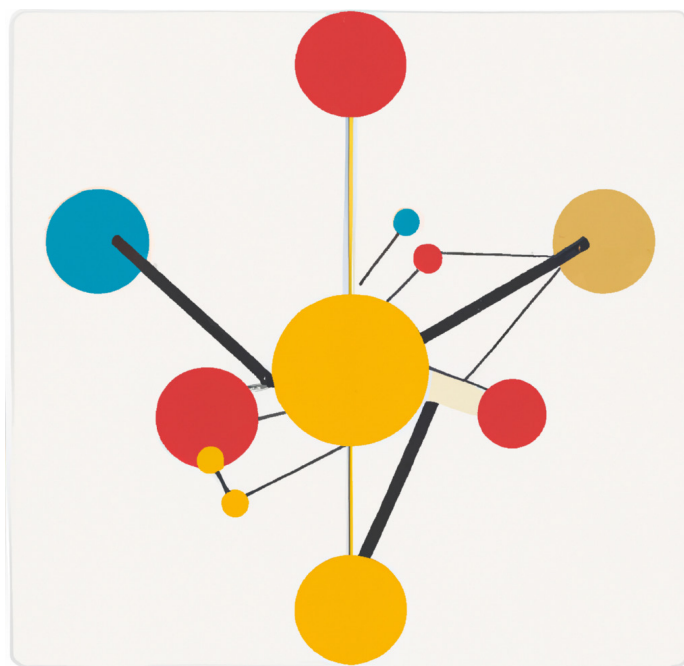


Doctoral Thesis in Physics

# Homology and machine learning for materials informatics

BART OLSTHOORN



# Homology and machine learning for materials informatics

BART OLSTHOORN

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Friday the 24th March 2023, at 3:00 p.m. in Hörsal 4, Hus 2, Albanovägen 18, Albano, Stockholm

Doctoral Thesis in Physics  
KTH Royal Institute of Technology  
Stockholm, Sweden 2023

© Bart Olsthoorn

ISBN 978-91-8040-505-8

TRITA-SCI-FOU 2022:08

Printed by: Universitetsservice US-AB, Sweden 2023

## Abstract

Materials informatics is the field of study where materials science is combined with modern data science. This data-driven approach is powered by the growing availability of computational power and storage capability. The development and application of these methods accelerates materials science and represents an effective way to study and model material properties. This thesis is a compilation of theoretical and computational works that can be divided into three key areas: materials databases, machine learning for materials, and homology for materials.

Machine learning and data mining rely on the availability of materials databases to test methods and models. The Organic Materials Database (OMDB), for example, contains a large number of organic crystals and their corresponding electronic structures. The electronic properties of the organic crystals are computed using atomic scale materials modelling, which is computationally expensive because organic crystals typically contain many atoms in the unit cell. However, the resulting data can be used in a variety of materials informatics applications. We demonstrate data mining for dark matter sensors as an example application.

Accurate machine learning models can capture the structure-property relationship of materials and accelerate the discovery of new materials with desired properties. This is explored by investigating the properties of the organic crystals in the OMDB. For example, we employ supervised learning on the electronic band gap, an important material property for technological applications. Unsupervised learning is used to construct a dimensionality-reduced chemical space that reveals interesting clusters of materials.

Finally, persistent homology is a relatively new method from the field of algebraic topology that studies the shapes that are present in data at different length scales. In this thesis, the method is used to study magnetic materials and their phase transitions. More specifically, in the case of classical models, we use persistent homology to detect the phase transition directly from sampled spin configurations. For quantum spin models, the shapes in the entanglement structure are captured and a sudden change reveals a quantum phase transition.

In summary, these three topics provide an overview on how to study material properties with modern data science methods. The tools can be used in combination with the traditional methods in materials science and accelerate materials design.

## Sammanfattning

Materialinformatik är ett forskningsområde där materialvetenskap kombineras med modern datavetenskap. Detta datadrivna tillvägagångssätt drivs av den växande tillgängligheten av beräkningskraft och lagringskapacitet. Utvecklingen och tillämpningen av dessa metoder accelererar materialvetenskapen och utgör ett effektivt sätt att studera och modellera materialegenskaper. Denna avhandling är en sammanställning av teoretiska och beräkningstekniska arbeten som kan delas in i tre nyckelområden: materialdatabaser, maskininlärning för material och homologi för material.

Maskininlärning och datautvinning är beroende av tillgången på materialdatabaser för att testa metoder och modeller. Organic Materials Database (OMDB) innehåller data för kristallin struktur och elektroniska egenskaper för ett stort antal organiska kristaller. De elektroniska egenskaperna hos de organiska kristallerna beräknas med hjälp av materialmodellering i atomskala, vilket är beräkningsmässigt dyrt då organiska kristaller vanligtvis innehåller många atomer i enhetscellen. Emellertid kan den resulterande datan användas i en mängd olika materialinformatikapplikationer. Vi demonstrerar datautvinning för att söka material till sensor för mörk materia som ett exempel på applikation.

Maskininlärningsmetoder kan fånga förhållanden mellan struktur och egenskap hos material, och därmed påskynda upptäckten av nya material med önskade egenskaper. Detta utforskas genom att undersöka egenskaperna hos de organiska kristallerna i OMDB. Till exempel använder vi övervakat lärande på elektroniska bandgap, en viktig materiell egenskap för tekniska tillämpningar. Övervakat lärande används för att konstruera en dimensionsreducerad kemisk rymd som avslöjar intressanta kluster av material.

Slutligen är ihållande homologi en relativt ny metod från området algebraisk topologi som studerar de former som finns i data i olika längdskalor. I denna avhandling används metoden för att studera magnetiska material och deras fasövergångar. Mer specifikt, när det gäller klassiska modeller, använder vi ihållande homologi för att detektera fasövergången direkt från samplade spin-konfigurationer. För kvantspinnmodeller fångas faserna i strukturen hos den kvantmekaniska sammanflätningen och en plötslig förändring avslöjar en kvantfasövergång.

Sammantaget utgör dessa tre ämnen ett bra exempel på hur materialegenskaper kan studeras med moderna datavetenskapliga metoder. Verktynen kan användas i kombination med traditionella metoder inom materialvetenskap och påskynda materialdesign.

# Contents

<b>Contents</b>	<b>iii</b>
<b>Preface</b>	<b>v</b>
Scientific papers and the author's contributions . . . . .	v
Acknowledgements . . . . .	viii
 <b>I Comprehensive summary</b>	 <b>1</b>
<b>1 Introduction</b>	<b>3</b>
1.1 Materials Informatics . . . . .	4
1.2 Topology . . . . .	5
1.3 Dirac materials . . . . .	7
1.4 Dark matter . . . . .	8
 <b>2 Materials databases</b>	 <b>11</b>
2.1 Organic Materials Database . . . . .	11
2.2 Atomic scale materials modelling . . . . .	14
2.3 Sensors for dark matter . . . . .	17
 <b>3 Machine learning for materials</b>	 <b>19</b>
3.1 Supervised learning . . . . .	20
3.2 Kernel Ridge Regression . . . . .	21
3.3 Neural networks . . . . .	24
3.4 Dimensionality-reduced chemical space . . . . .	29
 <b>4 Homology for materials</b>	 <b>33</b>
4.1 Simplicial homology . . . . .	33
4.2 Persistent homology . . . . .	39
4.3 Classical spin structures . . . . .	41
4.4 Quantum entanglement structures . . . . .	42
 <b>5 Summary</b>	 <b>47</b>

<b>Bibliography</b>	<b>49</b>
<b>II Included papers</b>	<b>61</b>

# Preface

This thesis is the result of my work as a graduate student under the supervision of Alexander V. Balatsky, which was carried out at the Department of Physics at KTH during the years 2018-2022. Additional supervision by Jens H. Bardarson at KTH.

Communications arising from the present thesis are gratefully accepted. My e-mail address is bartol@kth.se.

## Scientific papers and the author's contributions

The scientific papers that are included in this thesis are listed below, together with descriptions of my contributions. In all the collaborative works, the writing was very much a joint effort.

### Papers which are part of the thesis:

#### 1. Persistent homology of quantum entanglement

B. Olsthoorn, Preprint (Under review)

The idea for this paper arose naturally as a follow-up paper on the hidden order in classical spins (Paper 3). Extending to quantum phase transitions is a natural next step and I implemented the exact diagonalization and numerical analysis that produced the results.

#### 2. Shifting computational boundaries for complex organic materials

R. M. Geilhufe, B. Olsthoorn, A. V. Balatsky, Nature Physics, 1-3 (2021).

This comment article was a joint writing effort that describes our cumulative experience gained while running the Organic Materials Database (OMDB) and the vision for the future of structure-property prediction of large complex compounds. I also performed a scaling analysis to provide numerical evidence for this perspective.



### 3. Finding hidden order in spin models with persistent homology

B. Olsthoorn, J. Hellsvik, A. V. Balatsky, Phys. Rev. Research 2, 043308 (2020).

I conceived the idea of applying persistent homology to classical spin models while at a ERC HERO 2020 meeting in Switzerland with Alexander Balatsky and others. I constructed the Monte Carlo simulation and carried out the numerical analysis and produced the results. Johan Hellsvik joined the project later to assist in the understanding of the complex magnetic structures.

### 4. Identification of strongly interacting organic semimetals

R. M. Geilhufe, B. Olsthoorn, Phys. Rev. B 102, 205134 (2020).

The idea for this project came out of discussions with Matthias Geilhufe. I performed a large-scale search for potential organic semimetals using machine learning techniques developed in Paper 6. I mainly contributed on all sections that involve machine learning predictions.

### 5. Mass fluctuations and absorption rates in dark-matter sensors based on Dirac materials

B. Olsthoorn, A. V. Balatsky, Phys. Rev. B 101, 045120 (2020)

Alexander Balatsky provided the idea for this project and it fits in the general effort by our group to explore Dirac Materials as dark-matter sensors (see also Paper 7). I performed all analytical calculations and produced all figures and co-wrote the paper.

### 6. Band gap prediction for large organic crystal structures with machine learning

B. Olsthoorn, R. M. Geilhufe, S. S. Borysov, A. V. Balatsky, Advanced Quantum Technologies, 1900023 (2018)

For this paper, I trained two different machine learning models and tested their predictive power. I performed all the computational work in this paper and produced the figures.

### 7. Materials Informatics for Dark Matter Detection

R. M. Geilhufe, B. Olsthoorn, A. Ferella, T. Koski, F. Kahlhoefer, J. Conrad, A. V. Balatsky, Physica Status Solidi RRL (2018).

My main contribution to this paper are the sections on small-gap organic materials and organic dirac materials. I also performed a search in our database to identify tiny gap materials of which one was further analysed with ab initio calculations. This paper provided the foundation for the further work in Paper 5.

**8. Online search tool for graphical patterns in electronic band structures**

S. S. Borysov, B. Olsthoorn, M. B. Gedik, R. M. Geilhufe, A. V. Balatsky, npj Computational Materials 4 (1), 1-8 (2018)

The idea of searching through electronic band structure patterns arose out of the collaborative work on the Organic Materials Database. I implemented the algorithm into the Organic Materials Database website. The interface was extended with a drawing tool by M. B. Gedik.

**Papers which are not part of the thesis:****A. Periodogram-based detection of unknown frequencies in time-resolved scanning transmission X-ray microscopy**

S. Finizio, J. Bailey, B. Olsthoorn, J. Raabe, ACS Nano (2022).

This paper is focused on a specific experimental physics setup and I provided the data analysis method and theory. I performed the initial data analysis, but the final computation and the further development of the method was performed by S. Finizio. I mainly wrote the section on the theory of periodograms and what parameters to use for our data.

**B. LIDA - The Leiden Ice Database for Astrochemistry**

W. R. M. Rocha, M. G. Rachid, B. Olsthoorn, E. F. van Dishoeck, M. K. McClure, and H. Linnartz, Astronomy & Astrophysics (2022).

I developed the first version of LIDA launched in February 2015. In this work, the database is upgraded and both old and new functionalities (particularly relevant for the James Webb Space Telescope) are presented in an academic paper for the first time.

**C. Indoor radon exposure and its correlation with the radiometric map of uranium in Sweden**

B. Olsthoorn, T. Rönqvist, C. Lau, S. Rajasekaran, T. Persson, M. Månsson, A. V. Balatsky, Science of The Total Environment, 151406 (2021).

This work is the result of a collaboration between Radonova Laboratories AB, KTH and Nordita. Radonova Laboratories AB provided the indoor radon measurements that made it possible to do this large-scale survey of Sweden for the first time. I performed the statistical analysis and produced all the figures and tables. The writing and interpretation of the results was a joint effort.

## Acknowledgements

First and foremost, I wish to thank my main supervisor Alexander V. Balatsky for invaluable discussions and guidance throughout my doctoral research. Without his encouragement to explore new ideas this thesis would look very different. I would also like to thank my co-supervisor Jens H. Bardarson for being ready to help out when needed. Under their guidance I had the possibility to learn a lot and the freedom to study a wide range of topics over the years.

This thesis would not have been possible without the support of R. Matthias Geilhufe. His enthusiasm and perspective have shaped my approach to physics research from the beginning. This started already during his supervision of my master's project. Finally, I would like to thank all my colleagues at Nordita (Nordic Institute for Theoretical Physics) for the great academic and social events.

## **Part I**

# **Comprehensive summary**



# Chapter 1

## Introduction

Materials and their properties are of fundamental importance to human civilization. Their technological significance even leads to the naming of historic periods after materials, e.g. the silicon age. The importance of new materials is also highlighted in the Sustainable Development Goals (SDGs), adopted by the United Nations in 2015<sup>1</sup>. For example, SDG 12 aims to reduce our reliance on toxic chemicals, while SDG 9 aims to promote environmentally sound technologies in the industrial sectors. Given their importance, it is interesting to note that materials with useful properties are usually discovered serendipitously, even though the fundamental laws describing the interactions between atoms are known. This is because solving a quantum mechanical system of more than a few atoms is a very difficult task.

It turns out that a microscopic description (e.g. interaction between atomic nuclei and electrons) cannot always provide understanding of the phenomena of a macroscopic system. For example, the behaviour of living organisms is captured by a minimal theoretical framework that does not depend on the quantum mechanical behavior of the atom. This is an interesting fact about nature that P. W. Anderson presented in the “More is Different” 1972 paper [3]. The emergent behaviour of the system can be complex and warrant its own fundamental description. For example, in hydrodynamics, the Navier-Stokes equations describe how mass density, momentum and energy of a fluid behave, without keeping track of all its microscopic constituents. Inspired by this concept, one could imagine having an effective theory of material properties that would guide design choices.

This thesis is a compilation of works that aim to model and understand material properties with computational methods from the field of data science. This provides new ways to understand and predict properties, and search for new materials. For example, we have used machine learning models to search the vast chemical compound space of organic chemistry and identify semimetals – a rare electronic property for organic materials. The new methods are sometimes fast but approx-

---

<sup>1</sup>For a full list of the 17 Sustainable Development Goals, see <https://sdgs.un.org>.

imate (i.e. cost-accuracy tradeoff) and are complemented by traditional *ab initio* quantum chemistry computation.

The main topic of the works can be divided into three key areas, and are represented by three corresponding chapters in this thesis. Firstly, the development of the Organic Materials Database (OMDB), a large database containing electronic band structures and magnetic properties. This includes new search tools, such as pattern search, and an application with sensors for dark matter is discussed. Secondly, the application of machine learning methods to quantum chemistry. The models that capture the structure-property relationship of materials have been particularly successful. Thirdly, the relatively new field of topological data analysis (TDA) is used to study material properties. In particular, both classical and quantum phase transitions are discussed. In the following introduction sections, the overarching concepts are briefly presented, and these are relevant in multiple chapters.

## 1.1 Materials Informatics

The growing availability of computational power and data science methods has resulted in a new field called materials informatics. As the name implies, its practitioners apply methods from computer science and statistics to model and study materials. This is done to gain new insight into their properties and to accelerate the design of new materials. The importance of this is recognized globally and signified by, for example, the Materials Genome Initiative in the United States [53], which aims to accelerate materials development. Materials data is also of key importance when it comes to commercial companies in various industries. To maintain their competitive edge, this materials research (often experimental data) is often kept internal, and it has also lead to commercial materials data platforms such as Citrine [18] and Mat3ra (formerly Exabyte.io) [52].

However, data-driven materials science relies on the availability of theoretical and experimental data to test methods and models. There are many publicly available databases, with some well-known examples being the Materials Project [41], AFLOW [22], NOMAD Repository [60]. These databases also have an API (application programming interface), providing an interface to retrieve data with scripts. A more detailed discussion on materials databases is presented in Chapter 2. The growth of popularity materials informatics as a field also leads to more available data.

Materials science is a rich field with many different types of measurements and data. For example, a measurement might depend on strain, doping, magnetic field, temperature, synthesis conditions and so on. The result is a loosely structured set of heterogeneous data. In other words, it is challenging to find data that is formatted in a systematic way. This challenge is present in all scientific fields, and in 2016 the FAIR (Findable, Accessible, Interoperable and Reusable) guiding principles were introduced, setting out to improve the data infrastructure. Regarding interoperability and reusability, the development of a materials ontology is crucial.

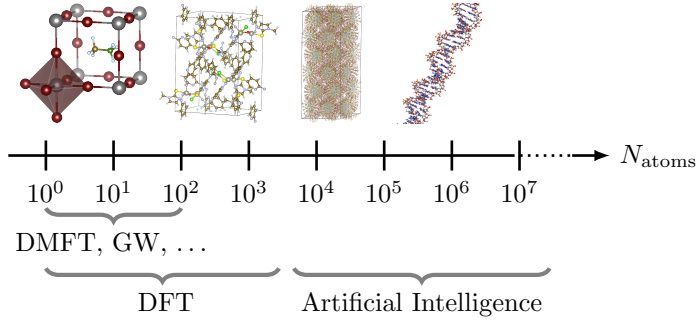


Figure 1.1: Conventional *ab initio* calculations scale up to  $10^3$  number of atoms, whereas *ab initio* calculations with strong correlations scale to smaller unit cells. More complex unit cells and large compounds. This is Fig. 1(a) in Paper 2.

This sets the terminology and framework for the materials informatics community. This is still an active effort within the community, with promising steps being made to establish a standardized ontology [6, 31].

One of the main goals within materials informatics is machine-learning guided design of materials. This is an interdisciplinary endeavor that combines data science, physics and chemistry knowledge. For example, the machine learning models that are introduced in this thesis are specifically designed to incorporate the known symmetries of the physical problem. The trained model has the potential to dramatically speed up the exploration of chemical space. A conventional *ab initio* quantum chemistry calculation scales with  $O(N^2 \log N)$  to  $O(N^3)$  in the number of atoms  $N$ , and can take many hours to complete on a supercomputer. In contrast, a neural network, once trained, can make a prediction in under a second, providing a speedup of over a million times. Moreover, depending on the machine learning model, it is possible to have linear scaling in the number of atoms, making more complex crystals and large biological systems within computational reach (see Fig. 1.1).

All works included in this thesis are connected to the field of materials informatics (except for Paper 5, which is analytical work on impurities).

## 1.2 Topology

Topology is the study of shapes and their properties under continuous deformations. Three common types of topological equivalence are *homeomorphism*, *homotopy equivalence* and *isomorphic homology groups*. Homeomorphism means that there exists a continuous bijective map  $f$  and continuous inverse  $f^{-1}$  between the two spaces. For example, a coffee mug and donut (filled torus) are famously homeomorphic. Using this concept, objects are classified by properties that stay the same



under homeomorphisms: topological invariants. Homeomorphism is a stronger case of the more general homotopy equivalence. This construction is based around the notion of contracting loops to points and is often easier to compute. Homeomorphic spaces are always homotopy equivalent but the opposite is not true. For example, the continuous deformation of a line into a point is not a homeomorphism, since it is not bijective (the line contains an infinite number of points). However, the line and the point are homotopy equivalent. Another example is that the cylinder and Möbius strip are homotopy equivalent but not homeomorphic. A related theory is homology, which assigns homology groups corresponding to a space. This aims to capture the number of  $k$ -dimensional holes in a space, and the topological invariant called the Betti number  $\beta_k$  counts the number of  $k$ -dimensional holes. Having isomorphic homology groups is a weaker condition than homotopy equivalence, but it is easier to compute. In summary, homeomorphism is a stronger condition than homotopy equivalence, which is stronger than isomorphic homology groups (see Fig. 1.2). In some special cases there are connections between the different levels of theory, such as the Hurewicz theorem providing a map from homotopy to homology [35]. In this compilation thesis, the focus is mainly on simplicial homology groups, since they represent a computationally simple choice that is based on linear algebra. Chapter 4 provides the definitions of this elegant theory from the field of algebraic topology, and a number of example computations.

Topological invariants also appear in many places in condensed matter physics. For example, the quantized conductance in the quantum Hall effect is a topological invariant that is independent of the sample geometry. Another example is the one-dimensional Su-Schrieffer-Heeger (SSH) model that has a integer topological invariant called the *winding number* that changes depending on the model parameters. This invariant describes the homotopy equivalence within the topological phase. However, in this compilation thesis the focus is on constructing completely

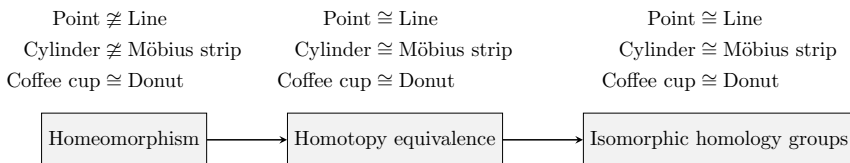


Figure 1.2: Three types of topological equivalence arranged in order of their differentiating power, from stronger (left) to weaker (right). The  $\cong$  symbol indicates topological equivalence at the level of theory shown in the box. For example, homeomorphism distinguishes a cylinder from a Möbius strip, whereas this is not detected at the other two levels of topological equivalence. Examples of non-homotopic spaces with isomorphic homology groups are more complicated, one such example being the Poincaré homology spheres [89]. The focus of the works presented in this thesis is at the level of homology groups.

new topological invariants using persistent homology. These invariants are then primarily used to detect phases in materials, but they may also prove useful in understanding materials within a new computational perspective.

### 1.3 Dirac materials

The Schrödinger equation governs the motions of electrons moving in a periodic lattice of nuclei. Solving this equation provides quantized energy levels as a function of momentum that can be plotted as a band structure (e.g. Fig. 2.2). This reveals the electronic properties of a materials, for example, whether a material is a metal, semiconductor or insulator. The band structure is often complicated, as is clear from its colloquial name: spaghetti diagrams. However, most of the behaviour of the system is governed by low-energy excitations. For this reason, this diagram is often replaced by an effective Hamiltonian with a small number of conductance and valence bands.

Dirac materials are a class of materials that are effectively modelled by the Dirac Hamiltonian, which in two dimensions takes the form

$$H = v_F (\sigma_x p_x + \sigma_y p_y) + M \sigma_z, \quad (1.1)$$

where  $\sigma_i$  are the Pauli matrices and  $v_F$  the Fermi velocity that controls the slope of the dispersion [90]. Materials in this class include graphene, topological insulators and honeycomb ferromagnets. Figure 1.3 shows the characteristic linear, rather than quadratic, band structure of Dirac materials. The introduction of the  $\sigma_z$  term causes a band gap in the dispersion that can be controlled by the size of  $M$ . Therefore, these materials are proposed as sensors for light particles, in particular for the hypothetical dark matter particle (see Section 1.4 and 2.3) [30, 37].

Dirac materials are relevant to a number of the works presented in this thesis. Two of the papers focus on finding new Dirac materials in the domain of the

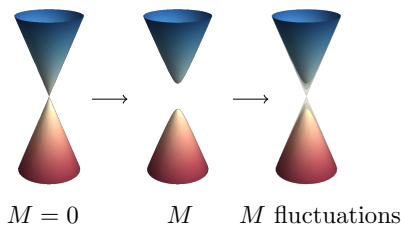


Figure 1.3: The characteristic linear dispersion of the Dirac fermions at zero mass  $M$ , and the opening of a gap introduced by the  $\sigma_z$  term. At the Dirac point (crossing point) the density of states  $\nu(E)$  vanishes. In general, the  $d$ -dimensional Dirac material has a density of states following  $\nu(E) \sim |E|^{d-1}$ . This is Fig. 1 in Paper 5.

organics. Paper 8 implements a graphical pattern search to efficiently find the characteristic band crossing in electronic structure databases. Paper 4 identifies the first three-dimensional organic Dirac material. Beyond the search of Dirac materials, we have also studied their utility as a dark matter sensor. Paper 7 outlines the potential of Dirac materials as dark matter sensors. The crucial advantage of Dirac materials is that the band gap can be of the order of meV. The small gap suppresses thermal excitations, while also fitting the requirements of candidate dark matter particles as presented in more detail in the next section. Paper 5 details the effects of impurities in gapped Dirac materials on their utility as a dark matter sensor. The impurities introduce so-called Lifshitz tails [49], which extend far into the gap of the sensor material.

## 1.4 Dark matter

The nature of most of the matter in the universe is unknown, representing one of the biggest puzzles in modern science. This elusive matter is only observed through gravitational effects, and it is therefore called *dark matter*. The total amount of dark matter is constrained by modern cosmological models, and it is estimated to be roughly 85% of all matter in the universe.

Historically, many astronomers in the early 20th century have searched for non-luminous astronomical objects. In the early days, dynamical mass measurements of orbiting stars lead to an estimate of the total non-luminous matter in our own galaxy, the Milky Way. In 1930, the Swedish astronomer Knut Lundmark discovered dark matter in five galaxies, including the Milky Way and Andromeda, by measuring its rotational velocities. Lundmark also pointed out that the galaxies required the presence of dark matter to be stable. Three years later, the Swiss astronomer Fritz Zwicky studied the Coma cluster, a structure of over a thousand galaxies, and found that the orbits of the galaxies at the edge provided evidence for dark matter. This is because the gravitational effect of the visible galaxies was too small to account for the fast orbits. In 1960s and 70s, a number of astronomers performed detailed observations of the rotational velocities of spiral galaxies. Considering all the baryonic matter present, it was expected that the velocities near the edge of the galaxy would decrease. This was not observed however, and the velocity rotation curves were in fact approximately constant far away from the galactic center, which ultimately lead to the modern understanding of galaxies embedded in a large spherical dark matter halo. More recently, dark matter has also been detected through gravitational lensing, where light is bent as described by general relativity. All the evidence combined implies a universe that is dominated by dark matter.

The Standard Model describes the particles and the known forces (except gravity) that exist in the universe. Gravity is not mediated by a particle and is instead explained with Einstein's general theory of relativity. None of the particles in the Standard Model fit the observational evidence for dark matter, which motivates the

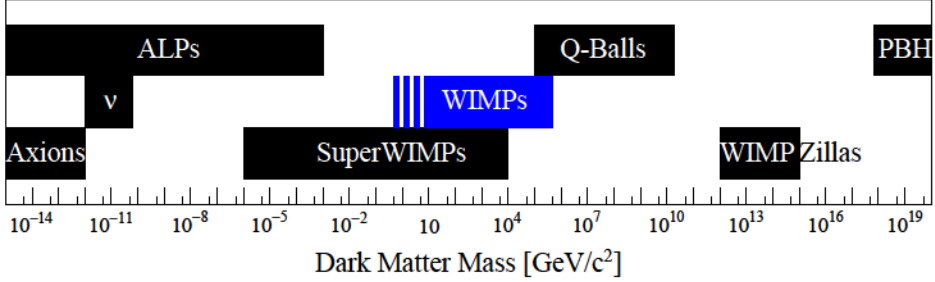


Figure 1.4: Dark matter candidates cover a wide range of mass magnitudes. Axions and axionlike particles (ALPs) constitute the lightest candidates. Sterile neutrinos  $\nu$  are another light candidate. The WIMPs are the focus of this thesis and the low-mass region ( $\leq 5 \text{ GeV}/c^2$ ) is shown by the dashed bars. Beyond particles, primordial black holes (PBH) are proposed to have been produced in the very early Universe and are referred to as a type of massive compact halo object (MACHO). *Figure reproduced from [77] with permission from IOP Publishing.*

search for new particles (see Fig. 1.4). The criteria for this dark matter particle are that it is weakly interacting, stable and cold (low kinetic energy). This last constraint is given by the cosmic microwave background (CMB) and the large-scale structure in the universe.

A popular dark matter candidate is the Weakly Interacting Massive Particle (WIMP). It is a hypothetical particle that interacts with other particles in the Standard Model only through the weak nuclear force (or via an undiscovered weaker force) and through gravity. There is an ongoing global effort to detect such particles non-gravitationally, through direct or indirect detection methods. Direct detection requires the dark matter particle to transfer energy to the detector that can then be measured as light, charge, or heat. For example, the XENONnT experiment uses almost 6000 kilograms of liquid xenon to search for WIMPs particles in the range of about 10 to  $10^4 \text{ GeV}$  [2, 5]. The detector measures light and charge signals caused by the excitation or ionisation of the xenon by dark matter particles. Indirect detection methods consider the possibility of dark matter annihilation or decay and search for observable matter particles from the Standard Model that are produced in this process. For example, a number of telescopes are searching for gamma rays and cosmic rays that are consistent with WIMP annihilation or decay [29].

Another well-established dark matter particle candidate is the axion. This is a boson with very low mass (meV or lighter). Finally, dark matter could also be a mix of multiple components. Material science plays a promising role in the detection of hypothetical dark matter particles. The case of directly detecting lighter (sub-MeV) particles is discussed in Section 2.3.



## Chapter 2

# Materials databases

The properties of crystalline materials are primarily governed by their electronic structure. Understanding and predicting the functional properties is of importance to create materials with useful technological applications. Even though the quantum mechanical nature and the fundamental laws of the electrons and nuclei are known, the computation is known to be challenging and demands large computational resources. Moreover, the possible chemical compound space to search for target properties is enormous. The growth in computational power has made high-throughput calculation of material properties more and more prominent within material science. This task is accelerated further by machine learning, which is the focus of Chapter 3.

Table 2.1 lists a number of well-known databases. In this chapter, we describe the computational tools that are relevant to the Organic Materials Database in more detail. We also discuss an example technological application of the OMDb, namely of dark matter sensors. The Papers 5, 7 and 8 are within this scope of this chapter.

### 2.1 Organic Materials Database

Motivated by the high demand for materials data and the potential for technological applications, the Organic Materials Database (OMDb) was launched in 2016 [13]. The database is open-access for academic users and available at <http://omdb.mathub.io>. It is an online platform that contains over 40,000 electronic structures, magnetic properties and search tools. The computation of the electronic properties are performed in the framework of density functional theory (DFT) with the Vienna Ab initio Simulation Package (VASP) [45]. The website itself is programmed using the PHP language and the Laravel web framework. Data processing and analysis is performed in Python.

The focus of this database is on organic and organometallic materials, which represent an understudied class compared to inorganic crystals. The main challenge

Database	Brief description	Ref.
AFLOW	General-purpose materials database	[22]
COD	Crystal structure information	[33]
Materials Project	General-purpose materials/molecules database	[41]
MatNavi	Collection of databases	[59]
Polymer Genome	Computational and experimental data on polymers	[43]
OMDB	Electronic/magnetic properties of organic crystals	[13]
OQMD	Computational materials database	[74]
QM	Datasets for machine learning	[65]
NoMaD	General-purpose materials/molecules database	[60]

Table 2.1: Selection of a few well-known materials databases with a brief description. These include computational data, experimental data, and molecules or materials data.

is that organic crystal structures typically contain hundreds of atoms in their unit cell. The time complexity of ab initio quantum chemistry codes typically scale with  $O(N^2 \log N)$  to  $O(N^3)$  in the number of atoms  $N$ . This means that organic materials require a lot of computational resources. However, the choice of organics also has its advantages. For example, the main constituents of carbon, hydrogen, nitrogen and oxygen are abundant and inexpensive. Inorganic crystals are typically hard and brittle (e.g. gemstones), whereas crystals constructed from organic molecules can be flexible, making flexible electronics an interesting application domain [50].

The Organic Materials Database includes a number of material properties and search capabilities. Crucially, each material is processed systematically in the same way to arrive at the electronic properties, see Fig 2.1. Therefore, even though the individual calculations are approximate (inherent to the quantum chemistry simulation), nevertheless the trends in the data are due to the different input materials and worth studying. There are two database functionalities that are particularly relevant to this thesis, and they are outlined below.

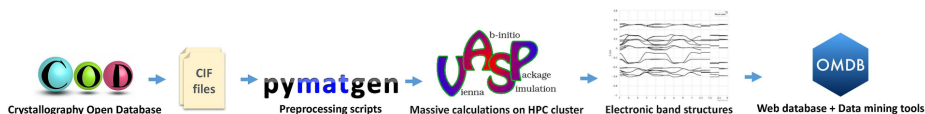


Figure 2.1: Process used for each material within the OMDB. The crystal structure is imported as a CIF file from the Crystallographic Open Database (COD). Pymatgen is used to convert these files into input files for the VASP DFT package. The electronic structure is stored in the OMDB as band structures. *Figure adapted from [13] (CC BY 4.0).*

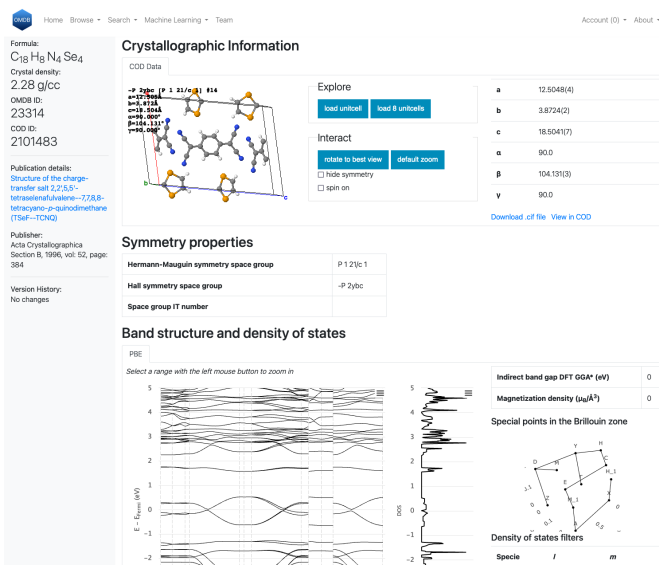


Figure 2.2: Example of a material page (TSeF-TCNQ) on the Organic Materials Database (OMDB) showing the crystallographic information, electronic band structure and density of states.

## Material properties

Figure 2.2 shows a screenshot of the material page for the compound TSeF-TCNQ. At the top of the page, we display the crystallographic information as collected from the Crystallographic Open Database (COD) and the symmetry information (e.g. space group). The computed Kohn-Sham band structure and density of states are displayed in interactive plots below. At the bottom of the page, similar materials based on crystal structure (see Section 3.1 on descriptors) and density of states are listed.

The electronic calculations are performed within the density functional theory (DFT) framework that is presented in more detail in Section 2.2. More specifically, the projector augmented wave method is used as implemented in the VASP code, and the exchange-correlation functional was approximated by the generalized gradient approximation (GGA) according to PBE [63].

## Graphical pattern search

The functionality of a material can often be characterized by its electronic band structure and density of states. Metals have finite density of states and Dirac materials are characterized by a linear crossing at the Fermi level. More complicated examples are topological insulators, that show a Mexican-hat structure.



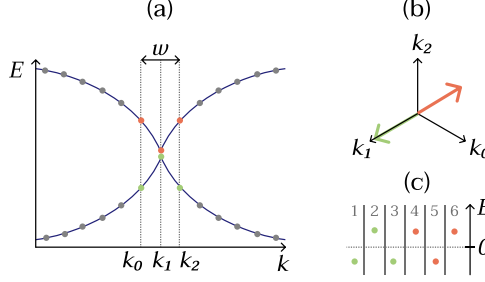


Figure 2.3: Sliding window of two electronic bands along a high symmetry path in the Brillouin zone of width  $w$  (a). The window can be represented by two vectors of each 3 points each (b) that is concatenated to a single vector (c). This is Fig. 1 in Paper 8.

The OMDB contains pattern search options for the Kohn-Sham band structures and density of states. This means that either a preset pattern or a user-specified drawing can be chosen to search for materials. First, a searchable index is created by sliding a window through the band structures and storing the resulting vectors, see Fig. 2.3. Next, the problem becomes a nearest neighbor search task for a given input query pattern. In order to do this online in a fast and efficient way, we use the approximate nearest neighbor algorithm called ANNOY, as implemented in the open-source library by Spotify [4]. Figure 2.3 shows how the algorithm works for a two-dimensional example, but it generalizes to high-dimensional data. In order to make the functionality of searching for electronic band pattern more broadly available, we have open-sourced the code along with the publication (see Paper 8).

## 2.2 Atomic scale materials modelling

The microscopic description of atomic nuclei and electrons is well known, as mentioned earlier in the introduction. For example, the Hamiltonian of a many-body system  $N$  electrons in the Born-Oppenheimer approximation can be written as

$$\begin{aligned}
 H(\mathbf{r}_1, \dots, \mathbf{r}_N) &= T + V_{\text{ext}} + W \\
 &= \underbrace{\sum_i^N \frac{\nabla_i^2}{2}}_{\text{electron kinetic energy}} + \underbrace{\sum_i^N V_{\text{ext}}(\mathbf{r}_i)}_{\text{electron-nuclei interaction}} + \underbrace{\sum_{i,j=1;i < j}^N \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}}_{\text{electron-electron interaction}}
 \end{aligned}$$

when using Hartree atomic units ( $e = \hbar = m = 1$ ). The solutions are given by the time-independent Schrödinger equation  $H\psi_i(\mathbf{r}_1, \dots, \mathbf{r}_N) = E_i\psi_i(\mathbf{r}_1, \dots, \mathbf{r}_N)$ . However, even a single eigenstate of  $H$  already quickly exceeds the storage space of

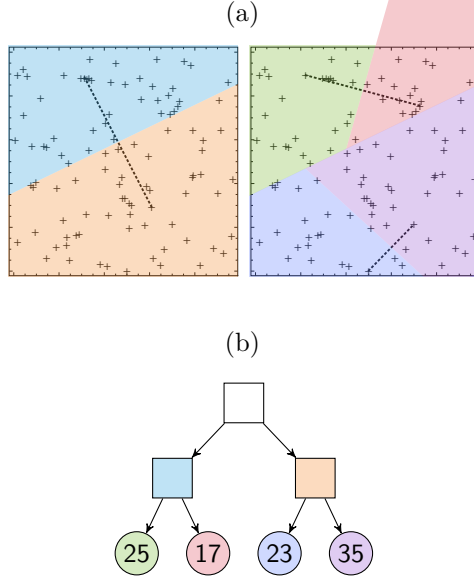


Figure 2.4: The approximate nearest neighbor algorithm ANNOY demonstrated for 100 random points in a 2D Euclidean space. (a) The algorithm starts by splitting the data into two subspaces using the equidistant hyperplane of two randomly selected points. This is done recursively until the number of points in the subspaces is below a certain threshold. (b) The recursive subspaces form a binary tree where nearest neighbors are accessible in logarithmic time. This is Fig. 2 in Paper 8.

a computer. The many-body wavefunction  $\psi_k$  maps the input  $(\mathbb{R}^3)^N$  to a complex number. Therefore, even for a simple  $10 \times 10 \times 10$  grid and only  $N = 6$  electrons, the wavefunction comprises  $10^{18}$  numbers, requiring roughly 4 exabytes of storage capacity. However, it turns out that it is possible to sidestep this problem through a theoretical framework that relies on the electron density  $n(\mathbf{r})$ , mapping  $\mathbb{R}^3$  to a real number. The theory is based on the variational principle.

The *variational principle* is an approach where solutions to a problem are found by minimizing or maximizing certain quantities. This approach is used in quantum mechanics where it can be shown that the ground state energy  $E_0$  is always less or equal than the expectation value given any wavefunction,

$$E_0 \leq \frac{\langle \psi | H | \psi \rangle}{\langle \psi | \psi \rangle}.$$

This is an essential ingredient for many methods in quantum chemistry, including variational Monte Carlo and density functional theory (DFT).

For the *variational quantum Monte Carlo* method it means an arbitrary ansatz wavefunction  $|\Psi(\dots)\rangle$  with many parameters can be used in a minimization scheme. For example, by calculating the derivative of the expectation value  $\langle H \rangle$  relative to the parameters, one can iteratively tweak the parameters to minimize  $\langle H \rangle$  towards  $E_0$  (this is gradient descent).

The foundations of *density-functional theory* are two theorems by Hohenberg and Kohn [38]. The proof of both theorems relies on the variational method. For the first theorem it starts by taking two systems  $H_A$  and  $H_B$  and using variational method. We know that,

$$E_A < \langle \Psi_B | H_A | \Psi_B \rangle \quad \text{and} \quad E_B < \langle \Psi_A | H_B | \Psi_A \rangle,$$

where  $E_A$  is the ground state energy of system  $A$  and  $\Psi_B$  the ground state of  $B$  (and the other way around). It turns out that by comparing these expressions it is impossible to have the same ground state density  $n_0$  for two different system  $A$  and  $B$ . This means there is a 1-to-1 mapping between a system  $H$  and the ground state density  $n_0$ , which is used in the second Hohenberg and Kohn theorem. The variational principle is used here again. Given a trial density  $n_{\text{trial}}$  and its corresponding ground state energy  $E_0$ , it must be greater or equal than the true ground state energy,

$$E_0[n_{\text{trial}}] \geq E_0[n_0].$$

A scheme for choosing  $n_{\text{trial}}$  was introduced by Kohn and Sham [44].

The main idea is to turn the problem into a system of  $N$  non-interacting electrons that has the same ground state electron density as the interacting system. This non-interacting system is solved through the Kohn-Sham equation,

$$\left[ \frac{\nabla^2}{2} + V_{\text{KS}}[n](\mathbf{r}) \right] \phi_i = \varepsilon_i \phi_i,$$

which follows a self-consistent method using multiple iterations, as shown in Fig. 2.5. The iteration starts with calculating the Kohn-Sham potential  $V_{\text{KS}}$  given an initial guess density  $n_{\text{start}}$ ,

$$V_{\text{KS}}[n](\mathbf{r}) = V_{\text{ext}}(\mathbf{r}) + V_{\text{Hartree}}[n](\mathbf{r}) + V_{\text{xc}}[n](\mathbf{r}). \quad (2.1)$$

Each of these terms is calculated separately. The external potential  $V_{\text{ext}}$  is the sum of nuclear potentials,

$$V_{\text{ext}}(\mathbf{r}) = \sum_i^{\text{sites}} V_i(\mathbf{r} - \mathbf{R}_i), \quad (2.2)$$

where  $i$  runs over the sites in the molecule or crystal structure. The potential  $V_\alpha$  is chosen to be the Coulomb attraction  $-Z_\alpha/r$  or a pseudopotential. The Hartree potential is computed using the electron density,

$$V_{\text{Hartree}}[n](\mathbf{r}) = \int d^3r' \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \quad (2.3)$$

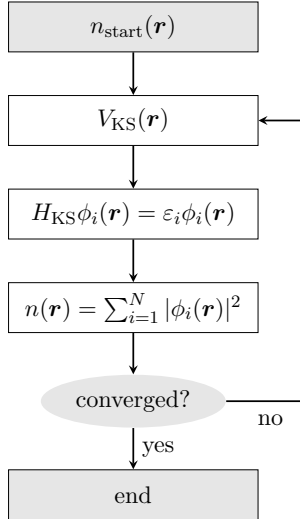


Figure 2.5: Procedure of a Kohn-Sham density functional theory calculation. *Figure adapted from [28] with permission from Springer Nature.*

The potential  $V_{\text{xc}}$  captures all the remaining exchange and correlation effects for the electron, which is, in general, a functional derivative, i.e.  $\delta E_{\text{xc}}/\delta n(\mathbf{r})$ . There are many choices available for the exchange-correlation energy  $E_{\text{xc}}$ , that can be either empirical or non-empirical. For example, in the OMDb we use the non-empirical Perdew-Burke-Ernzerhof (PBE) functional that belongs to the generalized gradient approximations (GGA). In general, the exchange-correlation methods can be ranked in accuracy following *Jacob's ladder* [64]. Finally, the iteration proceeds as listed in Fig. 2.5. Once the electron density has converged, the total energy (or other quantities of interest) can be calculated from it.

## 2.3 Sensors for dark matter

The observational evidence of dark matter suggests the existence of a particle beyond the Standard Model, as mentioned in the introduction. The search for this particle is ongoing, but the absence of the WIMP particle with a typical mass ranging from GeV to TeV, motivates to extend the search to lower mass (sub-GeV). The goal is to detect the dark matter particle non-gravitationally, and in our work we focus on the path of direct detection. This means detecting the recoil of the particle in a sensor material by an observable such as light, charge or heat/phonons. This requires ultrasensitive photon detectors or bolometers. For a suitable low-mass detector there are many constraints, including:

1. Small energy gap – A small energy band gap ensures that low-temperature background noise is excluded from the signal.
2. Shallow Slope – The slope of the Dirac cone is controlled by the Fermi velocity  $v_F$  (see Equation 1.1). A shallow slope is necessary for scattering to be allowed and it was found that the optimal slope is  $10^{-4}c$  to  $10^{-3}c$  where  $c$  is the speed of light [37].
3. Anisotropic / directional sensitivity – The rotation of the Earth combined with an anisotropic material leads to characteristic daily modulation of the signal [30].
4. Practical – A sensor material must be chemically stable, free of impurities that affect the signal, and economically viable.

Given these constraints, we searched for suitable sensor materials in the realm of organic materials (see Paper 7) in the Organic Materials Database and identified a number of candidates. A subsequent detailed study showed that one of those candidates is especially suitable, the quasi-two-dimensional organic molecular crystal bis(naphthoquinone)-tetrathiafulvalene (BNQ-TTF), which contains Dirac nodes that could potentially be gapped by applying strain [30].

Another important factor to consider is the effect of impurities on the signal. We have modeled this using both a discrete and a continuous random mass-term in the Dirac material (see Paper 5). In the case of the discrete mass-term, the presence of rare regions introduces Lifshitz tails [49] inside the band gap. The density of states of the Lifshitz tail decays exponentially, but remains finite throughout the gap. This is captured by the impurity concentration  $c$ , that we assume to be a few percent (consistent with metallurgical-grade silicon). For a normally-distributed mass-term, the density of states decays quadratically. Both of these cases lead to unwanted states within the gap, that is now sensitive to background noise. These models provide a theoretical framework to set even more constraints on potential Dirac sensors. This is still an active field of research and a full detection scheme based on Dirac materials remains to be realized.

## Chapter 3

# Machine learning for materials

The aim of machine learning (ML) is to detect patterns in data and to perform tasks using the uncovered patterns. A more precise definition of learning is provided by Tom Mitchell in 1997 [56]:

**Definition.** A computer program is said to *learn* from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ .

For example, a regression algorithm is tasked to predict the total energy of a molecule ( $T$ ) through experience obtained from traditional ab initio quantum chemistry calculations ( $E$ ), and it is improving as measured by its mean absolute error ( $P$ ) in its predictions.

Successful domains of application of machine learning methods include image processing, speech recognition, and robotics. For example, DeepMind’s AlphaFold models predict the structure of proteins given an amino acid sequence [42,81]. Another example is the OpenAI’s DALL-E model that generates a realistic image given an input text<sup>1</sup> [68]. As mentioned in the introduction, these type of methods have also been used to study materials. Accurate machine learning models can improve and accelerate the discovery of new materials. The exponential increase in computational power and storage capacity has lead to the creation of materials databases, such as the Organic Materials Database (OMDB) described in the previous chapter. Methodology adapted from computer science, and in particular machine learning, is a natural choice to take advantage of these newly emerged data.

Machine learning methods can be divided into three main categories. The *supervised learning* (or *predictive*) methods are given labeled data (i.e. input  $x$  and corresponding output  $y$ ) and construct a mapping between the two. Examples include linear regression and the more sophisticated regression models used in this thesis. The *unsupervised learning* (or *descriptive*) methods are only given a set

---

<sup>1</sup>The cover image of this thesis is the result of giving DALL-E 2 the input “Molecule in the style of Hilma af Klint”.

of inputs  $x$  and have the objective to find patterns. For example, clustering data points without knowing beforehand the desired cluster each point belongs to. Finally, *reinforcement learning* is the approach where a model is simply trying to maximize reward (or minimize punishments) while interacting with an environment. The model (or agent) aims to find a suitable action model based on the state of the environment. Here there is a *exploration-exploitation trade-off*, where random (and often poor) actions are necessary to explore new strategies, that prevent (in the short-term at least) the quick accumulation of more reward. In this thesis, the focus is on supervised and unsupervised learning techniques.

### 3.1 Supervised learning

Supervised learning is the category of machine learning where inputs (or features) are mapped to outputs (or labels). In materials science, it is particularly challenging to form appropriate inputs, sometimes called *feature vectors* or *descriptors*. There are many requirements for a suitable descriptor:

1. Bijective – one unique descriptor corresponds to one material
2. Invariant with respect to rotations and translations.
3. Invariant with respect to permutations of the atoms
4. Efficient – computationally cheap

For extensive properties such as total energy, there should be no invariance with respect to multiples of the unit cell. The representation of molecules and crystals is still an active field of research, but it also has a long history. Simply using the Euclidean coordinates of the atomic sites is not a good representation because it breaks all the required invariance rules. Internal coordinate space (referred to as Z-matrix) that describes a structure through bond lengths, bond angles, and so on, removes degrees of freedom, but still breaks the permutation symmetry and is not unique. Already in the 1980s the SMILES (Simplified molecular-input line-entry system) strings were introduced to encode molecules. This makes it possible to describe the structure of a molecule using a short combination of characters and parenthesis. However, this descriptor is also not unique (multiple SMILES strings encode the same molecule) and it does not capture the exact positions of the atoms<sup>2</sup>. However, there have been many seminal works that introduce descriptors for machine learning that do fit some, or even all the requirements above. These include the Coulomb matrix [72], Bag-of-Bonds [34], Sine Matrix [26], many-body tensor representation (MBTR) [39], and Atom-centered Symmetry Functions (ACSF) [9]. Combining these descriptors with a machine learning model makes it possible to capture the structure-property relationship.

---

<sup>2</sup>For example, the molecule ethanol can be written as C(O)C, CCO, or OCC.

In this thesis, the regression models that are used are kernel ridge regression (KRR) and a deep neural network specifically designed for atomistic systems (SchNet). These models are unique in that they do not require explicitly converting the molecule or crystal structure to a fixed-size input vector. KRR only relies on a pairwise similarity function (i.e. the kernel) that quantifies how similar two input structures are. SchNet uses fixed-size input vectors, but these are part of the learning scheme and not analytically computed. The next sections introduce these models in more detail and discuss their performance using the example of the structure-band gap relationship of materials in the OMDB.

### 3.2 Kernel Ridge Regression

Kernel Ridge regression (KRR) is a relatively simple, yet powerful method that combines *linear regression*, *regularization* and the *kernel trick*. The kernel trick refers to the use of a kernel function that computes the similarity between two inputs. This makes the regression non-linear, in a computationally efficient way, as will be shown in this section.

Ridge regression in  $d$  dimensions is a supervised method that makes predictions,

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x}, \quad (3.1)$$

where  $\mathbf{x}$  is a  $d$ -dimensional input vector. The coefficients  $\mathbf{w}$  minimize the squared error for a given set of labeled data points,

$$\arg \min_{\mathbf{w} \in \mathbb{R}^d} \sum_{i=1}^N \left( f(\mathbf{x}_i) - y_i \right)^2 + \lambda \|\mathbf{w}\|_2^2, \quad (3.2)$$

where  $y_i$  is the label and  $f(\mathbf{x}_i)$  the model prediction. The normalization coefficient  $\lambda$  controls the model complexity, introducing a *bias-variance trade-off*. This is a *hyperparameter*, meaning that it is not part of the model fitting but has to be set outside of the learning procedure. A higher value of  $\lambda$  favors smaller  $\mathbf{w}$  coefficients, which decreases the variance and increases the bias of the model (underfitting). A low value of  $\lambda$  leads to a model with high variance and low bias, potentially overfitting the data. A generalizable and accurate model aims for a balance between bias and variance. The solution of minimizing Equation 3.2 is

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_d)^{-1} \mathbf{X}^T \mathbf{y}, \quad (3.3)$$

where  $\mathbf{X}$  is the input matrix of size  $N \times d$ , containing the  $N$  data points [57].

In order to make the ridge regression non-linear, the input vectors in vector space  $\mathcal{X}$  are mapped to a higher-dimensional vector space  $\mathcal{Z}$  using a function  $\phi : \mathcal{X} \rightarrow \mathcal{Z}$ . However, the kernel trick makes it possible to perform regression without explicitly converting each data point to the higher-dimensional space [1, 57]. Instead, the



regression uses the kernel function  $k$ , that is defined as the inner product in the space  $\mathcal{Z}$

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}') = \mathbf{z}^T \mathbf{z}, \quad (3.4)$$

where  $\mathbf{z}$  and  $\mathbf{z}'$  represent the transformed points  $\mathbf{x}$  and  $\mathbf{x}'$ . Taking Equation 3.3, we define  $\Phi$  as the matrix where  $\phi$  is applied to each row of  $\mathbf{X}$ ,

$$\mathbf{w} = (\Phi^T \Phi + \lambda \mathbf{I}_z)^{-1} \Phi^T \mathbf{y}. \quad (3.5)$$

This can be rewritten using the push-through matrix identity<sup>3</sup>

$$\mathbf{w} = \Phi^T (\Phi \Phi^T + \lambda \mathbf{I}_N)^{-1} \mathbf{y} \quad (3.6)$$

from which we can define the kernel ridge regression coefficients  $\alpha$ ,

$$\alpha = (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{y}, \quad (3.7)$$

where  $\mathbf{K} = \Phi \Phi^T$  is the kernel matrix with elements  $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ . Finally, the prediction can also be expressed in terms of the kernel function,

$$f(\mathbf{x}') = \mathbf{w}^T \phi(\mathbf{x}') \quad (3.8)$$

$$= \phi(\mathbf{x}')^T \Phi^T \alpha \quad (3.9)$$

$$= \sum_i^N \phi(\mathbf{x}') \phi(\mathbf{x}_i) \alpha_i \quad (3.10)$$

$$= \sum_i^N \alpha_i k(\mathbf{x}', \mathbf{x}_i). \quad (3.11)$$

In summary, KRR involves computing the kernel matrix  $\mathbf{K}$ , the regression coefficients  $\alpha$ , and the kernel similarities  $k$  between the training data  $\mathbf{x}_i$  and a new input point  $\mathbf{x}$ . The use of the kernel function sidesteps the explicit conversion of each data point to the higher-dimensional space  $\mathcal{Z}$ . Due to the large kernel matrix and the computationally expensive matrix inversion, this method does not scale to large datasets. However, given an accurate kernel for the problem at hand (i.e. kernel engineering), it can provide state-of-the-art results, especially for smaller datasets, see Paper 6.

## Kernel for crystal structures: SOAP

The SOAP (Smooth Overlap of Atomic Positions) kernel provides a systematic way to compare two atomic structures [7, 8]. As depicted in Fig. 3.1, it is first defined for two atomic neighborhoods, which are subsequently aggregated into a final scalar similarity value.

---

<sup>3</sup>Push-through identity:  $(UV + I)^{-1}U = U(VU + I)^{-1}$  with  $U = \Phi^T$  and  $V = \Phi$ .

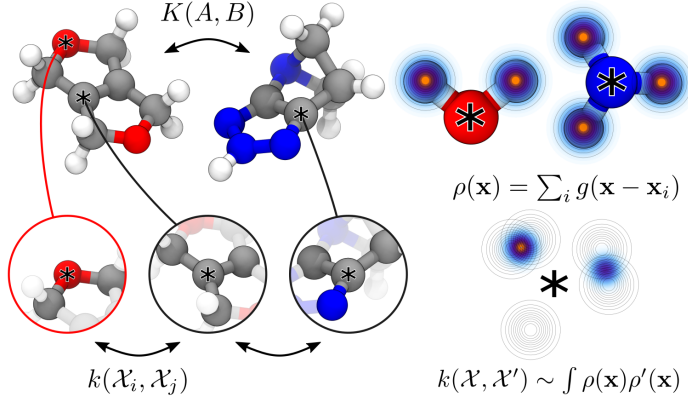


Figure 3.1: The Smooth Overlap of Atomic Positions (SOAP) kernel  $K(A, B)$  between two molecules  $A$  and  $B$  is an aggregate result of local environments comparisons  $k(\mathcal{X}_i, \mathcal{X}_j)$ . A Gaussian is placed at each atomic position and two environments are compared by integrating the overlap of the two environments. *Figure reproduced from [7] (CC BY-NC 4.0).*

A neighborhood is modelled by placing Gaussians at each atomic site,

$$\rho_{\mathcal{X}}(\mathbf{r}) = \sum_{i \in \mathcal{X}} \exp\left(-\frac{(\mathbf{x}_i - \mathbf{r})^2}{2\sigma^2}\right), \quad (3.12)$$

where  $\sigma$  is typically  $0.5 \text{ \AA}$ . Next, the similarity between the two environments  $\mathcal{X}_i$  and  $\mathcal{X}_j$  is computed by integrating over all 3D rotations,

$$k(\mathcal{X}_i, \mathcal{X}_j) = \int d\hat{R} \left| \int \rho_{\mathcal{X}_i}(\mathbf{r}) \rho_{\mathcal{X}_j}(\hat{R}\mathbf{r}) d\mathbf{r} \right|^2. \quad (3.13)$$

All these pairwise local kernels are collected in the similarity matrix,

$$C_{ij}(A, B) = k(\mathcal{X}_i, \mathcal{X}_j). \quad (3.14)$$

There are a number of options for how to combine this matrix into a single scalar kernel that can be used to model a structure-property relationship with kernel ridge regression. The simplest scheme is the average kernel,

$$K(A, B) = \frac{1}{N^2} \sum_{ij} C_{ij}(A, B), \quad (3.15)$$

which takes the average of all the matrix elements. Another popular option is the regularized-entropy match (RE-Match) kernel that aims to find the best pairwise

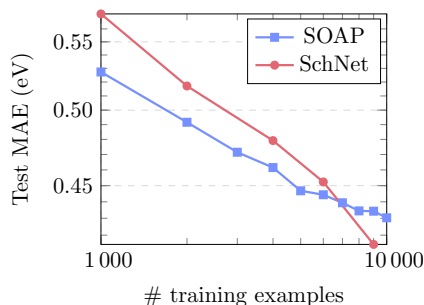


Figure 3.2: Benchmark of kernel ridge regression (KRR) with the Smooth Overlap of Atomic Positions (SOAP) kernel and the SchNet atomistic deep neural network predicting the OMDB band gap for increasing number of training examples. See Paper 6 for more detailed information on the model parameters. This is Fig. 3 in Paper 6.

match (between  $A$  and  $B$ ) for each atomic environment. The different chemical elements can be implemented in several ways, e.g. computing the kernel separately for each element and average the result.

Paper 6 uses the SOAP kernel to predict the band gap for organic crystals. Figure 3.2 shows that given 10 000 training examples, the model has a mean absolute error (MAE) of 0.430 eV on 2500 unseen materials (test set). In contrast, the deep neural network model (SchNet, see Section 3.3) has a MAE of 0.415 eV. However, for smaller number of training examples, the KRR model outperforms SchNet. This model is used to identify candidate solar cells with a band gap of  $(1.34 \pm 0.05)$  eV. This band gap corresponds to the Shockley-Queisser (SQ) limit, the maximum efficiency of a single p-n junction [70].

### 3.3 Neural networks

Neural networks, as the name suggests, are models that combine artificial neurons in a network architecture. In the 1950s, perceptrons were implemented as an artificial neuron [69]. The perceptron computes a single binary output  $y$  based on a weighted sum of its inputs, plus a bias term, i.e.  $y = w \cdot x + b$ . If the result is larger than a threshold value, the neuron outputs one, otherwise zero. A combination of connected perceptrons can model any function, similar to how a combination of NAND logic gates represent universal computation [58]. However, the challenge is to find the weights  $w$  and biases  $b$  that solve a problem. The key is to promote perceptrons to sigmoid neurons [71], where the output  $y$  is no longer binary, but instead just the weighted sum passed through the sigmoid function,

$$y = S(w \cdot x + b) \quad \text{where} \quad S(z) = \frac{1}{1 + e^{-z}}, \quad (3.16)$$

where  $S$  is also called the *activation function*. In comparison to the step function behaviour of perceptrons, the sigmoid neuron is essentially a smoothed step function. The benefit is that the sigmoid neuron is differentiable, allowing us to nudge the weights and biases to a desired output value. More recently, the rectifier linear unit (ReLU) is often used as an activation function to avoid the numerical problems with sigmoid functions (e.g. vanishing gradients at large inputs). In the fully connected neural networks, each neuron is connected to all the neurons in the next layer. The trainable parameters in a combination of neurons are the weights and biases. Given a loss function on the output, each individual parameter can be updated in a way that reduces the output loss, following the chain rule. The algorithm for updating the weights and biases efficiently is called *backpropagation*, because it involves computing the gradients iteratively from the output back to the input layer. Furthermore, the operations boil down to matrix multiplication, which is GPU-accelerated.

Vanilla neural networks, where all neurons in a layer are fully connected to the next layer, are unconstrained and have a large parameter space of possible weights and biases. However, when solving a real-world problem, we can expect a smaller subset of parameters that are sensible, reflecting the symmetries of the problem. Nature appears to favour symmetry, as is clear from the abundance of symmetric structures in biology. The reason is that symmetries represent an efficient way to encode a structure, requiring less information. This concept can also be applied to neural networks.

*Convolutional neural networks* (CNNs) are a neural network architecture where a kernel slides along a the input values and outputs a feature map [46]. This is an example of *weight sharing*, because the kernel has a relatively small number of

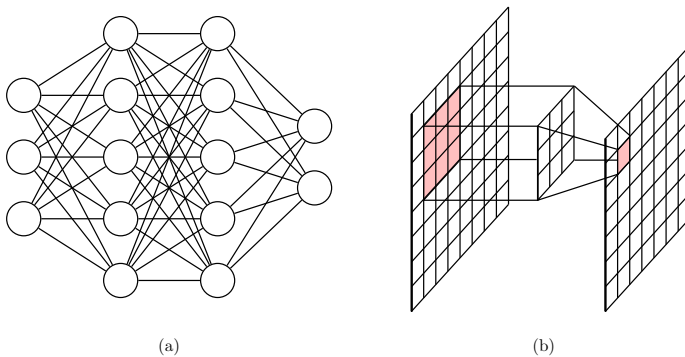


Figure 3.3: Schematic of two different neural network architectures. (a) Fully connected neural network with three inputs, two hidden layers and two outputs. (b) Convolutional neural network turning a  $8 \times 8$  input image into a  $7 \times 7$  feature map by sliding a  $3 \times 3$  kernel across the image.

weights that are utilized repeatedly along the input. The smaller number of total parameters leads to a reduction in computational cost. Additionally, the architecture guarantees that an input shift causes a shift in the activations, which is called translational *equivariance*. This leverages the fact that computer vision problems typically demand translational symmetry. For example, detecting whether an object is present in an image it does not matter exactly where it is located. There is typically a pooling layer to make the output layer invariant. Some problems explicitly require equivariance, where a shift in input must also shift the output in a certain controlled way. An example of such a problem is the prediction of non-scalar values (e.g. force/velocity of particles). An example of an invariant problem is the prediction of a scalar property of a molecule. Crystal structures and molecules have many symmetries that can be incorporated in a neural network, as discussed in the next subsections.

A common approach to address invariance in neural networks is to perform *data augmentation*. The idea is to artificially increase the size of the training dataset by applying random transformations. However, data augmentation is computationally expensive and the model does not guarantee equivariance (a more detailed discussion in Section 3.3). The unconstrained model has to learn both the symmetries of the data and solve the user-defined task at the same time. A small perturbation to the input could lead to wildly different results.

In the next section, a neural network architecture that is effective for molecules and materials is introduced.

### SchNet Atomistic Deep Neural Network

SchNet is a continuous-filter convolutional neural network with competitive accuracy that works both on molecules and crystal structures. For example, when predicting the total energy for the organic molecules in the QM9 dataset, SchNet achieves 0.31 kcal/mol with 110 000 training examples [78]. This is well below what is generally considered to be chemical accuracy of 1 kcal/mol. For the Organic Materials Database, the model reaches a mean-absolute error (MAE) of 0.415 eV (see Fig. 3.2).

Figure 3.4 shows the architecture of the model. The neural network is evaluated for each site in the atomic structure separately, and the atom-wise contributions are summed or averaged, depending on whether the output is an extensive or intensive property, respectively. This kind of weight-sharing achieves permutation equivariance (that becomes invariance after the sum/avg pooling). The use of local environments also makes the model translationally invariant. Finally, the way neighboring sites are introduced in the model achieves rotational invariance.

The  $N$  atoms are each represented by a vector  $\mathbf{x}_i^l \in \mathbb{R}^F$  where  $F$  is the number of feature maps in the model and  $l$  denotes the layer in the model. Initially, the vector is set to an embedding that is dependent on the atomic number  $Z_i$ ,

$$\mathbf{x}_i^0 = \mathbf{a}_{Z_i}. \quad (3.17)$$

The *embedding* is a  $F$ -dimensional vector that is initialized randomly and optimized during the training with backpropagation. This is a popular technique that is used in many machine learning applications when the input consists of discrete objects. For example, in the `word2vec` model in natural language processing, unique words are converted into embedding vectors [55]. This is denoted by the embedding block in Fig. 3.4.

Next, the interaction blocks introduce neighboring sites into the model

$$\mathbf{x}_i^{l+1} = \sum_{j=0}^N \mathbf{x}_j^l \odot W^l(d_{ij}) \quad (3.18)$$

where  $\odot$  represents element-wise multiplication,  $d_{ij}$  is the distance between site  $i$  and  $j$ , and  $W^l$  is a filter that maps  $\mathbb{R} \rightarrow \mathbf{R}^F$ . The interaction blocks update the vector  $\mathbf{x}_i$  that now incorporates the information of the interactions. There are a number of options for the filter  $W^l$  that weights the neighboring atoms, but the original SchNet model proposed a filter-generating neural network (cfconv block in Fig 3.4). For more detail on this implementation see [79].

The atom-wise layers layers are simply fully-connected layers that are applied to each atom  $i$ ,

$$\mathbf{x}_i^{l+1} = W^l \mathbf{x}_i^l + \mathbf{b}^l \quad (3.19)$$

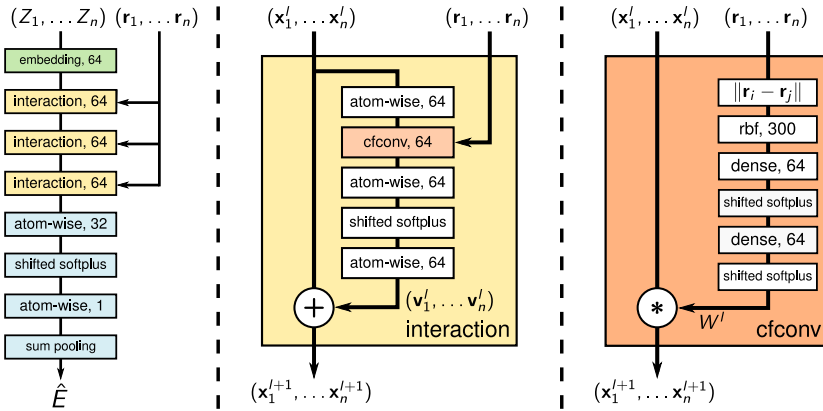


Figure 3.4: The architecture of SchNet predicts a single contribution for each atom, which is either averaged or summed in the end. Here we show the (extensive) total energy that is the sum of all atom-wise contributions. Starting from chemical embeddings (vector representations) where interaction layers with continuous-filter convolutions capture how different atoms interact with each other. The model parameters are trained by backpropagation. *Figure reproduced from [78] with permission from ACM.*

where  $W$  and  $\mathbf{b}$  are shared across all atoms  $i$ . This weight sharing makes the model efficient and permutation invariant.

In short, SchNet takes atom embeddings and update them at each layer to incorporate information about the interactions. SchNet is also differentiable. This allows for training using both forces and energies, since these are usually also available in the quantum chemistry codes. It also allows for inverse design by tuning the atomic positions to tune the functional property of the material. Moreover, the differentiability can be used to compute the forces on the  $N$  atoms,

$$\mathbf{F}_i(\mathbf{r}_1, \dots, \mathbf{r}_N) = -\frac{\partial E}{\partial \mathbf{r}_i}(\mathbf{r}_1, \dots, \mathbf{r}_N), \quad (3.20)$$

to find the equilibrium conformation of a molecule or material.

In the following section we show how the concept of equivariance constraints can be implemented in general, going beyond translations and rotations. This is also referred to as introducing an inductive bias to the model. By the choice of a suitable architecture, we reduce the search space of all possible neural networks to a realistic subset.

### Group equivariant neural networks

Group theory provides the necessary framework to implement symmetries in neural networks. Typical examples of groups of interest are the Euclidean group  $\mathbb{E}(n)$  (translations, rotations and reflections), the rotation group  $\text{SO}(n)$  (rotations in  $n$  dimensions) and cyclic group  $\mathbb{Z}_n$  (discrete translations). Different groups can also be combined to fit the symmetries of the problem at hand. A group element  $g$  is abstract, but in practice described by a *group representation*: an invertible matrix  $\rho(g)$  that represents the group element. Performing a group action to an input vector is then simply matrix multiplication.

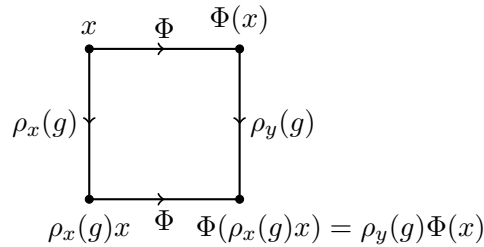


Figure 3.5: An operator (e.g. neural network)  $\Phi$  is equivariant if the input  $x$  that is transformed by  $\rho_x(g)$  and mapped to an output  $\Phi(\rho_x(g)x)$  is the same as mapping the input  $x$  to an output  $\Phi(x)$  and transforming the output  $\rho_y(g)\Phi(x)$ . In other words, the operator  $\Phi$  commutes with the group element  $g$ .

Figure 3.5 illustrates the definition of equivariance. A neural network is equivariant under a group  $G$  if it commutes with the group elements  $g \in G$ . Note that the input and output are transformed by the representation  $\rho$ , which is not necessarily the same for the input and output space. Convolutional neural networks are translation equivariant, typically with discrete grid translations, i.e.  $\mathbb{Z}_n \times \mathbb{Z}_n$ . However, ideally we want weight sharing and equivariance constraints that go beyond translations. For example, a CNN that detects people in an image might use features such as edges. The person might be in a different pose or further away in the image, indicating that rotation and scale equivariance are important.

In 2016, Cohen and Welling introduced group equivariant convolutional neural networks G-CNNs that included translations, reflections and 90° rotations [19]. In group theory notation, these G-CNNs have the symmetry group  $\mathbb{Z}_4 \ltimes (\mathbb{Z}_n \times \mathbb{Z}_n)$ . This group preserves the grid structure of the pixels, and a group action is simply a permutation of pixels. These representations allow the use of any element-wise activation function and the implementation is efficient. G-CNNs have also been shown to be robust to input noise and are more sample efficient [10]. In 2017 the Steerable CNNs were introduced with the aim of going beyond discretized equivariance (for example, including infinite rotational group actions) [20,91]. Recently, the escnn library was released that makes it easy to implement  $E(n)$ -equivariant CNNs. The Euclidean group  $E(n)$  includes rotations, translations and inversion in  $n$  dimensions, and is relevant for a wide range computer vision problems. This library will be used in the next section to demonstrate a biological example.

Finzi *et al.* introduced a method in 2021 to construct equivariant neural networks for any given group [27], where the weights and biases are projected to an equivariant subspace that is controlled by the group and the input and output representations.

$$y = \sigma(P_w W \cdot x + P_b b) \quad (3.21)$$

Note that  $x$  and  $y$  are vectors of representing a fully connected layer, and  $P_w$ ,  $W$  and  $P_b$  are matrices. The projectors  $P_w$  and  $P_b$  are found by solving the constraints given by the representations and group. This provides a unified, albeit computationally expensive, framework to implement any group equivariance. The approach of using projectors does not reduce the number of model parameters through explicit weight sharing, but the space of possible weights and biases is reduced. It is also important to implement the correct activation function  $\sigma$  (called gated nonlinearity) to not break equivariance. Pooling layers are used to achieve invariance along chosen group symmetries, similar to the case of translational equivariance in CNNs. For example, when predicting a location in the image, it is useful to make it rotational invariant.

### 3.4 Dimensionality-reduced chemical space

Dimensionality reduction aims to map high-dimensional data to a low-dimensional representation that preserves characteristic features of the original data. The out-



put data is often two or three-dimensional such that it can be plotted as a point cloud. A basic example is principal component analysis (PCA), where each data point is projected along axes that capture the most variance in the high-dimensional data [82]. In general, dimensionality reduction can be useful in many machine learning applications. For example, we can visualize high-dimensional vectors in neural networks to gain insight into their inner workings. In this compilation thesis, dimensionality-reduction is used in Paper 3 and 6. In this section, the focus is on the reduction of chemical space, which is used a number of times in the latter publication.

PCA is a linear dimensionality reduction technique, and its principal components are linear combinations of the input features. This may not be an effective way to characterize the data distribution. The method can be extended into kernel PCA, using the kernel trick as described for kernel ridge regression. Alternatively, there are a number of stochastic algorithms that preserve the local (and sometimes global) structures in the data. The t-distributed stochastic neighbor embedding (t-SNE) algorithm is incredibly popular in the machine learning community and it is used to make low-dimensional graphs where the local structure of very high dimensional data is preserved [87]. More recently, the Uniform Manifold Approximation and Projection (UMAP) algorithm is increasingly more popular because of improved performance on large datasets and the ability to preserve a balance of local and global structure of the data [54].

Dimensionality reduction is used in Paper 6 to understand chemical space based on similarities between crystal structures. First, the SOAP kernel is used to compute all pairwise distances between crystal structures. Next, this distance matrix is converted into a two-dimensional scatter plot using the t-SNE algorithm. Finally, each point is color-coded by the band gap that was computed using density-functional theory. Figure 3.6 shows the result of this analysis, and highlights a few interesting materials in this map. Materials with very distinct structures form

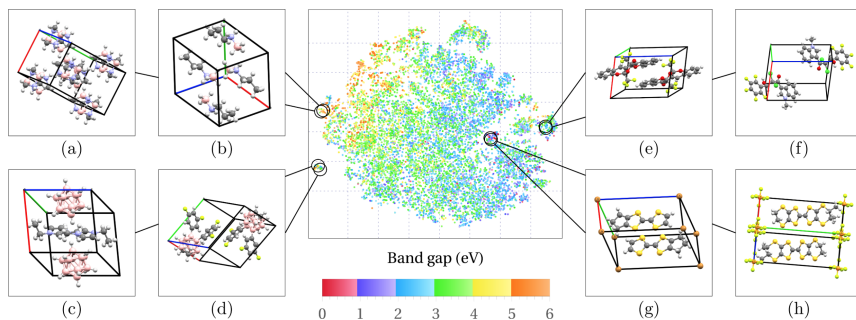


Figure 3.6: Dimensionality-reduced chemical space with the t-SNE algorithm using the average SOAP kernel ( $r_c = 4 \text{ \AA}$ ,  $n = 8$ ,  $l = 6$ ) as a distance metric. This is Fig. 4 in Paper 6.

well-isolated islands, such as the Boron clusters in Fig. 3.6 (c) and (d). Remarkably, clusters of materials appear that have been reported in the literature with certain applications. For example, chemical hydrogen storage devices appear in the vicinity of Fig. 3.6 (a) and (b). Similarly, Fig. 3.6 (g) and (h) are in a zero band gap region that contains organic metals and semiconductors. It is even possible to find structures that form a line in this dimensionality-reduced picture, revealing that these are a sequence of structures reported in pressure study. In short, the dimensionality-reduced chemical space provides an intuitive way to navigate the structure-property landscape of organic crystals.



## Chapter 4

# Homology for materials

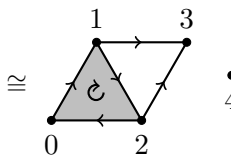
The field of *topological data analysis* (TDA) uses techniques from topology to extract information about the shapes present in data. While it is easy to see trends in two or three-dimensional data by eye, datasets are often high-dimensional. Cross-sections or dimensionality reduction (see Section 3.4) can be used to visualize high-dimensional data but usually loses some information. The main algorithm in TDA is persistent homology, which focuses on finding the shapes in discrete data that persist at different length scales. This is computed by first converting the data to a simplicial complex and filtering this complex using a chosen length scale. Next, homology groups describe the shapes present in the complex and this information is visualized in the form of persistence diagrams or barcodes. These concepts are introduced in detail in the sections below, along with examples.

Since persistent homology is relatively new, it has not seen much use in physics yet. However, this is rapidly changing and a wide range of applications exists with examples such as the filamentary structure present in the cosmic web [83] and the shapes in nanoporous materials [47]. In the domain of condensed matter physics, this method has been used to detect phase transitions in both classical [21, 23, 75, 80, 86] and quantum [36, 84, 86] lattice systems. We have contributed to this growing field by showing that persistent homology can fully capture a complex phase diagram of a classical spin model, even so-called hidden orders that are harder to characterize (see Paper 3). Beyond classical systems, we have also shown that persistent homology captures entanglement structures and quantum phase transitions in Paper 1.

### 4.1 Simplicial homology

The  $k$ -dimensional generalization of points, line segments and triangles are simplices. Each *simplex* can be oriented in two ways. Figure 4.1 shows examples of such simplices. The simplices can be glued together to form spaces, and their combinatorial nature makes the computation of their topological properties easier.

The definitions and concepts of simplicial homology are introduced here with an example simplicial complex  $K$ ,

$$K = \{ \underbrace{[012]}_{\text{2-simplex}}, \underbrace{[01], [12], [20], \dots}_{\text{1-simplices}}, \underbrace{[0], [1], \dots, [4]}_{\text{0-simplices}} \}$$

(4.1)

which consists of a 2-simplex, five 1-simplices and five 0-simplices. To define homology, we first need to introduce what chains, boundaries and cycles are. A combination of  $k$ -simplices (i.e. walking around on  $K$ ) forms a  $k$ -chain. For example, the set of 1-chains of  $K$  are,

$$C_1(K) = \{a[01] + b[12] + c[20] + d[13] + e[23]\}, \quad (4.2)$$

where the coefficients are typically integers  $\mathbb{Z}$  or booleans  $\mathbb{Z}_2$ . Typically, shorthand notation is used that makes the coefficients implicit,

$$C_1(K) = \langle [01], [12], [20], [13], [23] \rangle. \quad (4.3)$$

All the  $k$ -chains together form a *chain group*, and  $c \in C_k$  denotes an element, a specific  $k$ -chain, in this group. In other words, a  $k$ -chain group element  $c_i = \sum_i \alpha_i \sigma_i$  is a sum of simplices  $\sigma_i$  of the same dimension  $k$ . Adding two chains  $c_1$  and  $c_2$ , leads to another chain group element  $c_3$ , i.e.

$$\begin{aligned} c_3 &= c_1 + c_2 \\ &= \sum_i \alpha_i \sigma_i + \sum_i \beta_i \sigma_i \\ &= \sum_i (\alpha_i + \beta_i) \sigma_i, \end{aligned} \quad (4.4)$$

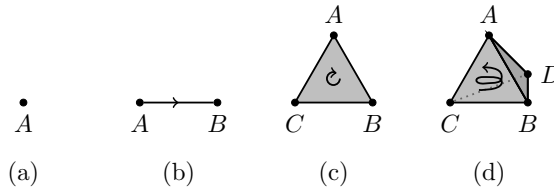


Figure 4.1: Examples of oriented simplices: (a) 0-simplex (point), (b) 1-simplex (line segment), (c) 2-simplex (filled triangle), (d) 3-simplex (filled tetrahedron). This is Fig. 3 in Paper 1.

where the addition  $\alpha_i + \beta_i$  is based on the choice of coefficients, and could include module 2 if  $\mathbb{Z}_2$  coefficients are chosen. Homology will be defined on these chains by looking for chains with specific properties.

The driving concept of homology is the *boundary operator*  $\partial$ . For a  $k$ -simplex, this is defined as

$$\partial S = \sum_{j=0}^k (-1)^j [v_0 v_1 \dots \hat{v}_j \dots v_k], \quad (4.5)$$

where  $\hat{v}_j$  is removed from the sequence. For example, the 2-simplex in  $K$  becomes the sum of its faces,

$$\partial[012] = [12] - [02] + [01] \quad (4.6)$$

where the sign captures the orientation of how to walk around the 2-simplex. Note that applying the boundary operator once more gives zero. This is the *Fundamental Lemma of Homology*,  $\partial\partial S = 0$ , that underlies homology [25]. We can define the *boundary group*  $B_k = \text{Im } \partial_{k+1}$ , for example

$$B_0 = \text{Im } \partial_1 \quad (4.7)$$

$$= \partial_1 C_1(K) \quad (4.8)$$

$$= \{a([1] - [0]) + b([2] - [1]) + c([0] - [2]) + d([3] - [1]) + e([3] - [2])\} \quad (4.9)$$

$$= \{(-a + c)[0] + (a - b - d)[1] + (b - c - e)[2] + (d + e)[3]\}. \quad (4.10)$$

This is a subgroup of all the possible 0-chains (combinations of points), i.e.  $B_0 \subset C_0$ .

Another important subgroup of the  $k$ -chain group is the *cycle group* of  $k$ -cycles, usually denoted by  $Z_k$  (from the German word Zyklus). This is the kernel of the boundary operator (all the elements that are mapped to the identity element),

$$Z_k = \text{Ker}(\partial_k) \quad (4.11)$$

$$= \{c \in C_k \mid \partial(c) = 0\}. \quad (4.12)$$

The  $\partial(c) = 0$  constrains the coefficients of the  $k$ -chains. The 1-cycles  $Z_1$  for the example  $K$  are given by solving for the coefficients in Equation 4.10. This is done by forming a boundary matrix,

$$\begin{matrix} & [01] & [12] & [20] & [13] & [23] \\ \begin{matrix} [0] \\ [1] \\ [2] \\ [3] \end{matrix} & \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{matrix}, \quad (4.13)$$

where the columns represent 1-simplices and the rows its faces. This is reduced to row echelon form,

$$\begin{array}{c} [0] \\ [1] \\ [2] \\ [3] \end{array} \begin{pmatrix} [01] & [12] & [20] & [13] & [23] \\ 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (4.14)$$

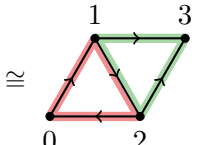
$\alpha \qquad \qquad \beta$

and identifying the space of solutions

$$\begin{pmatrix} [01] \\ [12] \\ [20] \\ [13] \\ [23] \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \alpha + \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 1 \end{pmatrix} \beta \quad (4.15)$$

Therefore the solutions are multiples of the cycles  $[01] + [12] + [20]$  and  $[12] - [13] + [23]$ , i.e.

$$Z_1 = \left\langle [01] + [12] + [20], [12] - [13] + [23] \right\rangle$$



(4.16)

Note that the larger cycle of  $0-2-3-1-0$  can be formed using a linear combination. Because  $\partial\partial S = 0$ , the boundary group elements are also cycles. In summary, the boundary operator leads to a structure of chains, cycles and boundaries,  $B_k \subset Z_k \subset C_k$ , see Fig. 4.2.

The aim of homology is to identify  $k$ -cycles that are not just simply boundaries of  $(k+1)$ -simplices. These cycles describe the interesting topological features (e.g. holes) of the simplicial complex. Therefore, the *homology group*  $H_k$  is defined as the quotient (or “cycles mod boundaries”),

$$H_k = \frac{Z_k}{B_k} = \frac{\text{Ker } \partial_k}{\text{Im } \partial_{k+1}}. \quad (4.17)$$

For the example  $K$ , the 1-homology group is

$$\begin{aligned} H_1(K) &= \frac{\langle [01] + [12] + [20], [12] - [13] + [23] \rangle}{\langle [12] - [02] + [01] \rangle} \\ &\cong \langle [12] - [13] + [23] \rangle \cong \mathbb{Z}, \end{aligned} \quad (4.18)$$

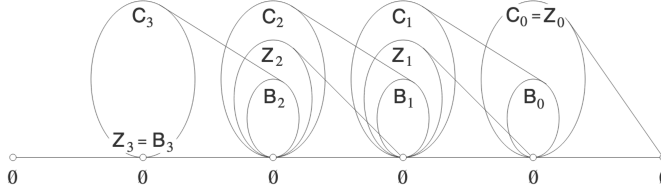


Figure 4.2: Chain  $C_k$ , cycle  $Z_k$  and boundary  $B_k$  groups and their connection to the  $k - 1$  groups through the boundary operator. This would be the structure for a 3-simplex. This also shows that  $B_k \subset Z_k \subset C_k$ . *Figure reproduced from [24] with permission from Springer Nature.*

After a similar calculation for other homology groups, we can conclude that the homology groups for  $K$  are  $H_0 \cong \mathbb{Z}^2$ ,  $H_1 \cong \mathbb{Z}$  and  $H_k = 0$  for  $k > 1$ . The simplicial complex is a topological space, and the zeroth homology group  $H_0$  captures the connected components in that space, similar to traditional clustering. The first homology group  $H_1$  detects the one-dimensional holes in the space, constructed from edges (1-simplices). The second homology group  $H_2$  measures the two-dimensional holes (voids) in the space. In general, the  $k$ -th homology group describes  $k$ -dimensional holes. In other words, the homology groups capture the topology of the topological space  $K$ . Often, we are only interested in the number of connected components, holes, or voids, which is given by the *Betti number*, defined as

$$\beta_k = \text{rank}(H_k). \quad (4.19)$$

In the case of our example  $K$ , the Betti numbers are  $\beta_0 = 2$ ,  $\beta_1 = 1$ ,  $\beta_k = 0$  for  $k > 1$ . These are the topological invariants which can also be used to compare two different (triangulated) spaces as described earlier in Section 1.2. Figure 4.3 shows some example topological spaces and their homology.

### Homology with different coefficients: torsion

The common practice of computing homology with  $\mathbb{Z}_2$  is efficient, but the more general case of homology with integer coefficients (i.e. integral homology) captures more information. In other words, topological spaces  $A$  and  $B$  with different integral homology groups ( $H_k(A, \mathbb{Z}) \neq H_k(B, \mathbb{Z})$ ) may have not be distinguished by the homology groups with  $\mathbb{Z}_2$ , i.e.  $H_k(A, \mathbb{Z}_2) = H_k(B, \mathbb{Z}_2)$ . This is connected to the notion of *torsion*. In this section we introduce this concept and show some examples.

Torsion means that a group element has *finite order* and maps back to identity after  $n$  times, i.e.  $g^n = e$  for a positive integer  $n$  and identity  $e$ . It turns out that non-orientable surfaces often have torsion in its integral homology. A surface is orientable if there is a consistent notion of clockwise rotation while moving around. Non-orientable surfaces, such as the Möbius strip or Klein bottle, turn clockwise



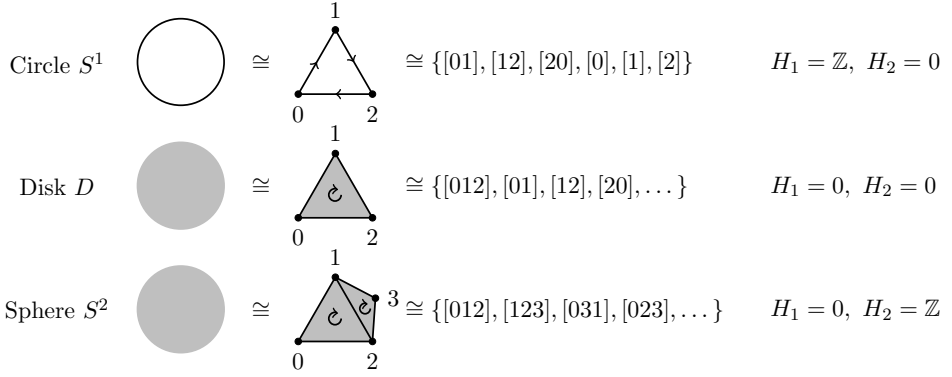


Figure 4.3: A number of common topological spaces, their simplicial complex and their homology groups  $H_k$  over the integers.

to counterclockwise. For example, the Möbius strip, a twisted band with only one side, can be triangulated and represented by the complex,

$$M \cong \begin{array}{ccccc} & 1 & & 4 & 2 \\ & \bullet & & \xrightarrow{\quad} & \bullet \\ 3 \uparrow & \nearrow & & \nwarrow & \downarrow 3 \\ & \bullet & & 5 & \bullet \\ & 2 & & & 1 \end{array} . \quad (4.20)$$

Here, there are 1-chains that are linear combinations of the simplices 3, 4, 5, and 6. Computing its reduced boundary matrix, we find the 1-homology for both integer and  $\mathbb{Z}_2$  coefficients to be,

$$H_1(M, \mathbb{Z}) \cong \mathbb{Z}, \quad H_1(M, \mathbb{Z}_2) \cong \mathbb{Z}_2. \quad (4.21)$$

In particular, we note that the 1-homology over the integers is torsion-free, and it is isomorphic to the group of integers  $\mathbb{Z}$ . This is the same result we would get for a cylinder, confirming that the Möbius strip is topologically equivalent (as listed in Fig. 1.2).

However, taking the Möbius strip and glueing its sides together (which is possible in  $\mathbb{R}^4$  without intersecting itself), we obtain the real projective plane,

$$\mathbb{RP}^2 \cong \begin{array}{ccccc} & 1 & & 4 & 2 \\ & \bullet & & \xrightarrow{\quad} & \bullet \\ 3 \uparrow & \nearrow & & \nwarrow & \downarrow 3 \\ & \bullet & & 4 & \bullet \\ & 2 & & & 1 \end{array} . \quad (4.22)$$

Note that the bottom edge (1-simplex that is numbered 4) is now the same as the top edge and that the arrow indicates the orientation of how these sides were glued together. This space has the 1-homology of

$$H_1(\mathbb{RP}^2, \mathbb{Z}) \cong \mathbb{Z}_2, \quad H_1(\mathbb{RP}^2, \mathbb{Z}_2) \cong \mathbb{Z}_2. \quad (4.23)$$

(For detailed computation of these homology groups, see the appendix of Paper 1.) Importantly, the 1-homology with  $\mathbb{Z}_2$  coefficients is not able to distinguish the Möbius strip and the real projective plane. This highlights the importance of integral homology.

Note that both orientable and non-orientable manifolds may have torsion in its integral homology. However, for closed connected non-orientable manifolds, there is a theorem that states that if  $M$  is a  $n$ -manifold, the homology group  $H_{n-1}$  contains torsion [35, Corollary 3.28]. Furthermore, the universal coefficient theorem provides a description of the integral homology groups change over different coefficients. However, the effect of torsion and homology over different coefficients for the applications in Paper 3 and Paper 1 has not been studied yet and is a topic for future work.

## 4.2 Persistent homology

Persistent homology (PH) is a way to find qualitative features of data with complex structure. This theoretical framework is flexible and the inputs that can be studied include point clouds, digital images, level sets of functions and networks [61]. PH provides a robust and compact description of the shapes that are present in the data, with a common representation being *barcodes*.

Historically, the algorithm of persistent homology was first described in 1994 [73], but only became popular when it was rediscovered in 2000 [24]. In 2004 the name *barcode* arose [17] and the algorithm was put on a mathematical footing and for general coefficients [92]. The basic framework is still generalized and refined in many ways. For example, algorithm for persistent homology finding integer homology groups by running with different coefficients [12].

Mathematically, persistent homology is the homology of a *filtration*. We construct a filtration by arranging a sequence of subcomplexes,

$$K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K, \quad (4.24)$$

which corresponds to a sequence of nested topological spaces [25]. There are different commonly used choices for filtration. For example, the Čech complex refers to a point cloud in a metric space and having the overlap of increasingly larger disks (or balls in 3D) decide the appearance of simplices. Once two disks overlap, a 1-simplex (line segment) connects the two associated points at the center of disks. Where three disks overlap, a 2-simplex (filled triangle) appears, and so on for higher dimensional simplices. It turns out that the union of disks at scale  $r$  is homotopy

equivalent to the simplicial complex at  $r$ , meaning that they share their topological features (this is known as the Nerve Theorem [25]). Another common choice is the Vietoris-Rips complex, where again, a simplex is formed as soon as disk overlap. However, here a  $(k - 1)$ -simplex is formed as soon as  $k$  points are connected. This subtle difference means that a complex of three simplices (like the circle in Fig 4.3) does not appear, and it immediately forms a disk (also shown in Fig 4.3). This is an approximation and the Nerve Theorem does not hold, but the Vietoris-Rips complex is more computationally efficient.

The homology is computed at every level of the filtration sequence. Figure 4.4 shows an example Vietoris-Rips complex for a small point cloud in two-dimensional Euclidean space. Homology elements appear (referred to as birth) and disappear (referred to as death), and their span is indicated by a bar in the persistence *barcode*. As the proximity parameter  $\epsilon$  (in this case Euclidean distance) changes, the simplicial complex grows. The zeroth homology group  $H_0$  tracks the number of connected components, similar to a dendrogram in traditional clustering algorithms. Note that there is always one bar surviving until infinity, reflecting the situation where all points are connected. The first homology group  $H_1$  detects the hole in point cloud, and tells us at which length scale it appears (birth) and when it closes again (death).

In summary, the barcode shows at what length scale topological features exist in the discrete input data. Usually, the bars with a long lifetime (i.e. death - birth) are considered important features, and the short bars are considered noise. However, the short lifetime bars can capture the curvature of the point cloud, which could be of interest depending on the application [15]. In practical applications, it is also common to use the Betti numbers as defined in Equation 4.19. The Betti numbers

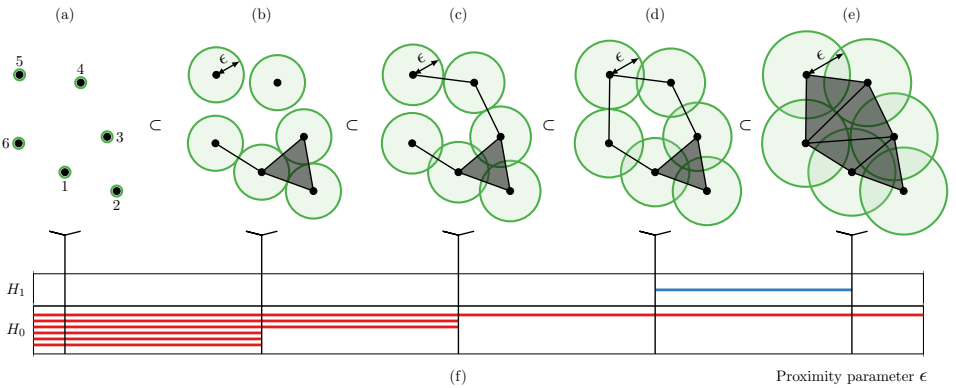


Figure 4.4: Vietoris-Rips (VR) complex of a point cloud in  $\mathbb{R}^2$  with six data points (a) at increasing proximity parameter  $\epsilon$  (b-e). The barcode (f) shows the persistent homology. This is Fig. 2 in Paper 1.

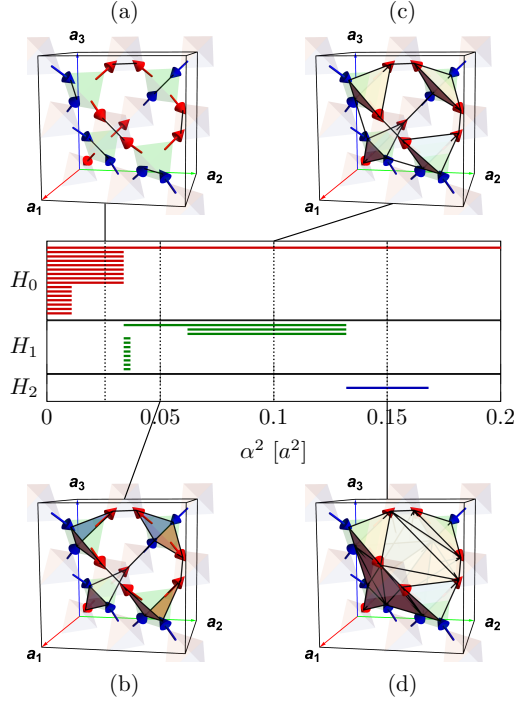


Figure 4.5: (a)-(d) Persistence barcode for a single unit cell of the spin ice phase, which has the spins pointing “two-in, two-out” (color-coded blue and red) aligned to the local  $z$  axis in each lattice tetrahedron. This is Fig. 2 in Paper 3.

can track when the topology of the point cloud changes. There are also distance metrics that compute a scalar distance value between complete barcodes, such as the Bottleneck and Wasserstein distance [14, 88].

In the next two sections, we demonstrate two applications of persistent homology for spin models. Starting with classical Heisenberg spins, where the spins are the point cloud and a change in barcode indicates a phase transitions. Afterwards, the intriguing case of quantum spins is studied.

### 4.3 Classical spin structures

Traditionally, phase transitions are detected using an order parameter, such as the total magnetization in the Ising model. However, finding the optimal order parameter can be challenging, which are usually constructed by hand. Especially in the presence of frustration, complicated phases that lack long range ordering can

appear. In Paper 3, we study the XXZ model,

$$H_{\text{XXZ}} = \sum_{\langle i,j \rangle} J_{zz} S_{i,z} S_{j,z} - J_{\pm} (S_i^+ S_j^- - S_i^- S_j^+), \quad (4.25)$$

with  $S_i^{\pm} = S_{i,x} \pm iS_{i,y}$  and  $\|\mathbf{S}_i\| = 1$  on a pyrochlore lattice (see Fig. 4.5). This model hosts six competing phases depending on the temperature  $T$  and the value of the exchange interaction  $J_{\pm}/J_{zz}$  [85].

In practice, one samples a small number (e.g. 32) spin configuration from a classical Monte Carlo simulation. The spin configurations are converted into persistence barcodes that show the presence of shapes at different length scales. For example, Fig. 4.5 shows the characteristic barcode for the spin ice phase. Abrupt changes in the barcode indicate a phase change, and the change in barcode is measured by the Sliced-Wasserstein distance. Finally, the space of barcodes can be dimensionality-reduced to produce a full phase diagram with any number of phases. For more details on these results, see Paper 3.

## 4.4 Quantum entanglement structures

One of the most profound features of quantum mechanics is *quantum entanglement*<sup>1</sup>. It means that two subsystems, such as spins, can be entangled with each other. The simplest example is the Bell state,

$$|\psi\rangle = \frac{1}{\sqrt{2}} (|00\rangle + |11\rangle), \quad (4.26)$$

which considers a system of two quantum bits (also called qubits). A *quantum bit* is a system which has two states that are labeled 0 and 1, for example a spin up or spin down of a particle. The two qubits form a *superposition*, where if the first qubit is measured to be in the 0 state, the second qubit is in the 0 state too.

This is remarkable, because the two subsystems can be very far apart, which lead Einstein to refer to this seemingly impossible behavior as “spooky action at a distance”. It has been experimentally verified, and in 2022 the Nobel Prize in Physics was awarded to A. Aspect, J. F. Clauser and A. Zeilinger for pioneering experiments with quantum entanglement. In 2017 this fact was used to perform quantum key distribution across 7600 kilometers (Graz, Austria – Xinglong, China) using the Chinese satellite Micius [48].

Quantum states do not always have entanglement, and can sometimes be separated and written as a tensor product, e.g.

$$|\psi\rangle = \frac{1}{\sqrt{2}} (|01\rangle + |11\rangle) = \frac{1}{\sqrt{2}} (|0\rangle + |1\rangle) \otimes |1\rangle. \quad (4.27)$$

---

<sup>1</sup>Erwin Schrödinger wrote this about quantum entanglement (in 1935): “I would not call that *one* but rather *the* characteristic trait of quantum mechanics, the one that enforces its entire departure from classical lines of thought.” [76].

A generalization of the Bell state to many qubits leads to a highly entangled state that is also referred to as the cat state (due to Schrödinger's cat, the thought experiment) or the Greenberger–Horne–Zeilinger (GHZ) state,

$$|\psi\rangle = |1111\dots\rangle + |0000\dots\rangle. \quad (4.28)$$

This macroscopic entanglement has been realized in quantum computers, but for a large number of spins, any small perturbation to a single site will cause the state to collapse to either all 1 or 0.

There are a number of ways to quantify entanglement between quantum subsystems. A more general way of describing a quantum state is the density matrix  $\rho$ ,

$$\rho = \sum_k p_k |\psi_k\rangle \langle \psi_k|, \quad (4.29)$$

where  $|\psi_k\rangle$  are pure states and  $p_i$  are probabilities that a system is in  $|\psi_k\rangle$ . Splitting the system in  $A$  and  $B$ , we can introduce the reduced density matrix associated to a subsystem  $A$ ,

$$\rho_A = -\text{Tr}_B \rho, \quad (4.30)$$

where  $\rho$  is the density matrix of the full system, and  $\text{Tr}_B$  is the partial trace over the basis of subsystem  $B$ . Now we can define *entanglement entropy* as the Von Neumann entropy,

$$S_A = -\text{Tr}_A (\rho_A \log \rho_A), \quad (4.31)$$

which ranges from 0 to  $\log(d)$  where  $d$  is the dimension of Hilbert space of  $A$  or  $B$  (whichever is smaller).

Quantifying entanglement is a general tool that shows up in many applications (e.g. quantum computing). In our case, we are especially interested in the case where Hilbert space is split spatially in a region  $A$  and its complement  $B$ . For most states in Hilbert space, the entanglement entropy scales with the volume of the region of  $A$ . This is because any two distant subsystems could be entangled. However, for the ground state of almost all Hamiltonian with local interactions, the scaling obeys an area law,

$$S_A \propto \Sigma, \quad (4.32)$$

where  $\Sigma$  is the boundary between subsystem  $A$  and its complement  $B$  [32]. This is a theorem for one-dimensional gapped systems, but it is also usually true in higher dimensions. Therefore, there is only a small corner of the Hilbert space (i.e. where states follow the area law), that is relevant when searching for the ground state. Additionally, these states have lower entanglement entropy, meaning that the ground state wavefunction is encoded by fewer complex parameters.

Entanglement is also proposed to be the source of the geometry of spacetime [16, 66, 67]. Analogous to the information in classical bits encoding the virtual 2D or 3D

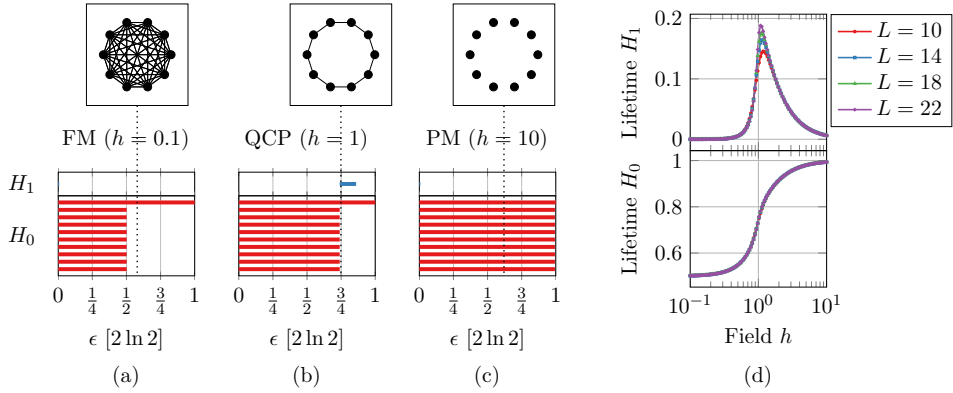


Figure 4.6: Persistence barcodes for the three phases in the transverse-field Ising model. (a) The quantum ferromagnet has all the spins equally entangled. (b) At the quantum critical point (QCP), neighboring spins have larger entanglement than next-nearest neighbors, and so on. This leads to a 1-homology element, capturing the hole in the center of the ring. (c) The quantum paramagnet does not have entangled spins, and therefore all the spins remain disconnected. (d) The lifetime (i.e. length) of the bars in  $H_0$  and  $H_1$  as a function of the transverse field  $h$ . This is Fig. 6 in Paper 1.

world in a video game, the quantum entanglement carries the information that leads to spacetime. This view is also consistent with the ER = EPR conjecture that states that two entangled particles (like the Bell state in Equation 4.26) are connected by Planck scale wormholes [51]. A connection between quantum entanglement and spacetime can also be referred to as quantum gravity. There is a large-scale effort to construct a theory of spacetime from quantum information, for example through the Simon's collaboration *It from Qubit* [40].

Paper 1 shows how persistent homology (PH) captures the geometry and topology of entanglement structures in quantum spin models. In this application of PH, the discrete data are the quantum subsystems and the distance metric is their degree of entanglement. We demonstrate this on quantum spin models, where the subsystems are the individual spins. The pairwise correlation is measured by the quantum mutual information (MI),

$$0 \leq M_{ij} = S_i + S_j - S_{ij} \leq 2 \ln 2. \quad (4.33)$$

This is large if two subsystems are entangled, whereas we want to bring subsystems that are strongly entangled close together. Therefore, the distance metric is the inverse MI,

$$2 \ln 2 \geq D_{ij} = 2 \ln 2 - M_{ij} \geq 0 \quad (4.34)$$

Another possible choice would be  $D_{ij} = -\ln(M_{ij}/2\ln 2)$ , which ranges from 0 to  $\infty$ . The result is that a quantum state is converted into a barcode (through the Vietoris-Rips complex). We implemented this by exact diagonalization (ED) of two different one-dimensional quantum spin systems. Figure 4.6 shows the result for the ground state of the transverse-field Ising model. The ferromagnetic ground state obeys the area law of entanglement. At the quantum critical point (QCP) at  $h = 1$ , the entanglement entropy diverges logarithmically with the subsystem size, i.e. volume law. The QCP is also captured by the formation of a hole in the entanglement structure (see Fig 4.6 (b) and (d)). The entanglement structure of this system is relatively simple due to the symmetries of the Hamiltonian, and therefore the higher homology groups ( $H_2$ ,  $H_3$ , and so forth) are not interesting.

In summary, the geometric and topological information of the quantum state is summarized in a persistence barcode. These barcodes can be used as an order parameter, since changes in the barcode indicate a phase transition. More fundamentally, since this approach constructs a geometrical object from entanglement, it is natural to ask whether a connection can be made to general relativity. In 1972, Roger Penrose has also stated: “*My own view is that ultimately physical laws should find their most natural expression in terms of essentially combinatorial principles, [...]. Thus, in accordance with such a view, should emerge some form of discrete or combinatorial spacetime.*” [11, 62]. This is not studied further in this compilation thesis, but it is the topic of recent work by Cao *et al.* [16]. It is still conjectural work that relies on a number of assumptions, but the prospect of connecting quantum entanglement and general relativity is exciting.





## Chapter 5

# Summary

This thesis provides an overview of how to study materials properties using materials informatics. Our modern society needs new materials across many sectors, not least considering the global energy budget. Discovering desired materials across the vast chemical space is an extremely difficult task. The data-driven approaches in this thesis aim to accelerate the search for new materials. Within this context, three lines of research are presented.

Materials databases are at the heart of materials informatics and enable the application of data mining and machine learning. We have developed the Organic Materials Database (OMDB), a database containing about 40 000 organic crystals and their electronic properties. This large dataset can be mined for functional properties. For example, we search for materials with a tiny gap (order of meV) for dark matter detection (see Paper 7). The effect of impurities on these gaps is studied in detail in Paper 5. Another example is the graphical pattern search that is able to find patterns in electronic band structures (see Paper 8).

Given the availability of materials databases, it is also possible to predict materials properties using machine learning. Conventional *ab initio* calculations scale up to  $10^3$  number of atoms, whereas machine learning models can typically scale to much larger compounds. This computational benefit opens the door towards large organic compounds as described in Paper 2. Using an atomistic machine learning model that predicts the electronic band gap given a crystal structure (see Paper 6), we have identified the first three dimensional organic semimetal (see Paper 4).

Modern data science methods also provide new ways to study classical and quantum materials directly. We have used persistent homology, a method that captures shapes in discrete data, to identify phase transitions. This was first demonstrated on a classical spin model with a complex phase diagram in Paper 3. Using a similar approach, but this time considering structures in quantum entanglement, we can detect quantum phase transitions (see Paper 1).

A common theme in this thesis is that problems in nature demand a certain symmetry. Efficient machine learning models that construct structure-property

maps should obey the symmetry relations of the problem (e.g. translational and rotational symmetry). In the case of quantum mechanical systems, the symmetry present in the quantum entanglement is used to look at a small corner of the Hilbert space that is sensible (e.g. states that obey the area law). This leads to the concept of studying the shape of entanglement structures to gain new insights. Data-driven approaches that capture the relevant symmetries of the problem at hand will continue to play an important role in materials informatics in the future.

# Bibliography

- [1] Yaser Abu-Mostafa. Lecture 15: Kernel methods. <http://work.caltech.edu/lectures.html>.
- [2] D. S. Akerib, P. B. Cushman, C. E. Dahl, R. Ebadi, A. Fan, R. J. Gaitskell, C. Galbiati, G. K. Giovanetti, Graciela B. Gelmini, L. Grandi, S. J. Haselschwardt, C. M. Jackson, R. F. Lang, B. Loer, D. Loomba, M. C. Marshall, A. F. Mills, C. A. J. OHare, C. Savarese, J. Schueler, M. Szydagis, Volodymyr Takhistov, Tim M. P. Tait, Y. D. Tsai, S. E. Vahsen, R. L. Walsworth, and S. Westerdale. Snowmass2021 cosmic frontier dark matter direct detection to the neutrino fog, 2022.
- [3] P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972.
- [4] ANNOY library. <https://github.com/spotify/annoy>. Accessed: 2017-08-01.
- [5] E. Aprile, J. Aalbers, F. Agostini, M. Alfonsi, L. Althueser, F.D. Amaro, V.C. Antochi, E. Angelino, J.R. Angevaare, F. Arneodo, D. Barge, L. Baudis, B. Bauermeister, L. Bellagamba, M.L. Benabderrahmane, T. Berger, A. Brown, E. Brown, S. Bruenner, G. Bruno, R. Budnik, C. Capelli, J.M.R. Cardoso, D. Cichon, B. Cimmino, M. Clark, D. Coderre, A.P. Colijn, J. Conrad, J.P. Cussonneau, M.P. Decowski, A. Depoian, P. Di Gangi, A. Di Giovanni, R. Di Stefano, S. Diglio, A. Elykov, G. Eurin, A.D. Ferella, W. Fulgione, P. Gaemers, R. Gaior, M. Galloway, F. Gao, L. Grandi, C. Hasterok, C. Hils, K. Hiraide, L. Hoetzsch, J. Howlett, M. Iacovacci, Y. Itow, F. Joerg, N. Kato, S. Kazama, M. Kobayashi, G. Koltman, A. Kopec, H. Landsman, R.F. Lang, L. Levinson, Q. Lin, S. Lindemann, M. Lindner, F. Lombardi, J. Long, J.A.M. Lopes, E. López Fune, C. Macolino, J. Mahlstedt, A. Mancuso, L. Marenti, A. Manfredini, F. Marignetti, T. Marrodán Undagoitia, K. Martens, J. Masbou, D. Masson, S. Mastroianni, M. Messina, K. Miuchi, K. Mizukoshi, A. Molinario, K. Morå, S. Moriyama, Y. Mosbacher, M. Murra, J. Naganoma, K. Ni, U. Oberlack, K. Odgers, J. Palacio, B. Pelssers, R. Peres, J. Pienaar, V. Pizzella, G. Plante, J. Qin, H. Qiu, D. Ramírez García, S. Reichard, A. Rochetti, N. Rupp, J.M.F. dos Santos, G. Sartorelli, N. Šarčević, M. Scheibelhut, J. Schreiner, D. Schulte, M. Schumann, L. Scotto Lavina, M. Selvi, F. Semeria,

- P. Shagin, E. Shockley, M. Silva, H. Simgen, A. Takeda, C. Therreau, D. Thers, F. Toschi, G. Trincherio, C. Tunnell, K. Valerius, M. Vargas, G. Volta, H. Wang, Y. Wei, C. Weinheimer, M. Weiss, D. Wenz, C. Wittweg, Z. Xu, M. Yamashita, J. Ye, G. Zavattini, Y. Zhang, T. Zhu, and J.P. Zopounidis. Projected WIMP sensitivity of the XENONnT dark matter experiment. *Journal of Cosmology and Astroparticle Physics*, 2020(11):031–031, nov 2020.
- [6] Toshihiro Ashino. Materials ontology: An infrastructure for exchanging materials information and knowledge. *Data Science Journal*, 9:54–61, 2010.
- [7] Albert P Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R Kermode, Gábor Csányi, and Michele Ceriotti. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.*, 3(12):e1701816, December 2017.
- [8] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87:184115, May 2013.
- [9] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics*, 134(7):074106, feb 2011.
- [10] Erik J. Bekkers, Maxime W. Lafarge, Mitko Veta, Koen A. J. Eppenhof, Josien P. W. Pluim, and Remco Duits. Roto-translation covariant convolutional networks for medical image analysis. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 440–448, Cham, 2018. Springer International Publishing.
- [11] Ginestra Bianconi. *Higher-Order Networks*. Elements in Structure and Dynamics of Complex Networks. Cambridge University Press, Cambridge, England, December 2021.
- [12] Jean-Daniel Boissonnat and Clément Maria. Computing persistent homology with various coefficient fields in a single pass. *Journal of Applied and Computational Topology*, 3(1-2):59–84, apr 2019.
- [13] Stanislav S. Borysov, R. Matthias Geilhufe, and Alexander V. Balatsky. Organic materials database: An open-access online database for data mining. *PLOS ONE*, 12(2):e0171501, feb 2017.
- [14] Peter Bubenik, Vin de Silva, and Jonathan Scott. Metrics for generalized persistence modules. *Found. Comput. Math.*, 15(6):1501–1531, December 2015.
- [15] Peter Bubenik, Michael Hull, Dhruv Patel, and Benjamin Whittle. Persistent homology detects curvature. *Inverse Problems*, 36(2):025008, jan 2020.

- [16] ChunJun Cao, Sean M. Carroll, and Spyridon Michalakis. Space from hilbert space: Recovering geometry from bulk entanglement. *Phys. Rev. D*, 95:024031, Jan 2017.
- [17] Gunnar Carlsson, Afra Zomorodian, Anne Collins, and Leonidas Guibas. Persistence Barcodes for Shapes. In Roberto Scopigno and Denis Zorin, editors, *Symposium on Geometry Processing*. The Eurographics Association, 2004.
- [18] Citrine Informatics. <https://citrine.io>. Accessed: 2022-11-14.
- [19] Taco Cohen and Max Welling. Group equivariant convolutional networks. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999, New York, New York, USA, 20–22 Jun 2016. PMLR.
- [20] Taco S. Cohen and Max Welling. Steerable CNNs. In *International Conference on Learning Representations*, 2017.
- [21] Alex Cole, Gregory J. Loges, and Gary Shiu. Quantitative and interpretable order parameters for phase transitions from persistent homology. *Physical Review B*, 104(10), September 2021.
- [22] Stefano Curtarolo, Wahyu Setyawan, Gus L.W. Hart, Michal Jahnatek, Roman V. Chepulskii, Richard H. Taylor, Shidong Wang, Junkai Xue, Kesong Yang, Ohad Levy, Michael J. Mehl, Harold T. Stokes, Denis O. Demchenko, and Dane Morgan. Aflow: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226, 2012.
- [23] Irene Donato, Matteo Gori, Marco Pettini, Giovanni Petri, Sarah De Nigris, Roberto Franzosi, and Francesco Vaccarino. Persistent homology analysis of phase transitions. *Physical Review E*, 93(5):052138, may 2016.
- [24] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 454–463, 2000.
- [25] Herbert Edelsbrunner and John Harer. *Computational Topology - an Introduction*. American Mathematical Society, 2010.
- [26] Felix Faber, Alexander Lindmaa, O. Anatole von Lilienfeld, and Rickard Armiento. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry*, 115(16):1094–1101, apr 2015.
- [27] Marc Finzi, Max Welling, and Andrew Gordon Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th*

- International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3318–3328. PMLR, 18–24 Jul 2021.
- [28] Carlos Fiolhais, Fernando Nogueira, and Miguel A.L. Marques. *A Primer in Density Functional Theory*. Springer, 06 2003.
  - [29] Stefan Funk. Indirect detection of dark matter with  $\gamma$  rays. *Proceedings of the National Academy of Sciences*, 112(40):12264–12271, 2015.
  - [30] R. Matthias Geilhufe, Felix Kahlhoefer, and Martin Wolfgang Winkler. Dirac materials for sub-mev dark matter detection: New targets and improved formalism. *Phys. Rev. D*, 101:055005, Mar 2020.
  - [31] 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: meta-data and data formats. *Npj Comput. Mater.*, 3(1), December 2017.
  - [32] Steven M. Girvin and Kun Yang. *Modern Condensed Matter Physics*. Cambridge University Press, feb 2019.
  - [33] Saulius Gražulis, Daniel Chateigner, Robert T. Downs, A. F. T. Yokochi, Miguel Quirós, Luca Lutterotti, Elena Manakova, Justas Butkus, Peter Moeck, and Armel Le Bail. Crystallography Open Database – an open-access collection of crystal structures. *Journal of Applied Crystallography*, 42(4):726–729, Aug 2009.
  - [34] Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O. Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The Journal of Physical Chemistry Letters*, 6(12):2326–2331, 06 2015.
  - [35] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, Cambridge, England, December 2001.
  - [36] Yu He, Shiqi Xia, Dimitris G. Angelakis, Daohong Song, Zhigang Chen, and Daniel Leykam. Persistent homology analysis of a generalized aubry-andré-harper model. *Phys. Rev. B*, 106:054210, Aug 2022.
  - [37] Yonit Hochberg, Yonatan Kahn, Mariangela Lisanti, Kathryn M. Zurek, Adolfo G. Grushin, Roni Ilan, Sinéad M. Griffin, Zhen-Fei Liu, Sophie F. Weber, and Jeffrey B. Neaton. Detection of sub-mev dark matter with three-dimensional dirac materials. *Phys. Rev. D*, 97:015004, Jan 2018.
  - [38] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

- [39] Haoyan Huo and Matthias Rupp. Unified representation of molecules and crystals for machine learning. *Machine Learning: Science and Technology*, 3(4):045017, nov 2022.
- [40] It from qubit: Simons collaboration on quantum fields, gravity and information. <https://www.simonsfoundation.org/mathematics-physical-sciences/it-from-qubit/>. Accessed: 2021-06-21.
- [41] 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. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.
- [42] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Židek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislaw Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.
- [43] Chiho Kim, Anand Chandrasekaran, Tran Doan Huan, Deya Das, and Rampi Ramprasad. Polymer genome: A data-powered polymer informatics platform for property predictions. *J. Phys. Chem. C Nanomater. Interfaces*, 122(31):17575–17585, August 2018.
- [44] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.
- [45] Georg Kresse and Jürgen Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16):11169, 1996.
- [46] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551, dec 1989.
- [47] Yongjin Lee, Senja D. Barthel, Paweł Dłotko, S. Mohamad Moosavi, Kathryn Hess, and Berend Smit. Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications*, 8(1), may 2017.
- [48] Sheng-Kai Liao, Wen-Qi Cai, Johannes Handsteiner, Bo Liu, Juan Yin, Liang Zhang, Dominik Rauch, Matthias Fink, Ji-Gang Ren, Wei-Yue Liu, Yang Li,



- Qi Shen, Yuan Cao, Feng-Zhi Li, Jian-Feng Wang, Yong-Mei Huang, Lei Deng, Tao Xi, Lu Ma, Tai Hu, Li Li, Nai-Le Liu, Franz Koidl, Peiyuan Wang, Yu-Ao Chen, Xiang-Bin Wang, Michael Steindorfer, Georg Kirchner, Chao-Yang Lu, Rong Shu, Rupert Ursin, Thomas Scheidl, Cheng-Zhi Peng, Jian-Yu Wang, Anton Zeilinger, and Jian-Wei Pan. Satellite-relayed intercontinental quantum network. *Phys. Rev. Lett.*, 120:030501, Jan 2018.
- [49] I M Lifshitz. Energy spectrum structure and quantum states of disordered condensed systems. *Soviet Physics Uspekhi*, 7(4):549, 1965.
- [50] Kai Liu, Bang Ouyang, Xiaojun Guo, Yunlong Guo, and Yunqi Liu. Advances in flexible organic field-effect transistors and their applications for flexible electronics. *npj flex. electron.*, 6(1), January 2022.
- [51] J. Maldacena and L. Susskind. Cool horizons for entangled black holes. *Fortschritte der Physik*, 61(9):781–811, 2013.
- [52] Mat3ra. <https://mat3ra.com>. Accessed: 2022-11-14.
- [53] Materials Genome Initiative. <https://www.mgi.gov/>. Accessed: 2022-11-11.
- [54] L. McInnes and J. Healy. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.
- [55] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv e-prints*, page arXiv:1301.3781, January 2013.
- [56] Thomas Mitchell. *Machine Learning*. McGraw-Hill series in computer science. McGraw-Hill Professional, New York, NY, March 1997.
- [57] Kevin P. Murphy. *Machine Learning, A Probabilistic Perspective*. MIT Press, 2012.
- [58] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.
- [59] NIMS Materials Database (MatNavi). <https://mits.nims.go.jp/en/>. Accessed: 2022-11-12.
- [60] The NOMAD (Novel Materials Discovery) Center of Excellence (CoE). <https://nomad-coe.eu>. Accessed: 2022-11-11.
- [61] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1), August 2017.
- [62] Roger Penrose. On the nature of quantum geometry. *Magic Without Magic: John Archbald Wheeler*, 1972. p. 333–354.

- [63] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.
- [64] John P. Perdew and Karla Schmidt. Jacob’s ladder of density functional approximations for the exchange–correlation energy. *AIP Conference Proceedings*, 577(1):1–20, 2001.
- [65] Datasets at quantum-machine.org. <http://quantum-machine.org/datasets/>. Accessed: 2022-11-11.
- [66] Mark Van Raamsdonk. Building up spacetime with quantum entanglement. *General Relativity and Gravitation*, 42(10):2323–2329, June 2010.
- [67] Mark Van Raamsdonk. Spacetime from bits. *Science*, 370(6513):198–202, 2020.
- [68] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.
- [69] F. Rosenblatt. The perceptron - a perceiving and recognizing automaton. Technical Report 85-460-1, Cornell Aeronautical Laboratory, Ithaca, New York, January 1957.
- [70] Sven Rühle. Tabulated values of the shockley–queisser limit for single junction solar cells. *Solar Energy*, 130:139–147, 2016.
- [71] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, oct 1986.
- [72] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108:058301, Jan 2012.
- [73] S. A. Barannikov. The Framed Morse complex and its invariants. In *Singularities and Bifurcations*, pages 93–116. Advances in Soviet Mathematics, American Mathematical Society, December 1994.
- [74] 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 (1989)*, 65(11):1501–1509, November 2013.
- [75] Nicholas Sale, Jeffrey Giansiracusa, and Biagio Lucini. Quantitative analysis of phase transitions in two-dimensional xy models using persistent homology. *Physical Review E*, 105(2), feb 2022.
- [76] E. Schrödinger. Discussion of probability relations between separated systems. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):555–563, oct 1935.

- [77] Marc Schumann. Direct detection of WIMP dark matter: concepts and status. *Journal of Physics G: Nuclear and Particle Physics*, 46(10):103003, aug 2019.
- [78] K. T. Schütt, P.-J. Kindermans, H. E. Saucedo, S. Chmiela, A. Tkatchenko, and K.-R. Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, page 992–1002, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [79] K T Schütt, H E Saucedo, P-J Kindermans, A Tkatchenko, and K-R Müller. SchNet - a deep learning architecture for molecules and materials. *J. Chem. Phys.*, 148(24):241722, June 2018.
- [80] Dan Sehayek and Roger G. Melko. Persistent homology of  $F_2$  gauge theories. *Phys. Rev. B*, 106:085111, Aug 2022.
- [81] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W R Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, January 2020.
- [82] Jonathon Shlens. A tutorial on principal component analysis, 2014.
- [83] T Sousbie. The persistent cosmic web and its filamentary structure - i. theory and implementation. *Mon. Not. R. Astron. Soc.*, 414(1):350–383, June 2011.
- [84] Daniel Spitz, Jürgen Berges, Markus Oberthaler, and Anna Wienhard. Finding self-similar behavior in quantum many-body dynamics via persistent homology. *SciPost Phys.*, 11:060, 2021.
- [85] Mathieu Taillefumier, Owen Benton, Han Yan, L. D. C. Jaubert, and Nic Shannon. Competing spin liquids and hidden spin-nematic order in spin ice with frustrated transverse exchange. *Phys. Rev. X*, 7:041057, Dec 2017.
- [86] Quoc Hoan Tran, Mark Chen, and Yoshihiko Hasegawa. Topological persistence machine of phase transitions. *Physical Review E*, 103(5), May 2021.
- [87] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.
- [88] Cédric Villani. *Optimal Transport*. Springer Berlin Heidelberg, 2009.
- [89] Klaus Volkert. Poincaré’s homology sphere. *Bulletin of the Manifold Atlas*, 2013.
- [90] T.O. Wehling, A.M. Black-Schaffer, and A.V. Balatsky. Dirac materials. *Advances in Physics*, 63(1):1–76, jan 2014.

- [91] Maurice Weiler and Gabriele Cesa. General  $E(2)$ -Equivariant Steerable CNNs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.
- [92] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, nov 2004.



# Index

- Activation function, 25
- Backpropagation, 25
- Barcode, 39, 40
- Betti number, 6, 37
- Bias-variance trade-off, 21
- Boundary group, 35
- Boundary operator, 35
- Chain group, 34
- Convolutional neural network, 25
- Cycle group, 35
- Data augmentation, 26
- Density-functional theory, 16
- Descriptor, 20
- Dirac material, 7
- Embedding, 27
- Entanglement entropy, 43
- Equivariance, 26
- Exact diagonalization, 45
- Exploration-exploitation trade-off, 20
- FAIR, 4
- Feature vector, 20
- Filtration, 39
- Fundamental Lemma of Homology, 35
- Group representation, 28
- Homeomorphism, 5
- Homology group, 36
- Homotopy equivalence, 5
- Hyperparameter, 21
- Jacob's ladder, 17
- Kernel trick, 21
- Nerve theorem, 40
- OMDB, 11
- Quantum entanglement, 42
- Quantum mutual information, 44
- Qubit, 42
- Regularization, 21
- Reinforcement learning, 20
- Simplex, 33
- SMILES, 20
- Supervised learning, 19
- t-SNE, 30
- Topological data analysis, 33
- Torsion, 37
- Unsupervised learning, 19
- Variational principle, 15
- Weight sharing, 25



**Part II**

**Included papers**





# Errata & corrigenda

Below are the known errata and typos in the publications.

- Paper 3

1. Equation 4, the distance metric for the spins  $i$  and  $j$  should be:

$$\mathcal{D}(i, j) = \left\| \mathbf{r}_i - \mathbf{r}_j + \frac{a}{2\sqrt{2}} (\mathbf{S}_i - \mathbf{S}_j) \right\|$$

It was however included correctly in the figure, text and open-source Python code.

- Paper 5

1. Typesetting problem in the caption of Fig. 5,

$$\langle R(\omega_{\text{th}}) \rangle \langle R(\omega_{\text{DM}}) \rangle$$

should be

$$\langle R(\omega_{\text{th}}) \rangle > \langle R(\omega_{\text{DM}}) \rangle$$

The same problem occurs in the main text. The version on the arXiv is typeset correctly.

