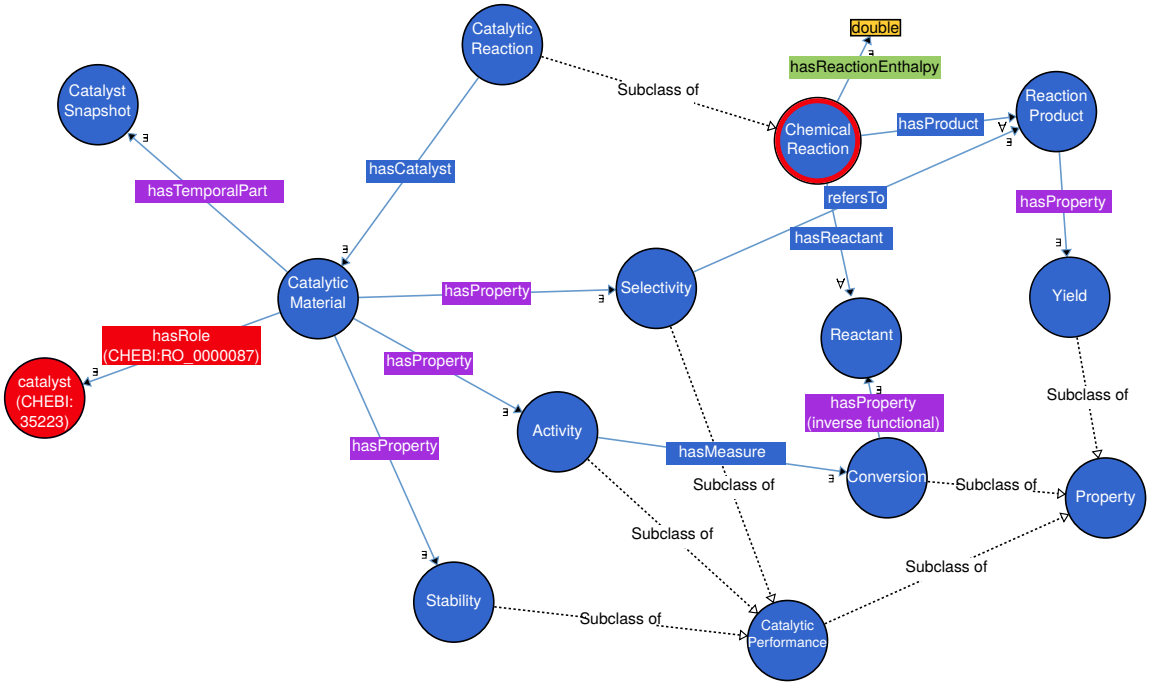


Figure 4.1: Excerpt from the Catalysis ontology developed by the author of this thesis showing the conceptual model of catalytic performance and how it is related to the underlying chemical reaction. All visualization specifics are as described in Section 1.4.1. Purple relations are imported from the EMMO and red color indicates imports from the ChEBI ontology (the blue-red node is a local class related to ChEBI via an equivalence statement).

these performance properties exist, the used formulas always have to be distributed along with the data values. In the catalysis ontology we develop, we will therefore only provide the concepts and relate them to measures that are used to quantify it. Used formulas for the calculations can be included at data level to describe specific data points. This means that data properties are provided that can be utilized to annotate the used performance measures as strings. Such strings are of course lacking any semantics. Offering semantic descriptions of multiple different established or even not so established ways to calculate and measure such properties go however beyond the scope of the current work.

Related concepts that are often used to evaluate the catalytic performance are the yield Y_j of a reaction product j and the conversion X_i of a reactant i. The conversion is used as a direct measure for the activity of the catalyst while the ratio of product yield and reactant conversion gives the selectivity of the catalyst towards this particular reaction product. Figure 4.1 illustrates that a catalytic material's performance can never be regarded isolated but is closely related to the reaction and its constituents. The concept of a chemical reaction is imported from the ChEBI ontology. As mentioned before, the ChEBI ontology defines "catalyst" as a role of a material – we also import that definition and use it in our ontology. The class `CatalyticMaterial` can have assigned all fixed properties such as the chemical formula. It is what we usually have in mind when we talk about a catalytic material. We also assign the catalytic performance properties to it. Its temporal parts, the `CatalystSnapshots`, represent the material at the different stages: before, during and after the catalytic reaction which are distinguished in the characterization.

Conditions of Catalysts

Whether a material shows its function as a useful catalyst often depends on various conditions. An example is the applied reactor technology which depends on the specifics of catalyst and reaction. In turn, different catalyst formulations are required for different reactor types or feed compositions [172]. Some process conditions are temperature, pressure, mass of the catalyst material, volume flow rate or feed compositions. The developed ontology includes the concepts of a gasfeed which flows through a chemical reactor. Specifying types of different reactors like fixed-bed or fluidized-bed reactors [176], and other experimental specifics require a dedicated ontology concerned with experimental instruments and other equipment. Such an instruments ontology called "Materials Science Lab Equipment" ontology is currently being developed within the STREAM project by our project partner at the KIT also with a focus on catalytic experiments.[1]

4.2.2 Knowledge Graph of Catalytic Propane Oxidation

Experimental data on the characterization and testing of different catalysts in propane (C_3H_8) oxidation reactions will serve as a first use-case. Representing these data with the help of the Catalysis ontology as a knowledge graph opens new possibilities on data handling and semantic enhancement. All data was produced according to the Clean Handbook [172] in the "Catalysis on Oxides" research group, led by Annette Trunschke at the Inorganic Chemistry Department of the Fritz Haber Institute. In particular the following three reactions occur simultaneously in such experiments

$$2\,C_3H_8 + O_2 \longrightarrow 2\,C_3H_6 + 2\,H_2O \tag{4.1}$$

$$C_3H_8 + 2\,O_2 \longrightarrow C_3H_4O_2 + 2\,H_2O \tag{4.2}$$

$$C_3H_8 + 5\,O_2 \longrightarrow 3\,CO_2 + 4\,H_2O\ . \tag{4.3}$$

In fact, propylene (C_3H_6) and acrylic acid ($C_3H_4O_2$) are only partial combustion products. Without a proper catalyst to control the reaction, they will eventually be further oxidized to carbon dioxide when temperatures are sufficiently high [177]. Carbon monoxide (CO) is another partial combustion by-product in these reactions. Obviously, the greenhouse gases CO and CO_2 are undesired products whereas propylene (C_3H_6) and acrylic acid ($C_3H_4O_2$) are valuable chemical substances and therefore the preferred outcomes.

Out of the thirteen different catalyst materials that have been investigated, nine vanadium-containing ones were chosen to populate the knowledge graph. Although vanadium compounds have wide applications as catalysts especially in alkane oxidation [178], the underlying mechanism are still being investigated, so that this group of materials is a good starting point for further studies. Only properties that are comparable across the different materials were considered, such as the normalized unit cell volume instead of the lattice constants, or the relative atomic contents of oxygen, carbon or vanadium. The original experimental data is given in tabular form (CSV format) and can be mapped to a graph format using the RDF Mapping Language (RML).

Given that only the catalyst stability, propane conversion, yields of different reaction products are given in the catalyst testing data tables, it is useful to calculate and add the selectivity to

[1] The ontology development can be followed in the Git repository: `https://github.com/stream-project/ontology`.

Query 4.1: SPARQL Update Request to add selectivity as catalytic performance property. It searches for yields that are properties of a reaction product or by-product and propane conversions that are properties of a catalyst of that same reaction. These quantities are then divided and bound to another variable for the selectivity value. For the new selectivity instance a new resouce identifier is also created. Refer to Section 1.3.6 for details of the SPARQL syntax.

```
prefix cat: <http://nomad-coe.eu/ontology/catalysis#>
prefix core: <http://nomad-coe.eu/ontology/core#>
prefix dat: <http://nomad-coe.eu/data/catalysis#>

INSERT{
    ?s a cat:Selectivity ;
        core:hasValue ?sval .
        cat:refersTo ?prod .
    ?cat cat:hasProperty ?s .
}
WHERE {
    ?rxn (cat:hasProduct | cat:hasByProduct ) ?prod ;
        cat:hasCatalyst ?cat .
    ?cat cat:hasProperty ?propconv .
    ?propconv a cat:PropaneConversion ;
        core:hasValue ?propconvval .
    ?prod cat:hasProperty ?y .
    ?y a cat:Yield ;
        core:hasValue ?yval .

    BIND(?yval / ?propconvval as ?sval)
    BIND(STRAFTER(STR(?y), "yield_") AS ?yield)
    BIND(IRI(CONCAT(str(dat:),"selectivity_", ?yield)) AS ?s )
}
```

the knowledge graph afterwards. With the yield Y_k of the product k and the propane conversion X_{propane}, the selectivity of the catalyst towards product k can be calculated as

$$S_k = Y_k / X_{\text{propane}} \ , \tag{4.4}$$

according to [172]. A SPARQL update request using the `INSERT` statement can be sent to the triplestore where the knowledge graph is stored. Query 4.1 calculates the selectivity on the fly and adds the new triples to the graph.

Now that the selectivity is added to the KG, let us work with the KG to grab some interesting information. To find out which product or by-product a catalyst is most selective towards, we run Query 4.2 (see below). The result is given in Table 4.1. As can be seen many of the catalysts are most selective towards the unwanted total combustion product CO_2 or the partial combustion by-product CO. The most favorable reaction product acrylic acid is only preferred by the MoVTeNbOx catalyst. Although this could be seen also in the original data, we have shown here a machine-actionable approach to answer such questions via SPARQL queries. Acrylic acid is formed after propane has transformed to propylene, but before the total oxidation product CO_2. The respective propane conversion can therefore only have intermediate values when selectivity towards acrylic acid is high, in accordance with [179].

Query 4.2: SPARQL Query for the reaction product a catalyst is most selective towards. A subquery is used to first find the highest selectivity value for each catalyst material. The associated reaction product is then identified in the outer query.

```
prefix cat: <http://nomad-coe.eu/ontology/catalysis#>
prefix core: <http://nomad-coe.eu/ontology/core#>
prefix dat: <http://nomad-coe.eu/data/catalysis#>

SELECT ?x ?prod ?maxsel
WHERE{
    ?x a cat:CatalystMaterial ;
        core:hasProperty ?y .
    ?y a cat:Selectivity;
        core:hasValue ?maxsel ;
        cat:refersTo ?prod .
    {
    SELECT ?x (MAX(?yval) as ?maxsel)
    WHERE {
        ?x a cat:CatalystMaterial ;
            core:hasProperty ?y .
        ?y a cat:Selectivity;
            core:hasValue ?yval .
    }
    GROUP BY ?x
    }
}
ORDER BY ?x
```

Table 4.1: Result to Query 4.2. Reaction product "prod" towards which a catalyst "x" is most selective ("maxsel"). For this table the values of "maxsel" were rounded to 3 decimals. The underlying knowledge graph was created from data tables from the group of Annette Trunschke at the Inorganic Chemistry Department of the FHI.

x	prod	maxsel
dat:A-VPP	dat:co_A-VPP	0.499
dat:MoVOx	dat:co_MoVOx	0.301
dat:MoVTeNbOx	dat:acrylacid_MoVTeNbOx	0.667
dat:V2O5	dat:co_V2O5	0.604
dat:VPP	dat:co2_VPP	0.457
dat:VWPOx	dat:co_VWPOx	0.511
dat:a-VOPO4	dat:propene_a-VOPO4	0.568
dat:a-VWOPO4	dat:co_a-VWOPO4	0.378
dat:b-VOPO4	dat:propene_b-VOPO4	0.607

4.3 Bridging the Data and Meta-data Levels

4.3.1 Materials Genes from Artificial Intelligence

Recently, the data from the Trunschke group at the FHI used in the previous section has been reported and used by Foppa *et al.* [179] to perform an AI study to investigate which characteristics of a catalyst material are most relevant for its catalytic performance. These key parameters can be called *materials genes* emphasizing the fundamental intrinsic role they play in catalytic processes and reactions. The symbolic regression method SISSO [180], in particular multi-task SISSO [181] was used which identifies descriptors for a target property. Such a descriptor can be a relatively complex non-linear analytical expression constructed from multiple input parameters, the so called primary features. The complexity of these expressions is beyond what a human could produce but still simple enough to be readable. Here, the target property is the selectivity of a catalyst towards acrylic acid and the primary features are all the characterizing properties of the different materials. The identified descriptor finally contains only ten of these materials properties and hence gives insight into which chemical and physical phenomena are most important for the catalytic selectivity.

Adding this newly created knowledge to our knowledge graph would increase the information density and connect the found formula with the original dataset. Optimally, a ready-to-use machine learning ontology would be used to include the SISSO study on heterogeneous catalysis. As SISSO is relatively new and used mainly in scientific context, such an ontology does not exist. In fact, available ontologies on Machine Learning algorithms (e.g. [182]) seem to be quite incomplete and can therefore only serve as a starting point to taxonomically embed more and newer algorithms such as SISSO. Detailed information about the method will be given in a text description as datatype property while the AI study itself is presented mainly as a black box. In other words, we neglect the information that SISSO creates candidate descriptors from multiple primary features to obtain specific target properties and then selects the best descriptors and identifies their coefficients. The only information we add to the knowledge graph is which characterization properties were selected as most relevant for the selectivity towards acrylic acid. It is the part of the study that is useful to gain scientific insight.

On the other hand, using this result, selectivities can likely also be predicted for other catalyst materials whose catalytic performances have not been studied experimentally. In computational materials science, this is an established way to screen a wide materials space for potential new candidates. A prerequisite is that any new material is similar enough to the ones used to build the model, so that the following found formula for the selectivity in [179] still holds:

$$S^{\text{(SISSO)}}_{\text{acrylic acid}}(T) = c_1^S(T)\left(\left(V_{\text{fr}}^{\text{pore}}\right)^2 \frac{W_{\text{rxn, wet}}}{E^{\sigma}_{\text{A, act}}} \frac{1}{\left(x^{\text{V}}_{\text{s, rxn, C3}} - x^{\text{V}}_{\text{s, act}}\right) a^{\text{C–O}}_{\text{act}}\, x^{\text{C}}_{\text{s,fr}}}\right) \tag{4.5}$$

$$+ c_2^S(T)\left(V_{\text{fr}}^{\text{pore}} V_{\text{act}}^{\text{pore}} \frac{1}{E^{\sigma}_{\text{A, act}}} \frac{x^{\text{V}}_{\text{s, rxn, wet}}}{\left(x^{\text{V}}_{\text{s, rxn, C3}} - x^{\text{V}}_{\text{s, act}}\right)\left(x^{\text{V}}_{\text{s, rxn, dry}} + x^{\text{C}}_{\text{s, fr}}\right)}\right) . \tag{4.6}$$

$c_1^S(T)$ and $c_2^S(T)$ are temperature-dependent coefficients and V^{pore} are the pore volumes where the indices $_{\text{fr}}$ and $_{\text{act}}$ refer to the fresh and activated catalysts. The work function of the catalyst is denoted by W and the E^{σ}_{A} is the activation energy of the conductivity. Wherever the subscripts $_{\text{rxn, wet}}$, $_{\text{rxn, dry}}$, $_{\text{rxn, C3}}$ occur, the quantity refers to the catalyst material

under reaction conditions where a wet, dry or C3 gas feed was used as is defined in the catalysis handbook [172]. x_s^{V} and x_s^{C} are the surface contents as atomic percentage of vanadium (V) and carbon (C). Finally, $a^{\mathrm{C–O}}$ is the fraction of surface carbon assigned to C–O bonds. It is obvious that these primary features from which the formula is built are of experimental nature and depend on the multiple different reaction conditions. When only catalytic characterization properties but no performance measures are available for a catalytic material, the selectivity can be calculated. Predicting selectivities was not done within [179], but we will here shortly sketch how future predictions can be added to the knowledge graph. While the measured selectivities are added directly as properties of the catalyst material, they should be distinguishable from such predicted selectivities. Several options are possible to indicate the predictive character:

(a) An additional class `PredictedProperty` may be defined and used to characterize all instances referring to predicted values. More relations to link the prediction model (machine learning study or formula) are also possible.

(b) One or more datatype properties may be ontologically defined and then used to distinguish between measured and predicted values including their prediction method.

(c) An object-type property can relate a predicted value via a relation of the form `wasPredictedBy` to the used model.

Option (a) is the one where the fact that this particular quantity was predicted is most obvious. The choice depends mainly on the purpose of a knowledge graph containing such information.

4.3.2 Breakdown of the Knowledge-Graph Approach

Modeling this identification of the most relevant properties is possible with different approaches:

1) Edge properties as already introduced in the last chapter can be used. The (inner) base triple statement would here include the information that a ML Study/SISSO has identified the most relevant characterization properties. The edge property (outer statement) would then specify that this refers to the selectivity towards acrylic acid.

2) Multiple relations that are specific for this SISSO approach can be defined: `identifiesDescriptor`, `hasTargetProperty`, and even `hasPrimaryFeature` if needed.

The use of edge properties is interesting because no machine learning specific ontology definitions exist yet and SISSO specific properties may be too specific for a real ontology. On the other hand, the traditional way of using two relations is more compatible with existing triplestores and other graph-based technology regarding the novelty of the RDF* syntax incorporating edge properties. For this exact reason, the second approach is chosen, even though this means that two very specific relations need to be created. Let us examine what this means in practice:

First, an instance `dat:mlstudy` of the ontology class `MachineLearningStudy` is created and related to another instance `dat:sisso` of the class `SISSO`. As primary features the characterizing property instances in the knowledge graph from Section 4.2.2 can be used and linked directly with `dat:mlstudy`. The target property is the selectivity towards acrylic acid. In Query 4.1, we have added selectivity instances to the knowledge graph for the differet catalyst materials. Each of these selectivities is an instance of the `Selectivity` class and refers

to an instance of the acrylic acid class in the ChEBI ontology. Choosing these instances as target properties is however misleading because SISSO only needs **one** target property. It is therefore important to emphasize that the target property is the *concept* of selectivity towards acrylic acid. It would be desirable to connect the `dat:mlstudy` instance directly with the class `Selectivity` with the restriction that it needs to refer to acrylic acid. Relating instances with classes is however not permitted within the OWL DL profile. Similarly, the selected final descriptor can be created as an instance of the `AnalyticalExpression` class in EMMO, but the quantities that it is created from, must be classes because they do not belong to a specific catalyst material anymore. The machine-learned conclusion, namely the found analytic expression, is meant to be valid for possibly more catalyst materials than only the nine ones in this dataset. Knowledge has been created using data with the ML approach; it has identified conceptual relations that do not belong to the data level anymore. These relations also don't relate instances with classes of the ontology directly, but do so only when certain other criteria are met. For instance, identified relevant quantities are the atomic percentage at the surface of vanadium atoms or the workfunction measure in a wet gas feed. Although this can be expressed through the means of OWL, it does not belong to the ontology level, because it is not consensual knowledge (yet) that was agreed upon in a community as the definition of an ontology requires.

This suggests that the strict separation between conceptual and data level is sometimes not appropriate. For our view on the world as citizens handling real-world objects this distinction works fine. More abstract knowledge created not by humans but by machines does not fit into this world view and therefore the approach breaks down. Hence, we suggest a different innovative perspective how to express information that is neither consensual concepts nor pure data. Another layer in between the ontology and the knowledge graph is added to represent these kinds of research results. Such a "concept" knowledge graph or "machine learning layer" can exist only in OWL Full, the non-decidable profile of OWL for which not only no reasoning tools exist today but it might never be possible to develop any.

Adding knowledge at this intermediate level is possible through SPARQL Update Requests as could already be seen for the selectivity above. However, restrictions to the classes create lots of nested statements which complicates the syntax of SPARQL requests very quickly. Here, it is therefore easier to create OWL statements with an ontology editor. Afterwards all ontology-specific headers are manually removed and the whole graph can be given a name indicating its intermediate nature. Such "named graphs" are like meta-statements, giving one or more triples a unique name and thereby allowing to distinguish where these triples came from or what their meaning is. Here, it contains in particular the information which level (ontology, knowledge graph or concept KG) the triples belong to. Storing all levels together in one large triplestore is still possible while maintaining provenance and the hierarchy.

Discussion and Outlook

In the first part of the thesis, a new relaxation scheme was presented which incorporates parametric constraints in a symmetry-reduced (or constrained) space. This algorithm can be used to relax meta-stable and unstable systems that are otherwise hardly addressable. As examples for such systems, zirconia and bismuth oxide were studied. Using a test-set of 359 different materials across different space groups and structure prototypes, the performance of the constrained relaxation was investigated. Strict symmetry preservation was shown for all materials and an average saving in the number of relaxation steps of about 50%. Finally, an electron hole polaronic distortion in rock-salt magnesium oxide was relaxed with parametric constraints demonstrating the unique advantage of the method to allow for local symmetry-breaking with known distortion patterns. For all calculations presented, the full input and output data are available in the NOMAD repository [2]. The methodology and results have already been published in [78] and cited several times of which at least two references use the constrained relaxation algorithm [53, 183]. This suggests a profound impact in the community of computational materials science. Monitoring how the symmetrized forces deviate from the full forces possibly allows the identification of new stable phases. The flexibility and generalizability of the constraints furthermore allow the extension and application to molecular systems, interfaces or transition states. Implementation in other electronic-structure theory codes is possible straightforwardly as well as including other types of coordinates.

It has been demonstrated that the new relaxation scheme can aid the search for novel materials in high-throughput studies by significantly reducing relaxation times and at the same time maintaining the correct symmetries and structures.

In part II of this thesis, we investigated how far semantic technology is able to further accelerate the search for novel materials. Multiple ontologies and knowledge graphs were built and used to annotate large amounts of data and store it in a linked data manner. Three newly developed top-down ontologies provide the semantic framework for crystal structures and materials' properties. Using the AFLOW Prototypes that have already been exploited in part I, a prototype knowledge graph was created revealing the distribution of well-known prototypes across the symmetry groups. Parallel, the comprehensive meta-data schema, the NOMAD Metainfo, was transformed to an ontology and semantically enhanced using a two-layer ontological structure ensuring extensibility. A dataset of hybrid organic-inorganic perovskites from the NOMAD Archive was converted to a knowledge graph of electronic-structure calculations using the Metainfo Ontology. Such a knowledge graph can easily be connected with new research results on the same data as discussed for the example of similarities between densities of states. Based on that, for two common photo-voltaic perovskites, similar materials were provided. As a second application, a completely different topic was

[2] https://doi.org/10.17172/NOMAD/2019.10.19-1

touched: Heterogeneous catalysis experiments are modeled in another ontology describing catalytic characterization and performance testing. For a single data table the experimental results of such a study were mapped to the ontology. Thus, a small knowledge graph of nine different Vanadium-containing catalysts for the propane oxidation reaction was created. Similarly to the perovskites dataset, additional information can be added straight-forwardly which is demonstrated at the example of selectivity. This latter case includes an AI study that outputs a formula enabling selectivity predictions. It became clear that the strict separation between ontology and knowledge graph layer is not always appropriate. For machine learning or AI studies where data is used to create new knowledge at a conceptual level, only the non-decidable variant OWL Full is able to express such research results. Another perspective is to consider a *machine community*. Let us assume that a machine-learned result has been confirmed by several different ML models or with different parameters sets so that it can be identified as robust. In this case there is a consensus within a machine community with respect to this result. It therefore qualifies as conceptual knowledge that a community has agreed upon and can go into an ontology. Such an approach could enable also automatic ontology creation or extension based on ML or even based on meta-studies on previous ML investigations.

All ontologies developed throughout this thesis are available on the Git repository of the STREAM project[3]. The knowledge graphs for the HOIPs including their enhancements are published at `https://edmond.mpdl.mpg.de/imeji/collection/mrPWUu1nzadKKUfY` in RDF/Turtle format.

We have seen how knowledge graphs based on domain ontologies can be used to link research results that would traditionally stay unconnected. This way enhanced linked-data sources of information with a high degree of connectivity are created. It is crucial to understand however, that information is not knowledge. Creating knowledge would mean to find new relationships or identify trends in data. This creation of knowledge does not work as expected though:

a) The often praised reasoning capabilities are insufficient to infer new logical consequences that are not trivial for a scientist. This is because existing reasoning software can only reliably work with description logic.

b) Exploiting tools and algorithms developed for complex networks is technically possible but usually not meaningful because such networks are assumed to have only one type of nodes and one type of edges (like social networks). For example, a cluster of nodes called a community can only be interpreted when all nodes represent the same concept, e.g. a person or a material. Knowledge graphs are by nature multi-dimensional (different edges) k-partite (different nodes) networks for which no useful analysis algorithms are known.

Hence, the procedure to gain interesting new insights on particular research questions is equivalently complex in knowledge graphs and traditional databases or datasets. Another layer of statistical analyses a.k.a. machine learning algorithms may be able to detect patterns in big data regardless of whether it is stored in relational or graph databases. Traditionally, a scientist has the necessary background knowledge to choose the right ML algorithm and input parameters. If an ontology contains that background knowledge in a formalized machine-understandable way, it can in principle act like a brain and an autonomous agent can utilize this "brain" to take these choices instead. The amount of this knowledge is however extremely vast, accumulating decades of experience and is often combined with intuition

[3] `https://github.com/stream-project/ontology/tree/nomad`

and vague notions. It would likely take many years and lots of top scientists to create such a comprehensive ontology, comparable to the development of large software packages like FHI-aims or AFLOW. The SISSO approach already implements the idea of letting the computer decide which concepts are most important to predict materials' properties. A scientist still has to choose a set of primary features first, but then relevant materials' descriptors are identified intelligently by the machine.

In the field of manufacturing, ontologies can be used to aid the design processes of construction lines and optimize several key industrial resources, e.g. the aerospace assembly lines in Airbus are modeled in an ontology [184, 185]. This is not transferable to the field of materials science where workflow design highly depends on the specific research question.

Furthermore, the stack of available semantic technologies that are ready-to-use today does not provide the necessary tools scientists in the natural sciences such as physics need to represent their research in an appropriate and efficient way. A crucial example is the lack of easy and efficient handling of numeric arrays. Storing each single array element as a unique resource is extremely storage-expensive and further neglects possible mathematical operations between arrays. Also, often the semantic annotation of each single element is unnecessary. Truly semantic relations between physical quantities are often encoded in complex mathematical formulas, operations or workflows. Re-formulating those within the OWL framework is impossible and including existing software codes or scripts does not provide the desired semantic connections. In general, there is no support for the inclusion of external scripts for example in the SPARQL query language nor are there basic frameworks for mathematical operations or rules within OWL. Since physics and chemistry are highly mathematical fields, they do not benefit enough from such linguistically motivated technologies. Indeed, ontologies can help automatically understand the context-dependent meaning of words and phrases in a running text. The already well-annotated and structured data in computational materials science databases however does not profit from that.

It can be concluded that the value semantic technologies bring to the materials sciences is very limited at the moment. This is also reflected in the prominent Gartner Hype Cycle where ontologies are currently (2020) in the through of disillusionment. Excessive expectations have grown in the last years about the idea of what an ontology is helpful for. With more realistic applications appearing now, this has led to a great disappointment. Eventually, ontologies will find their place in the technology stack at an intermediate level, e.g. as the brain behind AI technologies.

Bibliography

[1] Claudia Draxl and Matthias Scheffler. Big Data-Driven Materials Science and Its FAIR Data Infrastructure. In *Handbook of Materials Modeling*, pages 49–73. Springer International Publishing, 2020. arXiv:1904.05859, doi:10.1007/978-3-319-44677-6_104.

[2] Luca M. Ghiringhelli, Christian Carbogno, Sergey Levchenko, Fawzi Mohamed, Georg Huhs, Martin Lüders, Micael Oliveira, and Matthias Scheffler. Towards efficient data exchange and sharing for big-data driven materials science: Metadata and data formats, dec 2017. doi:10.1038/s41524-017-0048-5.

[3] 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. URL: https://www.sciencedirect.com/science/article/pii/S0927025612000687?via{%}3Dihub, doi:10.1016/j.commatsci.2012.02.002.

[4] 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), 2013. doi:10.1063/1.4812323.

[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, 2013. doi:10.1007/s11837-013-0755-4.

[6] Peter Murray-Rust and Henry S. Rzepa. CML: Evolution and design. *Journal of Cheminformatics*, 3(10):44, oct 2011. URL: https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-3-44, doi:10.1186/1758-2946-3-44.

[7] X. Gonze, C. O. Almbladh, A. Cucca, D. Caliste, C. Freysoldt, M. A.L. Marques, V. Olevano, Y. Pouillon, and M. J. Verstraete. Specification of an extensible and portable file format for electronic structure and crystallographic data. *Computational Materials Science*, 43(4):1056–1065, oct 2008. doi:10.1016/j.commatsci.2008.02.023.

[8] Casper Andersen, Rickard Armiento, Evgeny Blokhin, Gareth Conduit, Shyam Dwaraknath, Matthew L Evans, Adam Fekete, Abhijith Gopakumar, Saulius Gražulis, Vinay Hegde, Matthew Horton, Snehal Kumbhar, Nicola Marzari, Andrius Merkys, Fawzi Mohamed, Andrew Morris, Corey Oses, Giovanni Pizzi, Thomas Purcell, Gian-Marco

Rignanese, Matthias Scheffler, Markus Scheidgen, Leopold Talirz, Cormac Toher, Martin Uhrin, Donald Winston, and Chris Wolverton. The OPTIMADE Specification. Technical report, jul 2020. URL: https://zenodo.org/record/4195051, doi:10.5281/ZENODO.4195051.

[9] 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. Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):1–9, mar 2016. doi:10.1038/sdata.2016.18.

[10] Gerhard Goldbeck and Alexandra Simperler. Report on Workshop on Interoperability in Materials Modelling. Technical report, may 2018. doi:10.5281/ZENODO.1240229.

[11] Kirill Degtyarenko, Paula De matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan Mcnaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(SUPPL. 1), jan 2008. URL: https://pubmed.ncbi.nlm.nih.gov/17932057/, doi:10.1093/nar/gkm791.

[12] Huanyu Li, Rickard Armiento, and Patrick Lambrix. An Ontology for the Materials Design Domain. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12507 LNCS, pages 212–227. Springer Science and Business Media Deutschland GmbH, nov 2020. URL: https://doi.org/10.1007/978-3-030-62466-8{_}14, arXiv:2006.07712, doi:10.1007/978-3-030-62466-8_14.

[13] Toshihiro Ashino. Materials Ontology: An Infrastructure for Exchanging Materials Information and Knowledge. *Data Science Journal*, 9(0):54–61, jul 2010. URL: http://datascience.codata.org/articles/abstract/10.2481/dsj.008-041/, doi:10.2481/dsj.008-041.

[14] Kwok Cheung, John Drennan, and Jane Hunter. Towards an ontology for data-driven discovery of new materials. In *AAAI Spring Symposium: Semantic Scientific Knowledge Integration*, pages 9–14, 2008. URL: https://www.aaai.org/Papers/Symposia/Spring/2008/SS-08-05/SS08-05-003.pdf.

[15] Dennis G. Thomas, Rohit V. Pappu, and Nathan A. Baker. NanoParticle Ontology for cancer nanotechnology research. *Journal of Biomedical Informatics*, 44(1):59–74, feb 2011. doi:10.1016/j.jbi.2010.03.001.

[16] Janna Hastings, Nina Jeliazkova, Gareth Owen, Georgia Tsiliki, Cristian R. Munteanu, Christoph Steinbeck, and Egon Willighagen. eNanoMapper: Harnessing ontologies to enable data integration for nanomaterial risk assessment. *Journal of Biomedical Semantics*, 6(1):10, mar 2015. URL: http://www.jbiomedsem.com/content/6/1/10, doi:10.1186/s13326-015-0005-5.

[17] 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, nov 2009. doi:10.1016/j.cpc.2009.06.022.

[18] L. H. Thomas. The calculation of atomic fields. *Mathematical Proceedings of the Cambridge Philosophical Society*, 23(5):542–548, 1927. URL: https://www.cambridge.org/core/journals/mathematical-proceedings-of-the-cambridge-philosophical-society/article/abs/calculation-of-atomic-fields/ADCA3D21D0FACD7077B5FDBB7F3B3F3A, doi:10.1017/S0305004100011683.

[19] Enrico Fermi. Un Metodo Statistico per la Determinazione di alcune Prioprietà dell'Atomo. *Rend. Accad. Naz. Lincei.*, 6:602–607, 1927.

[20] G. Grosso and G.P. Parravicini. *Solid State Physics*. Elsevier Science, 2000. URL: https://books.google.de/books?id=-f7kenVZXlQC.

[21] Li Li and Kieron Burke. Recent Developments in Density Functional Approximations. In *Handbook of Materials Modeling*, pages 213–226. Springer International Publishing, 2020. doi:10.1007/978-3-319-44677-6_11.

[22] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865–3868, oct 1996. URL: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.77.3865, doi:10.1103/PhysRevLett.77.3865.

[23] John P. Perdew, Adrienn Ruzsinszky, Gábor I. Csonka, Oleg A. Vydrov, Gustavo E. Scuseria, Lucian A. Constantin, Xiaolan Zhou, and Kieron Burke. Restoring the density-gradient expansion for exchange in solids and surfaces. *Physical Review Letters*, 100(13):136406, apr 2008. URL: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.100.136406, arXiv:0711.0156, doi:10.1103/PhysRevLett.100.136406.

[24] Oleg A. Vydrov, Jochen Heyd, Aliaksandr V. Krukau, and Gustavo E. Scuseria. Importance of short-range versus long-range Hartree-Fock exchange for the performance of hybrid density functionals. *Journal of Chemical Physics*, 125(7):074106, aug 2006. URL: http://aip.scitation.org/doi/10.1063/1.2244560, doi:10.1063/1.2244560.

[25] T. Hahn. International Tables for Crystallography. Volume A: Space-Group Symmetry. *International Tables for Crystallography. Volume A: Space-Group Symmetry.*, 1987.

[26] U. Müller and H. Wondratschek. International Tables for Crystallography, Volume A1. *Acta Crystallographica Section A Foundations of Crystallography*, 61(a1), 2005. doi:10.1107/s0108767305094432.

[27] Mois Ilia Aroyo, Juan Manuel Perez-Mato, Cesar Capillas, Eli Kroumova, Asen Kirov, and Hans Wondratschek. Bilbao Crystallographic Server: I. Databases and crystallographic computing programs. URL: http://www.cryst.ehu.es, doi:10.1524/zkri.2006.221.1.15.

[28] Mois I. Aroyo, Asen Kirov, Cesar Capillas, J. M. Perez-Mato, and Hans Wondratschek. Bilbao Crystallographic Server. II. Representations of crystallographic point groups and

space groups. 62(2):115–128, mar 2006. URL: http://scripts.iucr.org/cgi-bin/paper?xo5013, doi:10.1107/S0108767305040286.

[29] 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. doi:10.1007/s11837-013-0755-4.

[30] Corey Oses, Cormac Toher, and Stefano Curtarolo. High-entropy ceramics, apr 2020. URL: www.nature.com/natrevmats, doi:10.1038/s41578-019-0170-8.

[31] 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, 2017. URL: https://www.sciencedirect.com/science/article/pii/S0927025617300241, arXiv:1806.07864, doi:10.1016/j.commatsci.2017.01.017.

[32] 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, 2019. URL: https://www.sciencedirect.com/science/article/pii/S0927025618307146, doi:10.1016/J.COMMATSCI.2018.10.043.

[33] 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, 2020. arXiv:2012.05961.

[34] David Hicks, Corey Oses, Eric Gossett, Geena Gomez, Richard H Taylor, Cormac Toher, Michael J Mehl, Ohad Levy, and Stefano Curtarolo. AFLOW-SYM : platform for the complete, automatic and self-consistent symmetry analysis of crystals. *Acta Crystallographica Section A Foundations and Advances*, 74(3):184–203, 2018. URL: http://scripts.iucr.org/cgi-bin/paper?S2053273318003066, doi:10.1107/S2053273318003066.

[35] R. P. Feynman. Forces in molecules. *Physical Review*, 56(4):340–343, aug 1939. URL: https://journals.aps.org/pr/abstract/10.1103/PhysRev.56.340, doi:10.1103/PhysRev.56.340.

[36] H. Hellmann. Zur Rolle der kinetischen Elektronenenergie für die zwischenatomaren Kräfte. *Zeitschrift für Physik*, 85(3-4):180–190, mar 1933. URL: https://link.springer.com/article/10.1007/BF01342053, doi:10.1007/BF01342053.

[37] Matthias Scheffler, Jean Pol Vigneron, and Giovanni B. Bachelet. Total-energy gradients and lattice distortions at point defects in semiconductors. *Physical Review B*, 31(10):6541–6551, may 1985. URL: https://journals.aps.org/prb/abstract/10.1103/PhysRevB.31.6541, doi:10.1103/PhysRevB.31.6541.

[38] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, second edition, 2006.

[39] Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint. *Trust-Region Methods*. SIAM, Philadelphia, PA, USA, 2000.

[40] C. G. Broyden. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, mar 1970. URL: https://academic.oup.com/imamat/article-lookup/doi/10.1093/imamat/6.1.76, doi:10.1093/imamat/6.1.76.

[41] R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322, mar 1970. URL: https://academic.oup.com/comjnl/article-lookup/doi/10.1093/comjnl/13.3.317, doi:10.1093/comjnl/13.3.317.

[42] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–23, jan 1970. URL: https://www.ams.org/journal-terms-of-use, doi:10.1090/s0025-5718-1970-0258249-6.

[43] D. F. Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–647, sep 1970. URL: https://www.ams.org/journal-terms-of-use, doi:10.1090/s0025-5718-1970-0274029-x.

[44] V. Havu, V. Blum, P. Havu, and M. Scheffler. Efficient O (N) integration for all-electron electronic structure calculation using numeric basis functions. *Journal of Computational Physics*, 228(22):8367–8379, dec 2009. doi:10.1016/j.jcp.2009.08.008.

[45] P. Pulay. Ab initio calculation of force constants and equilibrium geometries in polyatomic molecules. *Molecular Physics*, 17(2):197–204, 1969. URL: https://www.tandfonline.com/action/journalInformation?journalCode=tmph20, doi:10.1080/00268976900100941.

[46] Franz Knuth, Christian Carbogno, Viktor Atalla, Volker Blum, and Matthias Scheffler. All-electron formalism for total energy strain derivatives and stress tensor components for numeric atom-centered orbitals. *Computer Physics Communications*, 190:33–50, 2015. doi:10.1016/j.cpc.2015.01.003.

[47] N.W. Ashcroft and N.D. Mermin. *Solid State Physics*. Saunders College, Philadelphia, 1976.

[48] Stefano Baroni, Stefano de Gironcoli, Andrea Dal Corso, and Paolo Giannozzi. Phonons and related crystal properties from density-functional perturbation theory. *Rev. Mod. Phys.*, 73:515–562, Jul 2001. URL: https://link.aps.org/doi/10.1103/RevModPhys.73.515, doi:10.1103/RevModPhys.73.515.

[49] Paolo Giannozzi and Stefano Baroni. Density-Functional Perturbation Theory. In *Handbook of Materials Modeling*, pages 195–214. Springer Netherlands, 2005. URL: https://link.springer.com/chapter/10.1007/978-1-4020-3286-8_11, doi:10.1007/978-1-4020-3286-8_11.

[50] K. Parlinski, Z. Q. Li, and Y. Kawazoe. First-principles determination of the soft mode in cubic zro_2. *Phys. Rev. Lett.*, 78:4063–4066, May 1997. URL: https://link.aps.org/doi/10.1103/PhysRevLett.78.4063, doi:10.1103/PhysRevLett.78.4063.

[51] A Togo and I Tanaka. First principles phonon calculations in materials science. *Scr. Mater.*, 108:1–5, Nov 2015.

[52] Laurent Chaput, Atsushi Togo, Isao Tanaka, and Gilles Hug. Phonon-phonon interactions in transition metals. *Physical Review B - Condensed Matter and Materials Physics*, 84(9):094302, sep 2011. URL: https://journals.aps.org/prb/abstract/10.1103/PhysRevB.84.094302, arXiv:1103.0137, doi:10.1103/PhysRevB.84.094302.

[53] Florian Knoop, Thomas A. R. Purcell, Matthias Scheffler, and Christian Carbogno. Anharmonicity measure for materials. *Phys. Rev. Materials*, 4:083809, Aug 2020. URL: https://link.aps.org/doi/10.1103/PhysRevMaterials.4.083809, doi:10.1103/PhysRevMaterials.4.083809.

[54] Feliciano Giustino. Electron-phonon interactions from first principles. *Reviews of Modern Physics*, 89(1):015003, feb 2017. URL: https://journals.aps.org/rmp/abstract/10.1103/RevModPhys.89.015003, arXiv:1603.06965, doi:10.1103/RevModPhys.89.015003.

[55] Sebastian Kokott, Sergey V. Levchenko, Patrick Rinke, and Matthias Scheffler. First-principles supercell calculations of small polarons with proper account for long-range polarization effects. *New Journal of Physics*, 20(3):033023, 2018. URL: http://stacks.iop.org/1367-2630/20/i=3/a=033023?key=crossref.0b70ba9ef35bb1d1f6a37cfe8e36ab3d, doi:10.1088/1367-2630/aaaf44.

[56] Abanti Nag and V Shubha. Oxide thermoelectric materials: A structure-property relationship. *Journal of Electronic Materials*, 43(4):962–977, 2014. doi:10.1007/s11664-014-3024-6.

[57] Renier Arabolla Rodríguez, Eduardo L. Pérez-Cappe, Yodalgis Mosqueda Laffita, Armando Chávez Ardanza, Jaime Santoyo Salazar, Manuel Ávila Santos, Miguel A. Aguilar Frutis, Nelcy Della Santina Mohalem, and Oswaldo Luiz Alves. Structural defects in LiMn2O4 induced by gamma radiation and its influence on the Jahn-Teller effect. *Solid State Ionics*, 324:77–86, 2018. URL: https://www.sciencedirect.com/science/article/pii/S0167273817310986?via{%}3Dihub, doi:10.1016/j.ssi.2018.06.007.

[58] Joseph C.A. Prentice, Bartomeu Monserrat, and R J Needs. First-principles study of the dynamic Jahn-Teller distortion of the neutral vacancy in diamond. *Physical Review B*, 95(1):14108, 2017. URL: https://journals.aps.org/prb/pdf/10.1103/PhysRevB.95.014108, doi:10.1103/PhysRevB.95.014108.

[59] Robert Evarestov, Evgeny Blokhin, Denis Gryaznov, Eugene A Kotomin, Rotraut Merkle, and Joachim Maier. Jahn-Teller effect in the phonon properties of defective SrTiO 3 from first principles. *Physical Review B - Condensed Matter and Materials Physics*, 85(17):174303, 2012. URL: https://journals.aps.org/prb/pdf/10.1103/PhysRevB.85.174303, doi:10.1103/PhysRevB.85.174303.

[60] Alan G. Sharpe Catherine Housecroft. *Inorganic Chemistry*. Prentice Hall, 4 edition, 2012.

[61] Zdenek Dohnálek, Igor Lyubinetsky, and Roger Rousseau. Thermally-driven processes on rutile TiO2(1 1 0)-(1 $\times$ 1): A direct view at the atomic scale. *Progress in Surface Science*, 85(5-8):161–205, 2010. doi:10.1016/j.progsurf.2010.03.001.

[62] Michael A Henderson. A surface science perspective on TiO2 photocatalysis. *Surface Science Reports*, 66(6-7):185–297, 2011. URL: www.elsevier.com/locate/surfrep, doi:10.1016/j.surfrep.2011.01.001.

[63] Alexander J.E. Rettie, William D Chemelewski, David Emin, and C Buddie Mullins. Unravelling Small-Polaron Transport in Metal Oxide Photoelectrodes. *Journal of Physical Chemistry Letters*, 7(3):471–479, 2016. URL: https://pubs.acs.org/sharingguidelines, doi:10.1021/acs.jpclett.5b02143.

[64] Jürgen Hafner. Ab-initio simulations of materials using VASP: Density-functional theory and beyond. *Journal of Computational Chemistry*, 29(13):2044–2078, 2008. URL: http://doi.wiley.com/10.1002/jcc.21057, doi:10.1002/jcc.21057.

[65] X Gonze, J.-M. Beuken, R Caracas, F Detraux, M Fuchs, G.-M. Rignanese, L Sindic, M Verstraete, G Zerah, F Jollet, M Torrent, A Roy, M Mikami, Ph. Ghosez, J.-Y. Raty, and D C Allan. First-principles computation of material properties: the ABINIT software project. *Computational Materials Science*, 25(3):478–492, 2002. URL: https://www.sciencedirect.com/science/article/pii/S0927025602003257?via{%}3Dihub, doi:10.1016/S0927-0256(02)00325-7.

[66] Andris Gulans, Stefan Kontur, Christian Meisenbichler, Dmitrii 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-body perturbation theory. *Journal of Physics Condensed Matter*, 26(36):24, sep 2014. URL: https://iopscience.iop.org/article/10.1088/0953-8984/26/36/363202, doi:10.1088/0953-8984/26/36/363202.

[67] Leeor Kronik, Adi Makmal, Murilo L. Tiago, M. M.G. Alemany, Manish Jain, Xiangyang Huang, Yousef Saad, and James R. Chelikowsky. PARSEC - The pseudopotential algorithm for real-space electronic structure calculations: Recent advances and novel applications to nano-structures. *Physica Status Solidi (B) Basic Research*, 243(5):1063–1079, 2006. URL: http://doi.wiley.com/10.1002/pssb.200541463, doi:10.1002/pssb.200541463.

[68] Christoph Freysoldt. On-the-fly parameterization of internal coordinate force constants for quasi-Newton geometry optimization in atomistic calculations. *Computational Materials Science*, 133:71–81, 2017. URL: https://www.sciencedirect.com/science/article/pii/S0927025617301131?via=ihub, doi:10.1016/j.commatsci.2017.03.001.

[69] Chiara Panosetti, Konstantin Krautgasser, Dennis Palagin, Karsten Reuter, and Reinhard J Maurer. Global Materials Structure Search with Chemically Motivated Coordinates. *Nano Letters*, 15(12):8044–8048, 2015. URL: https://pubs.acs.org/sharingguidelines, doi:10.1021/acs.nanolett.5b03388.

[70] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, Andrea Dal Corso, Stefano de Gironcoli, Stefano Fabris, Guido Fratesi, Ralph Gebauer, Uwe Gerstmann, Christos Gougoussis, Anton Kokalj, Michele Lazzeri, Layla Martin-Samos, Nicola Marzari, Francesco Mauri, Riccardo Mazzarello, Stefano Paolini, Alfredo Pasquarello, Lorenzo Paulatto, Carlo Sbraccia, Sandro Scandolo, Gabriele Sclauzero, Ari P Seitsonen, Alexander Smogunov, Paolo Umari, and Renata M Wentzcovitch. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter*, 21(39):395502, 2009. URL: http://stacks.iop.org/0953-8984/21/i=39/a=395502?key=crossref.c21336c286fa6d3db893262ae3f6e151, doi:10.1088/0953-8984/21/39/395502.

[71] M. H. Bocanegra-Bernal and S Díaz de la Torre. Phase transitions in zirconium dioxide and related materials for high performance engineering ceramics. *Journal of Materials Science*, 37(23):4947–4971, 2002. URL: http://link.springer.com/10.1023/A:1021099308957, doi:10.1023/A:1021099308957.

[72] Christian Carbogno, Carlos G Levi, Chris G Van De Walle, and Matthias Scheffler. Ferroelastic switching of doped zirconia: Modeling and understanding from first principles. *Physical Review B - Condensed Matter and Materials Physics*, 90(14):144109, 2014. doi:10.1103/PhysRevB.90.144109.

[73] Akifumi Matsumoto, Yukinori Koyama, and Isao Tanaka. Structures and energetics of Bi_2O_3 polymorphs in a defective fluorite family derived by systematic first-principles lattice dynamics calculations. *Physical Review B*, 81(9):94117, 2010. URL: https://link.aps.org/doi/10.1103/PhysRevB.81.094117, doi:10.1103/PhysRevB.81.094117.

[74] Michel Drache, Pascal Roussel, and Jean-Pierre Wignacourt. Structures and Oxide Mobility in Bi-Ln-O Materials: Heritage of Bi_2O_3. *Chemical Reviews*, 107(1):80–96, 2007. URL: https://pubs.acs.org/doi/10.1021/cr050977s., doi:10.1021/cr050977s.

[75] K. Doll. Analytical stress tensor and pressure calculations with the CRYSTAL code. *Molecular Physics*, 108(3-4):223–227, 2010. doi:10.1080/00268970903193028.

[76] C Radhakrishna Rao and Sujit Kumar Mitra. Generalized inverse of a matrix and its applications. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*, pages 601–620, Berkeley, Calif., 1972. University of California Press. URL: https://projecteuclid.org/euclid.bsmsp/1200514113.

[77] 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, jun 2017. URL: https://iopscience.iop.org/article/10.1088/1361-648X/aa680e, doi:10.1088/1361-648X/aa680e.

[78] Maja Olivia Lenz, Thomas A.R. Purcell, David Hicks, Stefano Curtarolo, Matthias Scheffler, and Christian Carbogno. Parametrically constrained geometry relaxations for high-throughput materials science. *npj Computational Materials*, 5(1):1–10, dec 2019. doi:10.1038/s41524-019-0254-4.

[79] 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. URL: http://aip.scitation.org/doi/10.1063/1.4812323, doi:10.1063/1.4812323.

[80] David Hicks, Cormac Toher, Denise C. Ford, Frisco Rose, Carlo de Santo, Ohad Levy, Michael J. Mehl, and Stefano Curtarolo. AFLOW-XtalFinder: a reliable choice to identify crystalline prototypes, oct 2020. arXiv:2010.04222, doi:10.1038/s41524-020-00483-4.

[81] H Burzlaff and Y Malinovsky. A Procedure for the Clasification of Non-Organic Crystal Structures. I. Theoretical Background. *Acta Crystallographica Section A Foundations of*

Crystallography, 53(2):217–224, 1997. URL: http://scripts.iucr.org/cgi-bin/paper?S0108767396013852, doi:10.1107/S0108767396013852.

[82] Atsushi Togo and Isao Tanaka. *spglib*: a software library for crystal symmetry search. 2018. URL: http://arxiv.org/abs/1808.01590, arXiv:1808.01590.

[83] Tony Hey, Stewart Tansley, Kristin Tolle, et al. *The Fourth Paradigm: Data-Intensive Scientific Discovery*, volume 1. Microsoft research Redmond, WA, 2009.

[84] Luca M. Ghiringhelli, Patrick Lambrix, Javad Chamanara, Carsten Baldauf, Tatyana Sheveleva, Benjamin Regler, Alvin Noe Ladines, Christoph Koch, Christof Wöll, Stefano Cozzini, Astrid Schneidewind, Micael Oliveira, Sergey Levchenko, Claudia Draxl, Pasquale Pavone, Denis Usvyat, James Kermode, Tristan Bereau, Christian Carbogno, Omar Valsson, Markus Kühbach, Chuanxun Su, Ron Miller, Berk Onat, Stefano Curtarolo, Shyam Dwaraknath, Adam Michalchuk, Gian-Marco Rignanese, Jörg Schaarschmidt, Adam Fekete, Markus Scheidgen, Christoph Koch, Astrid Schneidewind, Maja-Olivia Lenz-Himmer, and Matthias Scheffler. Shared Metadata for Big-Data-Driven Materials Science. 2021. Unpublished.

[85] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, G Sherlock, and Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *NATURE GENETICS*, 25(1):25–29, MAY 2000. doi:{10.1038/75556}.

[86] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, jun 1993. URL: https://www.sciencedirect.com/science/article/pii/S1042814383710083, doi:10.1006/KNAC.1993.1008.

[87] Rudi Studer, V.Richard Benjamins, and Dieter Fensel. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1-2):161–197, mar 1998. URL: https://www.sciencedirect.com/science/article/pii/S0169023X97000566, doi:10.1016/S0169-023X(97)00056-6.

[88] Ontotext. What are Ontologies and What are the Benefits of Using Ontologies. https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies/. Accessed: 2020-08-26.

[89] M. Hepp. Ontologies: state of the art, business potential and grand challenges. In M. Hepp, P. De Leenheer, A. De Moor, and Y. Sure, editors, *Ontology management – semantic web, semantic web services and business applications*, pages 3–22. Springer International Publishing, 2008.

[90] Dan Brickley and R.V. Guha. RDF Schema 1.1 - W3C Recommendation 25 February 2014. Technical report, World Wide Web Consortium (W3C), February 2014. URL: http://www.w3.org/TR/rdf-schema/.

[91] Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah McGuinness, Peter Patel-Schneijder, and Lynn Andrea Stein. OWL Web Ontology Language Reference. Recommendation, World Wide Web Consortium (W3C), February 2004. See http://www.w3.org/TR/owl-ref/.

[92] Raymond Reiter. *Deductive Question-Answering on Relational Data Bases*, pages 149–177. Springer US, Boston, MA, 1978. URL: https://doi.org/10.1007/978-1-4684-3384-5{_}6, doi:10.1007/978-1-4684-3384-5_6.

[93] K. R. Chowdhary. *Networks-Based Representation*, pages 179–215. Springer India, New Delhi, 2020. URL: https://doi.org/10.1007/978-81-322-3972-7{_}7, doi:10.1007/978-81-322-3972-7_7.

[94] Olaf Hartig. What are Ontologies and What are the Benefits of Using Ontologies. https://blog.liu.se/olafhartig/2019/01/10/position-statement-rdf-star-and-sparql-star/. Accessed: 2020-11-17.

[95] Natalya F. Noy and Deborah L. McGuinness. Ontology development 101: A guide to creating your first ontology. *Knowledge Systems Laboratory*, 32, 01 2001.

[96] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001. URL: http://www.jstor.org/stable/26059207.

[97] Franck Michel, Catherine Faron-Zucker, and Johan Montagnat. Bridging the Semantic Web and NoSQL Worlds: Generic SPARQL Query Translation and Application to MongoDB. In Abdelkader Hameurlain, Roland Wagner, Franck Morvan, and Lynda Tamine, editors, *Transactions on Large-Scale Data-and Knowledge-Centered Systems XL*, pages 125–165. Springer, Berlin, Heidelberg, 2019. doi:https://doi.org/10.1007/978-3-662-58664-8.

[98] Stijn Heymans, Li Ma, Darko Anicic, Zhilei Ma, Nathalie Steinmetz, Yue Pan, Jing Mei, Achille Fokoue, Aditya Kalyanpur, Aaron Kershenbaum, Edith Schonberg, Kavitha Srinivas, Cristina Feier, Graham Hench, Branimir Wetzstein, and Uwe Keller. Ontology reasoning with large data repositories. In Martin Hepp, Pieter De Leenheer, Aldo De Moor, and York Sure, editors, *Ontology Management: Semantic Web, Semantic Web Services, and Business Applications*, pages 89–128. Springer US, Boston, MA, 2008. doi:10.1007/978-0-387-69900-4_4.

[99] Steffen Lohmann, Stefan Negru, Florian Haag, and Thomas Ertl. Visualizing ontologies with VOWL. *Semantic Web*, 7(4):399–419, 2016. URL: http://dx.doi.org/10.3233/SW-150200, doi:10.3233/SW-150200.

[100] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a Web of open data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4825 LNCS, pages 722–735. Springer, Berlin, Heidelberg, nov 2007. URL: http://www.mediawiki.org, doi:10.1007/978-3-540-76298-0_52.

[101] Marek Dudás, Steffen Lohmann, Vojtch Svátek, and Dmitry Pavlov. Ontology visualization methods and tools: A survey of the state of the art, 2018. doi:10.1017/S0269888918000073.

[102] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343. IEEE Comput. Soc. Press, 1996. URL: http://ieeexplore.ieee.org/document/545307/, doi:10.1109/VL.1996.545307.

[103] Alexander Veremyev, Laalitha Liyanage, Marco Fornari, Vladimir Boginski, Stefano Curtarolo, Sergiy Butenko, and Marco Buongiorno Nardelli. Networks of materials: Construction and structural analysis. *AIChE Journal*, 67(3):e17051, mar 2021. doi:10.1002/aic.17051.

[104] Aba Sah Dadzie and Matthew Rowe. Approaches to visualising Linked Data: A survey, jan 2011. doi:10.3233/SW-2011-0037.

[105] Josep Maria Brunetti, Sören Auer, Roberto García, Jakub Klímek, and Martin Nečaský. Formal linked data visualization model. In *ACM International Conference Proceeding Series*, pages 309–318, 2013. doi:10.1145/2539150.2539162.

[106] 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, feb 2015. arXiv:1412.4096, doi:10.1021/cm503507h.

[107] Muratahan Aykol, Vinay I. Hegde, Linda Hung, Santosh Suram, Patrick Herring, Chris Wolverton, and Jens S. Hummelshøj. Network analysis of synthesizable materials discovery. *Nature Communications*, 10(1):1–7, dec 2019. arXiv:1806.05772, doi:10.1038/s41467-019-10030-5.

[108] Vinay I. Hegde, Muratahan Aykol, Scott Kirklin, and Chris Wolverton. The phase stability network of all inorganic materials. *Science Advances*, 6(9):eaay5606, feb 2020. URL: https://advances.sciencemag.org/content/6/9/eaay5606https://advances.sciencemag.org/content/6/9/eaay5606.abstract, doi:10.1126/sciadv.aay5606.

[109] Mark Newman. *Networks: An Introduction*. Oxford University Press, 1 edition, 2010.

[110] M. E.J. Newman. The structure and function of complex networks, aug 2003. URL: http://www.siam.org/journals/sirev/45-2/42480.html, arXiv:0303516, doi:10.1137/S003614450342480.

[111] Frank Emmert-Streib. A brief introduction to complex networks and their analysis. In *Structural Analysis of Complex Networks*, pages 1–26. Birkhäuser Boston, 2011. URL: https://link.springer.com/chapter/10.1007/978-0-8176-4789-6{_}1, doi:10.1007/978-0-8176-4789-6_1.

[112] M. E.J. Newman. Mixing patterns in networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 67(2):13, feb 2003. URL: https://journals.aps.org/pre/abstract/10.1103/PhysRevE.67.026126, arXiv:0209450, doi:10.1103/PhysRevE.67.026126.

[113] M. E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 69(2 2):026113, feb 2004. URL: https://journals.aps.org/pre/abstract/10.1103/PhysRevE.69.026113, arXiv:0308217, doi:10.1103/PhysRevE.69.026113.

[114] Lauri Himanen, Amber Geurts, Adam Stuart Foster, and Patrick Rinke. Data-Driven Materials Science: Status, Challenges, and Perspectives, nov 2019. doi:10.1002/advs.201900808.

[115] Leopold Talirz, Snehal Kumbhar, Elsa Passaro, Aliaksandr V. Yakutovich, Valeria Granata, Fernando Gargiulo, Marco Borelli, Martin Uhrin, Sebastiaan P. Huber, Spyros Zoupanos, Carl S. Adorf, Casper Welzel Andersen, Ole Schütt, Carlo A. Pignedoli, Daniele Passerone, Joost VandeVondele, Thomas C. Schulthess, Berend Smit, Giovanni Pizzi, and Nicola Marzari. Materials Cloud, a platform for open computational science. *Scientific Data*, 7(1):1–12, dec 2020. URL: www.nature.com/scientificdata, arXiv:2003.12510, doi:10.1038/s41597-020-00637-5.

[116] Andrius Merkys, Nicolas Mounet, Andrea Cepellotti, Nicola Marzari, Saulius Gražulis, and Giovanni Pizzi. A posteriori metadata from automated provenance tracking: Integration of AiiDA and TCOD. *Journal of Cheminformatics*, 9(1):56, 2017. URL: https://jcheminf.springeropen.com/articles/10.1186/s13321-017-0242-y, arXiv:1706.08704v3, doi:10.1186/s13321-017-0242-y.

[117] C. Ortiz, O. Eriksson, and M. Klintenberg. Data mining and accelerated electronic structure theory as a tool in the search for new functional materials. *Computational Materials Science*, 44(4):1042–1049, feb 2009. arXiv:0808.2125, doi:10.1016/j.commatsci.2008.07.016.

[118] Open materials database. https://openmaterialsdb.se/. Accessed: 2021-05-31.

[119] 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 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. mar 2021. URL: http://arxiv.org/abs/2103.02068, arXiv:2103.02068.

[120] Micael J.T. Oliveira, Nick Papior, Yann Pouillon, Volker Blum, Emilio Artacho, Damien Caliste, Fabiano Corsetti, Stefano De Gironcoli, Alin M. Elena, Alberto García, Víctor M. García-Suárez, Luigi Genovese, William P. Huhn, Georg Huhs, Sebastian Kokott, Emine Küçükbenli, Ask H. Larsen, Alfio Lazzaro, Irina V. Lebedeva, Yingzhou Li, David López-Durán, Pablo López-Tarifa, Martin Lüders, Miguel A.L. Marques, Jan Minar, Stephan Mohr, Arash A. Mostofi, Alan O'Cais, Mike C. Payne, Thomas Ruh, Daniel G.A. Smith, José M. Soler, David A. Strubbe, Nicolas Tancogne-Dejean, Dominic Tildesley, Marc Torrent, and Victor Wen Zhe Yu. The CECAM electronic structure library and the modular software development paradigm. *Journal of Chemical Physics*, 153(2):204108, jul 2020. arXiv:2005.05756, doi:10.1063/5.0012901.

[121] Kyle Michel and Bryce Meredig. Beyond bulk single crystals: A data format for all materials structure-property-processing relationships. *MRS Bulletin*, 41(8):617–622, aug 2016. doi:10.1557/mrs.2016.166.

[122] E. B. Tadmor, R. S. Elliott, J. P. Sethna, R. E. Miller, and C. A. Becker. The potential of atomistic simulations and the knowledgebase of interatomic models, jul 2011. URL: www.tms.org/jom.html, doi:10.1007/s11837-011-0102-6.

[123] P. V. D. Vet, P. Speel, and N. Mars. The plinius ontology of ceramic materialspaul. 1994.

[124] Yingzhong Zhang, Xiaofang Luo, Yong Zhao, and Hong Chao Zhang. An ontology-based knowledge framework for engineering material selection. *Advanced Engineering Informatics*, 29(4):985–1000, oct 2015. doi:10.1016/j.aei.2015.09.002.

[125] Kwok Cheung, John Drennan, and Jane Hunter. Towards an Ontology for Data-driven Discovery of New Materials Introduction and Objectives. Technical report.

[126] Manoj Bhat, Sapan Shah, Prasenjit Das, Prabash Kumar, Nagesh Kulkarni, Smita S. Ghaisas, and Sreedhar S. Reddy. PREMΛP: Knowledge Driven Design of Materials and Engineering Process. pages 1315–1329. Springer, India, 2013. URL: https://link.springer.com/chapter/10.1007/978-81-322-1050-4{_}105, doi:10.1007/978-81-322-1050-4_105.

[127] Xiaoming Zhang, Changjun Hu, and Huayu Li. Semantic query on materials data based on mapping matml to an OWL ontology. *Data Science Journal*, 8(0):1–17, jan 2009. URL: http://datascience.codata.org/articles/10.2481/dsj.8.1/http://datascience.codata.org/articles/abstract/10.2481/dsj.8.1/, doi:10.2481/dsj.8.1.

[128] Martin Thomas Horsch, Silvia Chiacchiera, Youness Bami, Georg J. Schmitz, Gabriele Mogni, Gerhard Goldbeck, and Emanuele Ghedini. Reliable and interoperable computational molecular engineering: 2. Semantic interoperability based on the European Materials and Modelling Ontology. jan 2020. URL: http://arxiv.org/abs/2001.04175, arXiv:2001.04175.

[129] Olga Nevzorova, Nikita Zhiltsov, Alexander Kirillovich, and Evgeny Lipachev. $OntoMath^{\mathrm{PRO}}$ Ontology: A Linked Data Hub for Mathematics. *Communications in Computer and Information Science*, 468:105–119, jul 2014. URL: http://arxiv.org/abs/1407.4833, arXiv:1407.4833.

[130] Sydney Hall and Brian McMahon. International Tables For Crystallography Volume G: Definition and Exchange of Crystallographic Data. *International Union Of Crystallography*, 2005.

[131] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009. URL: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[132] ISO - ISO 80000-1:2009 - Quantities and units — Part 1: General. URL: https://www.iso.org/standard/30669.html.

[133] A. D. McNaught and A. Wilkinson. *IUPAC Compendium of Chemical Terminology (the "Gold Book")*. IUPAC Nomenclature Books Series ("Color Books"). Blackwell Science, Oxford, 2nd edition, 1997. Online version (2019-) created by S. J. Chalk. doi:https://doi.org/10.1351/goldbook.

[134] Fairsharing.org: Qudt; quantities, units, dimensions and types. https://doi.org/10.25504/FAIRsharing.d3pqw7. Accessed: Jun 25 2021 8:39 a.m.

[135] Leslie F. Sikos. Description Logics: Formal Foundation for Web Ontology Engineering. In *Description Logics in Multimedia Reasoning*, pages 67–120. Springer International Publishing, 2017. URL: https://link.springer.com/chapter/10.1007/978-3-319-54066-5{_}4, doi:10.1007/978-3-319-54066-5_4.

[136] Phil Archer. Data Catalog Vocabulary (DCAT) (W3C Recommendation). Online, January 2014. URL: https://www.w3.org/TR/vocab-dcat/.

[137] Stuart L Weibel and Traugott Koch. The dublin core metadata initiative. *D-lib magazine*, 6(12):1082–9873, 2000.

[138] Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. Prov-o: The prov ontology. 2013.

[139] Carl Boettiger. *rdflib: A high level wrapper around the redland package for common rdf applications*, 2018. doi:10.5281/zenodo.1098478.

[140] Mike Graves, Adam Constabaris, and Dan Brickley. Foaf: Connecting people on the semantic web. *Cataloging & classification quarterly*, 43(3-4):191–202, 2007.

[141] Renato Iannella and James McKinney. vcard ontology-for describing people and organizations. *W3C Group Note NOTE-vcard-rdf-20140522*, 2014.

[142] Alistair Miles, Brian Matthews, Michael Wilson, and Dan Brickley. Skos core: simple knowledge organisation for the web. In *International conference on dublin core and metadata applications*, pages 3–10, 2005.

[143] Chiho Kim, Tran Doan Huan, Sridevi Krishnan, and Rampi Ramprasad. A hybrid organic-inorganic perovskite dataset. *Scientific Data*, 4(1):170057, 2017. doi:10.1038/sdata.2017.57.

[144] Joseph S. Manser, Jeffrey A. Christians, and Prashant V. Kamat. Intriguing Optoelectronic Properties of Metal Halide Perovskites, nov 2016. URL: https://pubs.acs.org/sharingguidelines, doi:10.1021/acs.chemrev.6b00136.

[145] Thomas M. Brenner, David A. Egger, Leeor Kronik, Gary Hodes, and David Cahen. Hybrid organic - Inorganic perovskites: Low-cost semiconductors with intriguing charge-transport properties, 2016. doi:10.1038/natrevmats.2015.7.

[146] Jaeki Jeong, Minjin Kim, Jongdeuk Seo, Haizhou Lu, Paramvir Ahlawat, Aditya Mishra, Yingguo Yang, Michael A. Hope, Felix T. Eickemeyer, Maengsuk Kim, Yung Jin Yoon, In Woo Choi, Barbara Primera Darwich, Seung Ju Choi, Yimhyun Jo, Jun Hee Lee, Bright Walker, Shaik M. Zakeeruddin, Lyndon Emsley, Ursula Rothlisberger, Anders Hagfeldt, Dong Suk Kim, Michael Grätzel, and Jin Young Kim. Pseudo-halide anion engineering for α-FAPbI3 perovskite solar cells. *Nature*, 592(7854):381–385, apr 2021. doi:10.1038/s41586-021-03406-5.

[147] Jean-Baptiste Lamy. Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies. *Artificial intelligence in medicine*, 80:11—28, July 2017. doi:10.1016/j.artmed.2017.07.002.

[148] Liyang Yu. *A developer's guide to the semantic web*. 2011. doi:10.1007/978-3-642-15970-1.

[149] Stefano Ceri, Alessandro Bozzon, Marco Brambilla, Emanuele Della Valle, Piero Fraternali, and Silvia Quarteroni. *Semantic Search*, pages 181–206. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. doi:10.1007/978-3-642-39314-3_12.

[150] Marcelo Arenas, Bernardo Cuenca Grau, Evgeny Kharlamov, Šarunas Marciuška, Dmitriy Zheleznyakov, and Ernesto Jimenez-Ruiz. Towards semantic faceted search. In *WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web*, pages 219–220. Association for Computing Machinery, Inc, apr 2014. doi:10.1145/2567948.2577381.

[151] Sven Rühle. Tabulated values of the Shockley-Queisser limit for single junction solar cells. *Solar Energy*, 130:139–147, jun 2016. doi:10.1016/j.solener.2016.02.015.

[152] Corey Oses, Eric Gossett, David Hicks, Frisco Rose, Michael J. Mehl, Eric Perim, Ichiro Takeuchi, Stefano Sanvito, Matthias Scheffler, Yoav Lederer, Ohad Levy, Cormac Toher, and Stefano Curtarolo. AFLOW-CHULL: Cloud-Oriented Platform for Autonomous Phase Stability Analysis. *Journal of Chemical Information and Modeling*, 58(12):2477–2490, dec 2018. URL: https://pubs.acs.org/sharingguidelines, arXiv:1806.06901, doi:10.1021/acs.jcim.8b00393.

[153] Sebastian Siol. Accessing Metastability in Heterostructural Semiconductor Alloys. *physica status solidi (a)*, 216(15):1800858, aug 2019. URL: https://www.onlinelibrary.wiley.com/doi/full/10.1002/pssa.201800858, doi:10.1002/PSSA.201800858.

[154] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research*, 47(D1):D1102–D1109, 10 2018. arXiv:https://academic.oup.com/nar/article-pdf/47/D1/D1102/27437306/gky1033.pdf, doi:10.1093/nar/gky1033.

[155] Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, 44(D1):D1214–D1219, oct 2016. URL: http://www.ebi.ac.uk/chebi/, doi:10.1093/nar/gkv1031.

[156] Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen Ling Shaiu, and Lawrence W. Wright. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30–43, feb 2007. doi:10.1016/j.jbi.2006.02.013.

[157] Steven H. Brown, Peter L. Elkin, S. Trent Rosenbloom, Casey Husser, Brent A. Bauer, Michael J. Lincoln, John Carter, Mark Erlbaum, and Mark S. Tuttle. VA national drug file reference terminology: A cross-institutional content coverage study. *Studies in Health Technology and Informatics*, 107(Pt 1):477–481, jan 2004. URL: https://europepmc.org/article/med/15360858, doi:10.3233/978-1-60750-949-3-477.

[158] Joseph Berry, Tonio Buonassisi, David A. Egger, Gary Hodes, Leeor Kronik, Yueh Lin Loo, Igor Lubomirsky, Seth R. Marder, Yitzhak Mastai, Joel S. Miller, David B. Mitzi, Yaron Paz, Andrew M. Rappe, Ilan Riess, Boris Rybtchinski, Oscar Stafsudd, Vladan Stevanovic, Michael F. Toney, David Zitoun, Antoine Kahn, David Ginley, and David Cahen. Hybrid Organic-Inorganic Perovskites (HOIPs): Opportunities and Challenges. *Advanced Materials*, 27(35):5102–5112, sep 2015. URL: www.advmat.de, doi:10.1002/adma.201502294.

[159] Jongseob Kim, Sung Hoon Lee, Jung Hoon Lee, and Ki Ha Hong. The role of intrinsic defects in methylammonium lead iodide perovskite. *Journal of Physical Chemistry Letters*, 5(8):1312–1317, apr 2014. URL: https://pubs.acs.org/sharingguidelines, doi:10.1021/jz500370k.

[160] Jon M. Azpiroz, Edoardo Mosconi, Juan Bisquert, and Filippo De Angelis. Defect migration in methylammonium lead iodide and its role in perovskite solar cell operation.

Energy and Environmental Science, 8(7):2118–2127, jul 2015. URL: www.rsc.org/ees, doi:10.1039/c5ee01265a.

[161] Jason J. Yoo, Gabkyung Seo, Matthew R. Chua, Tae Gwan Park, Yongli Lu, Fabian Rotermund, Young Ki Kim, Chan Su Moon, Nam Joong Jeon, Juan Pablo Correa-Baena, Vladimir Bulović, Seong Sik Shin, Moungi G. Bawendi, and Jangwon Seo. Efficient perovskite solar cells via improved carrier management. *Nature*, 590(7847):587–593, feb 2021. doi:10.1038/s41586-021-03285-w.

[162] Martin Kuban, Santiago Rigamonti, Markus Scheidgen, and Claudia Draxl. Density-of-states similarity descriptor for unsupervised learning from materials data. jan 2022. URL: http://arxiv.org/abs/2201.02187, arXiv:2201.02187.

[163] Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jürgen Bajorath. Molecular Similarity in Medicinal Chemistry. *Journal of Medicinal Chemistry*, 57(8):3186–3204, apr 2014. doi:10.1021/jm401411z.

[164] Vincent D. Blondel, Jean Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008. URL: https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008, arXiv:0803.0476, doi:10.1088/1742-5468/2008/10/P10008.

[165] Marina R. Filip, Giles E. Eperon, Henry J. Snaith, and Feliciano Giustino. Steric engineering of metal-halide perovskites with tunable optical band gaps. *Nature Communications*, 5, 2014. doi:10.1038/ncomms6757.

[166] Swathi Ippili, Venkatraju Jella, Ji-Ho Eom, Jaegyu Kim, Seungbum Hong, Jin-Seok Choi, Van-Dang Tran, Nguyen Van Hieu, Yun-Jeong Kim, Hye-Jin Kim, and Soon-Gil Yoon. An eco-friendly flexible piezoelectric energy harvester that delivers high output performance is based on lead-free masni3 films and masni3-pvdf composite films. *Nano Energy*, 57:911–923, 2019. URL: https://www.sciencedirect.com/science/article/pii/S2211285519300023, doi:https://doi.org/10.1016/j.nanoen.2019.01.005.

[167] Swathi Ippili, Venkatraju Jella, Jaegyu Kim, Seungbum Hong, and Soon-Gil Yoon. Unveiling Predominant Air-Stable Organotin Bromide Perovskite toward Mechanical Energy Harvesting. *ACS Applied Materials & Interfaces*, 12(14):16469–16480, apr 2020. doi:10.1021/acsami.0c01331.

[168] Weijun Ke, Constantinos C. Stoumpos, Ioannis Spanopoulos, Lingling Mao, Michelle Chen, Michael R. Wasielewski, and Mercouri G. Kanatzidis. Efficient Lead-Free Solar Cells Based on Hollow {en}MASnI3 Perovskites. *Journal of the American Chemical Society*, 139(41):14800–14806, oct 2017. doi:10.1021/jacs.7b09018.

[169] Hongzhe Xu, Haiwen Yuan, Jialong Duan, Yuanyuan Zhao, Zhengbo Jiao, and Qunwei Tang. Lead-free ch3nh3snbr3-xix perovskite quantum dots for mesoscopic solar cell applications. *Electrochimica Acta*, 282:807–812, 2018. URL: https://www.sciencedirect.com/science/article/pii/S0013468618311897, doi:https://doi.org/10.1016/j.electacta.2018.05.143.

[170] Andrea D'Annibale, Riccardo Panetta, Ombretta Tarquini, Marcello Colapietro, Simone Quaranta, Alberto Cassetta, Luisa Barba, Giuseppe Chita, and Alessandro Latini. Synthesis{,} physico-chemical characterization and structure of the elusive hydroxylam-

monium lead iodide perovskite NH3OHPbI3. *Dalton Trans.*, 48(16):5397–5407, 2019. URL: http://dx.doi.org/10.1039/C9DT00690G, doi:10.1039/C9DT00690G.

[171] Olga Nazarenko, Martin R Kotyrba, Sergii Yakunin, Michael Wörle, Bogdan M Benin, Gabriele Rainò, Frank Krumeich, Mikaël Kepenekian, Jacky Even, Claudine Katan, and Maksym V Kovalenko. Guanidinium and Mixed Cesium–Guanidinium Tin(II) Bromides: Effects of Quantum Confinement and Out-of-Plane Octahedral Tilting. *Chemistry of Materials*, 31(6):2121–2129, mar 2019. doi:10.1021/acs.chemmater.9b00038.

[172] Annette Trunschke, Giulia Bellini, Maxime Boniface, Spencer J. Carey, Jinhu Dong, Ezgi Erdem, Lucas Foppa, Wiebke Frandsen, Michael Geske, Luca M. Ghiringhelli, Frank Girgsdies, Rania Hanna, Maike Hashagen, Michael Hävecker, Gregory Huff, Axel Knop-Gericke, Gregor Koch, Peter Kraus, Jutta Kröhnert, Pierre Kube, Stephen Lohr, Thomas Lunkenbein, Liudmyla Masliuk, Raoul Naumann d'Alnoncourt, Toyin Omojola, Christoph Pratsch, Sven Richter, Christian Rohner, Frank Rosowski, Frederik Rüther, Matthias Scheffler, Robert Schlögl, Andrey Tarasov, Detre Teschner, Olaf Timpe, Philipp Trunschke, Yuanqing Wang, and Sabine Wrabetz. Towards Experimental Handbooks in Catalysis. *Topics in Catalysis*, 63(19-20):1683–1699, dec 2020. doi:10.1007/s11244-020-01380-2.

[173] Stephen R. Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, 7(1):1–34, may 2015. URL: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-015-0068-4, doi:10.1186/s13321-015-0068-4.

[174] Guenter Grethe, Gerd Blanke, Hans Kraut, and Jonathan M. Goodman. International chemical identifier for reactions (RInChI). *Journal of Cheminformatics*, 10(1):22, dec 2018. doi:10.1186/s13321-018-0277-8.

[175] Kirsten T. Winther, Max J. Hoffmann, Jacob R. Boes, Osman Mamun, Michal Bajdich, and Thomas Bligaard. Catalysis-Hub.org, an open electronic structure database for surface reactions. *Scientific Data*, 6(1):1–10, dec 2019. URL: www.nature.com/scientificdata, doi:10.1038/s41597-019-0081-y.

[176] Freek Kapteijn and Jacob A. Moulijn. Laboratory Catalytic Reactors: Aspects of Catalyst TestingA list of symbols used in the text is provided at the end of the chapter. In *Handbook of Heterogeneous Catalysis*, pages 2019–2045. Wiley-VCH Verlag GmbH & Co. KGaA, mar 2008. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/9783527610044.hetcat0108, doi:10.1002/9783527610044.hetcat0108.

[177] Manhua Mandy Lin. Selective oxidation of propane to acrylic acid with molecular oxygen, feb 2001. doi:10.1016/S0926-860X(00)00609-8.

[178] Ryan R. Langeslay, David M. Kaphan, Christopher L. Marshall, Peter C. Stair, Alfred P. Sattelberger, and Massimiliano Delferro. Catalytic Applications of Vanadium: A Mechanistic Perspective, feb 2019. URL: https://pubs.acs.org/sharingguidelines, doi:10.1021/acs.chemrev.8b00245.

[179] Lucas Foppa, Luca M. Ghiringhelli, Frank Girgsdies, Maike Hashagen, Pierre Kube, Michael Hävecker, Spencer J. Carey, Andrey Tarasov, Peter Kraus, Frank Rosowski, Robert Schlögl, Annette Trunschke, and Matthias Scheffler. Materials genes of heterogeneous catalysis from clean experiments and artificial intelligence. feb 2021. URL: http://arxiv.org/abs/2102.08269, arXiv:2102.08269.

[180] 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, aug 2018. URL: https://journals.aps.org/prmaterials/abstract/10.1103/PhysRevMaterials.2.083802, arXiv:1710.03319, doi:10.1103/PhysRevMaterials.2.083802.

[181] Runhai Ouyang, Emre Ahmetcik, Christian Carbogno, Matthias Scheffler, and Luca M. Ghiringhelli. Simultaneous learning of several materials properties from incomplete databases with multi-task SISSO. *JPhys Materials*, 2(2):24002, apr 2019. arXiv:1901.00948, doi:10.1088/2515-7639/ab077b.

[182] Juliao Braga, Joaquim L R Dias, and Francisco Regateiro. A MACHINE LEARNING ONTOLOGY. URL: https://frenxiv.org/rc954/, doi:10.31226/OSF.IO/RC954.

[183] Prakriti Kayastha and Raghunathan Ramakrishnan. Machine Learning Modeling of Materials with a Group-Subgroup Structure, may 2021. arXiv:2012.15619, doi:10.1088/2632-2153/abffe9.

[184] Fernando Mas, Jesus Racero, Manuel Oliva, and Domingo Morales-Palma. Preliminary ontology definition for aerospace assembly lines in airbus using models for manufacturing methodology. In *Procedia Manufacturing*, volume 28, pages 207–213. Elsevier B.V., jan 2019. doi:10.1016/j.promfg.2018.12.034.

[185] Rebeca Arista, Fernando Mas, and Carpoforo Vallellano. Initial Approach to an Industrial Resources Ontology in Aerospace Assembly Lines. In *IFIP Advances in Information and Communication Technology*, volume 594, pages 285–294. Springer Science and Business Media Deutschland GmbH, jul 2020. URL: https://doi.org/10.1007/978-3-030-62807-9{_}23, doi:10.1007/978-3-030-62807-9_23.

[186] Mina Abd Nikooie Pour, Huanyu Li, Rickard Armiento, and Patrick Lambrix. A first step towards extending the materials design ontology. In *Domain Ontologies for Research Data Management in Industry Commons of Materials and Manufacturing*, 2021. URL: https://openreview.net/forum?id=v02wuy8Q89A.

Appendix A

Semantic technologies: Technical Background

Ontologies and knowledge bases use the same languages and data models, which will be explained here.

RDF

RDF – the Resource Description Framework – is a standard data model for expressing objects as resources and their relationships. There are multiple syntaxes for it, e.g. RDF/XML or Turtle. An RDF file consists of triples of the form `Subject Predicate Object`. Each of the three is a "resource" identified by a unique resource identifier (URI) or its extension, the internationalized resource identifier (IRI). For example the well known URLs are subsets of URIs.

RDF Schema

RDF Schema extends RDF by classes and properties and therefore serves as a data schema to describe RDF resources.

OWL

The Web Ontology Language (OWL) extends RDF Schema to express relations between classes, cardinality, equality, characteristics of properties and much more. Every OWL document is a RDF document. OWL exists in three variants of different complexity and expressivity. OWL Lite supports classification hierarchy and simple constraints but is too simple for most use cases. OWL DL provides maximum expressivity while staying computationally complete and decidable. One limitation is that a class cannot be at the same time an instance of another class. It is based on description logic, hence the name DL. This is the standard variant of OWL that is mainly used in ontologies nowadays. OWL Full has no computational guarantees but allows maximum freedom with RDF. Due to its complexity, full reasoning will most likely not be possible.

Turtle

Turtle is nowadays the most popular syntax to express RDF documents (and therefore also ontologies). It has a very compact textual form that makes it easy to read for humans.

Triple store

A triple store or RDF store is a database for the storage of triples. If a set of triples is annotated with a name it is called named graph or quad store.

SPARQL

SPARQL is a recursive acronym for the SPARQL Protocol And RDF Query Language. As the name suggests it is mainly used for querying RDF databases, i.e. triple stores. Syntax wise is it comparable with SQL.

RML/R2RML

The RDB to RDF Mapping Language (R2RML) maps data in relational databases (RDB) to the RDF data model. RML extends this capability to define mappings of data in other formats too.

Appendix B

Related Ontologies for Materials

Materials Design Ontology

Materials Design Ontology(MDO) [1], which defines concepts and relations to cover knowledge in the field of materials design. MDO is designed using domain knowledge in materials science (especially in solid-state physics), and is guided by the data from several databases in the materials design field. (OPTIMADE) It has recently been extended using text mining on thousands of journal articles. [186]

MatOnto

MatOnto [2] – an extensible ontology, based on the DOLCE upper ontology, that aims to represent structured knowledge about materials, their structure and properties and the processing steps involved in their composition and engineering. The primary aim of MatOnto is to provide a common, extensible model for the exchange, re-use and integration of materials science data and experimentation.

MatSeek

MatSeek [3] provides a federated search interface over the critical materials science databases. Based on an OWL ontology (MatOnto), it provides a single Web-based search interface to the Inorganic Crystal Structure Database (ICSD),3 the Ionic Radii database,4 and the US National Institute of Standards and Technology (NIST) Phase Equilibria Diagrams (PED) Database.

Materials Ontology

The Materials Ontology [4], which consists of several sub ontologies corresponding to substance, process, environment, and property, is developed using the ontology language of the

[1] https://arxiv.org/pdf/2006.07712.pdf
[2] https://aaai.org/Papers/Symposia/Spring/2008/SS-08-05/SS08-05-003.pdf
[3] https://ieeexplore.ieee.org/document/4763655
[4] https://datascience.codata.org/articles/abstract/10.2481/dsj.008-041/

Semantic Web, OWL, which enables the definition of a flexible and detailed structure of materials information. A versatile "materials data format" is built on the Materials Ontology as a component of the materials information platform and is applied to exchange data among three different thermal property databases, maintained by two major materials science research institutes in Japan.

Data Science Ontology

The Data Science Ontology [5] is a knowledge base about data science that aims to catalog the concepts of data science, semantically annotate popular software packages for data science, and power new AI assistants for data scientists. Written not in OWL but in Monocl (`https://arxiv.org/pdf/1807.05691.pdf`).

MatOwl

MatOWL [6] is an OWL ontology extracted from the MatML schema, the extensible markup language developed especially to facilitate the exchange of materials information.

PREMΛP

The Platform for Realization of Engineered Materials and Products (PREMΛP) [7] enables harnessing available knowledge, learning emerging knowledge and continually creating new knowledge. It consists of an ontology-based, knowledge-assisted method and platform to capture, structure, configure and reuse knowledge for designing materials and engineering systems. The PREMΛP ontology provides extensible representation of data and knowledge.

[5] `https://www.datascienceontology.org/`

[6] `https://www.researchgate.net/publication/220390027_Semantic_Query_on_Materials_Data_Based_on_Mapping_MatML_to_an_OWL_Ontology`

[7] `https://link.springer.com/chapter/10.1007%2F978-81-322-1050-4_105`

Appendix C

Software for Working with Ontologies and Knowledge Graphs

As a physicist/materials scientist a lot of new tools and software are necessary to work with ontologies. An overview of some useful software will be given here.

Protégé

Protégé is a graphical program to explore and develop ontologies and suitable for beginners. Lots of plugins allow to run a reasoner or to visualize the ontology or parts of it as well as a simple SPARQL queries can be used for validating. It has been used to develop ontologies.

owlready2

The python library owlready2 enables using and developing ontologies within Python. For more advanced ontology developers it might however be too restrictive. It has been used to convert the Metainfo to an ontology and to populate the developed ontologies with real data.

rdflib

Another python library is rdflib which allows handling RDF graphs including manipulating them. Because OWL is an RDF format it can be read in using rdflib but no ontology validation is happing. It has been used for converting between different formats (turtle, n3, rdf/xml).

Yarrrml

Yarrrml provides means to express RML-based mapping rules in the human-readable YAML format. Together with an RML processor like RMLmapper, it can be used to create linked data.

VOWL and WebVOWL

VOWL is a visualization tool for ontologies. It provides an online version accessible easily for everyone, WebVOWL.

Graphviz

Graphviz is simple open source graph visualization software supporting basic graph layouts.

OwlViz

OwlViz is a plugin for Protege with is able to easily display *is a* hierarchies on an ontology.

Gephi

Gephi [131] is a graphical network visualization and analysis tool and has a semantic web plugin to directly extract a graph/network via SPARQL queries from an ontology or knowledge graph. It supports various algorithms for layouting and has a Graphviz plugin.

Apache Jena

Apache Jena provides open source software for graph databases, SPARQL endpoints and more.

Stardog

Stardog is a commercial provider for RDF databases, knowledge graphs and SPARQL endpoints. Its Stardog Studio is an integrated development environment (IDE) with lots of capability and has been used to write and run SPARQL queries.

Selbstständigkeitserklärung

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 am 11.07.2018, angegebenen Hilfsmittel angefertigt habe.

Berlin, 3.10.2021

Maja-Olivia Lenz-Himmer