Predicting solid-state qubit material host

by

Oliver Lerstøl Hebnes

Thesis

for the degree of

Master of Science

Faculty of Mathematics and Natural Sciences
University of Oslo

April 22, 2021

Contents

Part I

Introduction

Chapter 1

Introduction

Discovery of current deep defects and material hots by serendipitet. (ref. data-mining artikkl)

This it The introduction. Another coffee.

Trengs det egentlig undertitler her? Tenketenketenk

1.1 Motivation

1.2 Holy grail

1.3 Structure of thesis

Part II

Theory

Chapter 2

Quantum technologies

This chapter will provide a brief overview of the current state-of-the-art in quantum technological advances. This will not only give us insights in how the technology is being used today, but also grant us the opportunity to discuss key concepts that are fundamental to understand for this thesis. Thereafter we will look into how materials are built up, and what kind of properties a material needs to exhibit to be an eligible host for quantum devices. Finally, we will giving a few specific examples of materials with promising point defects that have been comprehensively researched. Importantly, this will motivate the reasoning for finding new materials that might excel in areas where other materials falls short for utilization in quantum technology.

Quantum technology (QT) refers to practical applications and devices that utilize the principles of quantum physics as a foundation. Technologies in this spectrum are based on concepts such as *superposition*, *entanglement* and *coherence*, which are all closely related to one another.

A quantum superposition refers to that any two or more quantum eigenstates can be added together into another valid quantum state, such that every quantum state can be represented as a sum, or a superposition, of two or more distinct states. This is according to the wave-particle duality which states that every particle or another quantum entity may be described as either a particle or a wave. When measuring the state of a system residing in a superposition of eigenstates, however, the system falls back to one of the basis states that formed the superposition, destroying the original configuration.

Quantum entanglement refers to when a two- or many-particle state cannot be expressed independently of the state of the other particles, even when the particles are separated by a significant distance. As a result, the many-particle state is termed an entangled state [1].

Quantum coherence arises if two waves coherently interfere with each other and generate a superposition of the two states with a phase relation. Likewise, loss of coherence is known as *decoherence*.

Another concept that the reader should be familiar with is the famous Heisenberg uncertainty principle. It states that

$$\sigma_x \sigma_p \leqslant \frac{\hbar}{2}, \tag{2.1}$$

where σ_x is the standard deviation for the position and σ_p is the standard deviation in momentum. This means that we cannot accurately predict both the position and momentum of a particle at the same time. Thus, we often calculate the probability for a particle to be in a state which results in concepts such as an electron sky surrounding an atom core. However, remember that equation (2.1) is an inequality, which means that it is possible to create a state where neither the position nor the momentum is well defined.

2.1 Quantum computing

The start of the digital world's computational powers can be credited to Alan Turing. In 1937, Turing [2] published a paper where he described the *Turing machine*, which is regarded as the foundation of computation and computer science. It states that only the simplest form of calculus, such as boolean Algebra (1 for true and 0 for false), is actually computable. This required developing hardware that could handle classical logic operations, and was the basis of transistors that are either in the state ON or OFF depending on the electrical signal. Equipped with a circuit consisting of wires and transistors, commonly known as a computer, we could develop software to solve all kinds of possible applications.

Driven by the development of software, conventional computers have in accordance to Moore's law [3], doubled the amount of transistors on integrated circuit chips every two years as a result of smaller transistors. Furthermore, the clock frequency has enhanced with time, resulting in a doubling of computer performance every 18 months [4]. Alas, miniaturization cannot go on forever as transistors are mass-produced at 5 nm today and are expected to reach a critical limit of 3 nm in the following years [5].

To sustain the digital world's increasing computational demand, other alternatives than the conventional classical computer must be explored. This is where quantum computing comes into the picture. The term quantum computer is a device that exploits quantum properties to solve certain computational problems more efficiently than allowed by Boolean logic [6].

The idea is to pass information in the form of a quantum bit, or *qubit* for short. They are the building blocks of quantum computers, and as opposed to the conventional 0 or 1-bits that classical computers are based on, they can inhabit any superposition of the states 0 or 1. This is illustrated in figure 2.1.

The architecture of a gate-based quantum computer is dependent on a set of quantum logic gates that perform unitary transformations on sets of qubits [7, 8]. Other implementations of quantum computers exists, such as the adiabatic quantum computer. This approach is not based on gates, but on defining the answer of a problem as the ground state of a complex network of interactions between qubits, and then controlling the interactions to adiabatically evolve the system to the ground state [9].

It has been demonstrated that exponentially complex problems can be reduced to polynomially complex problems for quantum computers [4]. For example, a quantum search algorithm found by Grover [10] offers a quadratic speed-up compared to classical algorithms, while Shor's quantum integer factorization algorithm [11] presents an exponential speed-up. Intriguingly, Google reported in 2019 that they ran a random number generator algorithm on a superconducting processor containing 53 qubits in 200 seconds, which would most likely take several times longer for a classical supercomputer to solve [12]. It is anticipated that quantum computers will excel in exceedingly complex problems, while many simpler tasks may not see any speed-up at all compared to the classical regime. Hence, quantum- and classical computers are envisioned to coexist for each their purpose.

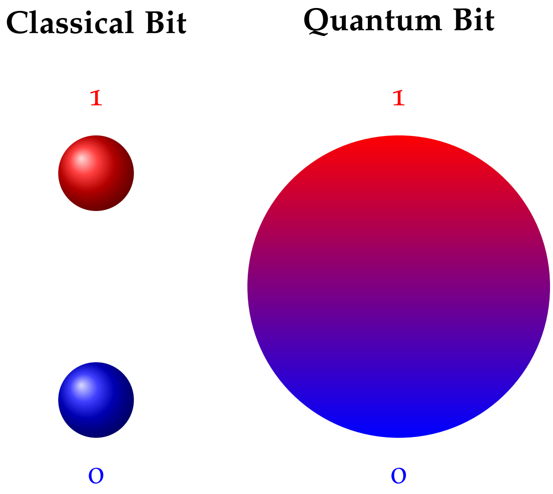


Figure 2.1: Conceptual illustration of the two-level classical bit, which are restricted to the boolean states 1 (true) or 0 (false), and the quantum bit that can be in any superposition of the states 0 or 1.

Quantum computing is a highly sought-after goal, but there are extensive challenges that needs to be adressed. Controlling a complex many-qubit system is difficult, since it is not always possible to establish interactions between qubits [7] and maintain entaglement over both time and distance. Additionally, decoherence and other quantum noise occurs as a result of the high volatility of quantum states, making quantum state manipulation prone to errors. The *quantum error correction* protocols and the theory of *threshold theorem* deals with this vulnerability, stating that noise most likely does not pose any fundamental barrier to the performance of large-scale computations [4].

2.1.1 Quantum computing requirements

As ever-promising the concepts of quantum technology are, the physical realizations are in the preliminary stage of development. Here we will concretize critical principles for a physical realisation of a quantum platform.

> "I always said that in some sense, these criteria are exactly the ones that you would teach to kindergarten children about computers, quantum or otherwise" DiVincenzo [13]

DiVincenzo formulated in the year of 2000 seven basic criteria for a physical qubit system with a logic-based architecture [7].

1. A scalable physical system with well characterized qubits
2. The ability to initialize the state of the qubits to a simple initial system
3. Have coherence times that are much longer than the gate operation time
4. Have a universal set of quantum gates
5. Have the ability to perform qubit-specific measurements

These five criteria must be met for a quantum platform to be considered a quantum computer.

2.2 Quantum communication

Quantum communication refers to the transfer of a state of one quantum system to another. Since information can be stored in qubits, we picture *flying qubits* that transfer information from one location to another [14]. The benefits of using flying qubits are in particular valued in quantum cryptography, since the quantum nature of qubits can be exploited to add extra layers of security [4].

Consider the example of encrypting a digitally transmitted conversation. It is difficult to avoid someone eavesdropping on a conversation, however, the problem is diminished if the eavesdropper does not speak the language, keeping the information in the conversation safe. This is the original idea of encryption, such that the information has been encrypted into something incomprehensible for any eavesdropper. A common practice is to encrypt information and share a public key, which everyone can read, and a private key, only known for the sender and receiver of information. This should be sufficient to keep the information secure, given that the complexity of the private key is impenetrable.

Importantly, we live in a digital world where most of our actions are increasingly being stored as information, and we could imagine that the eavesdropper in the latter example stored the conversation. Even if the content of the conversation was encrypted, it still presents a challenge, since encrypted information stored today could be deciphered in ten or twenty years' time . Consequently, finding an encryption method that could make information either impossible to eavesdrop on or make the security unbreakable forever is very desirable. This is the ultimate goal of quantum cryptography [4].

Consider the example of information encoded into a qubit as a superposition of two quantum states. Now, if a wild eavesdropper would try to measure the information, the nature of quantum physics tells us that the original configuration would be destroyed and the receiver would be alerted of the eavesdropper. Furthermore, if the eavesdropper would try to make a copy of the message, the copying itself would be limited of the no-cloning theorem [15] which declare that quantum states cannot be copied.

A clever approach to ensure confidentiality is to send the encryption key before sending the actual encrypted information. If the key is received unperturbed, the key remains secret and can be safely employed. If it turns out perturbed, confidentiality is still intact since the key does not contain any information and can be discarded. This approach is termed the *quantum key distribution* (QKD) [15, 16]. It should be noted that this requires both the sender and receiver to have access to methods for sending, receiving and storing qubit states, such as a quantum computer. Additionally, the sender and receiver will need to initially exchange a common secret which is later expanded, making quantum key *expansion* a more exact term for QKD [4, 16].

Most applications and experiments use optical fibers for sending information via photons, with the distance regarded as the main limitation. This is because classical repeaters are unable to enhance quantum information because of the no-cloning theorem, making photon loss in optical fiber cables inevitable. Thus, quantum communication must reinvent the repeater concept, using hardware that preserves the quantum nature [17] and are compatible with wavelengths used in telecommunication. Nonetheless, secure QKD up to 400 km has recently been demonstrated using optical fibres in academic prototypes [18].

2.3 Quantum sensing

Measurements are part of our digital world today to a great extent. There would be no way to exchange goods, services or information without reliable and precise measurements [17]. Thus, improving the accuracy of sensors for every measurement done is desirable. One method to improve measurement accuracy, resolution and sensitivity can be by utilizing quantum sensors.

Quantum sensors exploit quantum properties to measure a physical quantity [19]. This is possible because quantum systems are highly susceptible to pertubations to its surroundings, and can be used to detect physical properties such as either temperature or an electrical or magnetic field [19].

For a quantum system to be able to function as a quantum sensors, a few criterias needs to be met. Firstly, the quantum system needs to have discrete and resolvable energy levels. The quantum system also needs to be controllably initialised into a state that can be identified and coherently manipulated by time-dependent fields. Lastly, the quantum system needs to be able to interact with the physical property one wants to measure through a coupling parameter [19].

It is also possible to also exploit quantum entanglement to improve the precision of a measurement. This gain of precision is used to reach what is called the Heisenberg-limit, which states that the precision scales as the number of particles N in an idealized quantum system [17, 19], while the best classical sensors scale with $\sqrt{N}$.

2.4 Available quantum platforms

Many different quantum platforms have been physically implemented, and this section will serve as a brief overview of the current status. For a more thorough review of qubit implementations, the reader is directed to Refs. [8, 17].

Superconducting circuits can be used in quantum computing, since electrons in superconducting materials can form Cooper pairs via an effective electron-electron attraction when the temperature is lower than a critical limit. Below the limit, electrons can move without resistance in the material [20]. Exploiting this intrinsic coherence, qubits can be made by forming microwave circuits based on loops of two superconducting elements separated by an insulator, also known as Josephson tunnel junctions [17]. Today, superconducting Josephson junctions are the most widely used quantum platform, but they requires very low temperature (mK) to function, making them costly to use. Additionally, the current devices experience a relatively short coherence time, causing challenges in scaling up.

Single photons is an eligible quantum platform that can be implemented as qubits with one-qubit gates being formed by rotations of the photon polarization. Its use in fiber optics are less prone to decoherence, but faces challenges since the more complex photon-photon entanglement and control of multi-qubits is strenuous [8].

By fixing the nuclear spin of solid-state systems, it is possible to implement a quantum platform that experience long spin coherence. This enables the manipulation of qubits that utilize electromagnetic fields, making one-qubit gates realizable.

The isolated atom platform is characterized by its well-defined atom isolation. Here, every qubit is based on energy levels of a trapped ion or atom. Quantum entanglement can be achieved through laser-induced spin coupling, however scaling up to large atom numbers induce problems in controlling large systems and cooling of the trapped atoms or ions.

A quantum dot (QD) can be imagined as an artifical atom which is confined in a solid-state host. As an example, a quantum dot can occur when a hole or an electron is trapped in the localized potential of a semiconductor's nanostructure. QDs exhibit similar coherence potential as the isolated atom platform, but without the drawback of confining and cooling of the given atom or ion [17]. Moreover, it is possible to limit decoherence due to nuclear spins by dynamic decoupling of nuclear spin noise and isotope purification [8].

A QD can normally be defined litographically using metallic gates, or as self-assembled QDs where a growth process creates the potential that traps electrons or holes. The difference between them is a question of controllability and temperature, since the metallic gates is primarily controlled electrically and operate at < 1 K, while self-assembly QDs are primarily controlled optically at ~ 4 K [8]. Despite requiring very low temperatures, QDs have the potential for fast voltage control and opticial initialization. As with trapped ions, electrostatically defined quantum dots experience a short-range exchange interaction, imposing a limitation for quantum computing and quantum error correction protocols. A potential solution could include photonic connections between quantum dots. On the contrary, self-assembled quantum dots couple strongly to photons due to their large size in comparison to single atoms. However, the size and shapes of self-assembled quantum dots are decided randomly during the growth process, causing an unfavourable large range of optical absorption and emission energies [8].

Lastly, we will turn towards point defects in bulk semiconductors as a physical implementation of a quantum platform. Point defects shares many of the attributes of quantum dots, such as discrete optical transitions and controllable coherent spin states, but are vulnerable to small changes in the lattice of the semiconductor. Thus, it can be difficult to isolate a point defect from the surrounding environment. However, one can utilize the strength of the solid-state semiconductor host to isolate to some extent the point defect, yielding extended coherence times and greater optical homogeneity than other quantum dot systems. Before we dwell into the intricacies of point defect qubits as a building block for QT, we will provide the neccessary background for the crystal- and electronic structure of semiconductors.

2.5 Introduction to semiconductor physics

The interactions between atoms and characteristics of matter form the foundation of materials science. The applications of materials science are extensive, with examples such as a bottle of water or to a chair to sit in.

Solid materials, like plastic bottles, are formed by densely packed atoms. These atoms can randomly occur through the material without any long-range order, which would categorize the material as an *amorphous solid*. Amorphous solids are frequently used in gels, glass and polymers [21]. However, the atoms can also be periodically ordered in small regions of the material, classifying the material as a *polycrystalline solid*. All ceramics are polycrystalline with a broad specter of applications ranging from kitchen-porcelain to orthopedical bio-implants [22]. A third option is to have these atoms arranged with infinite periodicity, making the material a *crystalline solid* or more commonly named a *crystal*. The three options are visualised in figure 2.2. Hereon, we will focus on crystalline solids.

The periodicity in a crystal is defined in terms of a symmetric array of points in space called the *lattice*, which can be simplified as either a one-dimensional array, a two-dimensional matrix or a three dimensional vector space, depending on the material. At each lattice point we can add an atom to make an arrangement called a *basis*. The basis can be one atom or a cluster of atoms having the same spatial arrangement. Every crystal has periodically repeated building blocks called *cells* representing the entire crystal. The smallest cell possible is called a *primitive cell*, but such a cell only allows lattice points at its corners and it is often quite rigid to work with when the structure becomes complex. As a solution, we will consider the *unit cell*, which allows lattice points on face centers and body centers.

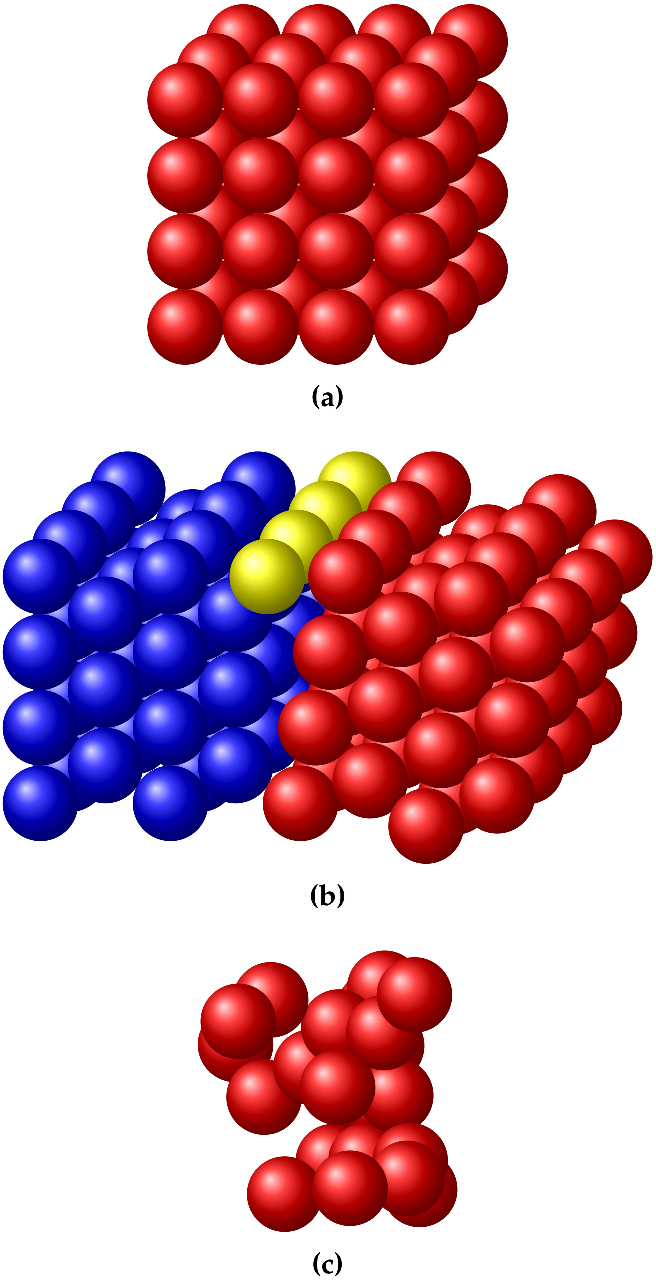


Figure 2.2: Schematic representation of different degrees of ordered structures, where (a) is a crystalline of a simple cubic lattice, (b) is a polycrystalline hexagonal lattice, and (c) is an amorphous lattice.

One example of a crystal structure is the perovskite structure. Compounds with this structure are characterized by having an ABX_3 stoichiometry whose symmetri belong to one of 15 space groups identified by Lufaso & Woodward [23], such as the cubic, orthorombic and tetragonal. For our purpose, we will be looking into when the X atom is oxygen, and refer to the oxygen-perovskite ABO_3. The A atom is nine- to 12-fold coordinated by oxygen, while the B atom is sixfold coordinated by oxygen, and the BO_6 octahedra are connected to the corners in all three directions as visualized in figure 2.3.

The motivation behind the research on perovskites is related to the large amount of available ABO_3 chemistries, where a significant portion of these take the perovskite structure. Perovskites have a broad specter of applications, ranging from high-temperature superconductors [24] and ionic conductors [25] to multiferroic materials [26]. Additionally, adding a perovskite-type compound to solar cells has reportedly resulted in higher performance efficiencies while being cheap to produce and simple to manufacture [27, 28]. However, this includes the use of hybrid organic-inorganic compounds and excludes the use of oxygen.

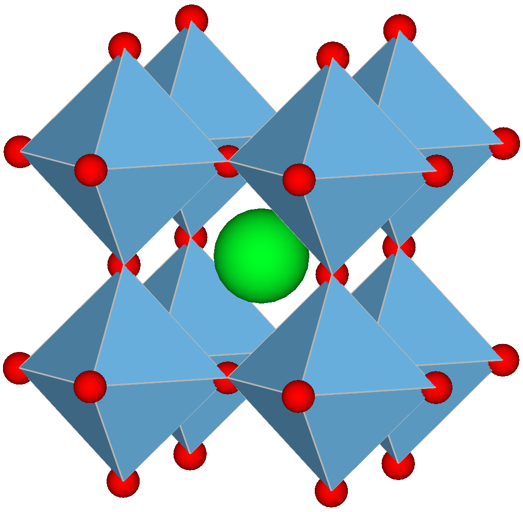


Figure 2.3: A crystal structure of $SrTiO_3$ which is a cubic perovskite. The red atoms are oxygen, whereas the green atom is strontium, and inside every corner-sharing BO_6 octahedral unit is a titanium atom.

Isolated atoms have distinct energy levels, where the Pauli exlusion principle [29] states for fermions that each energy level can at most accomodate two electrons of opposite spin. In a solid, the discrete energy levels of the isolated atom spread into continuous energy bands since the wavefunctions of the electrons in the neighboring atoms overlap. Hence, an electron is not neccessarily localized at a particular atom anymore. This is examplified as every material has a unique band structure, similar to every human having their unique fingerprint.

Knowing which energy bands are occupied by electrons is the key in understanding the electrical properties of solids. The highest occupied electron band at 0 K is called the valence band (VB), while the lowest unoccupied electron band is called the conduction band (CB). The energy gap of forbidden energy levels between the maximum VB and the minimum CB is known as the band gap, and its energy is denoted as E_g. If a material can be classified as a semiconductor depends on the band gap and the electrical conductivity. As an example, Silicon is commonly thought of as a semiconductor, and has a band gap of about 1.12 eV at 275 K [30].

To be able to accelerate electrons in a solid using an electrical field, they must be able to move into new energy states. At 0 K, the entire valence band of a semiconductor is full with electrons and there are no available states nearby, making it impossible for current to flow through the material. This can be solved by using either thermal or optical energy to excite electrons from the valence band to the conduction band, in order to *conduct* electricity. At room temperature, some semiconductors will have electrons excited to the conduction band solely from thermal energy matching the energy band gap [21].

In some scenarios, thermal or optical energy is not sufficient for an excitation since the energy bands are also dependent on the crystal momentum. A difference in the momentum of the minimal-energy state in the conduction band and the maximum-energy state in the valence band results in an *indirect bandgap* as seen in figure 2.4a. If there is no difference at all, the material has a *direct bandgap*, which is visualized in figure 2.4b.

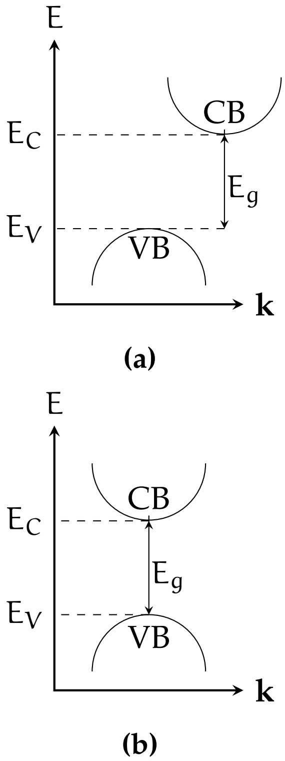


Figure 2.4: A schematic drawing of (a) an indirect- and (b) a direct bandgap.

Electrons in semiconductor materials can be described according to the Fermi-Dirac distribution

$$f(E) = \frac{1}{1 + e^{(E-E_F)/kT}},$$

where k is Boltzmann's constant, T is temperature, E is the energy and E_F is the Fermi level. The Fermi-Dirac distribution gives the probability that a state will be occupied by an electron, and at $T = 0$ K, every energy state lower than E_F is occupied by electrons while the opposite is true for energy states above E_F [21].

2.5.1 Point defects in semiconductors

In real life, a perfect crystal without any symmetry-breaking flaw does not exist. These flaws are known as defects and can occur up to three dimensions. An example one-dimensional defect is known as a *line defect*, while two dimensional defects can be *planar defects*, and in three dimensions we have *volume defects*. Lastly, defects can also occur in zero dimensions and are then termed *point defects*. Point defects normally occur as either vacancies, interstitial placement inbetween lattice sites or as substitution of another existing atom in the lattice.

Defects can greatly influence both the electronic and optical propertires of a material. A substitional defect can at first be regarded as an impurity or an antisite, but they can also be intentionally inserted, an approach known as *doping*. Doping can result in an excess of electrons or holes, making the semiconductor either an n- or p-type, respectively. Consequently, the semiconductor will have energy levels in the (forbidden) band gap that originates from the defects. If the energy levels introduced are closer than ~ 0.2 eV to the band egdes, they are termed *shallow* defects.

Shallow defects can contribute with either excess electrons to the conduction band, or excess holes to the valence band. However, the induced charge carriers (electrons or holes) interact strongly with the band egdes, resulting in a delocalized wavefunction regarding the position in the lattice.

For the opposite case, if the energy levels rests closer to the middle of the semiconductor's gap, the introduced defects are known as *deep level* defects. Deep levels normally occur due to either dangling bonds or impurities, and have highly localized electron wavefunctions. This might assure the isolation required for long coherence times, which is an appealing promise in quantum technological advances.

Deep levels can be unfortunate in semiconductors since they can interact with the charge carriers, potentially destroying the desired electronic or optical property of the material. Deep level defects can function as electron-hole recombination centers, or to trap charge carriers, yielding the commonly used name deep level *traps*. Both of the given situations results in a lower concentration of charge carriers, which showcase why deep levels can be unwanted in semiconductor devices. However, deep level defects show extraordinary properties in quantum technology due to their ability to ensure isolation and preserve coherence.

2.5.2 Optical defect transitions

Optical transitions refers to excitation of charge carriers due to either emission or absorption of electromagnetic radiation, and can be done with a laser light or electron beam. Figure 2.5 represents a configuration coordinate (CC) diagram of a defect transition. The y-axis is a function of the energy E, while the x-axis is a function of the configuration coordination Q. The lowest point in the lower parabola is known as the ground state (GS) configuration Q_{GS}, which is the most stable atomic position, while for the upper parabola it is known as the excited state configuration Q_{ES}. The dotted lines represent vibronic excitations to the energy of the ground state Q_{GS} for the lower parabola, while it represents Q_{ES} for the higher parabola.

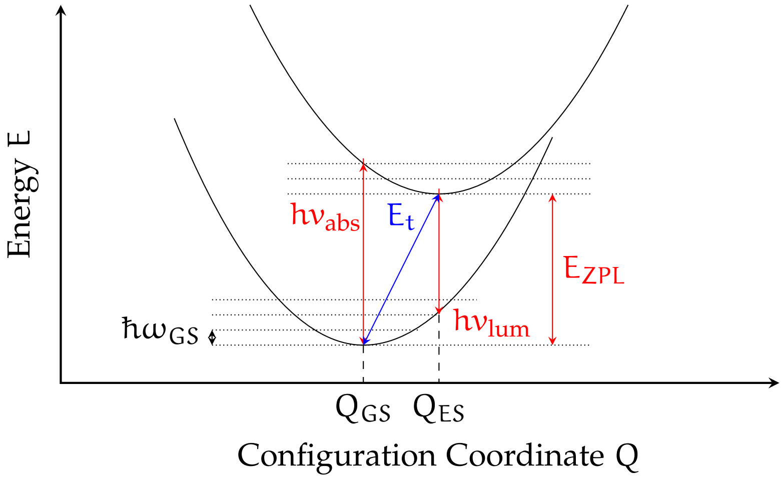


Figure 2.5: A schematic representation of a configuration coordination diagram based on Ref. [31].

The optical transitions in figure 2.5 are marked with red arrows. During slow transitions, such as during thermodynamic defect transitions, the original configuration have time to rearrange due to phonon vibrations. This is schematically drawn as the blue arrow, where the energy E_t equals the ionization energy or the position of the defect level. Optical transitions, on the other hand, are marked in red and occur in a short time range such that the original configuration does not change. They can appear in the exchange of charge carriers with the band egdes, and in a defect's internal excited state, with the latter scenario being most relevant for this thesis.

Consider a defect that rests in the ground state configuration Q_{GS}. Suddenly, it absorbs a photon with energy $h\nu_{abs}$ and occupies an excited vibronic state of the upper parabola after a vertical transition. Through lattice reconfigurations, the defect will move towards the bottom of the upper parabola, also known as Q_{ES}. Eventually, it will relax to the lower parabola by emitting a photon with energy $h\nu_{lum}$, also known as a zero-phonon line (ZPL) of energy E_{ZPL}. On the other hand, any transitions between vibronic excitation levels are phonon-related. How strong the electron-phonon interaction is can be quantified by the Huang-Rhys factor S [32]. If the two parabolas in figure 2.5 have the same configuration of Q, emission into the ZPL is enabled and $S \sim 0$. The stronger the coupling, the smaller amount of emission in the ZPL.

The optical properties of a host material can be greatly influenced by defects, in particular the ES to GS transition that can occur in a defect, as discussed for figure 2.5. If the defect were to fascilitate the emission of single photons with a detectable time inbetween together with a distinguishable ZPL, the defect would be referred to as a single photon source (SPS). The criteria for SPS are not met in many materials, since charge-state transitions often comprise interactions with either the VB or the CB. Thus, most SPSs' GS and ES levels are situated within the band gap of a host material. Consequently, mostly wide-band gap semiconductors are used as host materials for SPSs.

2.6 Semiconductor candidates for quantum technology

The properties of point defects are promising in a quantum technological perspective. We have seen that point defects fasciliate deep energy levels within the band gap of the semiconductor, and provide isolation in the solid-state matrix as a result from a high degree of localization of the defect orbitals. If the host material have a small spin-orbit coupling, it could provide long coherence times for a deep level trap in localized and high-spin states. Additionally, point defects have the potential to be single-photon sources, giving rise to sharp and distinguishable optical transitions, where a significant amount of the emission can be of the energy E_{ZPL}. This is in particular seen in wide-bandgap semiconductors, and combined with a weak electron-phonon interaction, can have the capacity to be fabricated as a high-fidelity SPS with a significant ZPL part.

In this section we will provide specific examples of a variety of promising candidates, and what properties they possess that makes them auspicious. Additionally, we will briefly mention what the challenges with the candidates are, and why it is important to explore other viable options.

2.6.1 Diamond - the benchmark material for QT

The most studied point defect system is the nitrogen-vacancy (NV^{-1}) in diamond. Figure 2.6 schematically shows the different stages of constructing the negative charge state. Panel 2.6a shows the electronic states that correspond to the difference for an isolated atom and a lattice of atoms, as a superposition of sp^3 orbitals that generates valence and conduction bands. In panel 2.6b, a vacancy has been created by removing a carbon atom, and the four orbitals interact with each other resulting in two new states with a_1 and t_2 symmetry due to dangling bonds. Substituting a carbon atom with a nitrogen atom further splits the t_2-states into two new states. The states $a(1)$ and e_x, e_y are of importance, as they are the GS and the ES of the qubit defects, respectively. Here, an optical spin-conserving transition can occur due to a laser light of correct wavelength [33], as exemplified from the discussion from the last section.

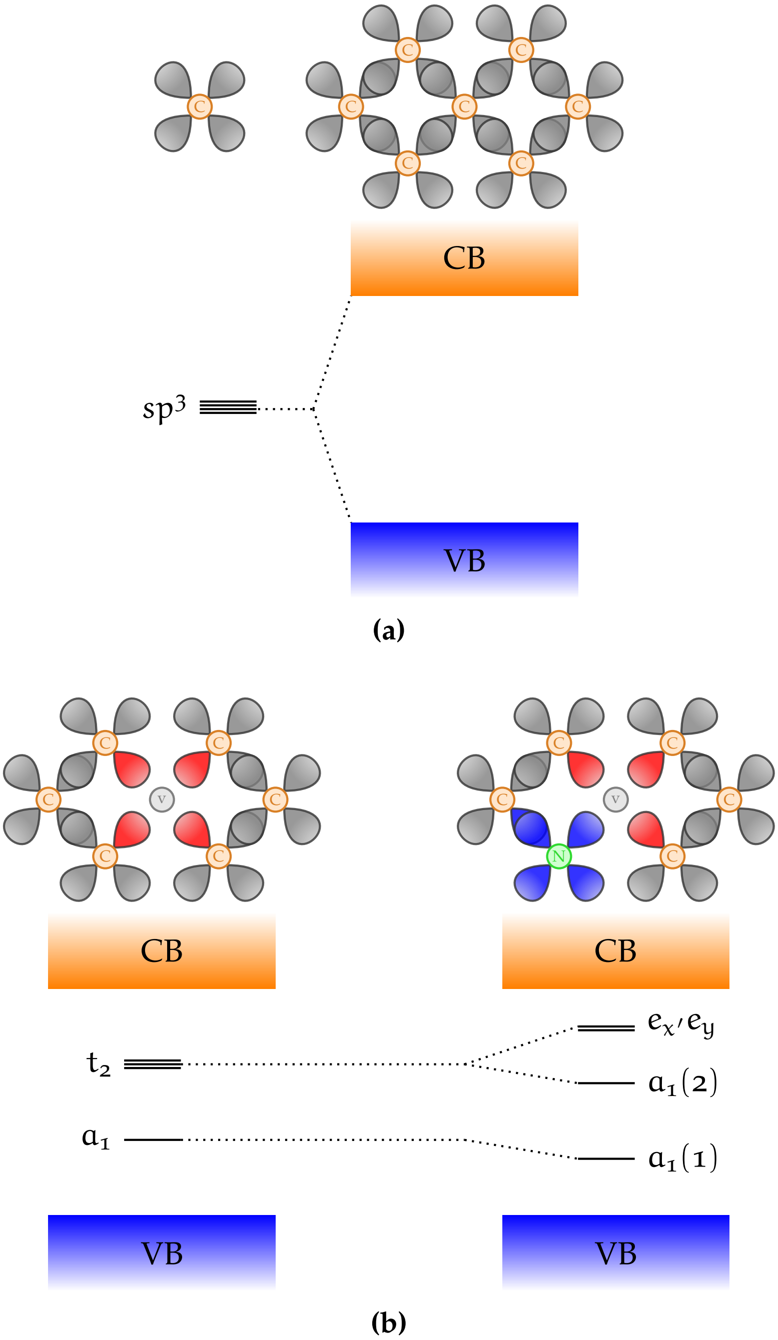


Figure 2.6: A schematic representation of the electronic structure of the NV^{-1} defect in a tetrahedrally coordinated semiconductor, exemplified by diamond. Figure adapted from Ref. [33].

The nitrogen-vacancy in diamond is a prominent single-photon source up to room temperatures. This involves initializing, manipulating and reading out of the qubit state using optical and electric excitations, and electric and magnetic fields [33]. The potential qubit system have promising applications in quantum- communication and computation, with a demonstrated entanglement between two NV center spins that are separated by 3 m [34]. Nevertheless, perhaps the most propitious application can be seen in quantum sensing as high-sensivity magnetometer with nanoscale resolution [35].

Unfortunately, the NV-center display restricted capababilities for quantum communication and computation. The amount of emission into the zero-phonon line is 4% at 6 K [36], which is low. The emission of the qubit center is not completely compatible with current optical fiber technologies, since the emission is in the red wave-length specter. Additionally, fabricating materials of diamond is far from unchallenging and serves as a signficant incentive to find other promising qubit candidates.

2.6.2 Qubit material host requirements

Therefore, we turn to the search of other QT compatible hosts that offers similar capabilities, but that are more user-friendly. In particular, we need to search for new promising materials that can host a potential point defect. Weber *et al.* [6] proposed in 2010 four criteria that should be met for a solid-state semiconductor material hosting a qubit defect, whereas some of the criteria has already been discussed. An ideal crystalline host should have [6]

(H1) A wide-band gap to accomodate a deep center.

(H2) Small spin-orbit coupling in order to avoid unwanted spin flips in the defect bound states.

(H3) Availability as high-quality, bulk, or thin-film single crystals.

(H4) Constituent elements with naturally occuring isotopes of zero nuclear spin.

Table (2.1) lists several material host candidates that exhibit promising band gap capable of accommodating a deep level defect. The spin-orbit splitting is an indication of the strength of the spin-orbit interaction, and is taken at the Γ point from the valence-band splitting. A smaller value may indicate less susceptibility to decoherence.

Criterion (H3) is important for scalability and further potential for a large-scale fabrication. The given candidate hosts provided in table (2.1) can all be grown as single crystals, but with varying quality and size.

Material	Band gap E_g (eV)	Spin-orbit splitting Δ_{so} (meV)	Stable spinless nuclear isotopes?
3C-SiC	2.39	10	Yes
4H-SiC	3.26 [37]	6.8	Yes
6H-SiC	3.02	7.1	Yes
AlN	6.13	19 [38]	No
GaN	3.44	17.0	No
AlP	2.45	50 [39]	No
GaP	2.27	80	No
AlAs	2.15	275	No
ZnO	3.44 [40]	-3.5	Yes
ZnS	3.72 [41]	64	Yes
ZnSe	2.82	420	Yes
ZnTe	2.25	970	Yes
CdS	2.48	67	Yes
C (Diamond)	5.5	6	Yes
Si	1.12	44	Yes

Table 2.1: Table taken from Gordon *et al.* [33] that lists a number of tetrahedrally coordinated hosts whose band gaps are larger than 2.0 (eV), and compares it to diamond and Si. All experimental values are from Ref. [30], except for where explicity cited otherwise.

Normally, nuclear spin is a major source of decoherence for all semiconductor-based quantum technologies. This would exclude the use of all elements in odd groups in the periodic table, since these elements exhibit nonzero nuclear spin. As a result, the spin-coherence time of a paramagnetic deep center [6] might increase. However, nuclear spin can also induce additional quantum degrees of freedom for applications in the right configuration [42]. Therefore, criterion (H4) is not a strict requirement but is a general recommendation for reducing decoherence time.

Weber *et al.* [6] use criteria $(H1)-(H4)$ to specifically find analogies to the NV^{-1} center in other material systems, thus leaving the discussion of other criteria out, such as the choice of crystal system. The atomic configuration and crystal structure of a material strongly influences the properties of a defect, since a defect's orbital and spin structure is dependent on its spatial symmetry [42]. In particular, it is the point group that decides which multiplicity a given energy level should have [43]. A higher defect symmetry group generally facilitates degenerate states, which may give rise to high spin states according to Hund's rules [42, 44]. Inversion symmetry in the host crystal can also be beneficial, resulting in reduced inhomogenous broadening and spectral diffusion of optical transitions as a consequence of being generally insensitive to external electric fields [42].

2.6.3 Silicon carbide

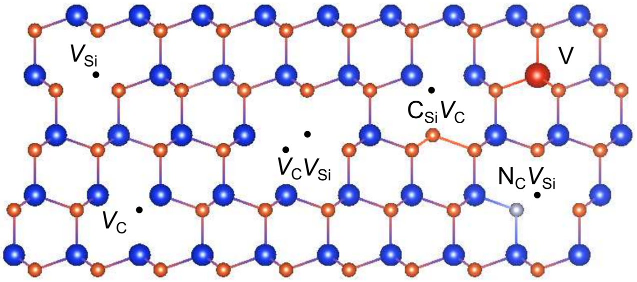

V_{Si}
V
$C_{\mathrm{Si}}V_{\mathrm{C}}$
$V_{\mathrm{C}}V_{\mathrm{Si}}$
$N_{\mathrm{C}}V_{\mathrm{Si}}$
V_{C}

Figure 2.7: Schematic illustration of various point defects in 4H-SiC, where Si atoms are blue while C atoms are orange. The illustration includes the point defects Si vacancy ($\mathrm{V_{Si}}$), C vacancy ($\mathrm{V_C}$), divacancy ($\mathrm{V_{Si}V_C}$), carbon antisite-vacancy pair ($\mathrm{C_{Si}V_C}$), nitrogen-vacancy ($\mathrm{N_CV_{Si}}$) and the vanadium impurity (V). Figure taken from Ref. [45].

Silicone carbide (SiC) is an emerging quantum platform that exists in a wide variety of polytypes, with 3C, 4H and 6H being the most prominent configurations. Several of the polytypes have been demonstrated to host SPEs with a slightly different emitter characteristic, which provides the opportunity to select the desired properties based on the variety of lattice configurations and point defects available [6, 46, 47]. While 3C has a cubic structure, we find 4H in a hexagonal structure with both hexagonal (h) and pseudo-cubic (k) lattice sites. 6H is found in a hexagonal structure with the three orientations that are labelled h, k_1 and k_2. Importantly, SiC in the three varieties experience wide-band gaps, low spin orbit coupling and stable spinless nuclear isotopes [6, 30, 37], as seen from table 2.1. Furthermore, SiC benefits from mature fabrication on the wafer-scale, which checks the last of the four (H1-H4) QT host requirements, marking it as a suitable quantum material platform.

The most studied emitters in SiC include the carbon antisite-vacancy pair $\mathrm{C_{Si}V_C}$ that emits in the red, the silicon vacancy $\mathrm{V_{Si}}$ that emits in the near infra-red, and the divacancy ($\mathrm{V_{Si}V_C}$) and the nitrogen-vacancy center ($\mathrm{N_CV_{Si}}$) that both emit at near-telecom wavelengths. Thus, the two latter emitters could potentially ease the integration with optic fiber technologies as compared to e.g. the $\mathrm{NV^-}$. Additionally, the four different point defects have all been identified as room-temperature SPEs with demonstrated coherent spin control [48]. Illustrations of several configurations of emitters in 4H-SiC are included in figure 2.7.

2.6.4 Alternative promising material hosts

Single photon emitters have been observed in other semiconductor materials, however most of the emitters are yet to be identified or are in an early stage of identification. Therefore, specific details about spin- or emission-related structure are yet to be implemented. In this section we will briefly mention recent promising materials for QT.

One immediate potential candidate is silicon, considering the favorable device fabrication processes that are available. It has demonstrated that phosphorous impurities at Si sites can store a quantum state for over 30 seconds, enabling their use in a potential Kane quantum computer [49]. Unfortunately, the P impurity lack any single photon source capabilities. Recently, however, the G-center arising from the carbon-interstitial carbon-substitutional (C_sC_i) complex was identified as an promising SPE candidate with single photon emissions at telecom wavelength [50].

Other materials that emits individual photons have been detected in other wide-band gap semiconductors, including ZnO, ZnS, GaAs, GaN and AlN [49, 51]. Unfortunately, challenges due to the specific materials complicate the implementation of defects for QT. ZnO and ZnS experience a broad emission due to a large photon involvement. GaAs is promising since it has been demonstrated as a SPS, but demonstration of spin manipulation is still in an early phase[51]. GaN and AlN, on the other hand, are more susceptible to a more narrow emission, where room-temperature SPE has been demonstrated for both GaN [52] and wurtzite AlN films [53]. The defect levels for AlN films have been tentatively assigned to the nitrogen-vacancy and divacancy complexes, but they tend to occur too close to the band edges for any SPE [49, 54].

Recent advances in material growth have enabled the use of hole spin-based semiconductors, such as SiGe quantum wells due to their low disorder and large intrinsic spin-orbit coupling strength [55]. Promising materials can also emerge from placing an impurity next to a vacancy. Cation vacancies in possible structures tend to be negatively charged, thus the impurities should act as donors. Therefore, the self-activation center in ZnSe can be a promising defect [6], but is still in the early stage of development.

Two-dimensional materials such as hexagonal boron nitride (h-BN), MoS_2, WSe_2 and WS_2 are also of interest as quantum platforms [56, 57]. The structure of h-BN exists in single- or multilayers, and it has been demonstrated a broad range of stable room-temperature single-photon emitters [58, 59]. In WSe_2, $MoSe_2$ and WS_2, there has been experimentally discovered optical excitation of defects, while also electrical excitation of defects for WS_2 [57]. However, secure identification for the source of the emission is yet to be established [57, 60, 61].

2.6.5 Associated challenges with material host discovery

The idea of finding new potential host candidates to utilise point defects in QT is of a challenging sort. Recall, we have made four criteria that deals with the required (H1) band gaps, (H2) spin-orbit coupling, (H3) availability and (H4) spin-zero isotopes, but we have no knowledge of if there should be more criteria or to what extent a criteron needs to be fulfilled. What we do know is that there are major advantages if materials exhibit properties such as isolation in the lattice and weak electron-photon interaction, however, the process to provide any quantity of measurements are through approximations and material-specific properties. These approximations does not neccessarily capture quantum properties well.

Furthermore, the identified candidates constitues an immensely selective group of only a handful potential hosts. As an example, most known potential hosts are unary or binary compounds. This is probably due to the increasing complexity dealing with an additional level of interactions in the lattice. Therefore, there are reasons to believe that many potential hosts are yet to be discovered, which serves as a motivation for studies involving exploratory research for new candidates.

Part III

Methodology and implementation

Chapter 3

Information flow

The information stream of this project can be regarded as many modular parts connected together in logical pieces, and is strongly influenced by the process that defines a *minimum viable product* (MVP) through iterative development. An MVP is commonly known (in the bussiness world) as a new product that enables the most learning out of the minimum effort possible. This method allows a product to be iteratively evolved by consistent feedback and development, which in return enables cooperation between cross-disciplinary fields.

Furthermore, by having several modules serving as the fundament of the project, it is possible to achieve a long-lasting and robust product that is simple to maintain yet straightforward to develop. Bugs can be tackled through a documented code simultaneously as visible future improvements can be adressed. Therefore, the product is not regarded as completed in any terms, but rather ready for a first release after iteratively finding the mimimum viable product.

The main project of this work can be found on the Github repository *predicting-solid-state-qubit-candidates* [62]. In this chapter we will look into the details and thoughts behind the extraction of data, building features, data preparation, data mining and eventually fabricating a generalized model that can predict unseen data with confidence.

3.1 Extraction and featurization of data

The initial step for gathering and building features can be visualised through the flowchart in figure 3.1. Initially, we start by extracting all entries in the Materials Project that matches a specific query. Thereafter, we apply Matminer's featurization tools to make thousands of features of the data. In a parallel step, entries that are deemed similar to the entries from the initial Materials Project query are extracted from AFLOW, AFLOW-ML, JARVIS-DFT, OQMD and Citrine Informatics. Finally, we combine the steps together as interim data that is ready for further analysis.

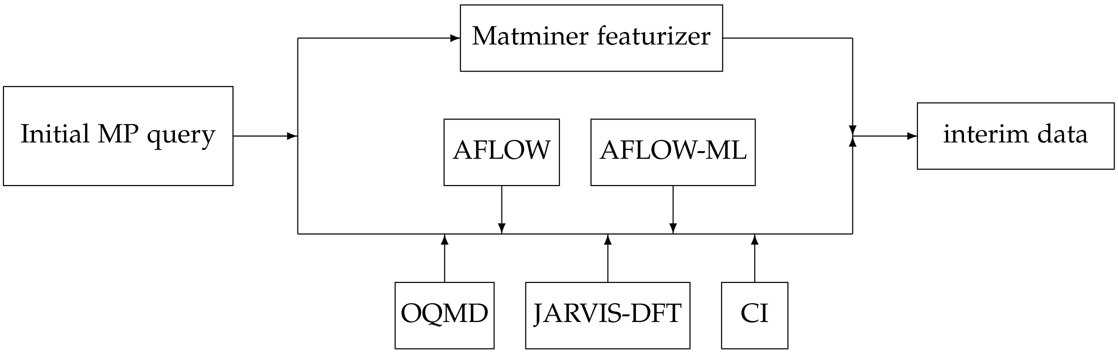


Figure 3.1: The data flow of the main project, starting from an initial MP-query, and ending with a featurized dataset with entries from several other databases. The matminer featurizer step is further visualized in-depth in figure 3.2.

The initial query has the requirement that all entries has to be derived from an experimental ICSD entry, and is reasoned by that we can identify equivalent entries in other databases. Furthermore, all entries in the Materials Project needs to have a band gap larger than 0.1eV. Recall that Materials Project applies the functional GGA in estimating the band gap, which is known to severely underestimate the given electronic property. Therefore, we have chosen a low value to not rule out any potential candidates but high enough to leave out all materials that can be considered metallic. Thus, out of a total of 139.367 entries in Materials Project, our initial requirement is satisfied by 25.352 of the entries.

From figure 3.1 we notice that by using many databases we do not add additional entries that exist in some databases but is not to be found in Materials Project. This is by design since it preserves the versatility of choosing a database to work with. Therefore, one can completely ignore steps such as the initial query of Materials Project or the featurization process, and rather focus on e.g. all the 400.000 entries existing in OQMD. The examples that follows will illustrate the ease of extracting data from several different databases, and can serve as the starting point for other research projects in computational material science.

3.1.1 Practical data extraction with Python-examples

For this section, we will show practical examples of how to extract data that might fulfill the criteria for a material to host a qubit candidate given in the theory part. We will begin with the database of Materials Project, and then search for entries in other databases that match entries from MP. This process is reproducable as a jupyter notebook[1] and the databases in question are the ones refered to in the previous section.

Instead of building multiple HTTP-methods from scratch, we will here take a look at the easiest method at obtaining data from each database. The range of data in a database can consist of data from a few entries up to an unlimited amount of entries with even further optional parameters, and has limitless use in applications. However, the amount of data in a database is irrelevant if the data is inaccessible. Therefore, we provide a toolbox in how to extract information in the easiest way possible. This includes looking into the APIs that supports data-extraction and that are recommended by each respective database.

Every data extraction class is based on an abstract parent class. The advantages of using a base parent class are many, such as improving the readability during code reviews, reducing the main barrier for understanding the underlying structure of a project and utilising reusable components. Yet, the main advantage of using a base parent class is the fact that it can effortlessly be extended for further implementations since it provides a code skeleton.

Materials Project

The most up-to-date version of Materials Project can be extracted using the python package pymatgen, which is integrated with Materials Project REST API. Other retrievel tools that is dependent on pymatgen includes matminer, with the added functionality of returning a pandas dataframe. Copies of Materials Project are added frequently to cloud services such as Citrine Informatics, but the latest added entries to Materials Project cannot be guaranteed in such a query.

Entries in Materials Project are characterized using more than 60 features[2], some features being irrelevant for some materials while fundamental for others. The data is divided into three different branches, where the first can be described as basic properties of materials including over 30 features, while the second branch describes experimental thermochemical information. The last branch yields information about a particular calculation, in particular information that's relevant for running a DFT script.

To extract information from the database, we will be utilising the module pymatgen. This query supports MongoDB query and projection operators[3], resulting in an almost instant query.

[1] add and insert DOI for JN 01-generateDataset-notebook.ipynb

[2] All features can be viewed in the documentation of the project: https://github.com/materialsproject/mapidoc/master/materials

[3] https://docs.mongodb.com/manual/reference/operator/query/

1. Register for an account[4], and generate a secret API-key.
2. Set the required critera.
3. Set the wanted properties.
4. Apply the query.

The code nippet in code listing 3.1 resembles steps $2-4$, and is filtered as the inital query.

```
from src.data.get_data_MP import data_MP

MAPI_KEY = ‘‘very_secret_key_here’’
MP = data_MP(API_KEY=MAPI_KEY)
df = MP.get_dataframe()
```

Listing 3.1: Practical example of extracting information from Materials Project using pymatgen, resulting in a Pandas DataFrame named entries that contains the properties given after performing a filter on the database. The criteria is given as a JSON, and supports MongoDB operators.

Citrine Informatics

Citrine Informatics is a cloud service, which means that the spectrum of stored information varies broadly. We will access research through open access for institutional and educational purposes. Information in Citrine can be stored using a scheme that is broken down into two sections, with private properties for each entry in addition to common fields that are the same for all entries.

In this example, we will gather experimental data using the module matminer. The following steps are required to extract information from Citrine Informatics.

1. Register for an account[5], and generate a secret API-key.
2. Set the required critera.
3. Set the wanted properties and common fields.
4. Apply the query.

The code listed in code listing 3.2 gives an easy example to steps $2-4$ with experimental data as filter, which results in an almost instant query.

[4] https://materialsproject.org

[5] https://citrination.com

```
from src.data.get_data_Citrine import data_Citrine

CAPI_KEY = ‘‘very_secret_key_here’’
citrine = data_Citrine(API_KEY=CAPI_KEY)
df = citrine.get_dataframe()
```

Listing 3.2: Practical example of extracting information from Citrine Informatics using matminer, resulting in a Pandas DataFrame named experimental_entries that contains the properties given after performing a filter on the database. The criteria is given as a JSON.

AFLOW

The query from AFLOW API [63] supports lazy formatting, which means that the query is just a search and does not return values but rather an object. This object is then used in the query when asking for values. For every object it is neccessary to request the desired property, consequently making the query process significantly more time-demanding than similar queries using APIs such as pymatgen or matminer for Citrine Informatics. Hence, the accessibility is strictly limited to either searching for single compounds or if the user possess sufficient time.

Matminer's data retrievel tool for AFLOW is currently an ongoing issue [64], thus we present in code listing 3.3 a function that extracts information from AFLOW and returns a Pandas DataFrame. In contrast to Materials Project and Citrine Informatics, AFLOW does not require an API-key for a query, which reduces the amount of steps to obtain data. The class searches for an stored AFLOW-data, and initialises a MP-query with the initial criteria if not successful. The resulting query will then be used as input to AFLOW.

```
from src.data.get_data_AFLOW import data_AFLOW

AFLOW = data_AFLOW()
df = AFLOW.get_dataframe()
```

Listing 3.3: Practical example of extracting information from AFLOW. The function can extract all information in AFLOW for a given list of compounds, however, it is a slow method and requires consistent internet connection.

Restricted by the available API, the resulting query of 25212 entries in Materials Project took place during the period from january to february 2021 and took in total 23 days. Unfortunately, less than 0.02% of the entries screened from Materials Project was present in AFLOW.

AFLOW-ML

In this part, we will be using a machine learning algorithm named AFLOW-ML Property Labeled Material Fragments (PLMF) [65] to predict the band gap of structures. This algorithm is compatible with a POSCAR of a compound, which can be generated by the CIF (Crystallographic Information File) that describes a crystal's generic structure. It is possible to download a structure as a poscar by using Materials Project front-end API, but is a cumbersome process to do so individually if the task includes many structures. Extracting the feature of POSCAR is yet to be implemented in the RESful API of pymatgen, thus we demonstrate the versatility of pymatgen with a workaround.

We begin with extracting the desired compounds formula, their Materials Project IDs (MPIDs) for identification, and their respectful structure in CIF-format from Materials Project. In an iterative process, each CIF-structure is parsed to a pymatgen structure, where pymatgen can read and convert the structure to a POSCAR stored as a Python dictionary. Finally, we can use the POSCAR as input to AFLOW-ML, which will return the predicted band gap of the structure. This iteratively process parsing and converting, but is an undemanding process. The function that handles this is presented in code listing 3.4. Similar to AFLOW-query, this code listing is dependent on MP-data and will apply for a query if the data is not present.

A significant portion of the process is tied up to obtaining the input-file for AFLOW-ML, and fewer structures will result in an easier process. Nevertheless, we present the following steps in order to receive data from AFLOW-ML.

1. Download AFLOWmlAPI[6].
2. Getting POSCAR from MP.
   (a) Apply the query from Materials Project with "CIF", "material_id" and "full_formula" as properties.
   (b) Insert resulting DataFrame into function defined in code listing 3.4.
3. Insert POSCAR to AFLOW-ML.

```
from src.data.get_data_AFLOWML import data_AFLOWML

AFLOWML = data_AFLOWML()
df = AFLOWML.get_dataframe()
```

[6] http://aflow.org/src/aflow-ml/ to the same directory as code listing 3.4

Listing 3.4: Practical example of extracting information from AFLOW-ML. The function will convert a CIF-file (from e.g. Materials Project) to a POSCAR, and will use it as input to AFLOW-ML. In return, one will get the structure's predicted band gap. It should be noted that this requires the AFLOW-ML library in the same directory.

The resulting ab-initio calculations used an average of 57s/compound, which in total sums up to 16.6 days. In contrast to AFLOW, 100% of the entries was present due to the fact that it is not based on a database but rather a model.

OQMD

To extract information from the OQMD, the easiest way was through the interface of Matminer. The difficulty of extraction are mostly regarded to column which are not assigned to a type, however, this is taken care of in the extraction class visualized in code listing 3.5.

```python
from src.data.get_data_OQMD import data_OQMD

OQMD = data_OQMD()
df = OQMD.get_dataframe()
```

Listing 3.5: Practical example of extracting information from OQMD through Matminer.

The query is done almost instantly, resulting in a DataFrame containing over 400.000 entries, where 40% of the entries are matching an entry of the initial MP query.

JARVIS-DFT

The newest version of the JARVIS-DFT dataset can be obtained by requesting an account at the official webpage, but with the drawback that an administrator has to either accept or deny the request. Thus, the accessibility of the database is dependent on if there is an active administrator paying attention to the requests, which is a limitation experienced during this work. Another approach is to download the database through matminer, however with the limitation of not neccessarily having the latest version of the database. A third approach is to download a version of JARVIS-DFT that have been made available for requests the 30.04.2020 at http://figshare.com by Choudhary *et al.* [66]. The author provides tools for extraction, yet not compatible with the latest version of Python (3.8) at the time writing (12.03.2021). Therefore, we provide a tool to extract this data through the use of our base class.

```
from src.data.get_data_JARVIS import data_JARVIS

JARVIS = data_JARVIS()
df = JARVIS.get_dataframe()
```

Listing 3.6: Practical example of extracting information from JARVIS-DFT. For this example, we exclude all metals by removing all non-measured band gaps.

We observe that there is no advanced search filter when loading the database from matminer. The author of matminer regards this as the user's task, and is indeed easily done through the use of the python library Pandas.

The resulting screening of 25212 entries from Materials Project was done almost instantly, and it was found 11% and 17.8% similar entries for the TBMBJ and OptB88 functionals with MP, respectively. Moreover, JARVIS-DFT contains information about spin-orbit splitting, but only 0.12% of the calculations was found as a match with the initial MP query.

3.2 Matminer featurization

Before applying any machine learning algorithm, raw data needs to be transformed into a numerical representation that reflects the relationship between the input and output data. This transformation is known as generating descriptors or features, however, we will in this work adapt the name *featurization*. The open source library of Matminer provides many tools to featurize existing features extracted from Materials Project. In this section we will describe how to extract the features from an initial Materials Project query result (see subsection. 3.1.1), and the resulting features. It is beyond the scope of this work to go in-depth of each feature since the resulting dataset contains a quantity of more than 4500 features, but we will here take the liberty to serve a brief overview of the features and refer to each respective citation for more information. The respective table with information regarding 39 distinct matminer featurizers is situated in the Appendix, table B.1.

The motivation behind the choice of featurizers is that we do not precisely know which features that describes a good potential host. A few potential candidates were briefly mentioned in section 2.6.4, while most candidates are probably yet do be discovered. If we had precise knowledge of what to look for, then there is a good chance that the list of hosts would be longer. Therefore, we strive to collect an achievable quantity of descriptors with the hope of getting wiser in terms of describing a potential material host.

To apply matminer's featurization tools, we extend an existing implementation by Breuck *et al.* [67] called the Materials Optimal Descriptor Network

(MODNet). The author Breuck *et al.* specifies that MODNet is a supervised machine learning framework for learning material properties based on either composition or crystal structure. To provide the training data for their model, MODNet featurizes (through matminer) structures either from Materials Project or in the form of a structure object made by pymatgen. Their current implementation provides featurization for compositions, structures and sites. However, matminer also provides featurization tools for density of states (DOS) and band structures, therefore we modify MODNet and extend it to fascilitate such featurizations.

One immediate limitation of our extension is that Matminer's tools is dependent on a pymatgen DOS- and bandstructure object. These objects contains information up to 10MB, and becomes a challenge when dealing with data containing several thousand such objects. This is solved by the required features for matminer's featurization for a subsample of the data, followed by a featurization process of the same subsample. When the feaurization is done, we store the new features and throw away the pymatgen features. This is done iteratively for the entire data set. Thus, a compromise between applying several queries and storing information has been done. The scheme can be visualised as the flow chart seen in figure 3.2.

In the extended version of the featurization process, we eliminate all columns that does not have any entries with physical meaning. This is beneficial for several reasons, such as to reduce memory allocated and to preprocess the data. If there are entries existing with both physical and non-physical for the same column, we replace the non-physical meanings with -1 for recognition in a later step. Additionally, we convert columns that are categorical or lacks a numerical representation into a categorical portrayal. Thus, we strive to limit the neccessary steps for further processing of data into a machine learning algorithm. Nevertheless, the featurization process results in 4876 descriptors.

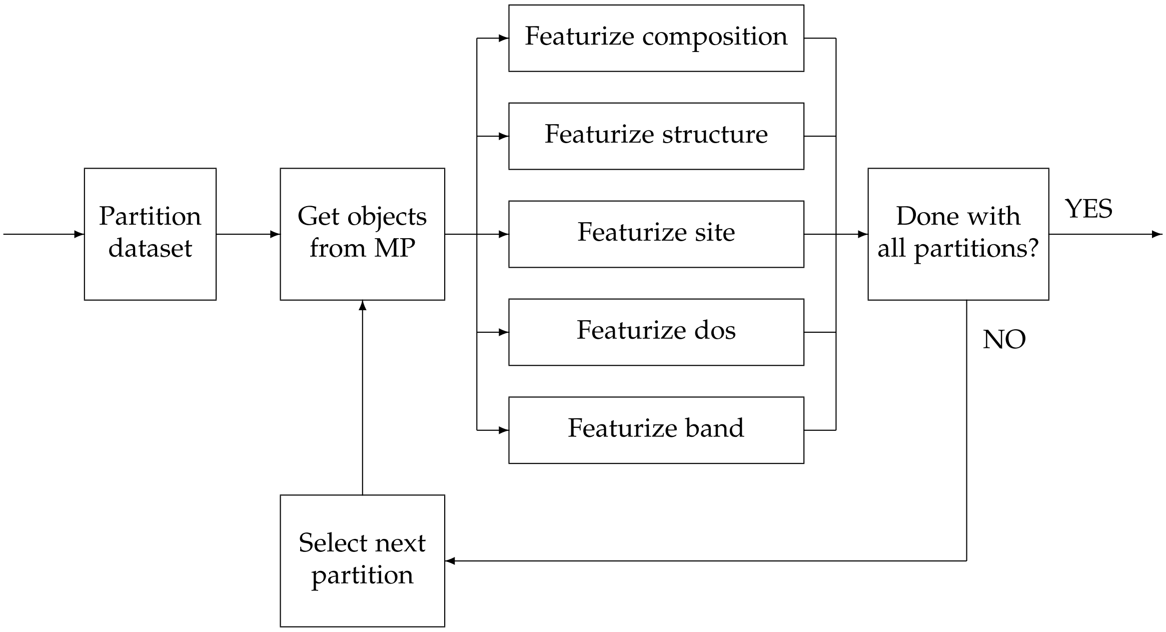


Figure 3.2: The process of the matminer featurizer step as seen in figure 3.1. To limit the memory and computational usage, the data is partioned into smaller subsets where the respective pymatgen objects are obtained through a query to be used in the following featurization steps. This is iteratively done until all the data has been featurized.

Even if the first version of Matminer was released in 2016, many issues concerning daily operational use are still present. During the featurization process in this work, we manually identified 14 (TODO: Update number) erroneous entries that are summarized in the Appendix, table B.2, which were excluded from the dataset. These entries were part of the reason why the featurization process is a time-consuming process, as there is currently no implementation in Matminer that can potential pick up and catch erroneous entries in Materials Project. The process of manually catching such an entry was identified by featurization of single entries causing one of two problems. The first problem could be that an entry could be causing a memory leak which leads to an exceedingly large memory allocation, or it could be that the featurization process needed days to calculate oxidation states for a structure.

3.3 Labelling of data

After selecting entries based on an initial query from Materials Project followed by a thorough featurization process using Matminer, we face a challenge in terms of defining a training set that we can train data on. This is not only challenging due to the lack of known candidates, but also due to the intricacy of defining materials as bad candidates. Therefore, in this section we describe three different approaches of finding a training set consisting of (1) good candidates and (0) bad candidates.

3.3.1 First approach; the Ferrenti approach

The first approach on defining a training set is based on the criteria from the paper "Identifying candidate hosts for quantum defects via data mining" of Ferrenti *et al.* [68], therefore we will name this approach *the Ferrenti approach*. They suggest a data mining process consisting of four stages by systematically evaluating the suitability of host materials from Materials Project. This procedure is referred to as *data mining*, and we will initially begin with looking at labelling good candidates.

Labelling good candidates

The first stage consists of the following steps to include materials that

Stage 1

- contains elements with a $>$ 50% natural abundance of zero spin isotopes.
- crystallize in nonpolar space groups.
- is present in the ICSD database.
- is calculated nonmagnetic.

The restriction of materials to only contain elements with at least 50% nuclear spin-free isotopes might help with reducing decoherence for all semiconductor-based quantum technologies, as discussed in section 2.6.2. The limit is chosen due to that elemental species with at least 50% nuclear spin free isotopes could likely be isotopically enriched to higher concentrations [68], which has been accomplished for carbon [69, 70] and silicon [71]. In particular, the restriction excludes the use of 53 elements from any species. Any magnetic noise or any presence of electric dipole moment could also potentially increase decoherence of defects. Therefore, we try to reduce any decoherence by restricting materials to possess highly symmetric structures which are nonmagnetic.

Stage two consists applying additional filtering due to practical reasons. This includes removing all materials containing radioactive or toxic elements, as well as removing noble gases because none exists as solids under standard conditions. Rare-earth metals were also excluded due to the difficulty of obtaining pure materials that are sufficiently free of nuclear spin. Lastly, we remove entries that occur mostly in very complex cluster structures (Ru, Os) or are not present in any identified phases (Fe, Ni). Therefore, the additional filter constitutes of obtaining materials that

Stage 2

- does not include Th, U, Cd or Hg.
- does not include any noble gases or rare-earth elements.
- does not include Ru, Os, Fe, Ni

Stage three consists of setting a lower band gap limit similar to that of silicon, but due to severe underestimation of bandgaps by PBE-GGA we set this restriction lower since we do not want to exclude any potential host candidates. The materials are required to have

Stage 3

- a bandgap larger than $0.5eV$ as calculated by MP PBE-GGA.

Finally, the last stage consists of identifying the thermodynamic stability of each compound. A large energy above hull (E Above Hull) is an indication of an unstable compound and would likely cause decomposition, therefore the last filter requires the materials to have

Stage 4

- a calculated E Above Hull $< 0.2eV/\text{atom}$.

The quantity of entries through the different stages have been visualized in table 3.1. The table compares our and their implementation of the same screen procedure with different results. In particular, we see that the remaining materials that have survived four stages of filtering are twice as many. This could be due to the date of extracting since it differs with 13 months, since over 14.000 new entries were added to Materials Project. However, another reason could be due to that they have done additional manual screening. Unfortunately, precise information of which entries that were excluded from the manual filtering were not included in neither the article or the supplementary information [68]. Yet, after doing a data mining procedure we have found 1046 potential candidates that exhibit promising features.

Table 3.1: A table that compares two different implementations of the same screen procedure. Ferrenti *et al.* extracted information March of 2020, while we did the extraction during April of 2021. The adjusted difference is given as our reported entries divided on their reported entries.

Stage	Good candidates based on Ferrenti *et al.* [68]	Good candidates based on approach 1	Adjusted difference
Total entries in Materials Project [72, 73]	125.223	139.367	11%
Stage 1	3363	4347	29%
Stage 2	1993	2226	12%
Stage 3	920	1181	28%
Stage 4	541	1046	93%

Labelling bad candidates

We have now defined good candidates, and turn our attention to defining bad candidates. This is perhaps the difficult part, since we do not know exactly which properties or combination of features a material needs to exhibit for it to be excluded from any use in quantum technology. Therefore, we try to find the opposite criteria of the four stages that defined good candidates.

If we were to turn around all criteria defined in the four stages above (except for energy above hull), it would result in only 52 entries which would make the combined data set very imbalanced. Instead, we try to provide a more general process that includes a larger variety of entries, which could potentially increase the predictor space for bad candidates. The screening procedure for stage 1 requires bad candidates to

Stage 1

- crystallize in polar space groups.
- be present in the ICSD database.
- be calculated as magnetic.

Stage 2

- have a bandgap larger than 0.1eV as calculated by MP PBE-GGA.

We include only ICSD entries and a lower band gap limit for consistency, since our data does not contain entries outside of these limits. The number of entries after stage 1 is 1520, while stage 2 reduces the entries to 684.

3.3.2 Second approach; the augmented Ferrenti approach

In the second approach we try to make adjustment of the first approach to improve the dataset. This approach is therefore named *the augmented Ferrenti Approach*.

Labelling good candidates

The first approach included unphysical criteria such as removing elements that are either radioactive, toxic, elements not occuring under standard conditions, or rare-earth elements that are difficult to obtain. In this approach, we remove those constraints since these are not criteria that neccessarily deem a material as either good or bad for QT, and it is eventually up to experimentalists for evaluation of such practicalities. Therefore, we remove stage 2.

We will in this approach not consider if a material is stable or not, since this is eventually up to experimentalists to evaluate. Additionally, we will include a few interesting elements that showed promising properties as discussed in section 2.6.4, and was originally excluded due to lack of spin zero isotopes. Thus, the second approach consists of the following steps to include materials that

Stage 1

- contains elements with a $>$ 50% natural abundance of zero spin isotopes except Al, P, Ga, As, B and N.
- crystallize in nonpolar space groups.
- is present in the ICSD database.
- is calculated nonmagnetic.

Stage 2

- have a bandgap larger than 1.5eV as calculated by MP PBE-GGA.

Stage 3

- have a calculated E Above Hull $<$ 0.2eV/atom.

Since we have removed restrictions, we can also infer a stronger one for the band gap. Therefore, we can be considerable more certain if a band gap can a accomodate deep defect due to an increasing amount of entries when removing restrictions.

Labelling bad candidates

For bad candidates, we implement the same strategy as defined for bad candidates in approach 1. The resulting table for both good and bad candidates is found in table 3.2. The table reveals a considerable imbalanced dataset with up to 75% being good candidates, while only 25% of the training data are labelled as bad candidates. However, the training set is 78% larger than in approach 1.

Table 3.2: A table showing the number of entries through the data mining process for good candidates in approach 2 and bad candidates in approach 1 and 2.

Stage	Good candidates approach 2	Bad candidates approach 1 and 2	Ratio
Total entries in Materials Project [72, 73]	139.367	139.367	-
Stage 1	7433	1520	83%/17%
Stage 2	2373	684	78%/22%
Stage 3	2141	—	75%/25%

3.3.3 Third approach; the insightful approach

The third approach is vastly different than the two first approaches in terms of labelling, therefore it is named *the insightful approach*.

Recall, in section 2.6.4 we discussed alternative promising material host candidates. The third approach for finding good candidates is to search our current data for any materials that overlap with known good candidates. Due to a concern of having a too small dataset, we will include materials that are promising and have shown suitable properties to accomodate deep defects that can exhibit quantum effects.

Labelling good candidates

Stage 1

- matches the formulas SiC [6, 30, 37, 46, 47], BN [56, 57], MoS_2[57], WSe_2[57], WS_2[57], GaN [52], GaAs [51], AlN [6, 53], ZnS [49], ZnSe [6], ZnO [49], AlP[6], GaP[6], AlAs[6], ZnTe[6], CdS[6], SiGe [55], C [33, 35, 36] or Si [49, 50].
- is present in the ICSD database.

Stage 2

- have a bandgap larger than 0.5eV as calculated by MP PBE-GGA.

Stage 3

- Manual screening of correct structures.

After stage 1, it was found 202 matching formulas which included 12 entries that had a bandgap lower than 0.5eV. These entries were structures that was reported as unstable in terms of energy above hull calculations, and would decompose into entries that were already present in the data after stage 1 with bandgap substantially larger than 0.5. We choose to exclude these entries with an additional band gap restriction due to the fact that the bandgap is not large enough to accomodate a deep defect. Therefore, these entries were instead labelled as bad entries.

Entries matching the formula C, SiC, BN, MoS_2, WSe_2 and WS_2 were manually screened to see if the entries have a matching structure to the respective candidates discussed in section 2.6.1 and 2.6.3 and 2.6.4, respectively. For C, we admit only three-dimensional diamond-like structures, as explicitly stated in the column tags at Materials Project. Two-dimensional graphite-like structures are labelled as bad candidates, while complex structures (eg. C_{28}, C_{48}, C_{60}) were moved to the test set. For SiC, we admit only entries with the polytypes 3C, 4H and 6H, while moving structures similar to 2H to the test set. Concerning the other materials, we only admit two-dimensional structures.

The materials AlP, GaP, AlAs, ZnTe and CdS were manually screened for tetrahedrally coordinated structures, and have been included since Weber *et al.* [6] has identified them as potential promising candidates due to acceptable properties defined in requirements (H1-H4) in section 2.6.2. We note that only tetrahedrally coordinated structures of the given formulas were present after the bandgap restriction of 0.5eV.

Since the number of elements in the good candidates are not containing more than two elements, we decide to remove the feature that explains how many elements due to that we do not want the model to discriminate based on this feature. After three stages, a total of 172 entries were labelled as good candidates.

Labelling bad entries

Since the training data that constitutes good candidates are few, we choose to add 400 random entries from the dataset of bad candidates used in approach 1 and 2, in addition to the entries stated above. We only add a subsample for increasing the potential dimensional space for predictions of candidates while avoiding having a too inbalanced dataset.

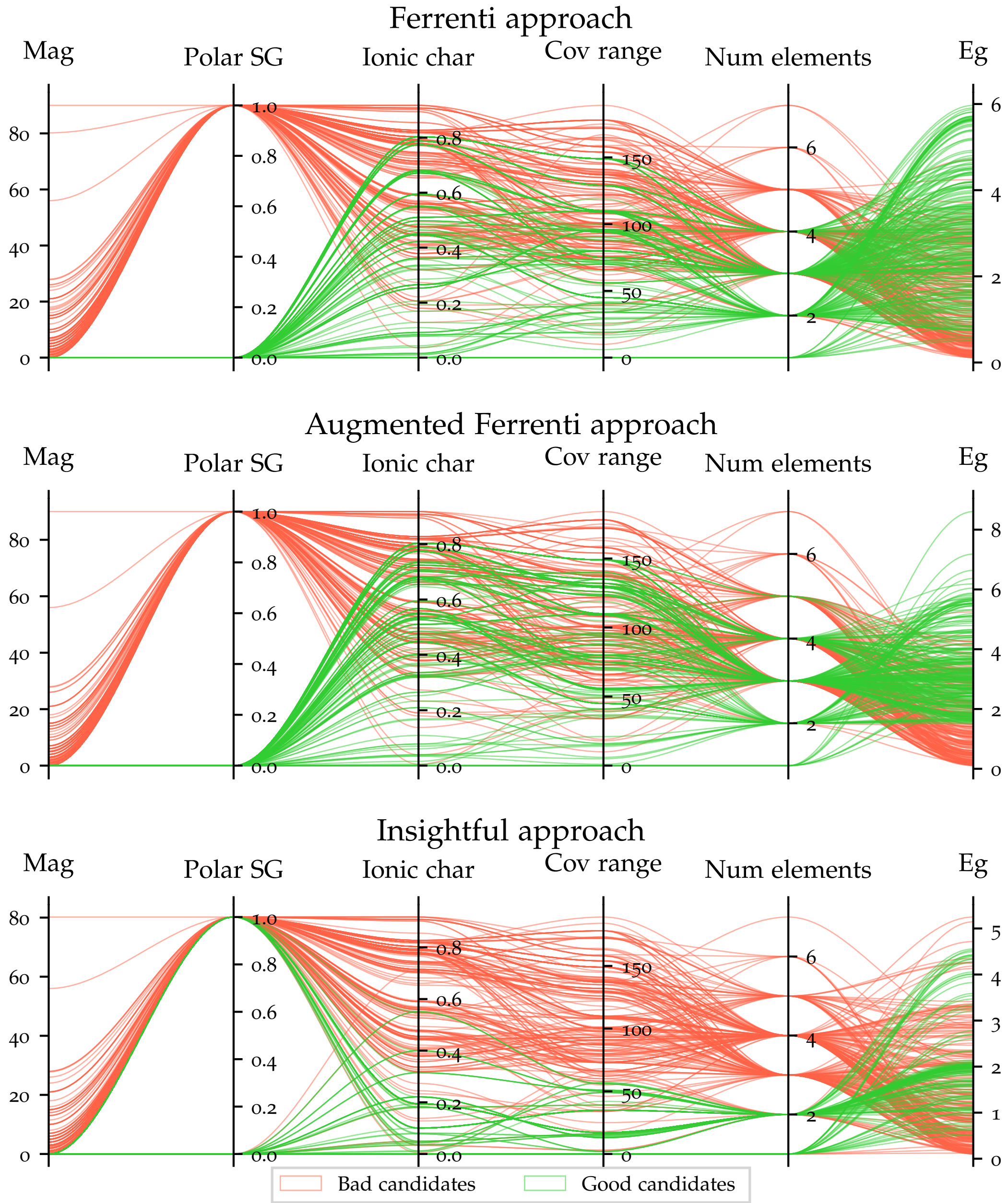


Figure 3.3: Parallel coordinate plots for the different approaches. To limit the data cluttering, we have random collected up to 250 entries for each class and made the lines transparent. For the insightful approach, we have used all entries 172 entries.

3.3.4 Comparison of the approaches

The three approaches provide special emphasis on each their goal. The Ferrenti approach is dependent on choosing only elements with zero spin isotopes together with practical filters, while the augmented Ferrenti approach allows a larger variety of elements and removes the practical reasons for excluding elements. Thus, the first approach targets a more narrow prediction space than the second approach does, and we would expect that the second approach will lead to more predicted candidates compared to approach one. However, perhaps the most restricted approach is the insightful approach. Since we only include known candidates, they should share the same properties and therefore provide a very narrow prediction space.

Unfortunately, the downside of including all the known candidates in one approach is that it becomes increasingly challenging to evaluate the approach or the resulting model. For the two first approaches, we can see if some of the known candidates are present in the predictions, while this is not possible for the latter approach.

We provide a visualization of each approach's training data as a parallel coordinate plot for a few selected features in 3.3. Parallel coordinate schemes [74, 75] represents a multi-dimensional data tuple as one polyline crossing parallel axis. The selected features are found on the x-axis, while the y-axis show the value of the data present. Thus, parallel coordinate plots can turn complex many dimensional data into a compact two-dimensional representation. However, due to data cluttering and that one entry can potentially reserve a large visual area of the figure, the utilization becomes limited when facing large datasets [76]. Therefore, we have chosen to plot a random sample of each class with an upper limit of 250 per class with transparent lines.

The Ferrenti approach and the augmented Ferrenti approach share similarities, such as only having polar space groups present and having an equal amount of upper limit for both ionic character and covalent range. Additionally, they share that entries constitute of up to five different elements. Interestingly, we can see that even if the augmented Ferrenti approach is less restricted, it appears that the entries map over the same dimension based on 3.3.

The biggest difference is seen for the insightful approach. The chosen entries do not posess any magnetization, even if there are both polar and nonpolar space groups present. The range of covalent radius and maximum ionic character is significantly lower than the two other approaches.

To visualize the complexity of the training sets, we have found the two largest eigenvalues of the covariance matrix of the initial data from Materials Project, and transformed the training sets according to the corresponding two eigenvectors. The resulting scatter plots is found in figure 3.4. In green squares, we find the good candidates for each approach, while the labelled

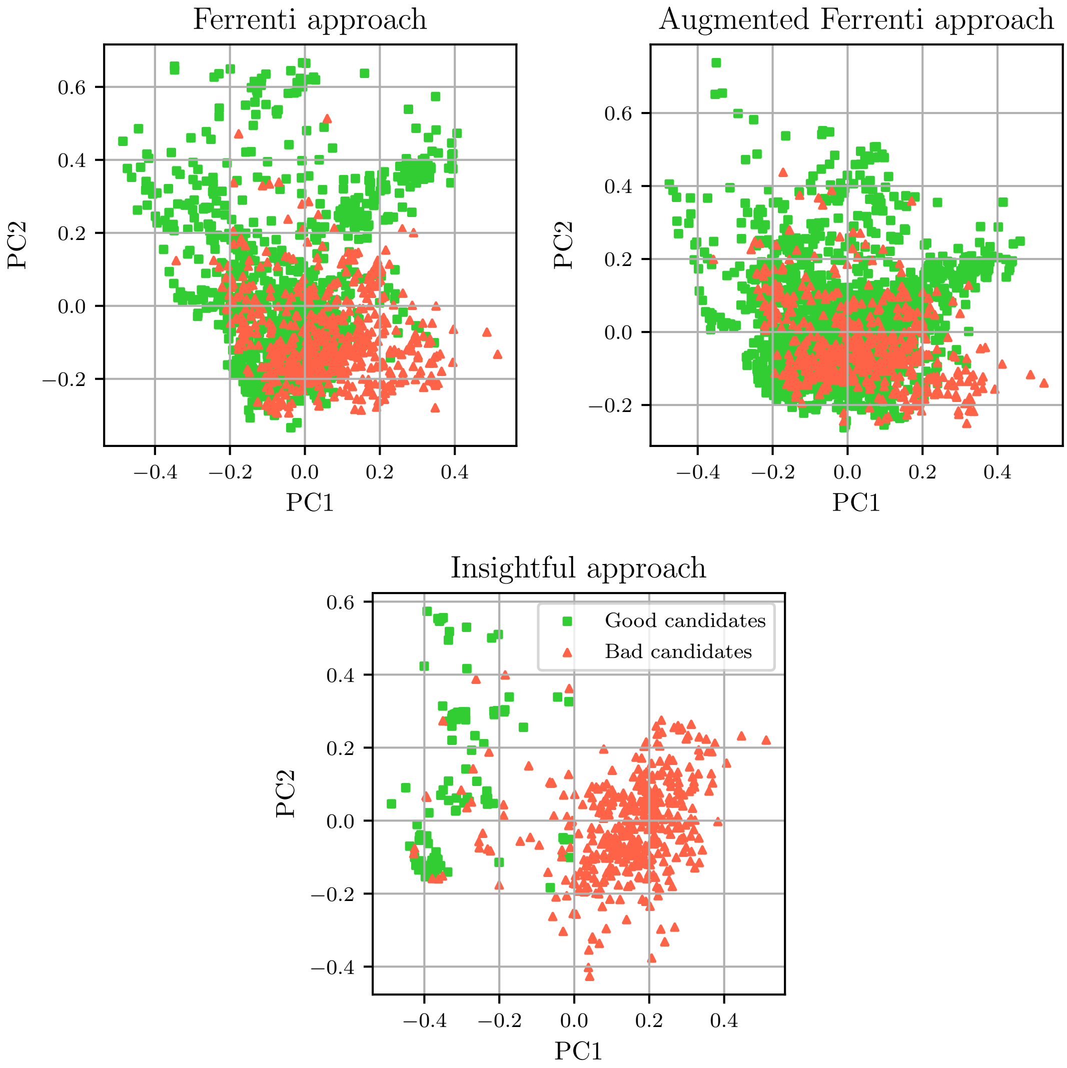


Figure 3.4: Two-dimensional scatter plots for the three different approaches. We have found the two eigenvectors corresponding to the two largest eigenvalues of the covariance-matrix, that is the two most important principal components, of the initial data from the Materials Project query. Then, we have transformed the three training sets resulting from the three approches and visualized it as a scatter plot for visualization purposes.

bad candidates are dressed in red triangles. Due to the simplicity of reducing the number of features down to 2 features, both good and bad candidates for the Ferrenti approach are overlapping which could be challenging for any model that would try to learn a clear-cut boundary. However, for the insightful approach, we can already start to see a trend where the upper left part of the figure is dominated by good candidates. Therefore, we can expect that the two Ferrenti approaches would need either supplementary dimensions for further distinguishment, or could be in trouble of finding a generalized model.

3.4 Model selection

After building a dataset through extraction, featurization and labelling, we turn our attention towards training a model. The flowchart is visualized in figure 3.5. After gathering and featurization, we achieve the interim data that goes through a final data preparation step to become preprocessed data. Then, we perform a data mining step using three different approaches as discussed in the last section. For each of the three approaches, we train and predict in the step called supervised learning. In the summary, we compare the different approaches and results.

In the data preparation step, we assess the quality of the data. Due to the large dimension of 25000 × 4500, we assume that there is a large amount of non-physical values present, therefore we fill all the missing values with zero and remove all columns with more than 70% containing only zeros. This value was chosen since all categorical features have at least 30 of an respective class present in the column, and a majority of the removed columns contained between 90% and 100% only zeros. This reduces the dimensionality substantially to only 679 features. It should be noted that other methods of data preparation resulted in equivalent preprocessed data due to the large amount of missing values in the data.

Four different supervised models has been selected for each of the three approaches defined in the previous section, resulting in a total of 12 unique models. As discussed in section **??**, models are unique and does not neccessarily perform optimal on all kinds of data. Therefore, the four models have been selected as a function of increasing complexity and ranges from the simplistic logistic regression and decision trees and up to random forest and gradient boost. We utilize the implementation of sklearn for all models [77].

Due to the fact that the current dimension of the entire dataset is still large, we apply the dimensionality reduction technique PCA to the dataset. This is benefical for several reasons, such as finding correlated features and reducing any dimensional. Additionally, it opens up for a visualized interpretation if we were to choose 3 or less principal components.

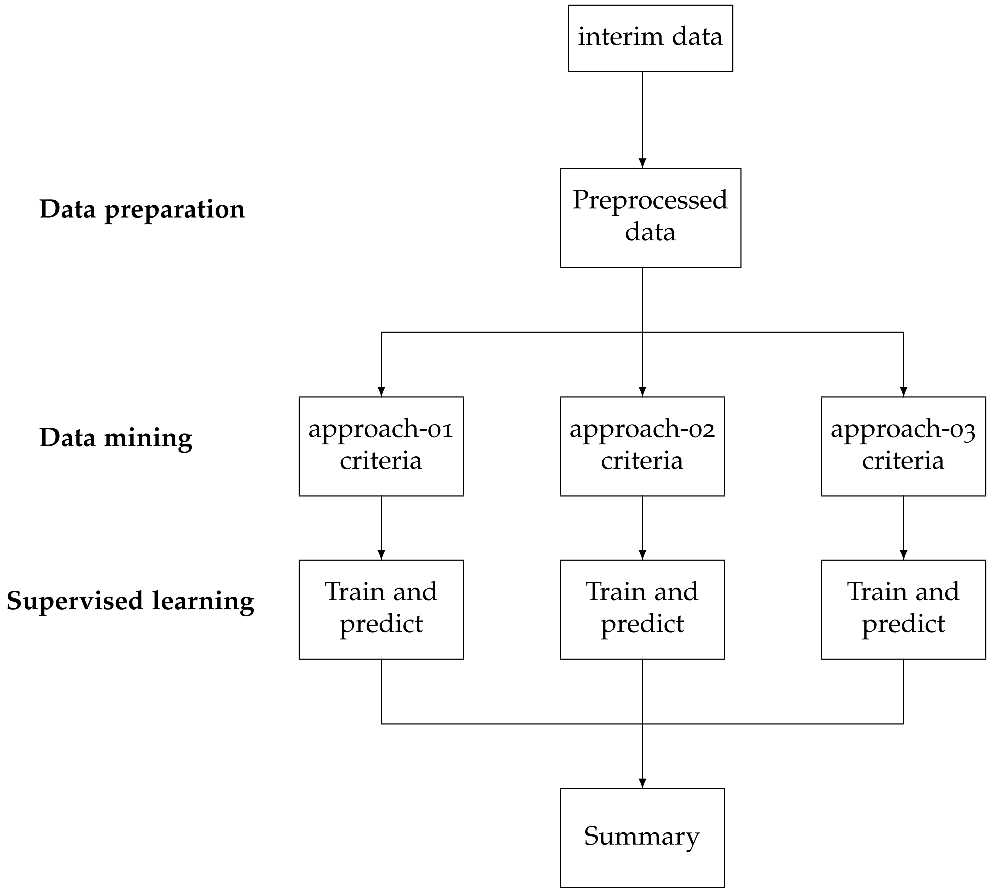


Figure 3.5: A continuity of the flowchart in figure 3.1 that visualise the steps of data preparation, the three approaches in the data mining step and the subsequent supervised learning. Finally, a summary will be provided. From a hierarchic perspective, we find the steps leading up to the preprocessed data as the top level, while each of the approaches are found one level down.

The optimal parameters are then searched for with the use of sklearn's gridsearch and imblearn's pipeline [77, 78]. Imblearn's pipeline enable the use of resampling methods, in contrast to sklearn's pipeline, but does not differ in any other way. In the pipeline, we provide a standardscaler that scales the data such that every feature will have a mean of 0 and a standard deviation of 1 [77]. Thereafter follows the dimensionality reduction and a supervised learning algorithm. It should be noted that the number of optimal principal components, up to an upper limit of accumulated explained variance of 95%, are also searched for in the grid-search scheme.

Part IV

Results and discussion

Chapter 4

Validation

A thorough testing procedure is important to find out if the code is working as intentionally. The procedure might reveal the presence or absence of bugs, and as a project grows, it can give an indication if a new implementation breaks the original project. Therefore, we present a test-case scenario to test if a few machine learning algorithms are able predict the correct label. It is the same algorithms that will be used in the following chapters, and it will provide us the opportunity to understand how the algorithm works and to draw parallells between the separate works. The entire work of the validation process can be found in the Github project *predicting-ABO3-structures* [79].

The validation process is a reproduction of Ref. [80]. To be able to draw any parallell to their work, we use the exact same dataset in the beginning phase. It should be noted that even if the computational aspects of the validation is closely related to Ref. [80], the work eventually diverges in terms of focus. In their work they include a stability analysis using convex hull analysis in DFT calculations from OQMD, however, we will in this thesis not decide whether a compound is considered stable or not in an atomic configuration.

4.1 The ABO3 dataset

The data used in the validation process is offered as supplimentary data from Ref. [80]. They provide the entire training data with both features and labels, but only provide the entries (compounds) of the test data. Therefore, it is neccessary to obtain the features for the test set ourself without knowledge if the resulting test set is identical with Ref. [80].

The training dataset in question contains 390 experimentally reported ABO_3 compounds. All compounds are charged balanced, and for every compound there is a feature explaining which structure the compound takes, either being a cubic perovskite, perovskite, or not a perovskite at all. Off the 390 compounds, there are 254 perovskites and 136 non-perovskites. Of the 254 perovskites, 232 takes a non-cubic perovskite structure while only 22 takes the cubic perovskite structure. Consequently, this will be visualized by two columns named Perovskite, which represents if a compound is either perovskite (1) or not perovskite (-1), and Cubic, which represents if a compound is cubic perovskite (1), non-cubic perovskite (-1), or not perovskite(0).

The original training dataset consists of 41 unique A atoms and 55 unique B atoms. To generate the test set, we implement all different combinations that are eligible with a total of (VI) oxidation number for the $A + B$ atoms. The resulting test data contains 625 entries and is considerable larger than the training data.

4.1.1 Features

There are in total 9 features we can train a model on. Many of the features are based on the Shannon ionic radii [81], which are estimates of an element's ionic hard-sphere radii extracted from experiment. They are dimensionless numbers, and are frequently used in studies involving perovskite structures of materials since they can be a measurement of the ionic misift of the B atom. This can be used to find the deviation of the structure from an ideal cubic geometry. The octahedral factor for an ABO_3 solid is known as

$$O = \frac{r_b}{r_O}, \tag{4.1}$$

where r_b and r_O are the Shannon radii for the B-atom and oxygen ($r_O = 1.4Å$), respectively. If the octahedral factor is $O = 0.435$, it corresponds to a hard-sphere closed-packed arrangement where B and O ions are touching, while a six-fold coordination appear to require $0.414 < O < 0.732$ according to empirical studies [82]. O, r_A and r_b are represented as features in our data set. We can also compute the Goldschmidt tolerance factor [83], which is defined as

$$t = \frac{r_A + r_O}{\sqrt{2}(r_A + r_O)}. \tag{4.2}$$

The tolerance factor favors the following structures in the interval:

- $t > 1$: Hexagonal nonperovskite.
- $0.9 < t < 1.0$: Cubic perovskite.
- $0.75 < t < 0.9$: Orthorombic perovskite.
- $t < 0.75$: Not a perovskite.

If the tolerance factor is exactly $t = 1$, the structure is known as perfectly cubic and is free for any structural alterations.

Furthermore, the Shannon radii r_A and r_B can be directly correlated with the structure. Perovskites require $r_A > r_B$, and that A-atoms are in a 12-fold coordinated site if $r_A > 0.9\text{Å}$. A-atoms also occur in a sixfold coordinated site if $r_A < 0.8\text{Å}$ and $r_B > 0.7\text{Å}$.

From bond valence theory we can find the valence of an ion to be the sum of valences, that is

$$V_i\& = \sum_i \nu_{ij} \tag{4.3}$$

$$\& = \sum_i \frac{\exp(d_o - d_{ij})}{b}, \tag{4.4}$$

where d_{ij} is the bond length while d_o and b are parameters from experimental data. The bond length can be found from 4.4 given the general value $b = 1.4\text{Å}$ and d_o, that can be found from Zhang *et al.* database [82]. The valence of an ion is associated with its neighboring ions and the chemical bonds, and therefore the band length d_{AO} and d_{BO} are included in the data set.

The two last features originates from the Mendeleev numbers of Villars *et al.* [84] for the A- and B atom, MA and MB, respectively. The given values positions the elements in structurally similar groups. This means that he groups the elements in the following interval.

- s-block $\in \{1, 10\}$.
- Sc $= 11$.
- Y $= 12$.
- f-block $\in \{13, 42\}$.
- d-block $\in \{43, 66\}$.
- p-block $\in \{67, 10\}$.

The dataset and its features have been visualized in the parallel coordinate **[Inselberga]** figure 4.1, and reveals several trends already. We can observe that an entry's A atom should preferably have a small Mendeley number (MA) and a large bond length d_{AO}. Yet, perhaps the most clear trend is the tolerance factor that should be around 1. A parallel coordinate plot can easily show trends, but becomes harder to interpret for many features and entries with a growing amount of overlapping. The trend for t values becomes harder to interpret when comparing with the distribution of entries for t-values in figure 4.2. From the distribution we learn that there is an overlap of perovskites or not for tolerance factor values in the interval 0.8 to 1.0, but the label perovskite is in general preferred. Additionally, we see that the interval $q1, q3$ for the label (1) completely overlaps with the corresponding interval for non-perovskites (-1), with very few entries outside of the intervals. This is presumably due to easy labelling for entries that rest outside of the intervals, but the exclusion of entries could potentially alter any model due to not enough entries.

4.2 Implementation

The machine learning classifiers that we will utilize are logistic regression, random forest and gradient boost. The implementation is optimized for adding new algorithms from libraries such as sklearn [77] or imblearn [78] with only few lines of code. This is in particular visualized through the implementation of the current algorithms in code listing 4.2, since a special emphasis on reuse and simplicity of code is in focus of this project.

```
InsertAlgorithms = [
  LogisticRegression(),
  RandomForestClassifier(),
  GradientBoostClassifier()
]
InsertAbbreviations = [
  "LOG", "RF", "GB"
]
InsertPrettyNames = [
  "Logistic regression",
  "Random forest",
  "Gradient boost"
]
```

The predictions are divided into two parts; perovskite predictions and cubic perovskite predictions. We apply the standard scaler of sklearn [77] to the training data, followed up by a search of optimal hyperparameters using a 10x10-stratified cross-validation. This ensures that the percentage of perovskites (cubic perovskites) or not are the same in every subsample in a cross validation as it is in the entire dataset. This is not neccessarily important for the perovskites predictions due to 65/35% of perovskites or not, but becomes significant for the cubic case where the ratio of cubic perovskites or not are 91/9%.

4.3 Results and discussion

Utilising three different classifiers on two different tasks, starting with prediction of perovskite and then prediction of cubic-perovskite, yields in total six different models. Here we present sets of representative results obtained by these models.

4.3.1 Technical details on ML classifiers

We first consider the ML classification of known ABO_3 into perovskite or nonperovskites. A search for optimal hyperparameters using scikit-learn's grid search scheme [77] reveals the following table with best parameters.

Add table of how many of the training process continues, and also for test (but in next section).

Then we consider the ML classification of known perovskites into cubic perovskites and noncubic perovskites.

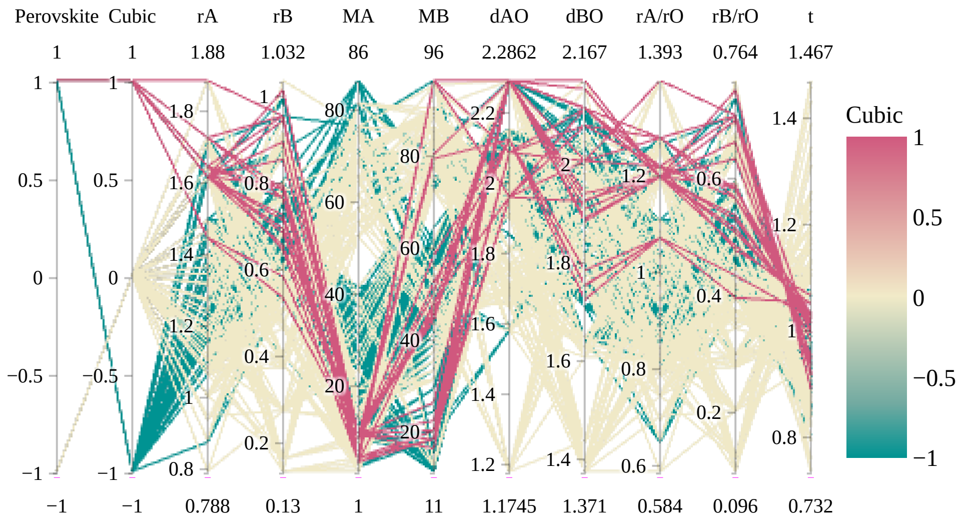


Figure 4.1: A parallel coordinate plot of the perovskite dataset, where the color is given by the Cubic label of an entry.

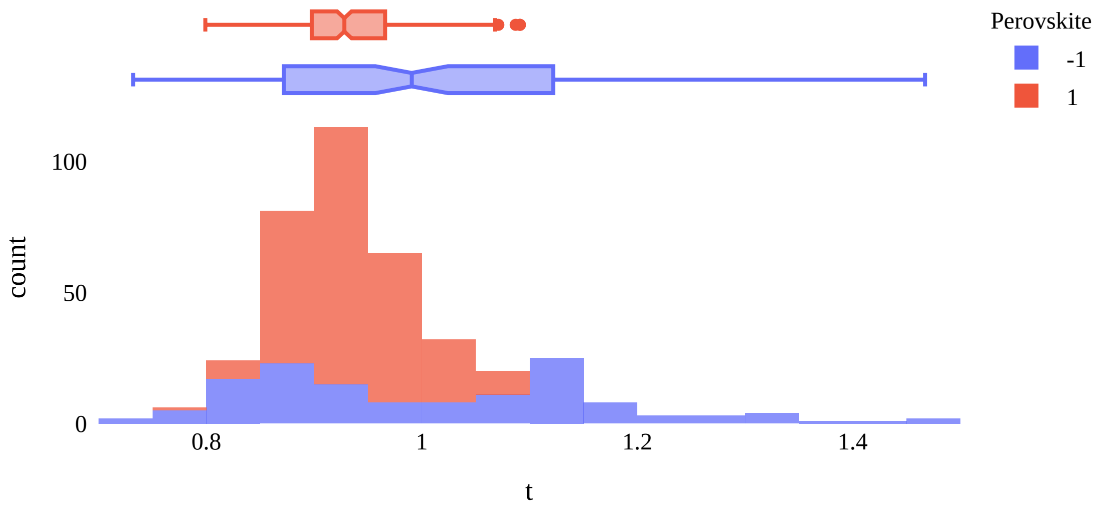
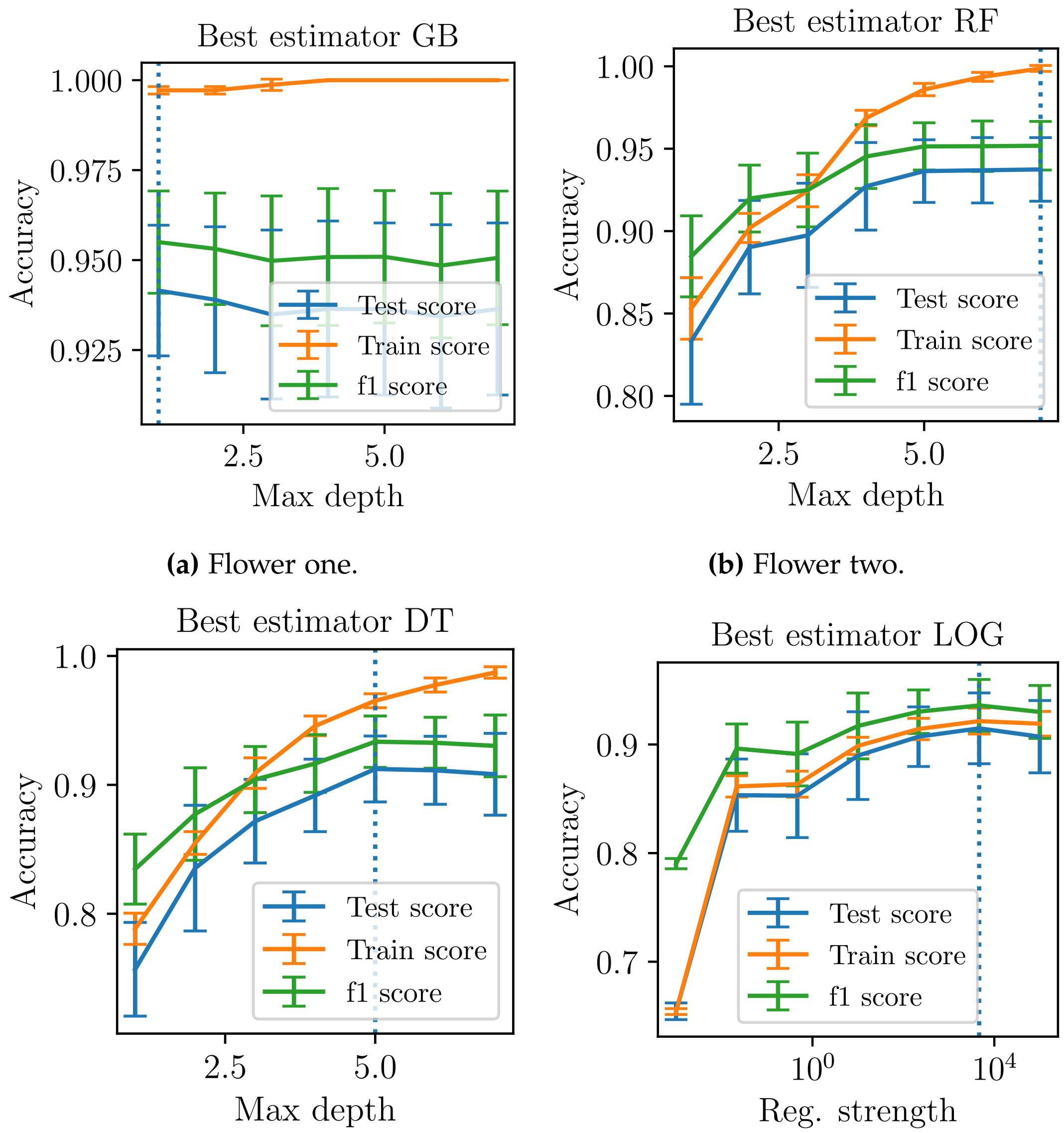


Figure 4.2: The t-distribution of entries in the dataset for perovskite (1) or not (-1). The upper part for perovskite (1) displays minimum value at 0.80, q1 at 0.90, median at 0.93, q3 at 0.97 and max at 1.10. For the non-perovskites (-1), the minimimum is at 0.73, q1 at 0.87, median at 0.99, q3 at 1.12 and max at 1.47.



(a) Flower one.

(b) Flower two.

(c) Flower one.

(d) Flower two.

Figure 4.3: Four figures displaying hyperparameter search for predicting perovskites or nonperovskites. The best estimator is visualized for all hyperparameters as a function of (a, b and c) max depth or (d) regularization strength during a grid search with a 5x5 stratified cross validation. The dotted lines marks the optimal hyperparameter-combination, while the error bars visualizes the standard deviation.

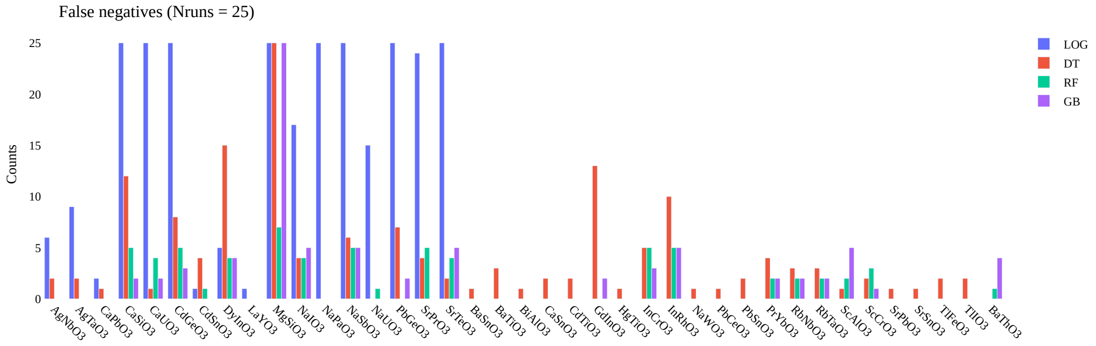


Figure 4.4: Dette kan vendes på 90 grader?

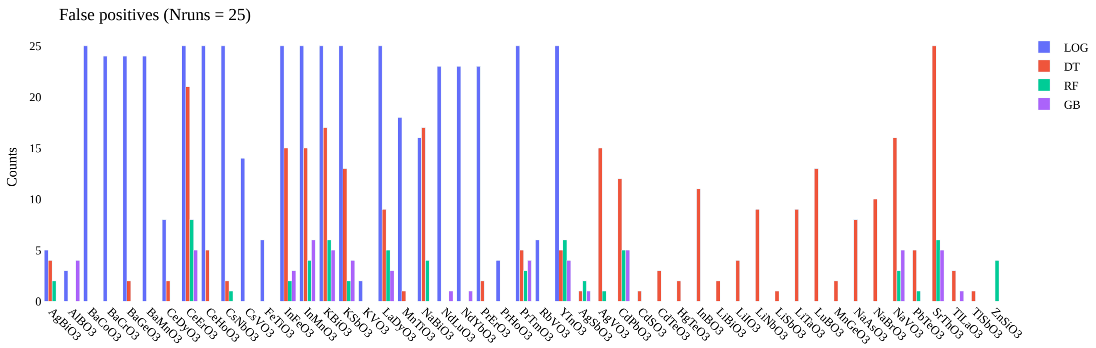
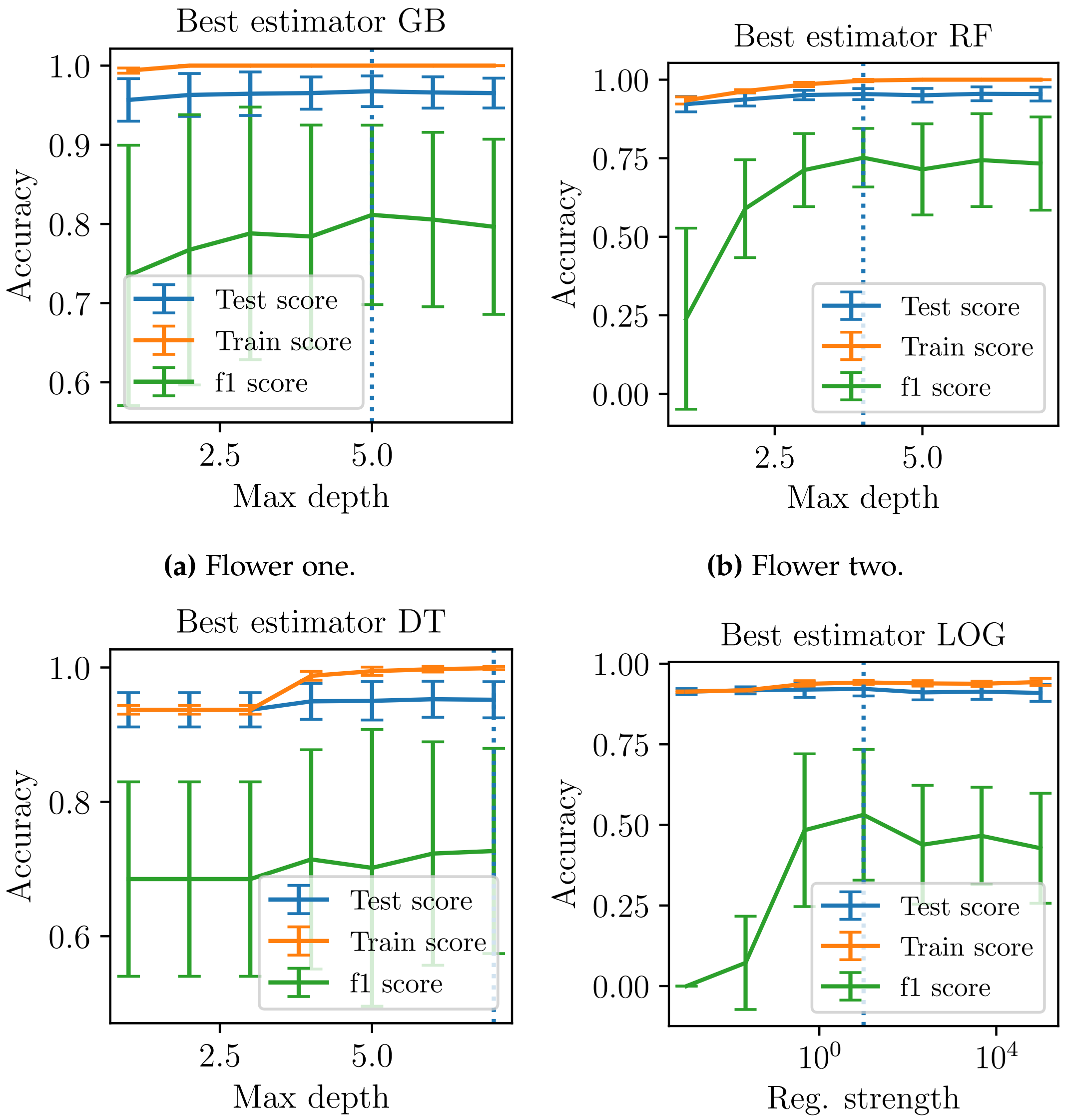


Figure 4.5: Dette kan vendes på 90 grader?



(a) Flower one.

(b) Flower two.

(c) Flower one.

(d) Flower two.

Figure 4.6: Four figures displaying hyperparameter search for predicting cubic perovskites or noncubic perovskites. The best estimator is visualized for all hyperparameters as a function of (a, b and c) max depth or (d) regularization strength during a grid search with a 5x5 stratified cross validation. The dotted lines marks the optimal hyperparameter-combination, while the error bars display the standard deviation.

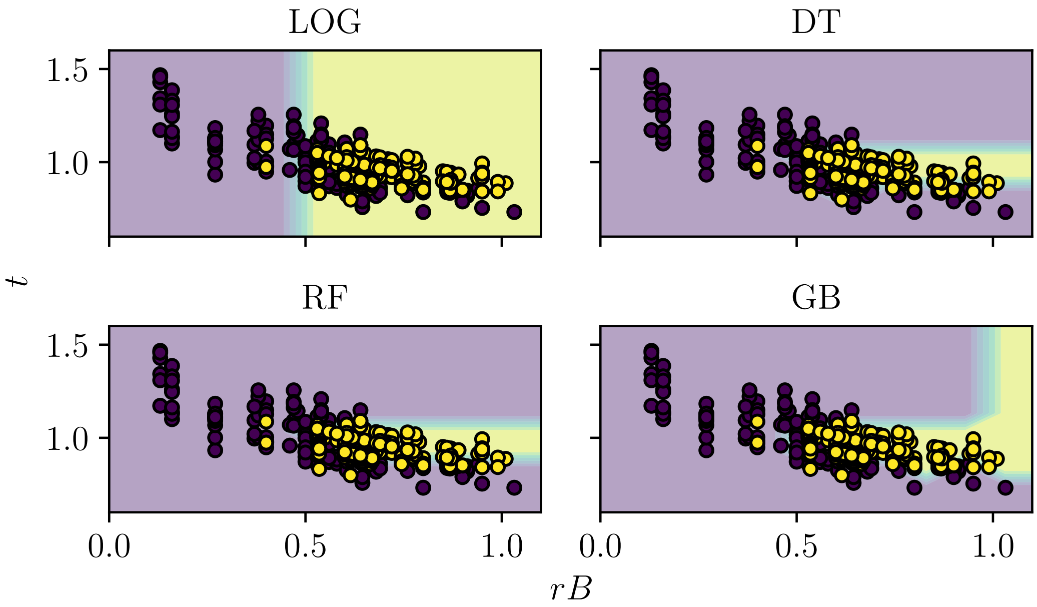


Figure 4.7: Decision boundaries of the four models from training on the feature pair r_B and t.

4.4 Predictions of new compounds

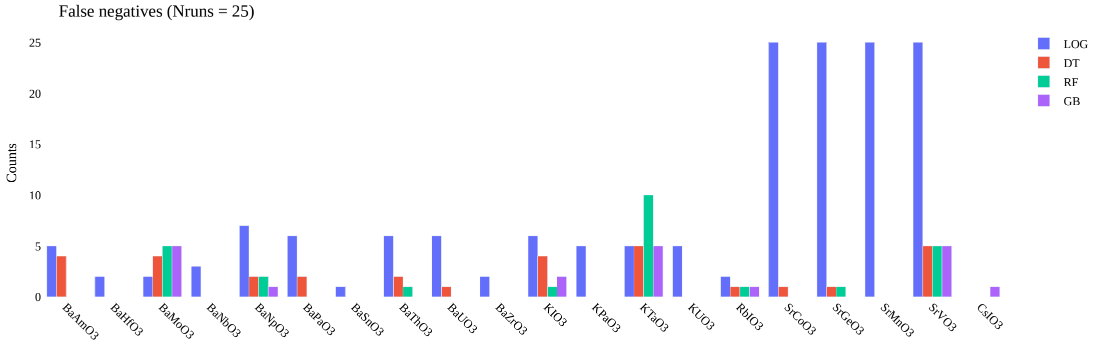


Figure 4.8: Dette kan vendes på 90 grader?

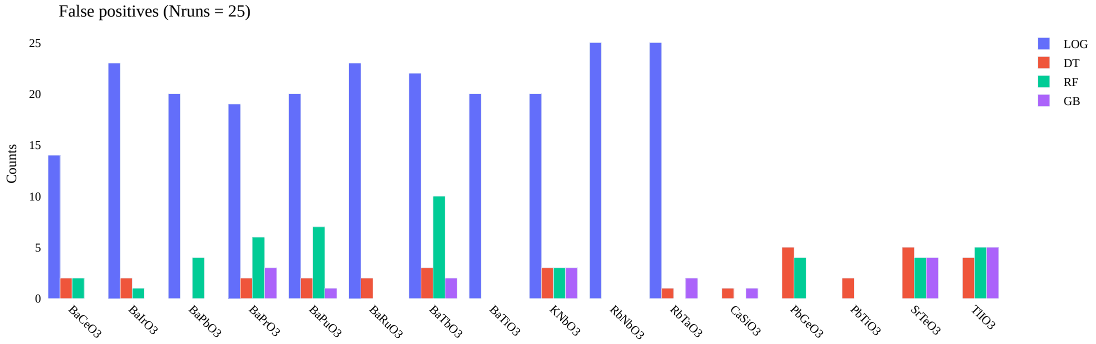


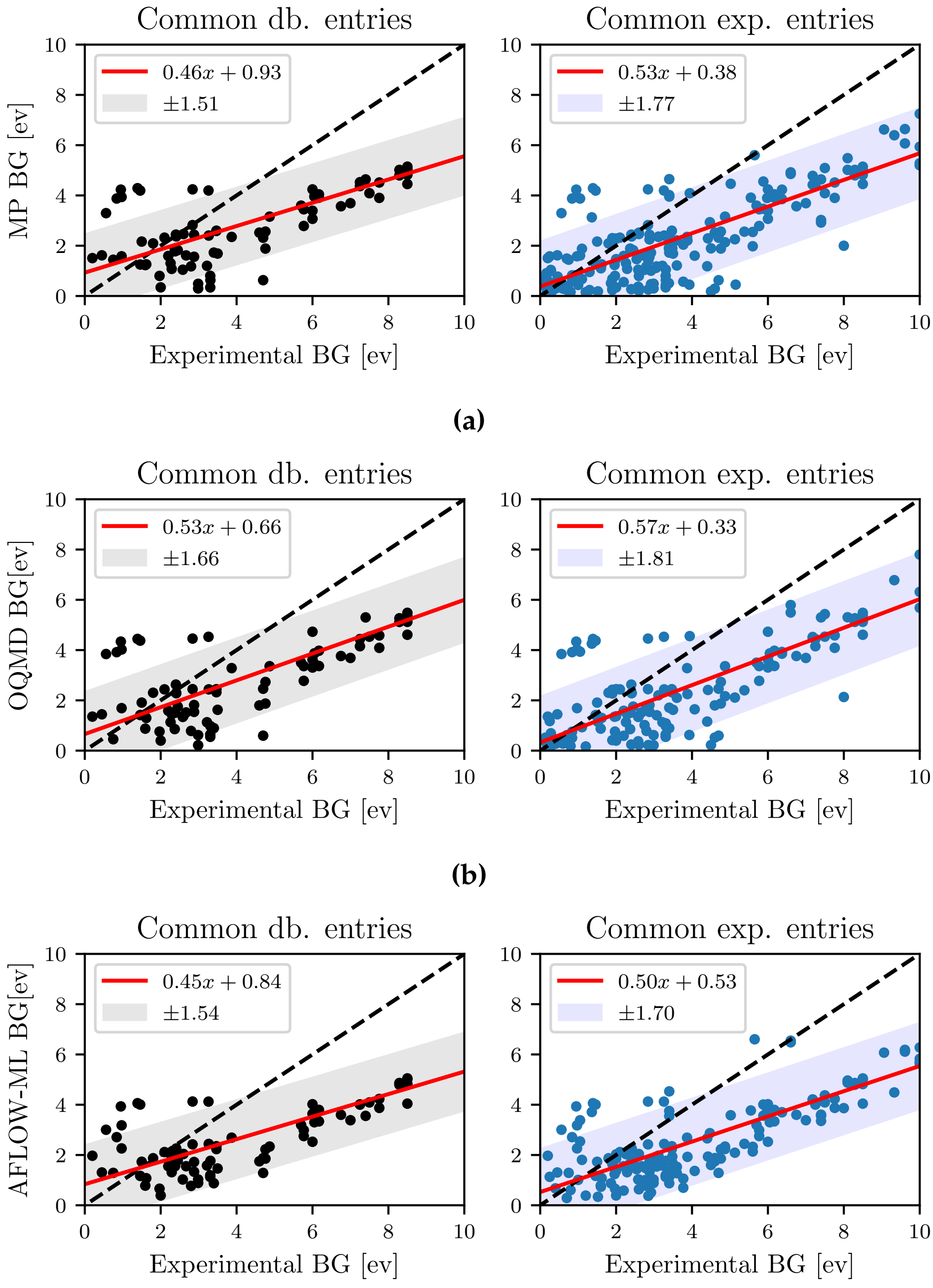
Figure 4.9: Dette kan vendes på 90 grader?

Chapter 5

Optimalization

This chapter is named optimalization due to its contents; here we will account for the choices we compose to optimize a potential machine learning algorithm. Initially, that involves finding what information is stored within the databases and the compromise of gathering the information, which further evolves to

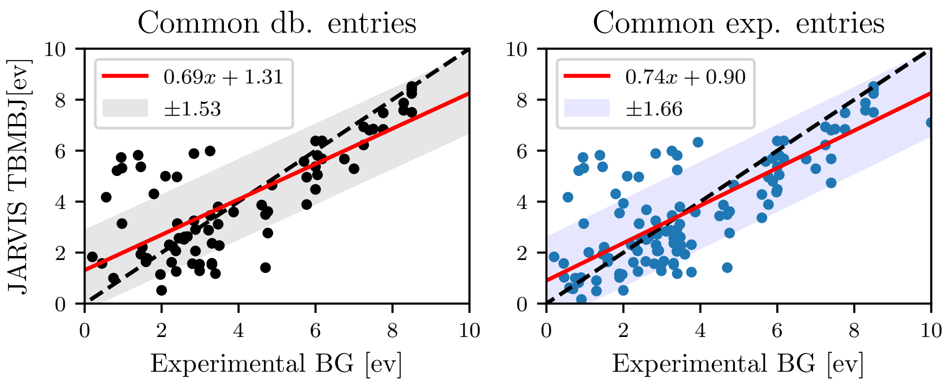
5.1 Comparing functionals for bandgaps


Common db. entries
Common exp. entries
MP BG [ev]
$0.46x + 0.93$
±1.51
$0.53x + 0.38$
±1.77
OQMD BG[ev]
$0.53x + 0.66$
±1.66
$0.57x + 0.33$
±1.81
AFLOW-ML BG[ev]
$0.45x + 0.84$
±1.54
$0.50x + 0.53$
±1.70
Experimental BG [ev]
0
2
4
6
8
10

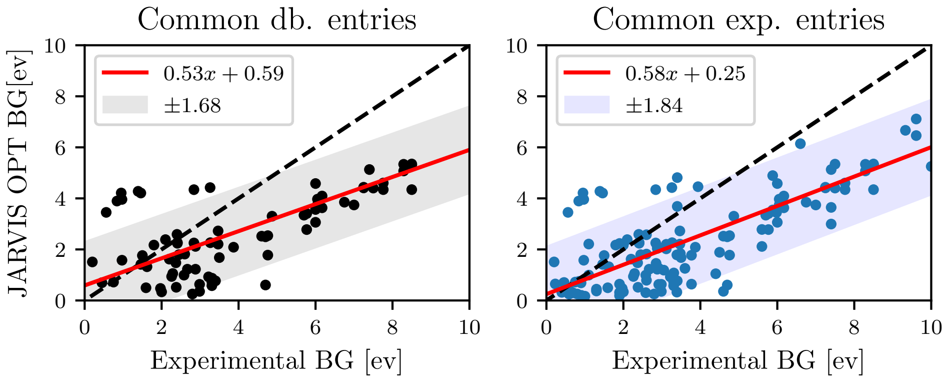
(a)

(b)

(c)


Common db. entries
Common exp. entries
JARVIS TBMBJ[ev]
Experimental BG [ev]
Experimental BG [ev]
$0.69x + 1.31$
±1.53
$0.74x + 0.90$
±1.66
0
2
4
6
8
10

(d)


Common db. entries
Common exp. entries
JARVIS OPT BG[ev]
Experimental BG [ev]
Experimental BG [ev]
$0.53x + 0.59$
±1.68
$0.58x + 0.25$
±1.84
0
2
4
6
8
10

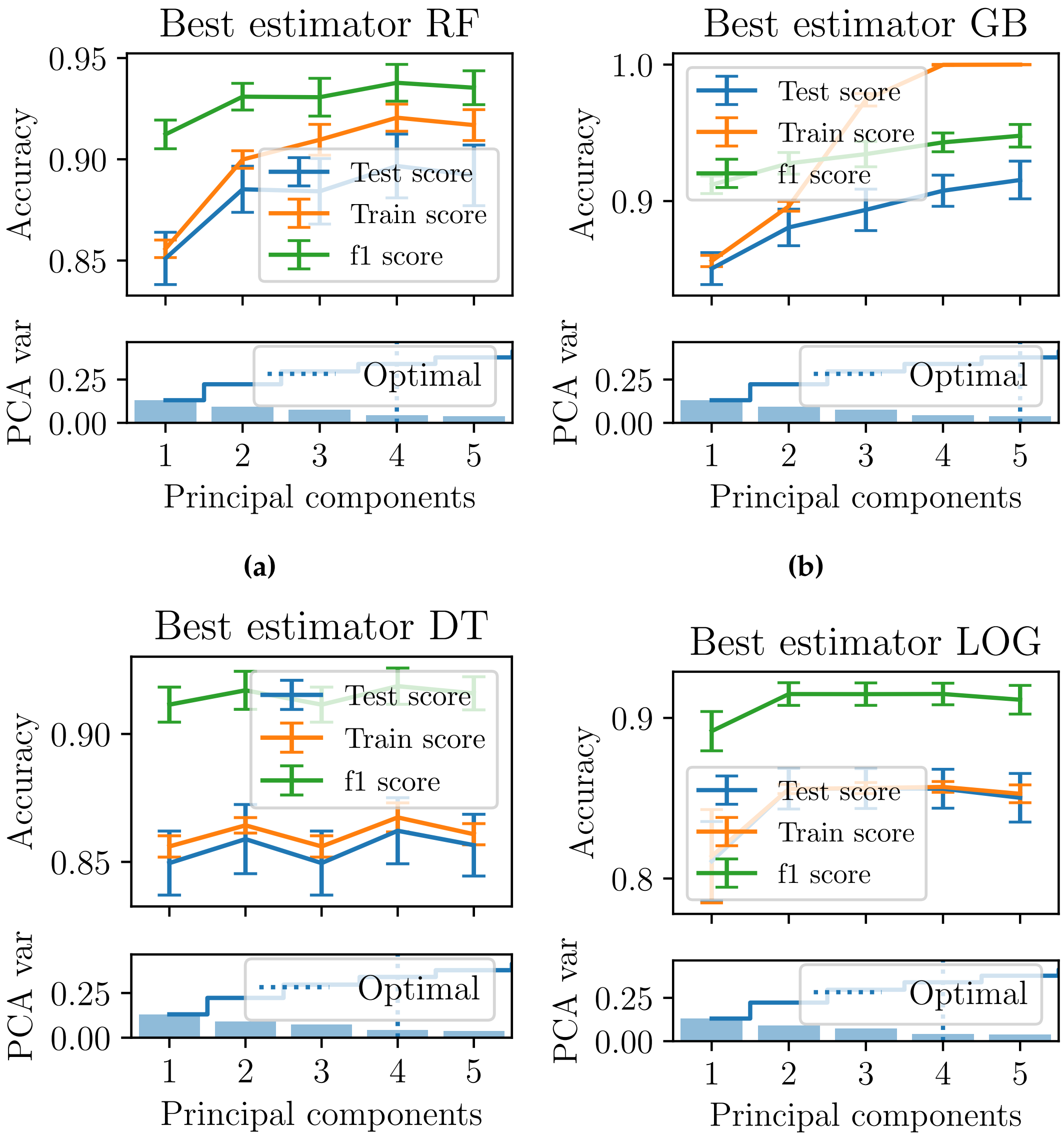
(e)

Figure 5.1

5.2 Technical details on ML classifiers

In this section we will provide technical details on the classifiers considering the training process. For each approach, we will apply combinations of principal components ranging from just one to several and look at the resulting implications. For each approach we can end up with over twenty different optimalization processes, which in total could potentially result in over sixty models in total. Therefore, we will not make an extensive analysis for every model, but emphasis important distinctions between the models and provide background for principal choices made. However, it should be noted that an an extensive automated analysis is distributed through the MIT license at the Github repository *predicting-solid-state-qubit-candidates* [62].

5.2.1 The Ferrenti approach



(a)

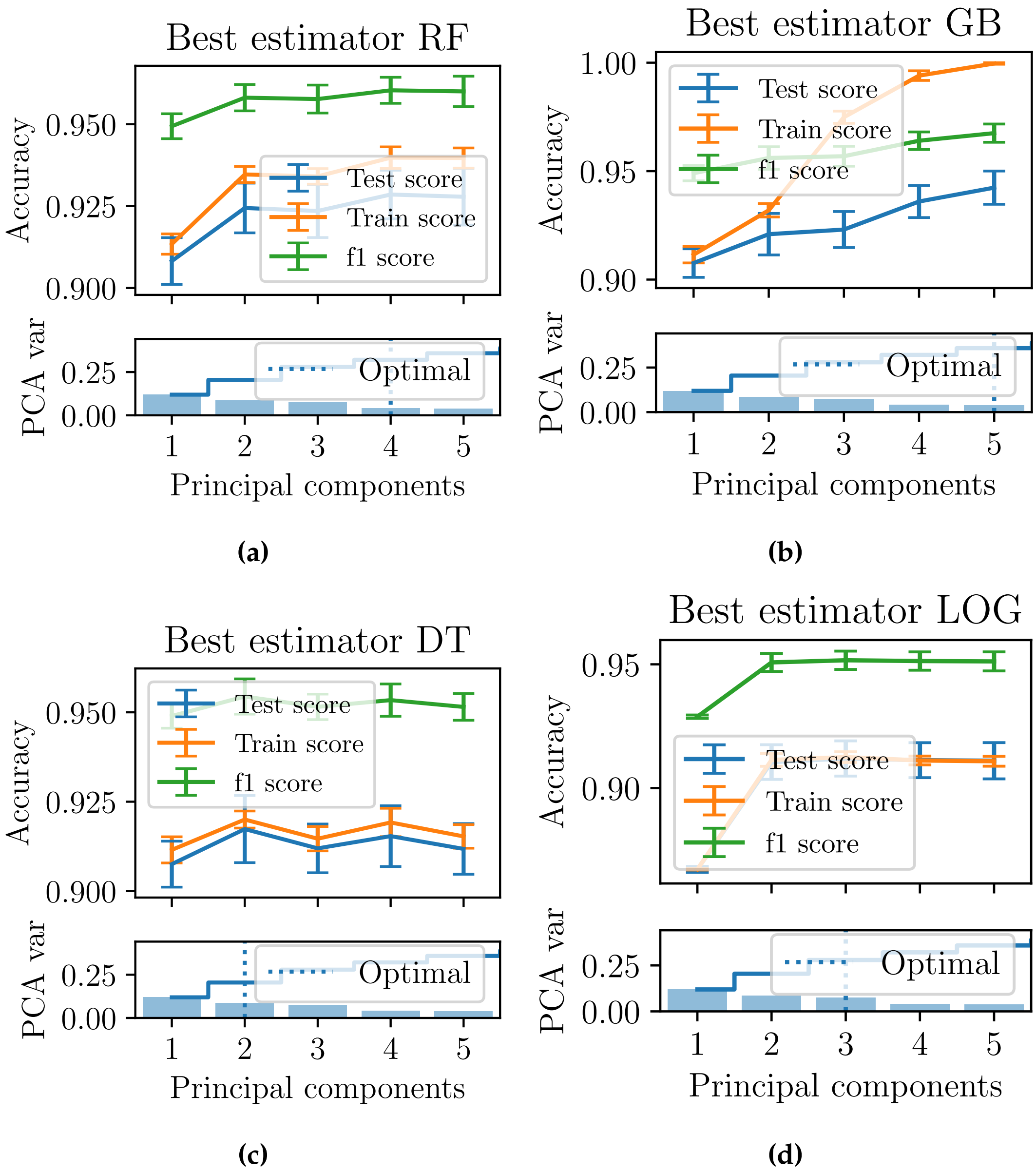
(b)

(c)

(d)

Figure 5.2: Four figures displaying hyperparameter search for the first approach. The best estimator is visualized for all hyperparameters as a function of principal components during a grid search with a 5x5 stratified cross validation. The lower plots visualizes the explained variance ratio, both accumulated and stepwise. The dotted lines marks the optimal hyperparameter-combination, while the error bars display the standard deviation.

5.2.2 The augmented Ferrenti approach



(a)

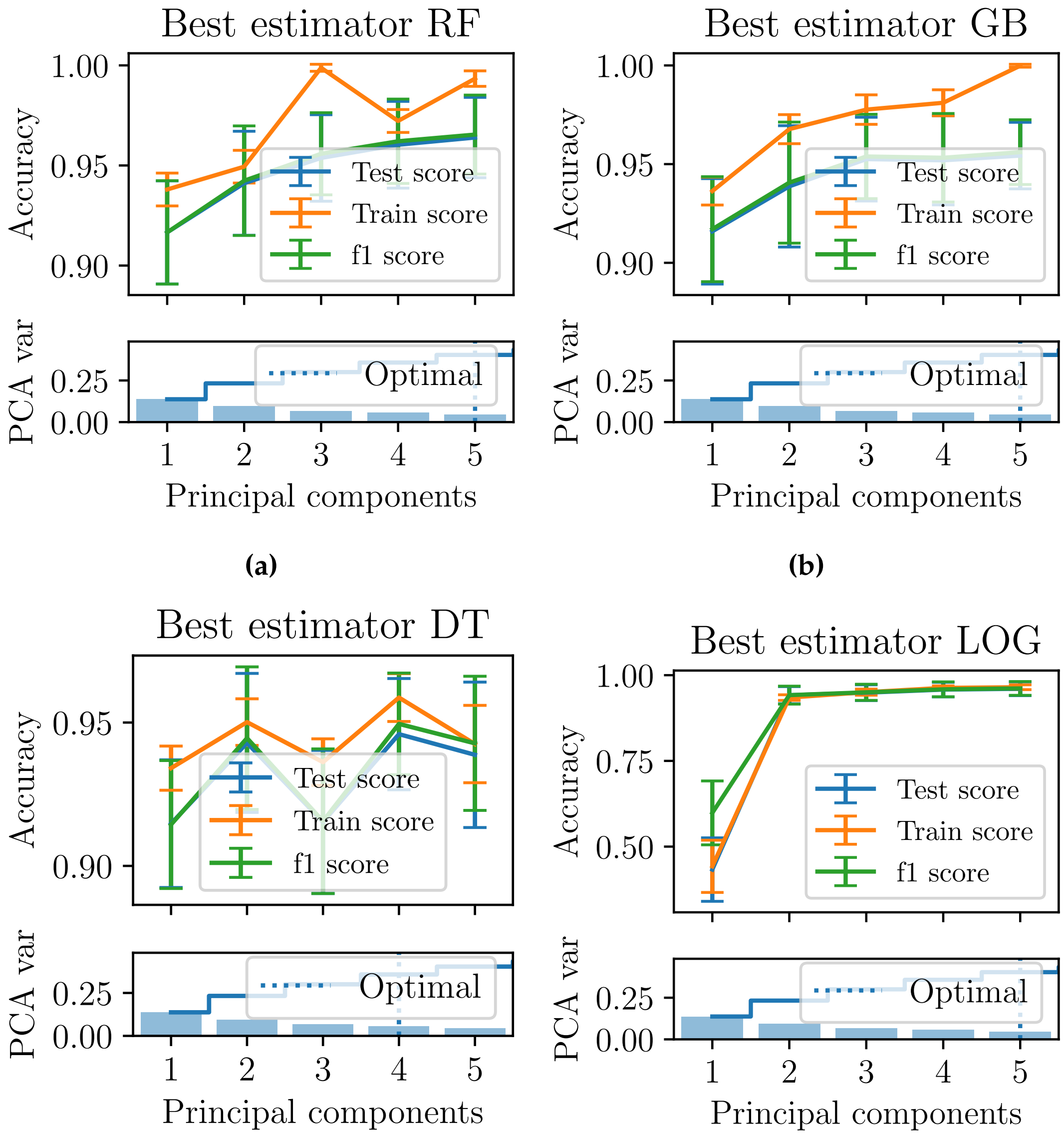
(b)

(c)

(d)

Figure 5.3: Four figures displaying hyperparameter search for the second approach. The best estimator is visualized for all hyperparameters as a function of principal components during a grid search with a 5x5 stratified cross validation. The lower plots visualizes the explained variance ratio, both accumulated and stepwise. The dotted lines marks the optimal hyperparameter-combination, while the error bars display the standard deviation.

5.2.3 The insightful approach



(a)

(b)

(c)

(d)

Figure 5.4: Four figures displaying hyperparameter search for the third approach. The best estimator is visualized for all hyperparameters as a function of principal components during a grid search with a 5x5 stratified cross validation. The lower plots visualizes the explained variance ratio, both accumulated and stepwise. The dotted lines marks the optimal hyperparameter-combination, while the error bars display the standard deviation.

Chapter 6

Predictions

6.1 The Ferrenti approach

6.2 The augmented Ferrenti approach

6.3 The insightful approach

Table 6.1: Table of the number of predictions made with the optimal model for the insightful approach.

Model	Optimal number PC	Number of predictions
Logistic regression	145	454
Decision trees	3	442
Random forest	10	325
Gradient boost	7	699

Part V

Appendices

Appendix A

Extracting information from ab-initio calculations

A.1 The Born-Oppenheimer approximation

The many-particle eigenfunction describes the wavefunction of all the electrons and nuclei and we denote it as Ψ_κ^{en} for electrons (e) and nuclei (n), respectively. The Born-oppenheimer approximation assumes that nuclei, of substantially larger mass than electrons, can be treated as fixed point charges. According to this assumption, we can separate the eigenfunction into an electronic part and a nuclear part,

$$\Psi_\kappa^{en}(\mathbf{r}, \mathbf{R}) \approx \Psi_\kappa(\mathbf{r}, \mathbf{R})\Theta_\kappa(\mathbf{R}), \tag{A.1}$$

where the electronic part is dependent on the nuclei. This is in accordance with the assumption above, since electrons can respond instantaneously to a new position of the much slower nucleus, but this is not true for the opposite scenario. To our advantage, we already have knowledge of the terms in the many-particle Hamiltionian, and we can begin by separating the Hamiltionian into electronic and nuclear parts:

$$\hat{H}^{en} = \overbrace{T_e + U_{ee} + U_{en}}^{\hat{H}^e} + \overbrace{T_n + U_{nn}}^{\hat{H}^n}. \tag{A.2}$$

Starting from the Schrödinger equation, we can formulate separate expressions for the electronic and the nuclear Schrödinger equations.

$$\hat{H^{en}}\Psi^{en}_{\kappa}(\mathbf{r},\mathbf{R}) = E^{en}_{\kappa}\Psi^{en}_{\kappa}(\mathbf{r},\mathbf{R}) \quad |\times \int \Psi^{*}(\mathbf{r},\mathbf{R})\mathrm{d}\mathbf{r} \tag{A.3}$$

$$\int \Psi^{*}_{\kappa}(\mathbf{r},\mathbf{R})(\hat{H}^{e}+\hat{H}^{n})\Psi_{\kappa}(\mathbf{r},\mathbf{R})\Theta_{\kappa}(\mathbf{R})\mathrm{d}\mathbf{r} = E^{en}_{\kappa}\underbrace{\int \Psi^{*}_{\kappa}(\mathbf{r},\mathbf{R})\Psi_{\kappa}(\mathbf{r},\mathbf{R})\mathrm{d}\mathbf{r}}_{1}\Theta_{\kappa}(\mathbf{R}). \tag{A.4}$$

Since $\Theta_{\kappa}(\mathbf{R})$ is independent of the the spatial coordinates to electrons, we get E_{κ} as the total energy of the electrons in the state κ.

$$E_{\kappa}(\mathbf{R})\Theta_{k}(\mathbf{R}) + \int \Psi^{*}_{k}(\mathbf{r},\mathbf{R})H^{n}\Psi_{k}(\mathbf{r},\mathbf{R})\Theta_{k}(\mathbf{R})\mathrm{d}\mathbf{r} = E^{en}_{k}\Theta_{k}(\mathbf{R}). \tag{A.5}$$

Now, the final integration term can be simplified by using the product rule, which results in

$$\Big(T_{n} + T_{n}^{'} + T_{n}^{''} + U_{nn} + E_{\kappa}(\mathbf{R})\Big)\Theta_{\kappa}(\mathbf{R}) = E^{en}_{\kappa}\Theta_{\kappa}(\mathbf{R}). \tag{A.6}$$

If we neglect T_{n}' and $T_{n}^{''}$ to lower the computational efforts, we obtain the Born-Oppenheimer approximation with the electronic eigenfunction as

$$(T_{e} + U_{ee} + U_{en})\,\Psi_{\kappa}(\mathbf{r},\mathbf{R}) = E_{\kappa}(\mathbf{R})\Psi_{\kappa}(\mathbf{r},\mathbf{R}) \tag{A.7}$$

and the nuclear eigenfunction as

$$\big(T_{n} + U_{nn} + E_{\kappa}(\mathbf{R})\big)\,\Theta_{\kappa}(\mathbf{R}) = E^{en}_{\kappa}(\mathbf{R})\Theta_{\kappa}(\mathbf{r},\mathbf{R}). \tag{A.8}$$

How are they coupled, you might ask? The total energy in the electronic equation is a potential in the nuclear equation.

A.2 The variational principle

So far, we have tried to make the time-independent Schrödinger equation easier with the use of an *ansatz*, but we do not neccessarily have an adequate guess for the eigenfunctions and the ansatz can only give a rough estimate in most scenarios. Another approach, namely the *variational principle*, states that the energy of any trial wavefunction is always an upper bound to the exact ground state energy by definition E_{0}.

$$E_{0} = \langle\psi_{0}|\,H\,|\psi_{0}\rangle \leqslant \langle\psi|\,H\,|\psi\rangle = E \tag{A.9}$$

The eigenfunctions of H form a complete set, which means any normalized Ψ can be expressed in terms of the eigenstates

$$\Psi = \sum_n c_n \psi_n, \quad \text{where} \quad H\psi_n = E_n \psi_n \tag{A.10}$$

for all $n = 1, 2, ...$ The expectation value for the energy can be calculated as

$$\begin{aligned}
\langle \Psi | H | \Psi \rangle &= \left\langle \sum_n c_n \psi_n \middle| H \middle| \sum_{n'} c_{n'} \psi_{n'} \right\rangle \\
&= \sum_n \sum_{n'} c_n^* c_{n'} \langle \psi_n | H | \psi_{n'} \rangle \\
&= \sum_n \sum_{n'} c_n^* E_n c_{n'} \langle \psi_n | \psi_{n'} \rangle
\end{aligned}$$

Here we assume that the eigenfunctions have been orthonormalized and we can utilize $\langle \psi_m | \psi_n \rangle = \delta_{mn}$, resulting in

$$\sum_n c_n^* c_n E_n = \sum_n |c_n|^2 E_n.$$

We have already stated that Ψ is normalized, thus $\sum_n |c_n|^2 = 1$, and the expectation value conveniently is bound to follow equation A.9. The quest to understand the variational principle can be summarized in a sentence - it is possible to tweak the wavefunction parameters to minimize the energy, or summed up in a mathematical phrase,

$$E_0 = \min_{\Psi \to \Psi_0} \langle \Psi | H | \Psi \rangle. \tag{A.11}$$

A.3 The Hohenberg-Kohn theorems

A.3.1 The Hohenberg-Kohn theorem 1

Proof. Assume that two external potentials $V_{ext}^{(1)}$ and $V_{ext}^{(2)}$, that differ by more than a constant, have the same ground state density $n_0(r)$. The two different potentials correspond to distinct Hamiltonians $\hat{H}_{ext}^{(1)}$ and $\hat{H}_{ext}^{(2)}$, which again give rise to distinct wavefunctions $\Psi_{ext}^{(1)}$ and $\Psi_{ext}^{(2)}$. Utilizing the variational principle, we find that no wavefunction can give an energy that is less than the energy of $\Psi_{ext}^{(1)}$ for $\hat{H}_{ext}^{(1)}$, that is

$$E^{(1)} = \left\langle \Psi^{(1)} \middle| \hat{H}^{(1)} \middle| \Psi^{(1)} \right\rangle < \left\langle \Psi^{(2)} \middle| \hat{H}^{(1)} \middle| \Psi^{(2)} \right\rangle \tag{A.12}$$

and

$$E^{(2)} = \left\langle \Psi^{(2)} \middle| \hat{H}^{(2)} \middle| \Psi^{(2)} \right\rangle < \left\langle \Psi^{(1)} \middle| \hat{H}^{(2)} \middle| \Psi^{(1)} \right\rangle . \tag{A.13}$$

Assuming that the ground state is not degenerate, the inequality strictly holds. Since we have identical ground state densities for the two Hamiltonian's, we can rewrite the expectation value for equation A.12 as

$$\begin{aligned}
E^{(1)} &= \left\langle \Psi^{(1)} \middle| \hat{H}^{(1)} \middle| \Psi^{(1)} \right\rangle \\
&= \left\langle \Psi^{(1)} \middle| T + U_{ee} + U_{ext}^{(1)} \middle| \Psi^{(1)} \right\rangle \\
&= \left\langle \Psi^{(1)} \middle| T + U_{ee} \middle| \Psi^{(1)} \right\rangle + \int \Psi^{*(1)}(\mathbf{r}) V_{ext}^{(1)} \Psi^{(1)}(\mathbf{r}) d\mathbf{r} \\
&= \left\langle \Psi^{(1)} \middle| T + U_{ee} \middle| \Psi^{(1)} \right\rangle + \int V_{ext}^{(1)} n(\mathbf{r}) d\mathbf{r} \\
&< \left\langle \Psi^{(2)} \middle| \hat{H}^{(1)} \middle| \Psi^{(2)} \right\rangle \\
&= \left\langle \Psi^{(2)} \middle| T + U_{ee} + U_{ext}^{(1)} + \overbrace{U_{ext}^{(2)} - U_{ext}^{(2)}}^{0} \middle| \Psi^{(2)} \right\rangle \\
&= \left\langle \Psi^{(2)} \middle| T + U_{ee} + U_{ext}^{(2)} \middle| \Psi^{(1)} \right\rangle + \int \left(V_{ext}^{(1)} - V_{ext}^{(2)} \right) n(\mathbf{r}) d\mathbf{r} \\
&= E^{(2)} + \int \left(V_{ext}^{(1)} - V_{ext}^{(2)} \right) n(\mathbf{r}) d\mathbf{r}.
\end{aligned}$$

Thus,

$$E^{(1)} = E^{(2)} + \int \left(V_{ext}^{(1)} - V_{ext}^{(2)} \right) n(\mathbf{r}) d\mathbf{r} \tag{A.14}$$

A similar procedure can be performed for $E^{(2)}$ in equation A.13, resulting in

$$E^{(2)} = E^{(1)} + \int \left(V_{ext}^{(2)} - V_{ext}^{(1)} \right) n(\mathbf{r}) d\mathbf{r}. \tag{A.15}$$

If we add these two equations together, we get

$$\begin{aligned}
E^{(1)} + E^{(2)} &< E^{(2)} + E^{(1)} + \int \left(V_{ext}^{(1)} - V_{ext}^{(2)} n(\mathbf{r}) d\mathbf{r} \right) \\
&\qquad + \int \left(V_{ext}^{(2)} - V_{ext}^{(1)} n(\mathbf{r}) d\mathbf{r} \right) \\
E^{(1)} + E^{(2)} &< E^{(2)} + E^{(1)},
\end{aligned} \tag{A.16}$$

which is a contradiction. Thus, the two external potentials cannot have the same ground-state density, and $V_{ext}(\mathbf{r})$ is determined uniquely (except for a constant) by $n(\mathbf{r})$. □

A.3.2 The Hohenberg-Kohn theorem 2

Proof. Since the external potential is uniquely determined by the density and since the potential in turn uniquely determines the ground state wavefunction (except in degenerate situations), all the other observables of the system are uniquely determined. Then the energy can be expressed as a functional of the density.

$$E[n] = \overbrace{T[n] + U_{ee}[n]}^{F[n]} + \overbrace{U_{en}[n]}^{\int V_{en} n(r) dr} \tag{A.17}$$

where $F[n]$ is a universal functional because the treatment of the kinetic and internal potential energies are the same for all systems, however, it is most commonly known as the Hohenberg-Kohn functional.

In the ground state, the energy is defined by the unique ground-state density $n_0(r)$,

$$E_0 = E[n_0] = \langle \Psi_0 | H | \Psi_0 \rangle . \tag{A.18}$$

From the variational principle, a different density $n(r)$ will give a higher energy

$$E_0 = E[n_0] = \langle \Psi_0 | H | \Psi_0 \rangle < \langle \Psi | H | \Psi \rangle = E[n] \tag{A.19}$$

Thus, the total energy is minimized for n_0, and so has to be the ground-state energy. □

A.4 Self-consistent field methods

So, the remaining question is, how do we solve the Kohn-Sham equation? First, we would need to define the Hartree potential, which can be found if we know the electron density. The electron density can be found from the single-electron wave-functions, however, these can only be found from solving the Kohn-Sham equation. This *circle of life* has to start somewhere, but where? The process can be defined as an iterative method, *a computational scheme*, as visualized in figure A.1.

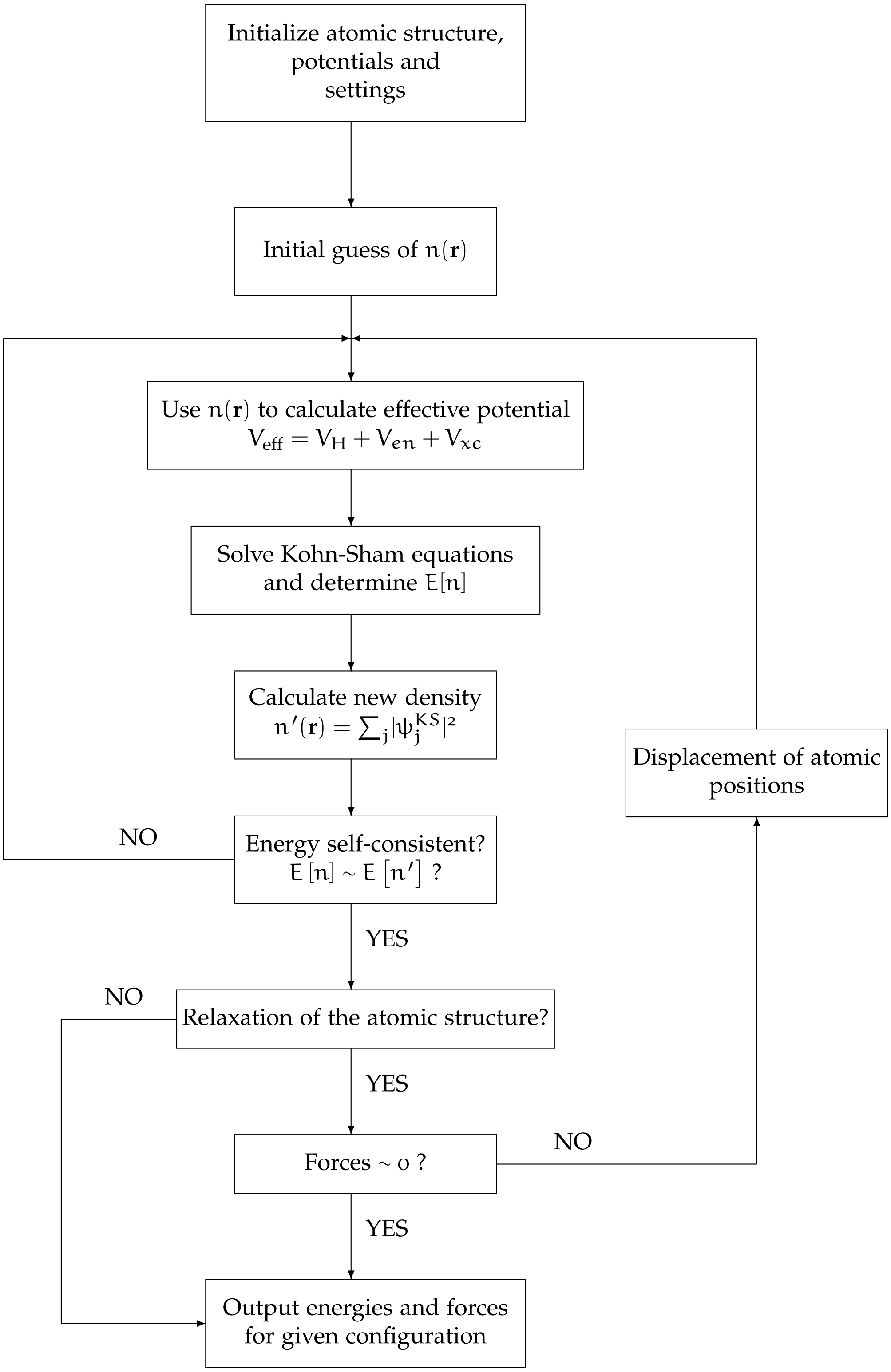

Initialize atomic structure, potentials and settings
Initial guess of $n(\mathbf{r})$
Use $n(\mathbf{r})$ to calculate effective potential $V_{\text{eff}} = V_H + V_{en} + V_{xc}$
Solve Kohn-Sham equations and determine $E[n]$
Calculate new density $n'(\mathbf{r}) = \sum_j |\psi_j^{KS}|^2$
Energy self-consistent? $E[n] \sim E[n']$?
NO
YES
Relaxation of the atomic structure?
NO
YES
Forces ~ 0 ?
NO
YES
Displacement of atomic positions
Output energies and forces for given configuration

Figure A.1: A flow chart of the self-consistent field method for DFT.

Appendix B

Featurizaton

B.1 Table of featurizers

Table B.1: This thesis' chosen 39 featurizers from matminer. Descriptions are either found from Ref. [85] or from the project's Github page.

Features	Description	Original reference
Composition features		
AtomicOrbitals	Highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO).	[86]
AtomicPacking-Efficiency	Packing efficiency.	[87]
BandCenter	Estimation of absolute position of band center using geometric mean of electronegativity.	[88]
ElementFraction	Fraction of each element in a composition.	-
ElementProperty	Statistics of various element properties.	[89–91]
IonProperty	Maximum and average ionic character.	[90]

Continued on next page

Table B.1 – continued from previous page

Features	Description	Original reference
Miedema	Formation enthalpies of intermetallic compounds, solid solutions, and amorphous phases using semi-empirical Miedema model.	[92]
Stoichiometry	L^p norm-based stoichiometric attributes.	[90]
TMetalFraction	Fraction of magnetic transition metals.	[91]
ValenceOrbital	Valence orbital attributes such as the mean number of electrons in each shell.	[90]
YangSolid-Solution	Mixing thermochemistry and size mismatch terms.	[93]
Oxid composition features		
Electronegativity-Diff	Statistics on electronegativity difference between anions and cations.	[91]
OxidationStates	Statistics of oxidation states.	[91]
Structure features		
DensityFeatures	Calculate density, volume per atom and packing fraction.	-
GlobalSymmetry-Features	Determines spacegroup number, crystal system (1-7) and inversion symmetry.	-
RadialDistribution-Function	Calculates the radial distribution function of a crystal system.	-
CoulombMatrix	Generate the Coulomb matrix, which is a representation of the nuclear coulombic interaction of the input structure.	[94]
PartialRadial-Distribution-Function	Compute the partial radial distribution function of a crystal structure	[95]

Continued on next page

Table B.1 – continued from previous page

Features	Description	Original reference
SineCoulombMatrix	Computes a variant of the coulomb matrix developed for periodic crystals.	[96]
EwaldEnergy	Computes the energy from Coulombic interactions based on charge states of each site.	[97]
BondFractions	Compute the fraction of each bond in a structure, based on nearest neighbours.	[98]
StructuralHeterogeneity	Calculates the variance in bond lengths and atomic volumes in a structure.	[99]
MaximumPackingEfficiency	Calculates the maximum packing efficiency of a structure.	[99]
ChemicalOrdering	Computes how much the ordering of species differs from random in a structure.	[99]
XRDPowderPattern	1D array representing normalized powder diffraction of a structure as calculated by pymatgen.	[89]
Site features		
AGNIFingerprints	Calculates the product integral of RDF and Gaussian window function	[100]
AverageBondAngle	Determines the average bond angle of a specific site with its nearest neighbors using pymatgens implementation.	[101]
AverageBondLength	Determines the average bond length between one specific site and all its nearest neighbors using pymatgens implementation.	[101]
BondOrientationalParamater	Calculates the averages of spherical harmonics of local neighbors	[102, 103]

Continued on next page

Table B.1 – continued from previous page

Features	Description	Original reference
ChemEnvSite Fingerprint	Calculates the resemblance of given sites to ideal environment using pymatgens ChemEnv package.	[104, 105]
Coordination-Number	The number of first nearest neighbors of a site	[105]
CrystalNN-Fingerprint	A local order parameter fingerprint for periodic crystals.	-
GaussianSymm-Func	Calculates the gaussian radial and angular symmetry functions originally suggested for fitting machine learning potentials.	[106, 107]
GeneralizedRadial-Distribution-Function	Computes the general radial distribution function for a site	[102]
LocalProperty-Difference	Computes the difference in elemental properties between a site and its neighboring sites.	[99, 101]
OPSite-Fingerprint	Computes the local structure order parameters from a site's neighbor environment.	[105]
Voronoi-Fingerprint	Calculates the Voronoi tessellation-based features around a target site.	[108, 109]
Density of state features		
DOSFeaturizer	Computes top contributors to the density of states at the valence and conduction band edges. Thus includes chemical specie, orbital character, and orbital location information.	[110]
Band structure features		

Continued on next page

Table B.1 – continued from previous page

Features	Description	Original reference
BandFeaturizer	Converts a complex electronic band structure into quantities such as band gap and the norm of k point coordinates at which the conduction band minimum and valence band maximum occur.	-

B.2 Erroneous entries

MPID	Full formula	Reference
mp-555563	$PH_6C_2S_2NCl_2O_4$	[111]
mp-583476	$Nb_7S_2I_{19}$	[112]
mp-600205	$H_{10}C_5SeS_2N_3Cl$	-
mp-600217	$H_{80}C_{40}Se_8S_{16}Br_8N_{24}$	-
mp-1195290	$Ga_3Si_5P_{10}H_{36}C_{12}N_4Cl_{11}$	-
mp-1196358	$P_4H_{120}Pt_8C_{40}I_8N_4Cl_8$	-
mp-1196439	$Sn_8P_4H_{128}C_{44}N_{12}Cl_8O_4$	-
mp-1198652	$Te_4H_{72}C_{36}S_{24}N_{12}Cl_4$	-
mp-1198926	$Re_8H_{96}C_{24}S_{24}N_{48}Cl_{48}$	-
mp-1199490	$Mn_4H_{64}C_{16}S_{16}N_{32}Cl_8$	-
mp-1199686	$Mo_4P_{16}H_{152}C_{52}N_{16}Cl_{16}$	-
mp-1203403	$C_{121}S_2Cl_{20}$	-
mp-1204279	$Si_{16}Te_8H_{176}Pd_8C_{64}Cl_{16}$	-
mp-1204629	$P_{16}H_{216}C_{80}N_{32}Cl_8$	-

Table B.2: A table of manually identified entries from Materials Project that experience issues concerning Matminer's featurization tools. These were excluded from the dataset.

Bibliography

1. Griffiths, D. *Introduction to quantum mechanics* ISBN: 9781107179868 (Cambridge University Press, Cambridge, 2017).
2. Turing, A. M. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London mathematical society* **2,** 230–265 (1937).
3. Moore, G. Cramming More Components Onto Integrated Circuits. *Proceedings of the IEEE* **86,** 82–85 (Jan. 1965).
4. Pavičić, M. *Quantum computation and quantum communication : theory and experiments* ISBN: 9786610743704 (Springer, New York, 2006).
5. Gwennap, L. Apple's 5 Nanometer Chip Is Another Signpost That Moore's Law Is Running Out. *Forbes.* <https://www.forbes.com/sites/linleygwennap/2020/10/12/apple-moores-law-is-running-out/> (Oct. 12, 2020).
6. Weber, J. R. *et al.* Quantum computing with defects. *Proceedings of the National Academy of Sciences* **107,** 8513–8518 (Apr. 2010).
7. DiVincenzo, D. P. The Physical Implementation of Quantum Computation. *Fortschritte der Physik* **48,** 771–783 (Sept. 2000).
8. Ladd, T. D. *et al.* Quantum computers. *Nature* **464,** 45–53 (Mar. 2010).
9. Mizel, A., Lidar, D. A. & Mitchell, M. Simple Proof of Equivalence between Adiabatic Quantum Computation and the Circuit Model. *Physical Review Letters* **99.** doi:10.1103/physrevlett.99.070502 (Aug. 2007).
10. Grover, L. K. A framework for fast quantum mechanical algorithms. arXiv: quant-ph/9711043v2 [quant-ph] (Nov. 20, 1997).
11. Shor, P. *Algorithms for quantum computation: discrete logarithms and factoring* in *Proceedings 35th Annual Symposium on Foundations of Computer Science* (IEEE Comput. Soc. Press, 1994). doi:10.1109/sfcs.1994.365700.
12. Martinis, J. M. *et al. Quantum supremacy using a programmable superconducting processor* en. 2019. doi:10.5061/DRYAD.K6T1RJ8.
13. Georgescu, I. The DiVincenzo criteria 20 years on. *Nature Reviews Physics* **2,** 666–666 (Nov. 2020).

14. Griffiths, R. B. Nature and location of quantum information. *Physical Review A* **66.** doi:10.1103/physreva.66.012311 (July 2002).

15. Gisin, N., Ribordy, G., Tittel, W. & Zbinden, H. Quantum cryptography. *Reviews of Modern Physics* **74,** 145–195 (Mar. 2002).

16. Gisin, N. & Thew, R. Quantum communication. *Nature Photonics* **1,** 165–171 (Mar. 2007).

17. Acín, A. *et al.* The quantum technologies roadmap: a European community view. *New Journal of Physics* **20,** 080201 (Aug. 2018).

18. Boaron, A. *et al.* Secure Quantum Key Distribution over 421 km of Optical Fiber. *Physical Review Letters* **121.** doi:10.1103/physrevlett.121.190502 (Nov. 2018).

19. Degen, C. L., Reinhard, F. & Cappellaro, P. Quantum sensing. *Reviews of Modern Physics* **89.** doi:10.1103/revmodphys.89.035002 (July 2017).

20. Kristian Fossheim, A. S. *Superconductivity: Physics and Applications* 442 pp. ISBN: 0470844523. <https://www.ebook.de/de/product/3608091/kristian_fossheim_asle_sudboe_superconductivity_physics_and_applications.html> (WILEY, 2004).

21. Ben Streetman, S. B. *Solid State Electronic Devices, Global Edition* 632 pp. ISBN: 1292060557. <https://www.ebook.de/de/product/30394493/ben_streetman_sanjay_banerjee_solid_state_electronic_devices_global_edition.html> (Pearson Education Limited, 2015).

22. Renganathan, G., Tanneru, N. & Madurai, S. L. in *Fundamental Biomaterials: Metals* 211–241 (Elsevier, 2018). doi:10.1016/b978-0-08-102205-4.00010-6.

23. Lufaso, M. W. & Woodward, P. M. Prediction of the crystal structures of perovskites using the software program SPuDS. *Acta Crystallographica Section B Structural Science* **57,** 725–738 (Nov. 2001).

24. Bednorz, J. G. & Müller, K. A. Perovskite-type oxides—The new approach to high-Tcsuperconductivity. *Reviews of Modern Physics* **60,** 585–600 (July 1988).

25. Boivin, J. C. & Mairesse, G. Recent Material Developments in Fast Oxide Ion Conductors. *Chemistry of Materials* **10,** 2870–2888 (Oct. 1998).

26. Cheong, S.-W. & Mostovoy, M. Multiferroics: a magnetic twist for ferroelectricity. *Nature Materials* **6,** 13–20 (Jan. 2007).

27. Ibn-Mohammed, T. *et al.* Perovskite solar cells: An integrated hybrid lifecycle assessment and review in comparison with other photovoltaic technologies. *Renewable and Sustainable Energy Reviews* **80,** 1321–1344 (Dec. 2017).

28. Chen, P.-Y. *et al.* Environmentally responsible fabrication of efficient perovskite solar cells from recycled car batteries. *Energy Environ. Sci.* **7,** 3659–3665 (2014).

29. Pauli, W. Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren. *Zeitschrift für Physik* **31,** 765–783 (Feb. 1925).

30. Martienssen, W. *Springer handbook of condensed matter and materials data* ISBN: 9786610625949 (Springer, Heidelberg New York, 2005).

31. Pelant, I. *Luminescence spectroscopy of semiconductors* ISBN: 0191738549 (Oxford University Press, Oxford, 2012).

32. Kun Huang, A. R. Theory of light absorption and non-radiative transitions in F -centres. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **204,** 406–423 (Dec. 1950).

33. Gordon, L. *et al.* Quantum computing with defects. *MRS Bulletin* **38,** 802–807 (Oct. 2013).

34. Bernien, H. *et al.* Heralded entanglement between solid-state qubits separated by three metres. *Nature* **497,** 86–90 (Apr. 2013).

35. Taylor, J. M. *et al.* High-sensitivity diamond magnetometer with nanoscale resolution. *Nature Physics* **4,** 810–816 (Sept. 2008).

36. Barclay, P. E., Fu, K.-M. C., Santori, C., Faraon, A. & Beausoleil, R. G. Hybrid Nanocavity Resonant Enhancement of Color Center Emission in Diamond. *Physical Review X* **1.** doi:10.1103/physrevx.1.011007 (Sept. 2011).

37. Neudeck, P. G. Progress in silicon carbide semiconductor electronics technology. *Journal of Electronic Materials* **24,** 283–288 (Apr. 1995).

38. Silveira, E., Freitas, J. A., Glembocki, O. J., Slack, G. A. & Schowalter, L. J. Excitonic structure of bulk AlN from optical reflectivity and cathodoluminescence measurements. *Physical Review B* **71.** doi:10.1103/physrevb.71.041201 (Jan. 2005).

39. Lawaetz, P. Valence-Band Parameters in Cubic Semiconductors. *Physical Review B* **4,** 3460–3467 (Nov. 1971).

40. Beckers, L. *et al.* Structural and optical characterization of epitaxial waveguiding BaTiO3 thin films on MgO. *Journal of Applied Physics* **83,** 3305–3310 (Mar. 1998).

41. Kumbhojkar, N., Nikesh, V. V., Kshirsagar, A. & Mahamuni, S. Photophysical properties of ZnS nanoclusters. *Journal of Applied Physics* **88,** 6260–6264 (Dec. 2000).

42. Bassett, L. C., Alkauskas, A., Exarhos, A. L. & Fu, K.-M. C. Quantum defects by design. *Nanophotonics* **8,** 1867–1888 (Oct. 2019).

43. James, W. J. Theory of defects in solidsby A. M. Stoneham. *Acta Crystallographica Section A* **32,** 527–527 (May 1976).

44. Togan, E. *et al.* Quantum entanglement between an optical photon and a solid-state spin qubit. *Nature* **466,** 730–734 (Aug. 2010).

45. Bathen, M. E. *Point defects in silicon carbide for quantum technologies: Identification, tuning and control* PhD thesis (The Faculty of Mathematics and Natural Sciences, University of Oslo).

46. Son, N. T. *et al.* Developing silicon carbide for quantum spintronics. *Applied Physics Letters* **116,** 190501 (May 2020).

47. Falk, A. L. *et al.* Polytype control of spin qubits in silicon carbide. *Nature Communications* **4.** doi:10.1038/ncomms2854 (May 2013).

48. Widmann, M. *et al.* Coherent control of single spins in silicon carbide at room temperature. *Nature Materials* **14,** 164–168 (Dec. 2014).

49. Zhang, G., Cheng, Y., Chou, J.-P. & Gali, A. Material platforms for defect qubits and single-photon emitters. *Applied Physics Reviews* **7,** 031308 (Sept. 2020).

50. Redjem, W. *et al.* Single artificial atoms in silicon emitting at telecom wavelengths. *Nature Electronics* **3,** 738–743 (Nov. 2020).

51. Wang, J. *et al.* Gallium arsenide (GaAs) quantum photonic waveguide circuits. *Optics Communications* **327,** 49–55 (Sept. 2014).

52. Berhane, A. M. *et al.* Photophysics of GaN single-photon emitters in the visible spectral range. *Physical Review B* **97.** doi:10.1103/physrevb.97.165202 (Apr. 2018).

53. Xue, Y. *et al.* Single-Photon Emission from Point Defects in Aluminum Nitride Films. *The Journal of Physical Chemistry Letters* **11,** 2689–2694 (Mar. 2020).

54. Varley, J. B., Janotti, A. & de Walle, C. G. V. Defects in AlN as candidates for solid-state qubits. *Physical Review B* **93.** doi:10.1103/physrevb.93.161201 (Apr. 2016).

55. Hardy, W. J. *et al.* Single and double hole quantum dots in strained Ge/SiGe quantum wells. *Nanotechnology* **30,** 215202 (Mar. 2019).

56. Toth, M. & Aharonovich, I. Single Photon Sources in Atomically Thin Materials. *Annual Review of Physical Chemistry* **70,** 123–142 (June 2019).

57. Atatüre, M., Englund, D., Vamivakas, N., Lee, S.-Y. & Wrachtrup, J. Material platforms for spin-based photonic quantum technologies. *Nature Reviews Materials* **3,** 38–51 (Apr. 2018).

58. Tran, T. T. *et al.* Robust Multicolor Single Photon Emission from Point Defects in Hexagonal Boron Nitride. *ACS Nano* **10,** 7331–7338 (July 2016).

59. Tran, T. T., Bray, K., Ford, M. J., Toth, M. & Aharonovich, I. *Quantum Emission from Hexagonal Boron Nitride Monolayers* in *Conference on Lasers and Electro-Optics* (OSA, 2016). doi:10.1364/cleo_qels.2016.ftu4d.1.

60. Weston, L., Wickramaratne, D., Mackoit, M., Alkauskas, A. & de Walle, C. G. V. Native point defects and impurities in hexagonal boron nitride. *Physical Review B* **97.** doi:10.1103/physrevb.97.214104 (June 2018).

61. Abdi, M., Chou, J.-P., Gali, A. & Plenio, M. B. Color Centers in Hexagonal Boron Nitride Monolayers: A Group Theory and Ab Initio Analysis. *ACS Photonics* **5,** 1967–1976 (Apr. 2018).

62. Ohebbi. *ohebbi/predicting-solid-state-qubit-candidates: v0.1-beta* 2021. doi:10.5281/ZENODO.4633959.

63. Curtarolo, S. *et al.* AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science* **58,** 227–235 (June 2012).

64. Rosenbrock, C. W. A Practical Python API for Querying AFLOWLIB. arXiv: 1710.00813v1 [cs.DB] (Sept. 28, 2017).

65. Isayev, O. *et al.* Universal fragment descriptors for predicting properties of inorganic crystals. *Nature Communications* **8.** doi:10.1038/ncomms15679 (June 2017).

66. Choudhary, K. *et al.* JARVIS: An Integrated Infrastructure for Data-driven Materials Design. arXiv: 2007.01831v1 [cond-mat.mtrl-sci] (July 3, 2020).

67. Breuck, P.-P. D., Evans, M. L. & Rignanese, G.-M. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODNet. arXiv: 2102.02263v1 [cond-mat.mtrl-sci] (Feb. 3, 2021).

68. Ferrenti, A. M., de Leon, N. P., Thompson, J. D. & Cava, R. J. Identifying candidate hosts for quantum defects via data mining. *npj Computational Materials* **6.** doi:10.1038/s41524-020-00391-7 (Aug. 2020).

69. Markham, M. *et al.* CVD diamond for spintronics. *Diamond and Related Materials* **20,** 134–139 (Feb. 2011).

70. Balasubramanian, G. *et al.* Ultralong spin coherence time in isotopically engineered diamond. *Nature Materials* **8,** 383–387 (Apr. 2009).

71. Tyryshkin, A. M. *et al.* Electron spin coherence exceeding seconds in high-purity silicon. *Nature Materials* **11,** 143–147 (Dec. 2011).

72. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1,** 011002 (July 2013).

73. Ong, S. P. *et al.* The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles. *Computational Materials Science* **97,** 209–215 (Feb. 2015).

74. Inselberg, A. The plane with parallel coordinates. *The Visual Computer* **1,** 69–91 (Aug. 1985).

75. Inselberg, A. & Dimsdale, B. *Parallel coordinates: a tool for visualizing multi-dimensional geometry* in *Proceedings of the First IEEE Conference on Visualization: Visualization 90* (IEEE Comput. Soc. Press, 1990). doi:10.1109/visual.1990.146402.

76. Ericson, D., Johansson, J. & Cooper, M. *Visual Data Analysis using Tracked Statistical Measures within Parallel Coordinate Representations* in *Coordinated and Multiple Views in Exploratory Visualization (CMV'05)* (IEEE). doi:10.1109/cmv.2005.21.

77. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research (2011).* arXiv: 1201.0490v4 [cs.LG] (Jan. 2, 2012).

78. Lemaitre, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. arXiv: 1609.06570v1 [cs.LG] (Sept. 21, 2016).

79. Ohebbi. *ohebbi/predicting-ABO3-structures: v0.1-alpha* 2021. doi:10.5281/ZENODO.4633968.

80. Balachandran, P. V. *et al.* Predictions of new ABO3 perovskite compounds by combining machine learning and density functional theory. *Physical Review Materials* **2.** doi:10.1103/physrevmaterials.2.043802 (Apr. 2018).

81. Shannon, R. D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallographica Section A* **32,** 751–767 (Sept. 1976).

82. Zhang, H., Li, N., Li, K. & Xue, D. Structural stability and formability of ABO3-type perovskite compounds. *Acta Crystallographica Section B Structural Science* **63,** 812–818 (Nov. 2007).

83. Goldschmidt, V. M. Die Gesetze der Krystallochemie. *Die Naturwissenschaften* **14,** 477–485 (May 1926).

84. Villars, P., Cenzual, K., Daams, J., Chen, Y. & Iwata, S. Data-driven atomic environment prediction for binaries using the Mendeleev number. *Journal of Alloys and Compounds* **367,** 167–175 (Mar. 2004).

85. Ward, L. *et al.* Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **152,** 60–69 (Sept. 2018).

86. Kotochigova, S., Levine, Z. H., Shirley, E. L., Stiles, M. D. & Clark, C. W. Local-density-functional calculations of the energy of atoms. *Physical Review A* **55,** 191–199 (Jan. 1997).

87. Laws, K. J., Miracle, D. B. & Ferry, M. A predictive structural model for bulk metallic glasses. *Nature Communications* **6.** doi:10.1038/ncomms9123 (Sept. 2015).

88. Butler, M. A. & Ginley, D. S. Prediction of Flatband Potentials at Semiconductor-Electrolyte Interfaces from Atomic Electronegativities. *Journal of The Electrochemical Society* **125,** 228–232 (Feb. 1978).

89. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68,** 314–319 (Feb. 2013).

90. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2.** doi:10.1038/npjcompumats.2016.28 (Aug. 2016).

91. Deml, A. M., O'Hayre, R., Wolverton, C. & Stevanović, V. Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression. *Physical Review B* **93.** doi:10.1103/physrevb.93.085142 (Feb. 2016).

92. Weeber, A. W. Application of the Miedema model to formation enthalpies and crystallisation temperatures of amorphous alloys. *Journal of Physics F: Metal Physics* **17,** 809–813 (Apr. 1987).

93. Yang, X. & Zhang, Y. Prediction of high-entropy stabilized solid-solution in multi-component alloys. *Materials Chemistry and Physics* **132,** 233–238 (Feb. 2012).

94. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters* **108.** doi:10.1103/physrevlett.108.058301 (Jan. 2012).

95. Schütt, K. T. *et al.* How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Physical Review B* **89.** doi:10.1103/physrevb.89.205118 (May 2014).

96. Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry* **115,** 1094–1101 (Apr. 2015).

97. Ewald, P. P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik* **369,** 253–287 (1921).

98. Hansen, K. *et al.* Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *The Journal of Physical Chemistry Letters* **6,** 2326–2331 (June 2015).

99. Ward, L. *et al.* Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Physical Review B* **96.** doi:10.1103/physrevb.96.024104 (July 2017).

100. Botu, V. & Ramprasad, R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *International Journal of Quantum Chemistry* **115,** 1074–1083 (Dec. 2014).

101. De Jong, M. *et al.* A Statistical Learning Framework for Materials Science: Application to Elastic Moduli of k-nary Inorganic Polycrystalline Compounds. *Scientific Reports* **6.** doi:10.1038/srep34256 (Oct. 2016).

102. Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. & Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Physical Review B* **95.** doi:10.1103/physrevb.95.144110 (Apr. 2017).

103. Steinhardt, P. J., Nelson, D. R. & Ronchetti, M. Bond-orientational order in liquids and glasses. *Physical Review B* **28,** 784–805 (July 1983).

104. Waroquiers, D. *et al.* Statistical Analysis of Coordination Environments in Oxides. *Chemistry of Materials* **29,** 8346–8360 (Sept. 2017).

105. Zimmermann, N. E. R., Horton, M. K., Jain, A. & Haranczyk, M. Assessing Local Structure Motifs Using Order Parameters for Motif Recognition, Interstitial Identification, and Diffusion Path Characterization. *Frontiers in Materials* **4.** doi:10.3389/fmats.2017.00034 (Nov. 2017).

106. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics* **134,** 074106 (Feb. 2011).

107. Khorshidi, A. & Peterson, A. A. Amp: A modular approach to machine learning in atomistic simulations. *Computer Physics Communications* **207,** 310–324 (Oct. 2016).

108. Peng, H. L., Li, M. Z. & Wang, W. H. Structural Signature of Plastic Deformation in Metallic Glasses. *Physical Review Letters* **106.** doi:10.1103/physrevlett.106.135503 (Mar. 2011).

109. Wang, Q. & Jain, A. A transferable machine-learning framework linking interstice distribution and plastic heterogeneity in metallic glasses. *Nature Communications* **10.** doi:10.1038/s41467-019-13511-9 (Dec. 2019).

110. Dylla, M. T., Dunn, A., Anand, S., Jain, A. & Snyder, G. J. Machine Learning Chemical Guidelines for Engineering Electronic Structures in Half-Heusler Thermoelectric Materials. *Research* **2020,** 1–8 (Apr. 2020).

111. None Available. *Materials Data on PH6C2S2N(ClO2)2 by Materials Project* en. 2020. doi:10.17188/1268877.

112. None Available. *Materials Data on Nb7S2I19 by Materials Project* en. 2014. doi:10.17188/1277059.