Faculty of Sciences

Department of Physics and Astronomy

Boosting crystal discovery in quaternary space: from brute force to intelligent screening

Titus Crepain

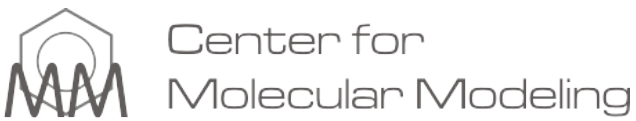
Supervisor: Prof. Dr. Stefaan Cottenier,

Counsellor: Ir. Michael Sluydts

Thesis submitted in partial fulfilment of the requirements for the degree of

Master of Science in Physics and Astronomy

Academic year 2015–2016



This research was conducted at the Center for Molecular Modeling.

Preface

What this preface is for you, dear reader, is for me a postface of a year of hard labour. The work that lays in front of you would not have been there without the valuable support and advice of some people I want to address here.

First of all I want to thank Michael Sluydts, who was ready every moment of the day for answering all my questions and troubles. He taught me the worth of a good data backup and often could not resist making a light-hearted joke.
Furthermore I would like to thank Stefaan Cottenier, for giving me the chance to do this research under his supervision. Being supported by his constructive and encouraging ideas was really comforting for me as a thesis student.
My parents, who always gave me the confidence to do what I wanted. Helping me with my mathematical homework has already stopped more than a decade ago. Luckily, they were able to adapt their support over time by giving me lovely words, coffee and freshly baked pie at the right moments.
Jan, Broud and Wim, fellow citizens of our own microstate the Brandstaat (possibly 'nanostate' or 'picostate' are more appropriate terms in the physical context of this work). Without their support, the laughs and adventures we had together, the past years would not have been as memorable. It is time for me to leave the Brandstaat, take care!
This is also the right place to thank my fellow thesis students Pieter, Klaas, Michiel, Yentl and Senne. We spent a year of passing each other LaTeX tricks and complaining about our hard knock thesis lives, which resulted for me in getting to know five beautiful people.
Finally, I need to name the heroes who had the courage to proofread my writings and gave me helpful feedback. Thibault, Wim, Michael and Stefaan, many thanks!

I would like to end this preface with a motivational quote. Not a quote from a famous enlightened philosopher but from a lesser known young man. The quote can only be expressed in my beloved West-Flemish mother tongue and means as much as 'In the end, everything will be alright'.

Zie kalm.

— My younger brother, Loewie Crepain

Titus Crepain

Titus Crepain,
Ghent, June 1st 2016

Boosting crystal discovery in quaternary space: from brute force to intelligent screening

Titus Crepain

Supervisor: Prof. Dr. Stefaan Cottenier
Counsellor: Ir. Michael Sluydts

Thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Physics and Astronomy

Department of Physics and Astronomy
Faculty of Sciences
Academic year 2015-2016

Summary

How many crystalline solids are left to be discovered? Elementary, binary and ternary crystals are already known in large numbers. Although quaternary crystals have more combinatorial possibilities to be constructed, there are comparatively not very many quaternaries known. Is this because mankind did not discover them yet, or rather because Nature limits the existence of more than a few stable quaternary crystals? This thesis aims at contributing to the answer to this question, by examining computationally how many quaternary crystals in a given 'family' are intrinsically stable.

Starting from an experimentally known quaternary crystal one can create a 'family' by substituting the elements. A family is characterized by the space group of the original crystal and can be submitted to a high-throughput DFT computation, to determine its possible stability against decomposition into a combination of elementary, binary or ternary crystals. Such a high-throughput process could be sped up if not all members of the family would have to be calculated explicitly by DFT. Can one restrict the expensive DFT calculations to the most relevant members of the family? In this work we examine how a neural network can be trained to learn throughout the screening process how to identify and prioritize the most relevant members. We show that all stable quaternaries in the family can be found, while making DFT calculations for less than one half of the set.

Three different families were studied in this work, for which 11 experimental quaternary crystals were known. We suggest 427 yet undiscovered candidates. These results suggest that the total number of stable quaternaries is by no means small.

Keywords: Quaternary crystals, density functional theory, high-throughput, neural networks, crystal stability prediction

Boosting crystal discovery in quaternary space: from brute force to intelligent screening

Titus Crepain

Promoter: Prof. Dr. Stefaan Cottenier
Begeleider: Ir. Michael Sluydts

Masterproef ingediend tot het behalen van de academische graad van Master in de Fysica en Sterrenkunde

Vakgroep Fysica en Sterrenkunde
Faculteit Wetenschappen
Academiejaar 2015-2016

Samenvatting

Hoeveel kristallijne materialen zijn er nog te ontdekken? Elementaire, binaire en ternaire kristallen zijn reeds gekend in groten getale. Alhoewel er voor quaternaire kristallen meer combinatorische mogelijkheden zijn om te vormen, is het aantal gekende quaternairen in vergelijking laag. Dienen er dan nog veel quaternairen ontdekt te worden of worden er slechts enkele stabiele quaternairen toegelaten door de natuur om te bestaan? Deze thesis tracht een bijdrage te leveren tot het antwoord op deze vraag, door computationeel na te gaan hoeveel quaternaire kristallen er in een gegeven 'familie' intrinsiek stabiel zijn.
Vertrekkend van een experimenteel gekend quaternair kristal kan men een 'familie' creëren door substitutie van de elementen. Een familie wordt gekenmerkt door de ruimtegroep van het originele kristal en kan in zijn geheel worden berekend via dichtheidsfunctionaaltheorie (DFT), met als doel de stabiliteit te achterhalen ten opzichte van elementaire, binaire en ternaire decompositie producten. Zo een high-throughput screening kan worden versneld indien niet alle leden van de familie expliciet met DFT hoeven worden te berekend. Is het mogelijk om de tijdrovende DFT berekeningen te beperken tot de meest relevante leden van de familie? In dit werkstuk onderzoeken we hoe een neuraal netwerk doorheen het proces kan worden getraind om de meest relevante kristallen te identificeren en te prioriteren. We tonen aan dat alle stabiele quaternairen in de familie kunnen worden gevonden, terwijl slechts de helft van de DFT berekeningen nodig zijn.
Drie verschillende families werden bestudeerd, waarbij van 11 experimenteel gekende quaternaire kristallen vertrokken werd. 427 onontdekte kandidaten worden voorgesteld. Deze resultaten suggereren het totaal aantal stabiele quaternairen veel hoger ligt dat tot nu toe geweten is.

Trefwoorden: Quaternaire kristallen, dichtheidsfunctionaaltheorie, high-throughput, neurale netwerken, voorspellen van kristalstabiliteit

Contents

1 Crystal discovery **3**

1.1 Stable crystals . . . 3
1.1.1 Crystals . . . 3
1.1.2 Decomposition energy . . . 5
1.2 How much is left to discover? . . . 5
1.3 Quaternary Zintl phases . . . 8
1.3.1 Zintl phase definition . . . 8
1.3.2 Suitable QZP elements . . . 9
1.4 QZPs in the Pnma space group . . . 10

2 High-throughput DFT **13**

2.1 Crystal modelling: the many-electron problem . . . 13
2.1.1 The Born-Oppenheimer approximation . . . 13
2.1.2 Density Functional Theory . . . 14
The Hohenberg-Kohn theorems . . . 14
The Kohn-Sham equations . . . 15
Ascertain the exchange-correlation functional . . . 16
2.1.3 Solving the equations . . . 17
Basis sets . . . 17
k-point sampling . . . 18
Pseudopotentials . . . 18
2.2 The Vienna Ab initio Simulation Package . . . 18
2.2.1 What is VASP? . . . 18
2.2.2 The calculation workflow and VASP settings . . . 19
2.3 High-throughput computing . . . 21
2.3.1 Definition . . . 21
2.3.2 Examples . . . 22
2.3.3 The high-throughput framework . . . 22
External database . . . 23
Data selection . . . 23
Data generation . . . 24
Data storage . . . 24

Data analysis . . . 25

3 Machine learning **27**
3.1 The scope of ML: an overview . . . 27
3.1.1 General ML applications . . . 27
3.1.2 ML in computational material physics . . . 28
3.2 Theoretical background . . . 29
3.2.1 The perceptron . . . 29
3.2.2 The Neural Network model . . . 31
3.3 The MATLAB Neural Network Toolbox . . . 33
3.3.1 Adjustable parameters . . . 33
3.3.2 Pre- and postprocessing . . . 34
3.3.3 Outcome of the network . . . 35

4 Calibrating the Neural Network **37**
4.1 The optimal Neural Network parameters . . . 37
4.1.1 Composing a comfortable starting set . . . 37
4.1.2 Varying the number of nodes . . . 39
4.1.3 Subsequent modifications . . . 40
4.2 Intelligent screening . . . 41
4.2.1 Exploiting the Neural Network . . . 41
4.2.2 Results of the method . . . 42

5 Screening of a straightforward crystal set **45**
5.1 Pre-analysis . . . 45
5.2 Results . . . 48
5.2.1 Intelligent screening . . . 48
5.2.2 Crystal properties . . . 48

6 Screening of a challenging crystal set **55**
6.1 Pre-analysis . . . 55
6.2 Results . . . 57
6.2.1 Intelligent screening . . . 57
6.2.2 Crystal properties . . . 58

7 Global results **63**

Bibliography **68**

A Crystallographic information of the I-42m family **77**

B Crystallographic information of the $P6_3/mmc$ family **81**

List of abbreviations

BO	Born-Oppenheimer
CMM	Center for Molecular Modeling
COD	Crystallography Open Database
DFT	Density Functional Theory
DOS	Density Of States
GGA	Generalized Gradient Approximation
HF	Hartree-Fock
ICSD	Inorganic Crystal Structure Database
IS	Intelligent Screening
KS	Kohn-Sham
ML	Machine Learning
MP	Materials Project
OQMD	Open Quantum Materials Database
PAW	Projector Augmented-Wave
PES	Potential Energy Surface
PLA	Perceptron Learning Algorithm
QM	Queue Manager
QZP	Quaternary Zintl Phase
STD	Standard Deviation
VASP	Vienna Ab initio Simulation Package
ZP	Zintl Phase

List of Figures

1.1 The NaCl crystal . . . 4
1.2 The zinc-blende ZnS crystal. . . . 4
1.3 The wurtzite ZnS crystal . . . 4
1.4 Evolution of crystal discovery over the years . . . 6
1.5 Potential crystals versus their decomposition paths . . . 7
1.6 A pragmatic ZP definition . . . 10
1.7 The discovery rate of QZPs . . . 11
1.8 Results of a brute force calculation in Pnma space group . . . 12
1.9 Examples of relaxed structures in the Pnma space group . . . 12

2.1 The DFT self-consistency cycle . . . 16
2.2 A high-throughput data flow diagram . . . 23
2.3 The Queue Manager workflow . . . 24

3.1 The 2D perceptron . . . 29
3.2 Logical operations with perceptrons . . . 30
3.3 Stochastic Gradient Descent . . . 32
3.4 Backward Propagation Algorithm . . . 33
3.6 NN error histogram . . . 35
3.7 NN regression plot . . . 36

4.1 NN input vectors . . . 37
4.2 Linear regression for the unaltered input set, with E_{pure} as the target energy. . . . 38
4.3 Linear regression for the unaltered input set without Polonium compounds and E_{decomp} as target. . . . 38
4.4 Standard deviation for the unaltered input set, with E_{pure} as target energy. . . . 38
4.5 Standard deviation for the input set without the Polonium compounds and E_{decomp} as target data. . . . 38
4.6 Regression for all points of the data set without the Polonium structures, with the attributes in the input vectors ordered as in structure formula and with E_{decomp} as target. . . . 40
4.7 Standard deviation for the data set without the Polonium structures, with the attributes in input vectors ordered as in structure formula and with E_{decomp} as target. . . . 40
4.8 Linear regression calculated for NNs from 1 to 20 nodes for the unaltered data set. . . . 40

4.9 Linear regression calculated for NNs from 1 to 40 nodes for the data set with E_{decomp} as target and ordered along the elements in the structure formula. 40
4.10 NN performances of modifications on the reference input set 41
4.11 *Intelligent screening* algorithm 42
4.12 Results of on the Pnma the space group 43

5.1 The originally published $K_2BaSnTe_4$ crystal structure 46
5.2 The $K_2BaSnTe_4$ crystal structure in the I-43m space group. 47
5.3 The $K_2BaSnTe_4$ crystal structure in the I-42m space group without partial occupancies. . . . 47
5.4 Periodic table separations for the I-42m space group 48
5.5 Results of *intelligent screening* on the I-42m the space group 49
5.6 The crystal structure of $K_2RbSiSe_4$ (I-I-IV-VI). 50
5.7 The flatter crystal structure of $Rb_2MgSiSe_4$ (I-II-IV-VI). 50
5.8 DOS of $K_2RbSiSe_4$ (I-I-IV-VI). 52
5.9 DOS of $K_2MgSiSe_4$ (I-II-IV-VI). 52
5.10 DOS of $Ba_2RbSiTe_4$ (II-I-IV-VI). 52
5.11 DOS of K_2RbPSe_4 (I-I-V-VI). 52
5.12 BANDS of $K_2RbSiSe_4$ (I-I-IV-VI). 53
5.13 BANDS of $K_2MgSiSe_4$ (I-II-IV-VI). 53
5.14 BANDS of $Ba_2RbSiTe_4$ (II-I-IV-VI). 53
5.15 BANDS of K_2RbPSe_4 (I-I-V-VI). 53

6.1 The $K_6AlSb_4Na_3$ crystal structure 56
6.2 Periodic table separations for the $P6_3/mmc$ space group 57
6.3 Results of the *intelligent screening* algorithm on the $P6_3/mmc$ the space group 58
6.4 DOS of $Cs_6InSb_4Rb_3$ (I-III-V-I). 60
6.5 DOS of $Cs_6AlBi_4K_3$ (I-III-V-I). 60
6.6 DOS of $Cs_6BiSb_4Rb_3$ (I-V-V-I). 60
6.7 DOS of $Cs_6BiSb_4K_3$ (I-V-V-I). 60
6.8 BANDS of $Cs_6InSb_4Rb_3$ (I-III-V-I). 61
6.9 BANDS of $Cs_6AlBi_4K_3$ (I-III-V-I). 61
6.10 BANDS of $Cs_6BiSb_4Rb_3$ (I-V-V-I). 61
6.11 BANDS of $Cs_6BiSb_4K_3$ (I-V-V-I). 61
6.12 DOS of $Cs_6GaBi_4Na_3$ (I-III-V-I). 61
6.13 BANDS of $Cs_6GaBi_4Na_3$ (I-III-V-I). 61

7.1 The *intelligent screening* algorithm applied on the three crystal families. The vertical axis represents the fraction of stable crystals that has already been found, while the horizontal axis is the fraction of the total data set that has been calculated. The yellow dots indicate the NN updates. 64
7.2 Overview of the Pnma, I-42m an $P6_3/mmc$ crystal families 66

List of Tables

2.1 VASP settings for relaxation. 20
2.2 VASP settings for DOS and band structures. 21

5.1 The Wyckoff positions in the $K_2BaSnTe_4$ crystal structure that adopts the I-42m space group. 46

6.1 The Wyckoff positions in the $K_6AlSb_4Na_3$ crystal structure. 55
6.2 The subfamilies of the $P6_3/mmc$ space group. 56

7.1 The total number of consumed node days for the I-42m and $P6_3/mmc$ families on the clusters Muk and Delcatty. 67
7.2 Interesting QZP families for further research. 67
7.3 Overview of the stable crystals in the three studied QZP families. A crystal is defined stable if $E_{decomp} \leq 0.05 eV/atom$. 67

A.1 Crystallographic information of the 3 smaller subfamilies. 77
A.2 Crystallographic information of the I-I-IV-VI subfamily. 77
A.3 Crystallographic information of the I-II-IV-VI subfamily. 78
A.4 Crystallographic information of the I-I-V-VI subfamily. 78
A.5 Crystallographic information of the II-I-IV-VI subfamily. 80

B.1 Crystallographic information of the I-III-V-I subfamily. 81
B.2 Crystallographic information of the I-V-V-I subfamily. 84

Chapter 1

Crystal discovery

Crystals come in many different shapes and sizes and become more exotic as the number of elements inside the structure increases. One can easily say that all possible binary combinations of elements have at least been studied, but for ternary, quaternary, ... crystals this statement is far from true. Recent developments in the field of quaternary crystals[1;2;3;4] prove that this topic is still very active. This chapter indicates the enormous scale of materials left to be discovered and the possibilities of an extensive study of (especially) quaternary crystals. Section 1.1 delimits the main concepts of this thesis whereafter Section 1.2 treats its main question. Section 1.3 confines the crystal search space to the feasible set of Quaternary Zintl Phases (QZPs) and finally Section 1.4 summarizes the work that has been done on this topic by Dumon[4].

1.1 Stable crystals

The main concepts of this research have variating interpretations in literature and thus it is necessary to clarify the terminology used in this work. The terms that are highlighted here are crystals in Section 1.1.1 and their stability in Section 1.1.2.

1.1.1 Crystals

All solids in this world are made from atoms, which themselves contain only two types of particles: nuclei and electrons. Different combinations of atoms and the various types of bonds between them result in the wide diversity of solids we see around us. In the context of material science one is especially interested in crystalline solids, because of the typical properties their long-range periodicity involves. Crystals are therefore the main topic of this work, therefore a good notion of the usage of this word is needed.

Consider NaCl which is constructed by arranging Na^+ and Cl^- ions alternately at the lattice points of a simple cubic lattice. Thence each ion is surrounded by six ions with opposite charge, see Figure 1.1. Many binary crystals adopt the same structure as NaCl. Kittel[5] reports for instance LiH, MgO, MnO, AgBr, PbS, KCl and KBr.

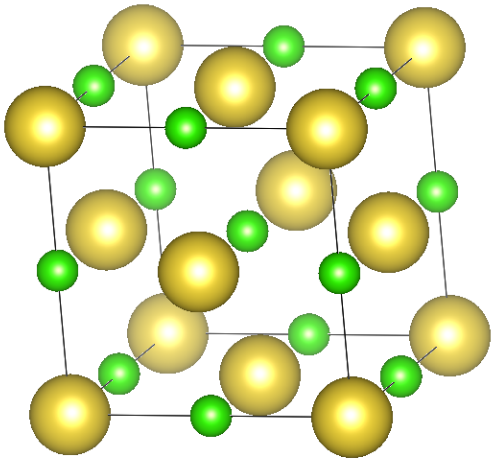


Figure 1.1: The well-known NaCl crystal, whereby the Na^{+} ions are coloured yellow, the Cl^{-} ions green.

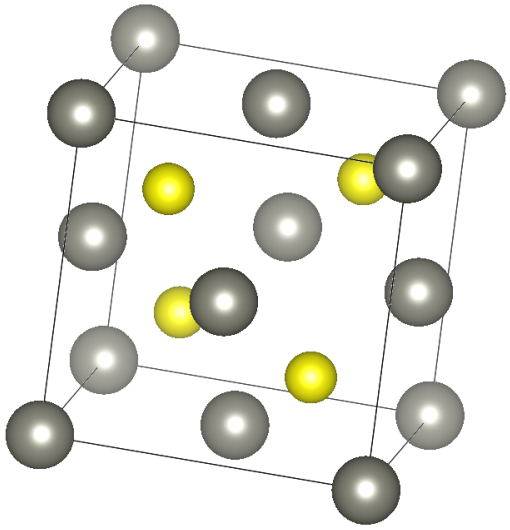


Figure 1.2: A ZnS crystal in the cubic zinc-blende structure, whereby Zn^{+2} ions are colored gray and S^{-2} ions yellow.

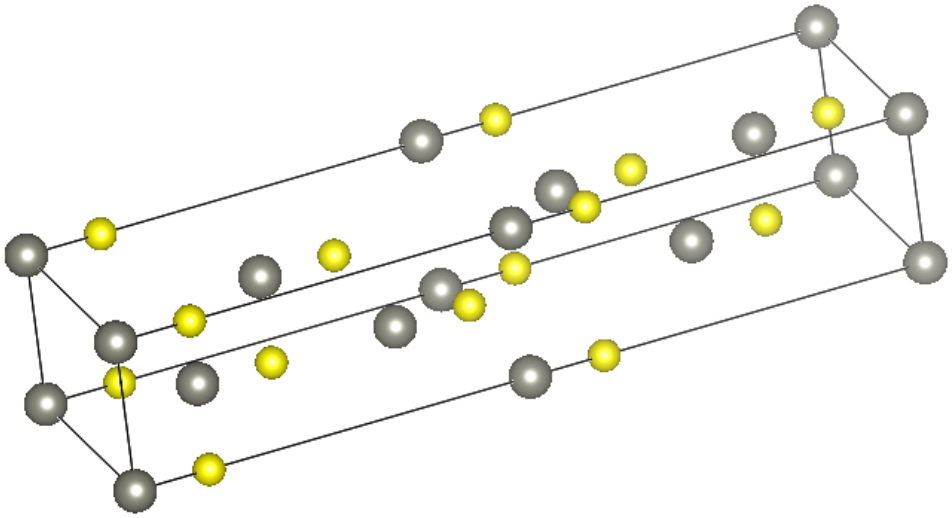


Figure 1.3: A ZnS crystal in the hexagonal wurtzite structure, whereby Zn^{+2} ions are colored gray and S^{-2} ions yellow.

On the contrary, chemical formulas do not tend to have one unique structure associated with it. Figure 1.2 and 1.3 show that crystals can have the same chemical formulae, but different structures.

During this research, the terms 'crystal type' and 'crystal structure' will be used as synonyms, referring to the combination of a specific structure and chemical formula.

1.1.2 Decomposition energy

Not all combinations of chemical elements that are placed in an arbitrary chosen structure are blessed to be stable. Unstable combinations decompose to one or more stable phases, whether it be elementary crystals or phases with more than one element. Equation (1.1) shows all possible decay paths for a ternary crystal ABC where neither the stoichiometry, nor the structure were taken into account.

$$\begin{aligned} ABC \quad \Rightarrow \quad & AB + C \qquad AC + B \\ & BC + A \qquad A + B + C \end{aligned} \tag{1.1}$$

The stability can be calculated from the total energy of a crystal. Primarily, define the pure formation energy per atom E_{pure} on the basis of the total compound energy E_{tot} and the energies per atom E_{el} of the elementary crystals:

$$E_{pure} = \frac{1}{N_{atoms}}\Big(E_{tot} - \sum_i n_i E_{el,i}\Big) \tag{1.2}$$

where n_i is the number of occurences of the ith element in the compound and N_{atom} the total number of involved atoms. The pure formation energy is computable when the total energy of the compound is known, since the total energy of each elementary crystal can be found in literature. A negative value for E_{pure} is then a first indication that the material could be absolutely stable. The pure formation energy is a necessary yet not a sufficient condition for a crystal to be stable. Not sufficient, because it does not take into account all possible decomposition products. These decomposition products can take many forms and may not be known in advance. Thus, including them all is far from evident. The Materials Project[6;7] (MP) is a good attempt to catalog as many as possible decomposition products of compounds. One can extend equation (1.2) to a decomposition energy E_{decomp} that takes into account all possible decomposition products:

$$E_{decomp} = \frac{1}{N_{atoms}}\Big(E_{tot} - \sum_i E_{dp,i}\Big) \tag{1.3}$$

where E_{dp} is the total energy of a stable decomposition product which is extracted from the MP. A negative E_{decomp} indicates an absolutely stable crystal, a positive E_{decomp} means that the compound is absolutely unstable or metastable. The decomposition path with the most positive E_{decomp} will then be the most favourable path to follow.

The pure formation energy and the decomposition energy will both be used in this work.

1.2 How much is left to discover?

The Inorganic Crystal Structure Database[8;9] (ICSD, version 2015-1) contains almost 180 000 experimentally observed crystals.The Open Quantum Materials Database[10] (OQMD) on the other hand, is a database with DFT calculated thermodynamic and structural properties and at this moment it contains around 100 000 crystals more than the ICSD. The Crystallographic Open Database[11;12] (COD) is an open-access collection of crystal structures and holds at this moment more than 360 000 crystal structures.

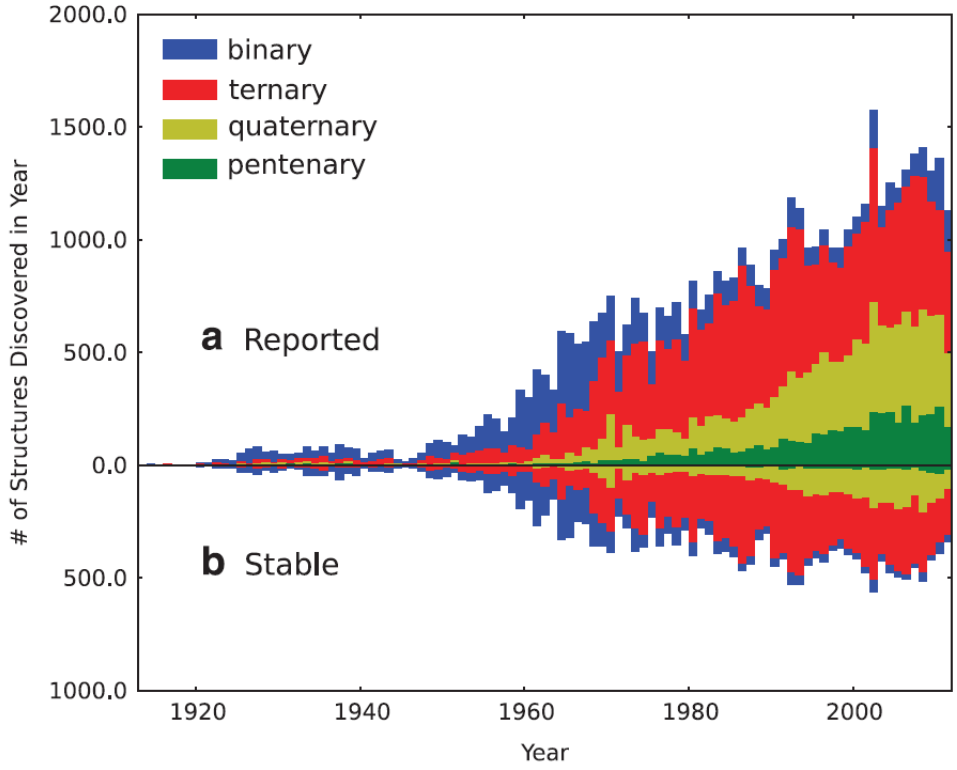


Figure 1.4: **a** gives the total number of crysal structure discovery within the ICSD by year. **b** depicts how many of these new entries are stable at $0K$, where the stabilities are provided by the OQMD.[14]

The aforementioned databases contain an enormous amount of crystal structures and from the growth rate in Figure 1.4 one can infer that they are still far from complete. For the time being there is no indication that the growth rate will deminish. Glotzer[13] performed a research on the challenges and opportunities for the future of solid-state chemistry. He discusses the rate of discovery and chances of finding new binary, ternary, quaternary,... compounds. For binary valence compounds the composition can be predicted based on oxidation states and at the moment essentially all of these binary compounds have been synthesized. This statement is evidenced by Figure 1.4, as the number of newly found stable binaries declines steadily over the years. The crystal structure (and composition) of ternary and quaternary compounds on the other hand take more effort to predict, as oxidation state rules often do not help. Attempts at extrapolating knowledge from binary compounds are hindered by the significantly higher number of potential crystal types. Some of these structure types have a lot of representatives (approximately 500 solids adopt the $ThCr_2Si_2$ structure I4/mmm[13]), while the most of the structures have only few representatives.

So, how many crystals are left to be discovered? Answering this question is not evident, but one can make an educated guess. To make this guess some simplifications are needed. The first of which is that all crystals will be assumed to follow the 1:1:1:...:1 stoichiometry. Secondly, the different kinds of crystal structures will be ignored: it is assumed that every chemical formula corresponds with just one crystal structure. Armed with these two approximations, the graphs on Figure 1.5 can be created. The blue line is the binomial co-

efficient $\binom{90}{N}$ where N unique elements are chosen from the first 90 elements on the periodic table. In fact, this is exactly how many crystals with N different elements could potentially exist. One would expect to find significantly more crystals with ~45 unique elements than any other crystal, as it has over 10^{26} possible combinations. In reality however, especially crystals with a very low number of elements are found to be common. This observation suggests that there is another factor playing which disadvantages the bountiful materials so the simple ones become more abundant.

Until now, decompositions of potential crystals were not taken into account. Any crystal of N unique elements has a number of possible decomposition paths to crystals with a lower number of elements. This was already discussed in Section 1.1.2. This number of decomposition paths is called the Bell number B_N, where the initial undecayed crystal is included. An example of the Bell number can be given for the ternary crystal in equation (1.1), where $B_3 - 1 = 4$. If one of these decomposition paths has a positive E_{decomp}, the original crystal will naturally decay to this path.

The red graph on Figure 1.5 gives the Bell number for each number of unique elements in a crystal. $B_N - 1$ rises dramatically to enormous numbers: a fictive crystal containing the first 90 elements of the periodic table would have more than 10^{101} ways to disintegrate. As soon as one of these many possibilities leads to a lower energy, the initial crystal will not be stable. Glotzer[13] did a more concise but similar exercise on this topic as the one above.

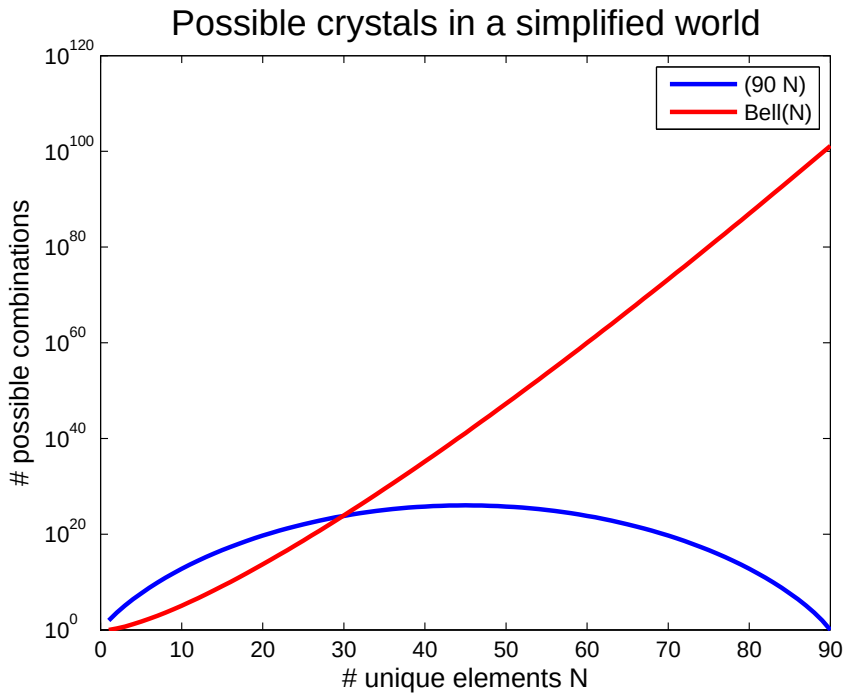


Figure 1.5: The blue line gives the number of possible combinations with N unique elements chosen from the first 90 elements on the periodic system. The red line gives the number possible decomposition products with less unique elements.

1.3 Quaternary Zintl phases

The quaternary phase space is immense. Using the same approximations as in Section 1.2, 2 555 190 materials can be found. Allowing other indices than 1:1:1:1 and adding all possible crystal structures for each chemical formula leads to an even bigger number. This defines a large search space of possible quaternary crystals and legitimates the choice for a smaller, well-defined feasible set to start the exploration. Finding a innovative unexplored set leads us past the popular sets of the quaternary Heusler phases[15;16;17] and the quaternary MAX phases[18;19;20;21] and brings us to the QZPs. These materials are interesting yet have only undergone limited exploration, leaving us a lot of freedom to search. The Zintl phases (ZPs) are defined formally in Section 1.3.1, whereafter a more handy and pragmatic definition is chosen in Section 1.3.2.

1.3.1 Zintl phase definition

ZPs inherit their name from the German chemist Eduard Zintl who pioneered the detection of this large group of materials in the 1930s with the publication of a series of articles on materials such as NaTl[22], LiIn[23] and LiAl[24]. The materials that Zintl investigated did not follow the Hume-Rotery rules[25] for crystal structures in the intermetallic phase, and so their properties differed from other previously known crystal structures. A representative ZP contains at least one alkali or alkaline-earth metal and one or more post-transition elements or metalloids, together called metametals. These two categories of elements bond together in an anion-cation compound by transferring electrons from the electropositive group I-II elements to the electronegative metametals.

Sevov[26] describes the classic ZPs as chemical compounds that:

1. contain an alkali or alkaline-earth metal and a p-block element (or elements) that is a metal, semimetal, or small-gap semiconductor
2. are electronically balanced or closed-shell compounds, i.e. the number of electrons provided by the constituting elements equals the number of electrons needed for covalent bonding in the structure
3. have very narrow or no homogeneity width, i.e. they are line compounds
4. are semiconductors or poor conductors
5. are diamagnetic or show very weak, temperature- independent paramagnetism
6. are brittle

Criteria 2 and 6 are characteristic properties of normal valence compounds like NaCl, KCl and Al_2O_3. The number of valence electrons of anions (e_A) and cations (e_C) for such a normal valence compound C_mA_n (C = Cation, A = Anion) obey the simple rule $me_C = n(8 - e_A)$[27] , where all atoms follow the octet rule.

The main difference between normal valence compounds and ZPs is that the ZP anions do not achieve an octet structure by the full electron transfer from the cations. Moreover, the anions have to share their valence and received electrons with each other in order to do so[26;28]. Because of the full electron transfer from the electropositive alkali or alkaline-earth metal in a ZP C_mA_n, the cations $(C^{\alpha+})_m$ obtain an octet structure through oxidation to the electronic configuration of the nearby lighter noble gas. This is typical for ionic bonds. The anions $(A^{\frac{m\alpha}{n}-})_n$ on the other hand group together via covalent bonds to reach the octet structure.

Each shared pair of electrons provides one additional electron for the anions. This necessary combination of covalent and ionic bonding properties is the reason why criterium 1 is such an important requisite of being a ZP and it bridges the gap to the general class of intermetallics.

The foregoing digression on electron exchange and the different types of bonds in a ZP is known as the Zintl concept[29]. Klemm[30] generalized this idea with the proposition of treating the charged $A^{\frac{m\alpha}{n}-}$ as a pseudo-atom. Such pseudo-atom would adopt the structure characteristics of the isoelectronic main group element. The combination of Zintl's original concept and Klemm's pseudo-atom description, is called the Zintl-Klemm concept. To illustrate this, a general binary ZP *CA* is chosen. The cation *C* donates its valence electron(s) to the anion *A*, which gets negatively charged. The anion is thus converted into a pseudo-P atom ψ-P (where the P stands for the p-block). ψ-P then behaves like an atom from one group to the right on the periodic table, with the corresponding structural properties of that group.

1.3.2 Suitable QZP elements

Good agreements make good friends; unfortunately good agreements are hard to find when determining the seperation of different ZP components, named the 'Zintl border' or 'Zintl line'. Several definitions of the Zintl line were proposed during history, including the ones from Zintl and Klemm themselves[28]. Nevertheless, the qualifications for belonging to the ZPs are often taken in a somewhat broader interpretation. Compounds that include non-metals (KBa_5P_4[31]) or transition metals instead of alkali and alkaline-earth metals (CsAu[28]) are frequently considered being part of the Zintls.

As a result of these doubtful borderline cases, one is tempted to look for an appropriate practical definition. The periodic table in Figure 1.6 highlights two zones of typical Zintl elements. The yellow zone contains the alkali and alkaline-earth metals, the blue zone the metametals. The materials that will be discussed further on in this research will only be constructed from these marked elements, where at least one element has to come from both coloured regions.

The selection does not include Polonium or heavier elements as the MP[6] does not hold enough information on binary and ternary compounds containing these elements.

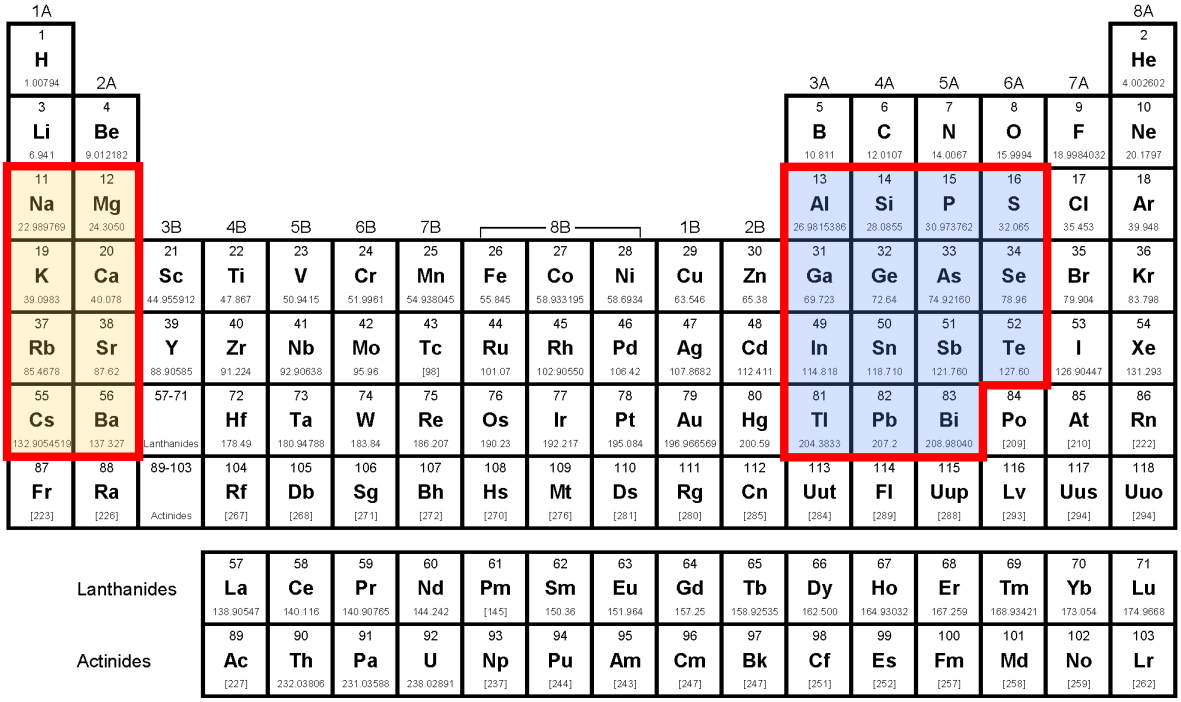


Figure 1.6: A periodic system showing two regions of suitable elements according to the Zintl definition. The alkali and alkaline-earth metals are marked yellow while the metals, semi-metals and small-gap semiconductors are marked blue.

1.4 QZPs in the Pnma space group

Opposite to the binary and ternary ZPs, the QZPs are almost totally undiscovered. Figure 1.7 depicts the rate of discovery of QZPs from the beginning of the 70s until 2014. Up to 1990 only 10 QZPs were experimentally known, after that a linear trend started to become visible. Investigating this phase space with high-throughput calculations could give a global idea of the stable crystals and in that way speed up the experimental discoveries.

The precursor[4] of this work discussed the brute force calculation of the orthorhombic space group Pnma, with the experimentally known $CsPbPSe_4$ as flagship. This brute-force computation of 4224 QZPs was performed by a combinatorial exchange of the elements in the $CsPbPSe_4$ template crystal with other elements from the same or nearby groups of the periodic table.

A summary of the results of Dumon's work[4] can be seen in Figure 1.8, where a grid is made of the necessary alkali and earth-alkaline metals in the vertical direction and the group VI elements in the other direction. So, on each tile of this grid two elements of the QZP are fixed and two are still undetermined. These pairs of undetermined elements are chosen from group III, IV and V and have 132 possible combinations. Each tile on Figure 1.8 then shows the minimum pure formation energy of 132 calculated solids. Apparently, the Po-possessing structures are much more unstable than any other analogous material.

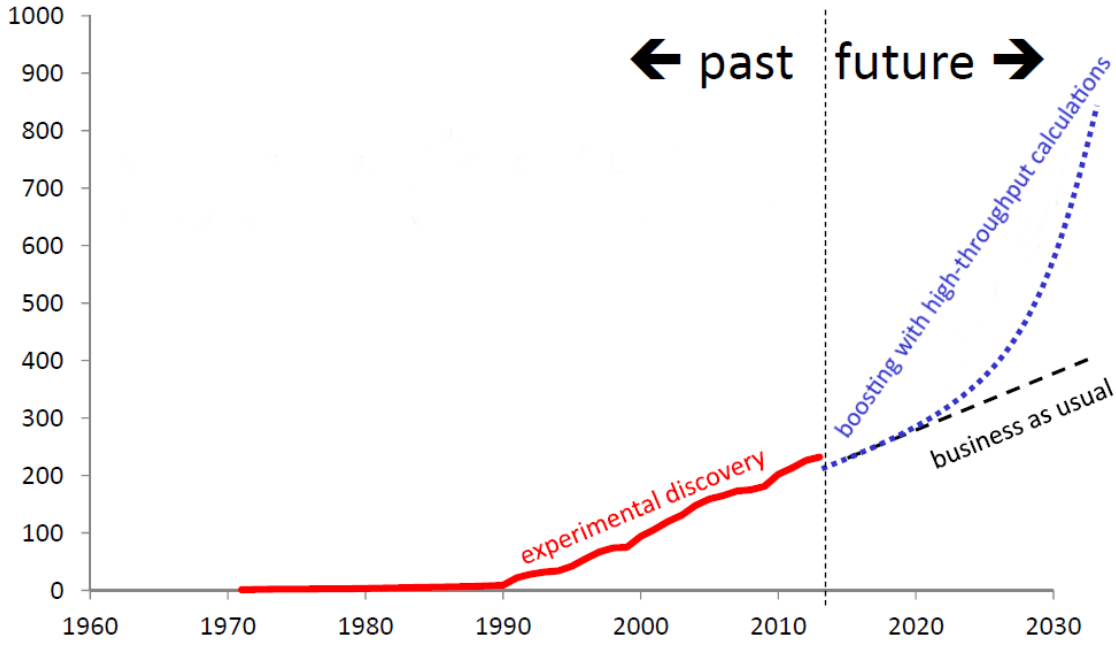


Figure 1.7: Discovery rate of the QZPs, as registered in the ICSD up to 2014[4]. The linear trend that started in the early 90s could be bent with high-throughput calculations.

Figure 1.9 shows two resulting structures from this elaboration. $BaAlGaS_4$ reveals a zigzag pattern for the Al and Ga atoms, while the equivalent Al and Si atoms in $CsAlSiS_4$ are placed collinear. Because the zigzag pattern corresponds with the template crystal structure of $CsPbPSe_4$, $CsAlSiS_4$ seems to have a new set-up that has never been experimentally observed before. Thus, it may be clear that the approach of this research made room for multiple different structures.

With the brute force calculation of the Pnma QZP family, Dumon[4] paved the way for the work done in this research. This completed data set will offer valuable assistance in the pursuit of a more intelligent method of running through other large QZP sets.

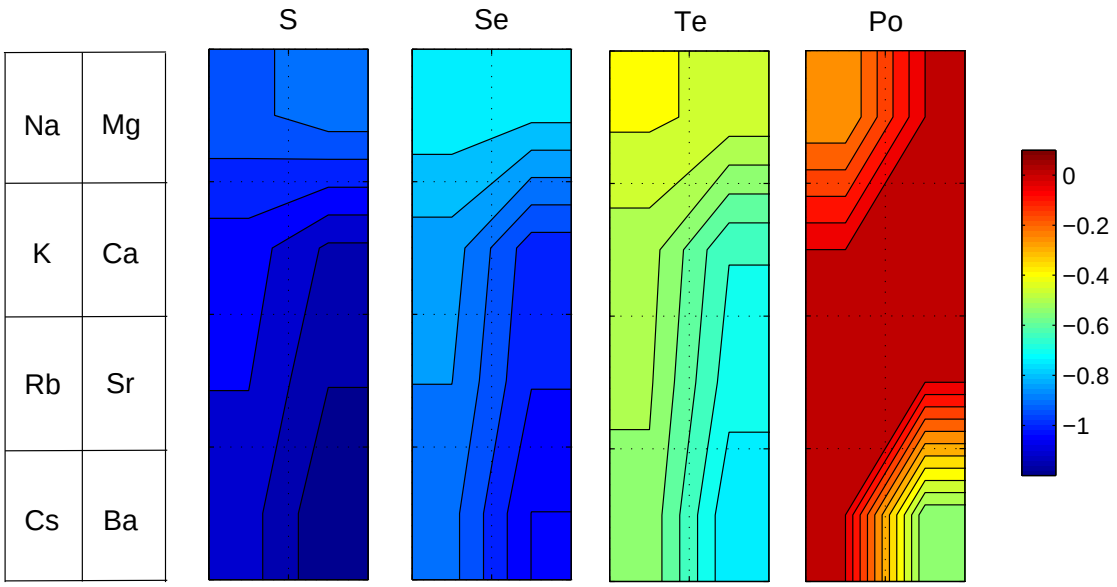


Figure 1.8: This contour plot gives an overview of Dumon's results[4] of the Pnma space group. The averages of the minimum pure formation energies (Ev/atom), are placed on the contour plot according to the alkali or alkaline-earth metals and per 4th element in the QZP (S,Se,Te,Po). Each tile represents the minimum of 132 QZP pure formation energies.

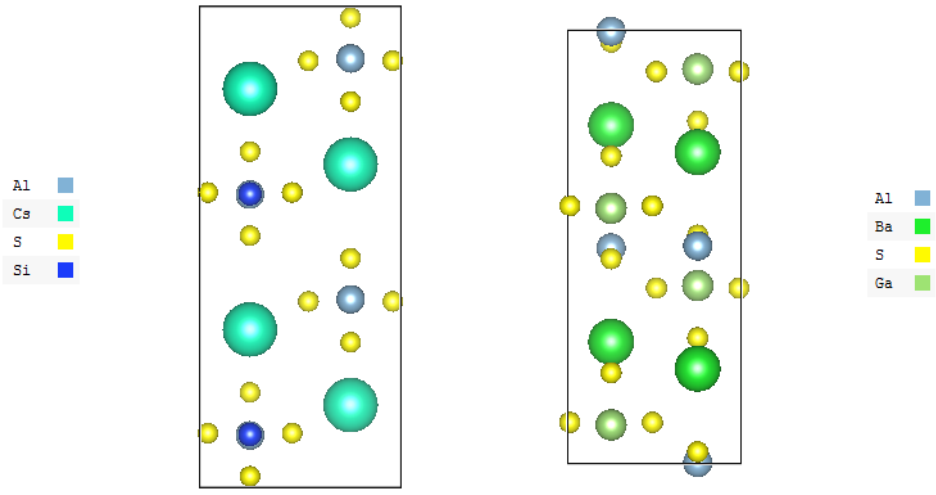


Figure 1.9: Two QZP examples (left: $CsAlSiS_4$, right: $BaAlGaS_4$) of relaxed structures from the Pnma space group, looking down the a-axis[4].

Chapter 2

High-throughput DFT

High-throughput DFT, these two terms hide an enormous load of scientific research that can easily fill more than one chapter. This chapter elaborates briefly on the theoretical background and the practical implementations of the computational methods used in this research. Density Functional Theory (DFT) will be explained in Section 2.1. Thereafter, in Section 2.2, the Vienna Ab initio Simulation Package (VASP) is proposed as the software package for the necessary DFT calculations. Lastly, the high-throughput concept is followed in Section 2.3.

2.1 Crystal modelling: the many-electron problem

There are three levels of approximation when handling a many-body problem like the QZP crystals discussed in this work. On the basis of these three levels, the necessary computational methods will be explained.

The first approximation to be done is the Born-Oppenheimer (BO) approximation (Section 2.1.1), an often used technique in molecular dynamics. After that, one has to choose a good approach for the many-electron problem which will be DFT in this case. Section 2.1.2 gives a concise summary of DFT. Finally, implementing these two theoretical tricks requires some typical numerical approximations that are treated in Section 2.1.3.

2.1.1 The Born-Oppenheimer approximation

Digressions on the BO approximation can be found in various computational textbooks. Thijssen's disquisition[32] will followed here.

The Hamiltonian of a system consisting of electrons e and nuclei n reads:

$$\hat{H} = \hat{T}_e + \hat{T}_n + \hat{V}_{ee} + \hat{V}_{ne} + \hat{V}_{nn} \tag{2.1}$$

The first two terms of $\hat{H}$ represent the kinetic energies of the electrons and nuclei respectively. The third term describes the Coulomb repulsion between the electrons and the fourth term the Coulomb attraction between the electrons and nuclei. Lastly, $\hat{V}_{nn}$ contains the Coulomb repulsion between the nuclei. Writing out this Hamiltonian results in an expression that is impossible to solve if the number of electrons and nuclei is not extremely small (typically smaller than four).

Separating the degrees of freedom connected with the motion of the nuclei from those of the electrons is the first step of the BO approximation. This idea can be justified with the fact that electrons move much faster than nuclei. Any change in nuclear position leads to an immediate adjustment of electron motion. Therefore,one can state that the Hamiltonian for the electrons is generated by a static configuration of the nuclei. The total energy is the sum of the energies of the electrons and the energy resulting from the Schrödinger equation satisfied by the nuclei. In a further approximation, the motion of the nuclei is neglected and only the electrostatic energy of the nuclei should be added to arrive at the total energy. The BO Hamiltonian for the electrons can be derived from equation (2.1) and reads:

$$\begin{aligned} \hat{H}_{BO} &= \hat{T}_e + \hat{V}_{ee} + \hat{V}_{ne} \\ &= \hat{T} + \hat{V} + \hat{V}_{ext} \end{aligned} \tag{2.2}$$

$\hat{T}$ and $\hat{V}$ are purely electronic here, whereas the electronic interaction with the nuclei is interpreted as a external field $\hat{V}_{ext}$. It is important to mention that the nuclei positions can be varied in order to find the minimum of the total energy, which is the ground state of the system. But even with the nuclei positions fixed, the simplified Hamiltonian (2.2) is far from solvable, since too many degrees of freedom are involved.

2.1.2 Density Functional Theory

Several methods exist for reducing equation (2.2) to an even more executable form. Hartree-Fock formalism (HF) and DFT are the two most common ways to perform this second level approximation. The family of post-HF methods accomplishes the needs very well for atoms and molecules, but disappoints for extended systems like solids. However it is not impossible[33]. A better option to examine the QZP crystals in this work is the above-mentioned DFT. Two important theorems presented by Hohenberg and Kohn in 1964 initiated DFT, which meant the beginning of an enormous growth in computational material physics.

The Hohenberg-Kohn theorems

Consider a many-electron system, enclosed in a large box and moving under the influence of an external potential V_{ext} and their mutual Coulomb repulsion. It may be clear that the ground-state electron density $\rho(\mathbf{r})$ is a functional of this external potential. With their first theorem, Hohenberg and Kohn[34;35] showed that this relation also applies the other way around. There is a unique correspondence between $\rho(\mathbf{r})$ and V_{ext} and thus the ground-state energy from the Hamiltonian (2.2) can be rewritten as a functional of the ground-state electron densities:

$$E[\rho] = \hat{T}[\rho] + \hat{V}[\rho] + \hat{V}_{ext}[\rho] \tag{2.3}$$

In the same general many-electron system it is known that the wavefunction Ψ is a functional of $\rho(\mathbf{r})$, and so is the kinetic and interaction energy:

$$F_{HK}[\rho(\mathbf{r})] \equiv \langle \Psi | (T + V) | \Psi \rangle \tag{2.4}$$

$F_{HK}[\rho(\mathbf{r})]$ is called the Hohenberg-Kohn density functional. Because this functional does not contain any information on the nuclei and their positions, it is a universal functional: valid for any number of particles and any external potential. This means that a general expression exists which can be used for every atom, molecule or solid which can be imagined. Unfortunately, it should be stressed that no explicit expression for $F_{HK}[\rho(\mathbf{r})]$ is known. With the aid of equations (2.3) and (2.4) one can define the energy functional for a given external potential V_{ext}:

$$E_{V_{ext}}[\rho] \equiv \int V_{ext}(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} + F_{HK}[\rho(\mathbf{r})] \tag{2.5}$$

Hohenberg and Kohn show, in their second theorem, that equation (2.5) reaches the absolute minimum when ρ equals the ground-state density.

The Kohn-Sham equations

From the Hohenberg-Kohn theorems, Kohn and Sham[35;36] developed approximation methods for treating inhomogeneous many-electron systems, which resulted in a practical formulation of DFT. These methods are exact for systems of slowly varying or high density.

The first step leading to the Kohn-Sham (KS) equations is to rewrite the Hohenberg-Kohn density functional F_{HK} as[35] :

$$F_{HK} = T_0 + V_H + V_{xc} \tag{2.6}$$

where T_0 is the is the functional for the kinetic energy of a non-interacting electron gas and V_H is the Hartree contribution. The third term V_{xc} contains contributions from the quantum mechanical correlation and exchange interactions between the electrons and is formally unknown. The energy functional (2.5) becomes:

$$E_{V_{ext}}[\rho] = T_0[\rho] + V_H[\rho] + V_{xc}[\rho] + V_{ext}[\rho] \tag{2.7}$$

The trick to find a useful expression is to interpret this expression as the energy functional of a non-interacting classical electron gas, subject to two external potentials. These potentials are generated by the nuclei on the one hand, and by exchange and correlation effects on the other hand. The corresponding KS Hamiltonian $\hat{H}_{KS}$ is then:

$$\begin{aligned} \hat{H}_{KS}[\rho] &= \hat{T}_0[\rho] + \hat{V}_H[\rho] + \hat{V}_{xc}[\rho] + \hat{V}_{ext}[\rho] \\ &= -\frac{\hbar^2}{2m_e}\nabla_i^2 + \frac{e^2}{4\pi\epsilon_0}\int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}d\mathbf{r}' + V_{xc} + V_{ext} \end{aligned} \tag{2.8}$$

Applying this Hamiltonian on the single-particle wave functions $\phi(\mathbf{r})$, the intended KS equation appears:

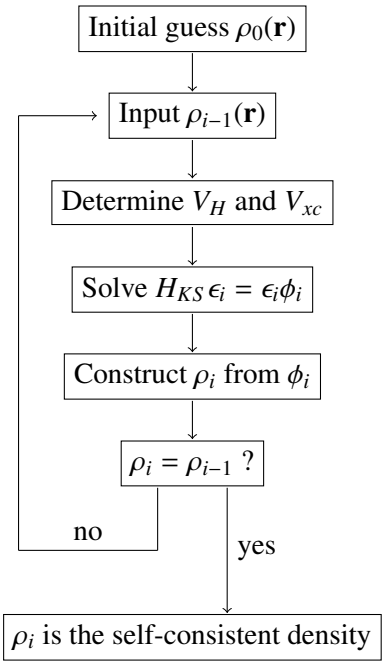


Figure 2.1: This scheme shows the iterative DFT workflow starting from the initial guess of the ground-state density to the final self-consistent density[35].

$$\hat{H}_{KS}\,\phi_i = \epsilon_i\phi_i \tag{2.9}$$

The N lowest energy solutions of this equation contribute to the ground-state density of the whole N-electron system:

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} \phi_i^*(\mathbf{r})\phi_i(\mathbf{r}) \tag{2.10}$$

It should be noted that the mentioned single-particle wave functions are not the wave functions of the physical electrons. These are wave functions of mathematical quasi-particles whose overall density is equal to the true electron density.

The attentive reader sees that a new paradoxal problem has emerged. The Hartree operator V_H and the exchange operator V_{ext} depend on the density $\rho(\mathbf{r})$, while this depends on the wave functions ϕ_i that determine V_H and V_{ext}. An iterative scheme as in Figure 2.1 provides relief to this self-consistent problem.

Ascertain the exchange-correlation functional

Thus far, the deductions of the Hohenberg-Kohn theorems and the KS equations did not assume any approximations. But as the exchange-correlation functional is still unknown, approximations for this functional have to be made. Several good attempts made their entry during the past decades[37]. The local spin density (LSD) approximation from Kohn and Sham[36] was a good standard since the early 1970s, whereafter generalized gradient approximation (GGA)[38;39;40] took this role from the 1990s. The GGA depends both on the local

density $\rho(\mathbf{r})$ and has a factor $|\nabla\rho(\mathbf{r})|$, which gives information on the non-local behaviour of this density. Perdew, Burke and Ernzerhof developed a simple but accurate derivation of GGA[41;42] and this GGA PBE functional will be used for all DFT calculations in this work.

2.1.3 Solving the equations

Now that DFT is introduced one can think of numerical approaches to implement this theory. This includes the choice of an approximating basis set, a **k**-point grid and a pseudopotential that replaces the real potential of the inner electrons.

Basis sets

The full expression of the Schrödinger equation (2.9) can be written down as:

$$\Big(-\frac{\hbar^2}{2m_e}\nabla_i^2 + \frac{e^2}{4\pi\epsilon_0}\int\frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + V_{xc} + V_{ext}\Big)\phi_i(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r}) \tag{2.11}$$

After defining a basis set ϕ_p^b, all solutions of the Schrödinger equation will have the form

$$\phi_m = \sum_{p=1}^{P} c_{mp}\phi_p^b \tag{2.12}$$

Due to obvious limitations of computers, it is impossible to define infinite basis sets and therefore the calculated solutions ϕ_m can never be precisely the aimed wave functions of the real system. The next task is then to find the best matching finite basis set to this system.

In a periodic solid, one can use Bloch's theorem to show that the orbital wave functions can be expressed as the product of a planewave and a function with the periodicity of the lattice[43]. Considering a crystal with reciprocal lattice vectors **G**, one can express the wave function as a sum of planewaves:

$$\phi_i(\mathbf{r}) = \sum_{\mathbf{G}} c_{i,\mathbf{k}+\mathbf{G}}\, e^{(i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r})} \tag{2.13}$$

Where $c_{i,\mathbf{k}+\mathbf{G}}$ are the coefficients that have to be adapted to find the the lowest energy solution. This planewave approach changes equation (2.9) from an integral equation to a set of algebraic equations.

One can limit the number of basis components by only including the wave functions in the basis set that have an energy beneath a certain cutoff energy. Thereby, all new crystal structures will need to be subjected to a convergence test to obtain the optimal cutoff energy. It is important to know that whenever DFT calculations for multiple systems are compared to calculate energy differences, the same energy cutoff should be used in all calculations[44]. With this thought in mind an energy of 400 eV is preferred, as this was the cutoff energy that was used and tested in the master thesis of Dumon[4].

k-point sampling

When modelling periodic systems like crystals, it is often assumed that this crystal reaches to infinity in all three dimensions. One can map the quantum numbers of such lattices to a continue and finite part of the momentum or reciprocal space. It turns out that wave functions at k-points that are close together are similar, thus an interpolation scheme can be used with a finite number of k-points[43]. The most common schemes for generating k-points are the Chadi-Cohen scheme[45] and the Monkhorst-Pack scheme[46]. The second scheme is the one that will be used during this research. Convergence tests will determine the optimal number of k-points in each direction, because their number depends on the specific crystal structures.

Pseudopotentials

The core electrons of an atom are extremely localized and their orbitals are therefore rapidly fluctuating. Trying to to expand these fluctuations with a planewave basis set requires high-frequency components in this set[43]. Furthermore, the effect of these core electrons on the bonding between atoms is negligible in comparison with the effect of the valence electrons. It is preferable to substitute the potential due to the core electrons with a pseudopotential that has a similar effect on the remaining valence electrons. A good choice to deal with this is the Projector Augmented-Wave method (PAW)[47;48]. These are the pseudopotentials that will be used during this research.

2.2 The Vienna Ab initio Simulation Package

The software package that will be used to perform the necessary DFT calculations is VASP[49]. This section elaborates on the specific aspects of this package (Section 2.2.1), the workflow of the calculations that will be performed later in this work and the corresponding VASP settings (2.2.2).

2.2.1 What is VASP?

The VASP Guide[49] describes VASP as a complex package for performing ab-initio quantum-mechanical molecular dynamics (MD) simulations using ultrasoft pseudopotentials or the PAW method and a planewave basis set. The size of this basis set can be kept very small, in most cases 50 planewaves per atom are sufficient. The approach implemented in VASP is based on the local-density approximation with the free energy as variational quantity and an exact evaluation of the instantaneous electronic ground state at each MD time step.

The self-consistency cycle that is implemented is rather traditional. The combination of this scheme with very efficient numerical algorithms leads to a robust and fast way of evaluating the KS equations.
VASP includes a full-featured symmetry code which determines the symmetry of arbitrary configurations automatically and which is used to set up the Monkhorst Pack special k-points. The integration of the band-structure energy over the Brillouin zone is performed with smearing or tetrahedron methods. For the tetrahedron method, Blöchl's corrections, which remove the quadratic error of the linear tetrahedron method, can be used resulting in a fast convergence speed with respect to the number of special points.

For a system with N valence electrons, the execution time scales like N^3 for some parts of the code. For systems with 2000 valence electrons, the N^3 term becomes comparable to other terms and on account of this fast growing scaling relation, VASP is expected to be useful for systems smaller than 4000 valence electrons. In this work N will be typically of the order of 200 electrons, which is far below the performance inflection point.

2.2.2 The calculation workflow and VASP settings

In order to determine whether a crystal is stable, a series of calculations will be performed. The first step of this series will be a volume relaxation: VASP searches for the minimal total energy of the system where only the cell volume can be varied. Thereafter a general relaxation of all degrees of freedom (cell volume, cell shape and ion positions) takes place, starting from the result of the first step. The third step is a Birch-Murnaghan equation of state (EOS) fit[50] on 7 DFT-calculated data points. The volumes of these data points differ from 96% to 106% of the volume from the obtained volume in the previous step. The Birch-Murnaghan fit returns a minimal energy at a given cell volume whereafter the final step is to calculate the exact total energy at this minimum. After the last step, more other interesting information can be retrieved with a Density Of State (DOS) or bands calculation.

VASP features more than 60 different input tags. For the major part of these tags, a default value exists that is appropriate for most calculations. The settings that are interesting in the context of this research are discussed here on the basis of the VASP Guide[49] and the settings for each step can be found in Table 2.1 and Table 2.2.

EDIFF - This is the global condition for the energy to break out of the self-consistency loop. If the total free energy and the band structure energy between two iterations differs less than EDIFF, the total energy has been found.

ENCUT - This is the cutoff energy for the planewave basis set in eV. It has already been mentioned above that the ENCUT of different calculations must be the same in order to compare the results of these calculations.

IBRION - This tag indicates how the ions are updated and moved. IBRION = 2 is preferred for difficult relaxation problems. This activates the conjugate gradient algorithm, which presently possesses the most reliable backup routines. For IBRION = -1, no updates are done and the ions stay at their initial location.

ICHARG - This tag gives information on the intial charge density. ICHARGE = 1 reads the charge densities in from a file. ICHARGE = 2 takes a superposition of the atomic charge densities. Adding 10 to the ICHARGE (10, 11, 12) means that the charge density will be held constant during calculation.

ISIF - This tag controls wether the time-consuming calculation of the stress tensor has to be performed. If so, ISIF determines which degrees of freedom (ions, cell volume, cell shape) are allowed to change. ISIF = 3 allows changes in all degrees of freedom, ISIF = 7 only allows changes in cell volume.

ISTART - A calculation with ISTART = 0 starts from scratch. ISTART = 1 starts the continuation of a job.

LCHARG - This tag determines if the charge densities are written to a file.

Table 2.1: The settings for volume relaxation (left), general relaxation (mid) and the single points for the EOS-fit (right)

Volume relaxation	
EDIFF	1.00E-06
ENCUT	400
IBRION	2
ICHARG	2
ISIF	7
ISMEAR	1
ISPIN	1
ISTART	0
LCHARG	.TRUE.
LWAVE	.FALSE.
NPAR	1
NSW	500
PREC	Accurate
SYSTEM	ISIF7

General relaxation	
EDIFF	1.00E-06
ENCUT	400
IBRION	2
ICHARG	2
ISIF	3
ISMEAR	1
ISPIN	1
ISTART	0
LCHARG	.TRUE.
LWAVE	.FALSE.
NPAR	1
NSW	500
PREC	Accurate
SYSTEM	ISIF3

EOS-fit	
EDIFF	1.00E-06
ENCUT	400
IBRION	-1
ICHARG	2
ISMEAR	-5
ISPIN	1
ISTART	0
LCHARG	.TRUE.
LWAVE	.FALSE.
NPAR	1
PREC	Accurate
SYSTEM	single point

LWAVE - This tag determines if the orbitals are written to a file.

NPAR - This tag organizes the parallelization over a number of nodes. For calculations running on a small number of nodes, NPAR = 1 gives the best performance.

NSW - This is the maximum number of ionic steps.

PREC - Changing the PREC parameter, adapts the default value of some other tags. Possible values for the PREC tag are: Normal, Single, Accurate, Low, Med, High.

ISMEAR - This tag determines how the partial occupancies are set for each wavefunction: ISMEAR = 0 causes a Gaussian smearing, ISMEAR = -5 the tetrahedron method with Blöchl corrections.

NEDOS - For a calculations where the density of states (DOS) is calculated, NEDOS gives the number of grid points in the DOS.

SIGMA - This tag determines the width of the smearing.

SYSTEM - This can be interpreted as a comment or title tag. The value of this tag is for the user only, as a reminder of the kind of calculation one wants to execute.

Table 2.2: The settings for DOS calculations (left) and Bands calculations (right).

DOS	
EDIFF	1.00E-06
ENCUT	400
IBRION	-1
ICHARG	11
ISMEAR	-5
ISPIN	1
ISTART	0
LCHARG	.TRUE.
LWAVE	.FALSE.
NPAR	1
PREC	Accurate
NEDOS	3001
SYSTEM	DOS

Bands	
EDIFF	1.00E-06
ENCUT	400
IBRION	-1
ICHARG	11
ISMEAR	0
ISPIN	1
ISTART	0
LCHARG	.TRUE.
LWAVE	.FALSE.
NPAR	1
PREC	Accurate
SIGMA	0.01
SYSTEM	BANDS

2.3 High-throughput computing

The demand for new materials with very specific properties is high and classic methods of experimentally composing crystals cannot satisfy this growing need. Experimental methods are not only time consuming, but also very expensive and inefficient. Thankfully, an alternative method is right under our nose. DFT proved its value and precision and supercomputer architectures are available for scientific research. Combining these two elements in an automated framework can drastically speed up the discoveries in material design. Section 2.3.1 defines the concept of high-throughput computing and bridges the gap with combinatorial computing. Section 2.3.2 gives a brief historical overview of high-throughput experiments in (computational) material science. At the end of the chapter, in Section 2.3.3, the high-throughput framework that will be used in this work is illustrated.

2.3.1 Definition

The European Grid Infrastructure[51] defines high-throughput Computing as "a computing paradigm that focuses on the efficient execution of a large number of loosely-coupled tasks. Given the minimal parallel communication requirements, the tasks can be executed on clusters or physically distributed resources using grid technologies. High-throughput computing systems are typically optimised to maximise the throughput over a long period of time and a typical metric is jobs per month or year."

The use of the terms combinatorial and high-throughput in literature is confusing and Maier *et al.*[52] determine the difference between these. According to Maier *et al*. the term combinatorial should refer to experiments in which groups or elements of different materials or components of a recipe, such as solvents, additives, or other components, are combined. Combinatorial thus refers to a change in the nature of the parameters, not to the change in the value of the parameters. The systematic variation of given compositions, temperatures, pressures, or other single parameters to explore a wide parameter space is not a combinatorial, but a high-throughput experiment.

It is clear that the work done in this research is of a combinatorial nature. The QM (see Section 2.3.3) that directs the processing is capable of performing both combinatorial and high-throughput calculations and since the small difference between both concepts, they will be seen as synonyms.

2.3.2 Examples

High-throughput research is not a recent phenomenon. During history several successful high-throughput experiments have been executed. The famous Edison systematically tested over 1600 different earths, minerals and ores to identify carbonized cotton thread as the optimal material for his light bulb[53]. In 1912, the Italian photochemist Ciamician exposed hundreds of potentially photoactive materials to the sun for his research on photochemical processes for batteries[53]. Both examples count as the first combinatorial approaches in material research.

Another example that is more related to current study is the work of Curtarolo *et al*[54]. This work presents the results of a first principles study of 14 080 computed total energies on 176 crystal structures in 80 binary alloys. Only 3 on 14 080 compounds are found to have a significant disagreement between experiment and ab initio DFT calculations. The scale and accuracy of this research is encouraging and establishes clearly the potential benefits of extensive high-throughput calculations.

The MP[6] , which was already mentioned in Section 1.1.2, is a high-throughput computational approach to screen up to tens of thousands of compounds for potential new technological materials. Because of the web-based dissemination of the computed materials datasets, many other researchers can benefit from their databases.

A high-throughput DFT research in the same context of present work and that also relies on the MP is performed by Cerqueira *et al*[55]. They tested the stability of ternary Si clathrates $C_8X_xSi_{46-x}$, whereby C is a group I cation and X a tested transition metal in a cubic Si-cage. 20 new phases are found likely to be produced experimentally.

2.3.3 The high-throughput framework

Jain[56] describes the calculation infrastructure used to compute some properties of approximately 80,000 compounds, encompassing the majority of unique compounds in the ICSD[8]. For a huge dataset like this, an efficient framework is needed to maintain an oversight. In the same paper Jain proposes a schematic process for a general high-throughput crystal computation, see Figure 2.2. The diagram starts from an external crystal structure database and leads to an iterative process whereby information from earlier calculations is used to select the following calculations. This data flow diagram will be the guidance for the QZP high-throughput DFT computation. Each step on the scheme is extensively discussed below for the specific needs of this research.

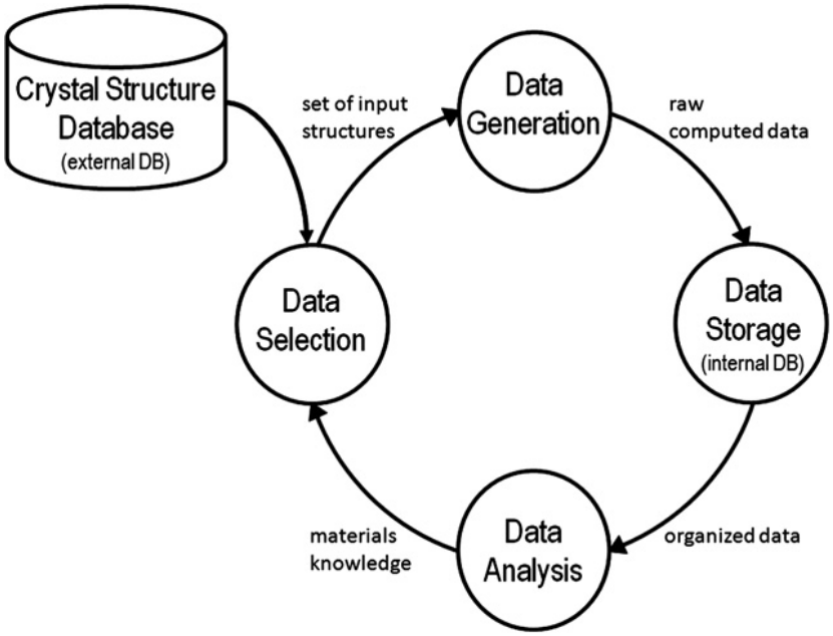


Figure 2.2: Data flow diagram highlighting the important steps in high-throughput computational screening and how these steps relate to each other.[56]

External database

The diagram starts from an external database that contains crystallographic parameters of experimentally known materials. The ICSD will take the place of the external database in this research. In Chapter 1, QZPs were found to be a good limited set to explore en thus these will be extracted from the ICSD. Once a good QZP in a new space group has been found, a family of similar QZPs will be created by chemical substitutions.

One QZP family will be of the order of several thousands materials. For managing all computational steps in such large sets, M. Sluydts developed the Queue Manager (QM) at the Center for Molecular Modeling (CMM). The QM allows to add calculation workflows that are specific to any project. Materials can then be coupled to a workflow and ordered in different queues. Furthermore, each entry has a priority that can be adapted manually during the progress of the research. Highly prioritized crystals will be calculated first.

Data selection

In the data selection phase, QZPs will be picked on the basis of information of earlier calculated QZPs. In the first iteration of the cycle there is no information available on previously calculated crystal structures and therefore materials are randomly selected to be generated.

The data selection phase is differently implemented in a later iteration. All members of the studied QZP family have a priority now and the QM selects the materials in descending order of priorities.

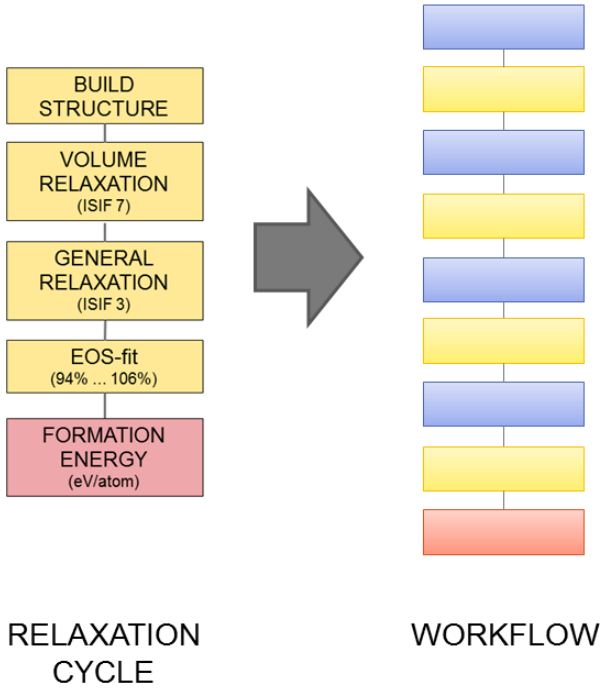


Figure 2.3: This figure shows the relaxation workflow in the same graphical interface as the QM. Yellow labels are jobs that are ready and waiting to be started. A yellow label turns blue when the job starts and yellow again when succeeded. When the whole workflow is finalized, the label turns red.[4]

Data generation

In the data generation phase, the structural data from the data selection phase are transformed into appropriate VASP input files and distributed over available computational resources. These computational resources were provided by the VSC (Flemish Supercomputer Center) and access was granted to the Flemish tier-1 supercomputer facilities for 1000 node days.

The QM shows the status of all running and waiting jobs as depicted in Figure 2.3 and the details of each calculation step is given in Section 2.2.2. In the first stage of the research only the relaxation workflow is fulfilled. Later, when all materials in one queue of the QM are relaxed, the properties like DOS and bands structures are calculated.

Data storage

In the data storage step, the raw output files from the computations are sent to an efficient database system. This internal database contains the results of all computations. The data that was generated can be accessed through the browser-based phpMyAdmin's interface[57] , whereby each material is identified by a unique filenumber and each calculation by its own id number. Materials from different structural families or projects can easily be distinguished from each other thanks to the queue subdivision, whereby every material belongs to one specific queue.

Data analysis

In the data analysis phase, the already generated data is examined in order to produce new information that may guide future data selection. The important information that will be extracted from these calculated crystal structures is the decomposition energy, using the binary and ternary energy delivered by the MP[6] and equation (1.3). This process will be performed for each material that reaches the end of the workflow. Subsequently, a scheme is needed in order to make a prediction about the other decomposition energies. Examples of such schemes published in the literature comprise evolutionary and genetic algorithms, data mining of spectral decompositions and Bayesian probabilities, refinement and optimization by cluster expansion and structure map analysis. Neural networks and support vector machines have also been utilized in a few cases[58]. For the goal of this research Neural Networks should perform well and Chapter 3 focusses further on this and other forms of machine learning.

Chapter 3

Machine learning

Machine Learning (ML) is a research field within computer sciences that concentrates on algorithms that can be trained with large amounts of data. These algorithms then make predictions for new data, according to how they are trained. Section 3.1 demonstrates a wide range of ML applications and focusses on the Neural Network (NN) algorithm. Subsequently, the theoretical background of the NN algorithm is examined in Section 3.2. In the last section of the chapter, Section 3.3, the Matlab toolbox for NNs is presented.

3.1 The scope of ML: an overview

ML algorithms generally map input vectors into different classes, where each input vector is described by a set of (statistical) features or attributes and their associated values. As there are several possible ML algorithms, the classification accuracy and computational performance are two important quantities on which the algorithms can differ. Of course, one wants to do as much and as accurate predictions as possible[59].

In this section some types of algorithms will be reported, where there has been made a distinction between applications within (Section 3.1.1) and outside (Section 3.1.2) the context of this thesis. The emphasis lies on the NNs, as this is the algorithm that will be used further on.

3.1.1 General ML applications

The Naive Bayes algorithm[59] is based on the Bayesian theorem. These classifiers estimate the probabilities of a feature having a certain feature value. After estimating this for the whole input vector, a conditional probability for the relationships between the feature values and the class can be made. Wang *et al*[60] developed a Naive Bayesian classifier for classifying bacterial rRNA sequences into bacterial taxonomies.

A C4.5 Decision Tree[59] creates a model based on a tree structure. Nodes in the tree represent features, with branches representing possible values connecting other features. A leaf representing the class terminates a series of nodes and branches. Determining the class of an instance is a matter of tracing the path of nodes and branches to the terminating leaf.

Support Vector Machines[61] (SVM) map the training vectors $\mathbf{x}_i$ into a higher dimensional space by a function ϕ. SVM finds then a linear separating hyperplane with the maximal margin in this higher dimensional space. The function $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ is called the kernel function and determines the nature of the mapping, which can be linear, polynomial, radial or sigmoid.

NNs are a set of algorithms that are designed to recognize patterns in large, complex datasets. The patterns they recognize can contain many kinds of real-world data like images, sound, text or in the context of this research: materials[62]. The possibility to describe all those quantities comes with a wide variety of applications. NNs made themselves useful for speech recognition[63;64;65] , 3D objects recognition from 2D photographic images[66], time-forecasting of financial markets[67], face recognition[68] and many more. NNs are also embedded in Google's image search, where learning word representations and learning image representations are combined in order to associate text with images[69;70].

3.1.2 ML in computational material physics

Meredig *et al*[71;72] combined two prediction models for the pure formation energy of ternary crystals against DFT calculations. The first model is a heuristic rule, which states that the formation energy in a ternary system can be estimated from its constituent binary formation energies. The second model is a rotation forest[73] ML model. Meredig *et al* trained both models on several thousand DFT calculations. The combination of the heuristic and the ML model results in a model that is six orders of magnitude faster than a DFT calculation and has a similar thermodynamic predictive accuracy. As an example application the evaluation of the stability of 1.6 million ternary compositions has been performed and ranked by their probability to exist as a stable compound. This search revealed a total of about 4500 until then unknown materials.

NNs become a more and more familiar tool in the context of computational material physics. Molecular Modeling (MD) and Monte Carlo (MC) simulations need a good underlying Potential Energy Surface (PES) for accurate results. PESs of simple chemical systems are typically provided by DFT calculations. For large clusters of atoms or bulk materials, these DFT calculations become computationally very demanding and ask for a better alternative. A generalized NN representation of high-dimensional PESs has been performed by Behler & Parrinello[74]. They represent PESs based on NNs taking into account the positions of all atoms of arbitrary sized systems and show that the accuracy of all investigated properties are very close to the DFT data. Specific eleborations of this general approach are the NN representations of the ab initio PES for the graphite-diamond phase coexistence[75] , the sodium PES at high pressures[76;77] and the PES for bulk TiO_2[78]. With this last example, Artrith and Urban[78] released the source code under an open license to advance the NN potential method in general.

However, the reason why NNs are used in this research is not because of exacting calculations. In this case it is the scale of the group of materials to be examined that is extraordinary. Brute force calculations of all the materials in one QZP family take a huge amount of time and can be speeded up with the help of NNs.

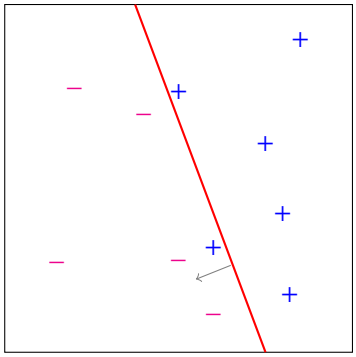


Figure 3.1: A 2D example, where the red perceptron line iteratively gets nudged to its final position. The gray arrow indicates in what direction the line is pushed.

3.2 Theoretical background

The goal of a ML algorithm is to approximate an unknown function $f : \mathbf{x} \mapsto y$. The vector $\mathbf{x}$ is here an input vector of all quantities that can possibly be related to the (in this case scalar) solution y. Relations that are not easy to pin down mathematically can be reconstructed by ML on the condition that a pattern exists and that there are enough data points $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)$ available. For the ease of the algorithm implementation a hypothesis set $\mathcal{H}$ has to be chosen, which is the set of all possible formulas. Two hypothesis sets that will be discussed here are the perceptron in Section 3.2.1 and the NN Model in Section 3.2.2, whereby Y. Abu-Mostafa's course[79] will be the guidance.

3.2.1 The perceptron

The Perceptron Learning Algorithm (PLA) is a straightforward algorithm for the classification of linearly separable data. For 2D data, the perceptron $h(\mathbf{x})$ is a line. For N-dimensional data this is a hyperplane. The data is binary classified by putting the dot product of the input vector $\mathbf{x}$ with its weight vector $\mathbf{w}$ across a treshold:

$$h(\mathbf{x}) = sign\Big(\sum_{i=1}^{d} w_i x_i - threshold\Big) \tag{3.1}$$

The threshold can be treated as a constant weight w_0 for a constant input $x_0 = 1$. This simple change in notation simplifies equation 3.1 to

$$h(\mathbf{x}) = sign\Big(\sum_{i=0}^{d} w_i x_i\Big) = sign(\mathbf{w} \cdot \mathbf{x}) \tag{3.2}$$

The PLA then goes as follows: given the data set $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)$, one picks a misclassified point $(h(x) \neq y_n)$ and updates its weight in this manner $\mathbf{w}_{new} \leftarrow \mathbf{w}_{old} + y_n \mathbf{x}_n$. This step has to be repeated until there are no misclassified points left. An intuitive example of one iteration step of the PLA is given in Figure 3.1.

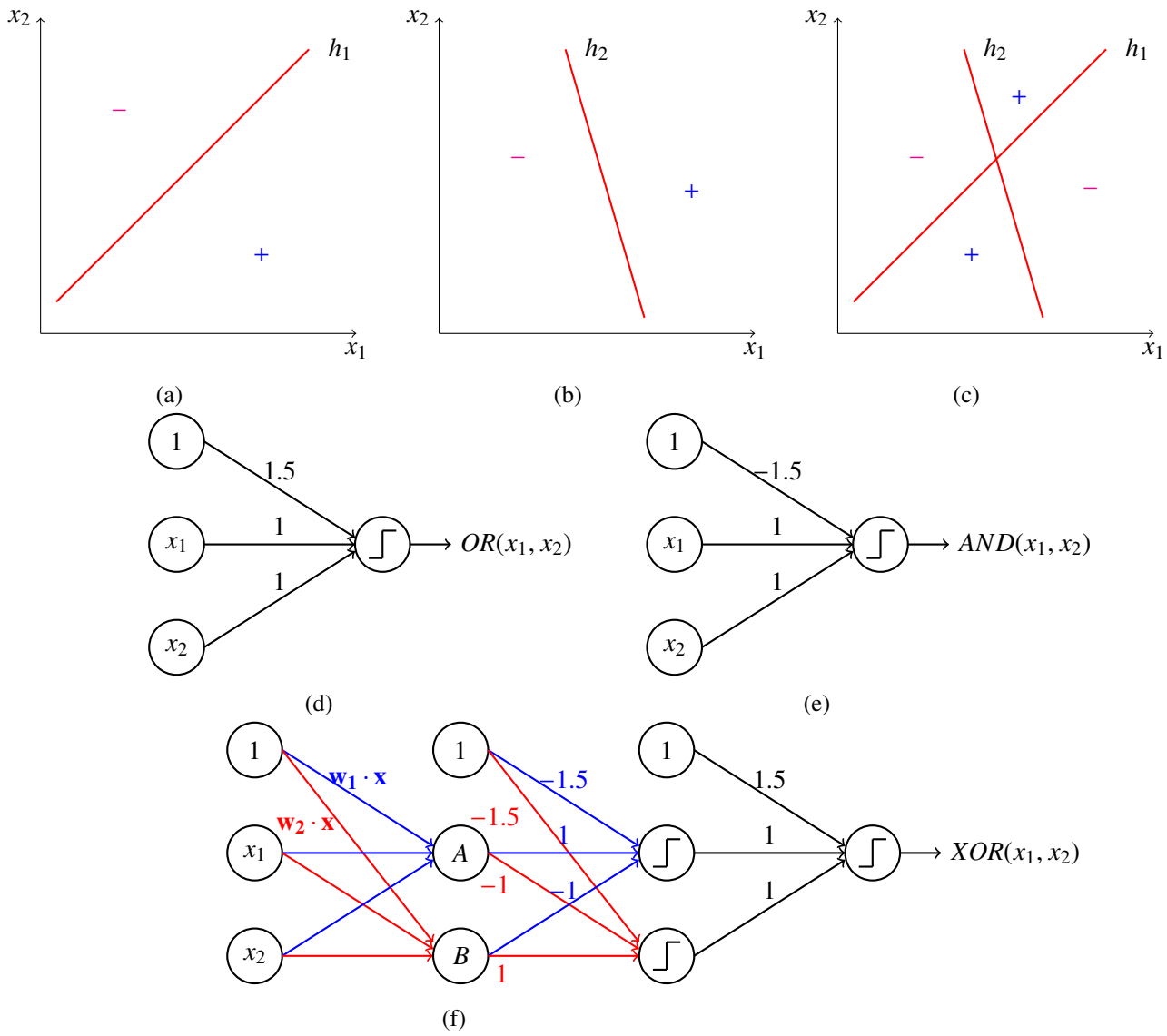


Figure 3.2: (a) and (b) show two simple two-dimensional perceptrons h_1 and h_2. As an example, these two perceptrons are used in (c) in an XOR combination. (d) and (e) show how the basic logical functions AND and OR can be implemented with the help of just one extra perceptron. The AND perceptron takes -1.5 as the threshold value so that x_1 and x_2 have to be 1 for a positive output. The OR perceptron on the other hand uses a $+1.5$ threshold, so that just one of the input values has to be 1 to give a positive output. As a final example (f) shows how the XOR function from (c) can be implemented with OR and AND functions by constructing the logical operation ($\bar{A}$ OR B) AND (A OR $\bar{B}$).[79]

3.2.2 The Neural Network model

The earlier introduced perceptron model plays an important role in the formation of the NN model. Because of the easy interpretable binary classification of the perceptrons, it is possible to create simple logical AND and OR functions. Subsequently, more complex logical functions can be created with the help of these basic functions. Figure 3.2 (a)-(f) shows the whole implementation of AND, OR and XOR functions starting from 2D seperators. Generalizing the 2D seperators to ND seperators and expanding the number of layers gives the possibility of creating multi-dimensional logical hyperbodies. These hyperbodies indicate the division of seperatable data, but for the aim of this research a NN is needed that can train on non-seperatable data.

The solution comes with a soft threshold $\theta(s)$, where s is the signal due to the projection of $\mathbf{w}$ on $\mathbf{x}$. This is a non-linear function with a range $]-1,1[$ where the limit of $\theta(s)$ equals the discrete threshold of the seperatable data. A good and typically chosen soft threshold is the hyperbolic tangent:

$$\theta(s) = tanh(s) = \frac{e^{s} - e^{-s}}{e^{s} + e^{-s}} \tag{3.3}$$

The signal $s_j^{(l)}$ of each layer l in a NN depends on the previous layer $l-1$ as it is a linear combination of the input vector $\mathbf{x}$ and weight $\mathbf{w}$ of the previous layer. The weights $w_{ij}^{(l)}$ of the lth layer out of L layers can thus be defined as follows:

$$w_{ij}^{(l)} \begin{cases} 1 \leq l \leq L & layer \\ 0 \leq i \leq d^{l-1} & inputs \\ 1 \leq j \leq d^{(l)} & outputs \end{cases} \tag{3.4}$$

where $d^{(l)}$ is the dimension of the lth layer. Note that $i = 0$ is included for the inputs because of the constant threshold x_0 which is always present. And so the recursive formula for the feed-forward NN becomes:

$$x_j^{(l)} = \theta(s_j^{(l)}) = \theta\Big(\sum_{i=0}^{d^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)}\Big) \tag{3.5}$$

The error $e(\mathbf{w})$ of each layer is determined by the weights. As one wants to minimize the error of each layer in the NN, an iterative Stochastic Gradient Descent algorithm (SGD) is used. A brief explanation of this SGD is shown in Figure 3.3. The gradient with respect to the weights is given by

$$\begin{aligned} \nabla e(\mathbf{w}) &= \frac{\partial e(\mathbf{w})}{\partial w_{ij}^{(l)}} \\ &= \frac{\partial e(\mathbf{w})}{\partial s_j^{(l)}} \frac{\partial s_j^{(l)}}{\partial w_{ij}^{(l)}} \\ &= \delta_j^{(l)} \frac{\partial s_j^{(l)}}{\partial w_{ij}^{(l)}} \\ &= \delta_j^{(l)} x_i^{l-1} \end{aligned} \tag{3.6}$$

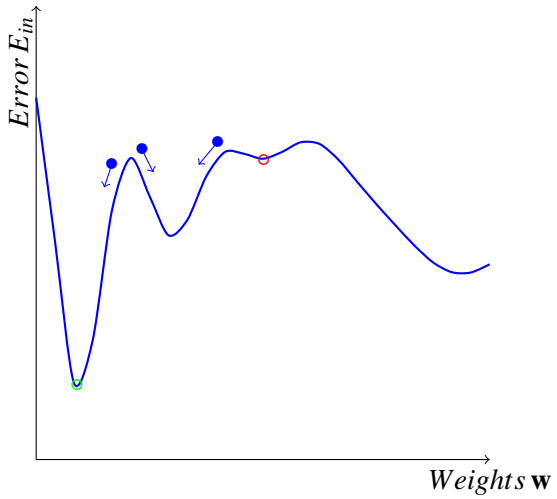


Figure 3.3: Stochastic Gradient Descent (SGD) typically minimizes functions by iteratively taking the gradient of a point and following it along the descending direction. The three blue dots in the figure show this for one iteration at three different points. Depending on where the algorithm starts, it ends up in a local minimum. The green and red circles on the graph show respectively an optimal and inferior local minimum. The stochastic aspect adds a random element to the GD, which makes it possible to escape from shallow minima like the one at the red circle.

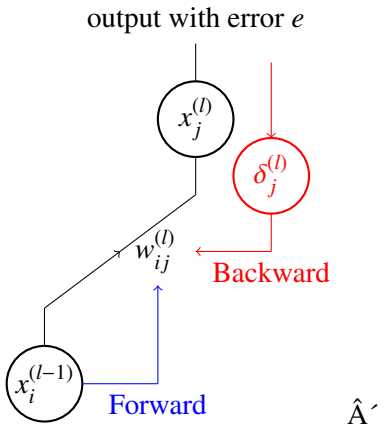
The factors $x_i^{(l)}$ are constructed from the outputs from the previous layer, as seen on the right panel of Figure 3.4. The factor $\delta_j^{(l)}$ has to be computed by backward propagation from the error on the outer layer. On this final layer the indices become $l = L$ and $j = 1$, as the output function is a scalar. When using a mean square error function, $\delta_1^{(L)}$ can be calculated as:

$$\begin{aligned} \delta_1^{(L)} &= \frac{\partial e(\mathbf{w})}{\partial s_1^{(L)}} \\ &= \frac{\partial}{\partial s_1^{(L)}} (x_1^{(L)} - y_n)^2 \\ &= \frac{\partial}{\partial s_1^{(L)}} (\theta(s_1^{(L)}) - y_n)^2 \end{aligned} \tag{3.7}$$

Knowing that $\theta'(s_1^{(L)}) = 1 - \theta^2(s_1^{(L)})$ for the hyperbolic tangent, the general expression for $\delta_i^{(l-1)}$ becomes:

Backward Propagation Algorithm

Initialize all weights $w_{ij}^{(l)}$ at random (not zero)
while $\nabla \mathbf{e}$ not close to zero **do**
  FORWARD Compute all $x_j^{(l)}$
  BACKWARD Compute all $\delta_j^{(l)}$
  Update the weights $w_{ij}^{(l)} \leftarrow w_{ij}^{(l)} - \eta x_i^{(l-1)} \delta_j^{(l)}$
end while
return Final weights



Â´

Figure 3.4: Calculation of the inner-layer weights using the Backward Propagation Algorithm. The $\delta_j^{(l)}$'s flow backwardly from the outer-layer $\delta_1^{(L)}$, while the $x_i^{(l)}$'s are constructed in the natural forward direction.

$$\begin{aligned}
\delta_i^{(l-1)} &= \frac{\partial e(\mathbf{w})}{\partial s_j^{(l-1)}} \\
&= \sum_{j=1}^{d^{(l)}} \frac{\partial e(\mathbf{w})}{\partial s_j^{(l)}} \frac{\partial s_j^{(l)}}{\partial x_i^{(l-1)}} \frac{\partial x_i^{(l-1)}}{\partial s_j^{(l-1)}} \\
&= \sum_{j=1}^{d^{(l)}} \delta_j^{(l)} w_{ij}^{(l)} \theta'(s_i^{(l-1)} \\
&= (1 - (x_i^{(l-1)})^2) \sum_{j=1}^{d^{(l)}} w_{ij}^{(l)} \delta_j^{(l)}
\end{aligned} \tag{3.8}$$

And so the error gradients of each layer of the NN can be calculated, using (3.6). The Backward Propagation Algorithm (BPA) on Figure 3.4 shows how this can be achieved.

3.3 The MATLAB Neural Network Toolbox

Matlab provides a Neural Network Toolbox[80;81] with the possibility to fit large amounts of input data to corresponding target data. The function call *nftool* initiates the creation of a NN. This creation can be customized by setting the appropriate parameters, see Section 3.3.1. Although it happens automatically, it is important to consider the pre- and postprocessing of the network, which is done in Section 3.3.2. After running the NN, one can analyse the results as explained in Section 3.3.3.

3.3.1 Adjustable parameters

The input data for the training of the NN has to be a $m \times n$ matrix, where m is the number of input vectors and n the number of features each vector contains. For the case of this study scalar targets are sufficient and the target data is a $m \times 1$ matrix. After selecting the two associated data sets, the whole coupled collection

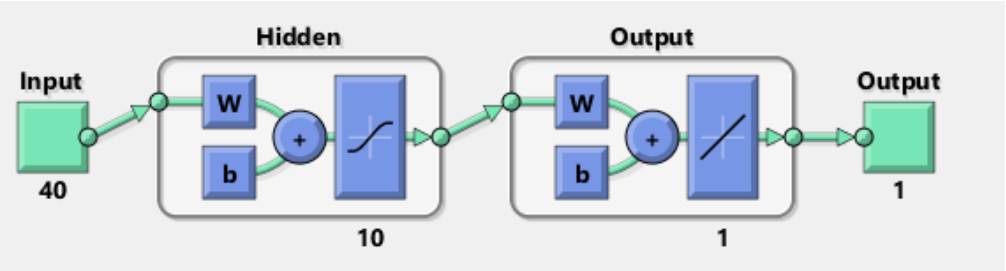


Figure 3.5: A NN as a graphical diagram.[81]

of data is then randomly divided in three subsets of which the relative sizes can be set as parameter:

Training set - This is obviously expected to be the largest subset, because this is used for computing the SGD and updating the weights of the network.

Validation set - In each iteration of the algorithm, the training is tested with this subset. The validation error gets smaller as the network is better trained, but increases when the network begins to overfit. After a specified number of consecutive steps of increasing validation error, the algorithm stops. The weights and biases from the last validation error minimum are given as output.

Test set - This set is not really used during the training phase with the intention to have an uncorrelated set afterwards to compare different models. However, in practice the test set is also used in the exact same way as the validation set, but does not trigger a stop of the algorithm. If the iteration numbers of both error minima differ significantly, this indicates a deficient division of data sets.

Another parameter the MATLAB Neural Network Toolbox allows to adjust is the number of hidden layers (nodes) the NN contains. The default NN function call *fitnet* returns a feedforward network with the default tan-sigmoid function $\theta(s)$ in the hidden layers and a linear transfer function in the output layer[82]. Figure 3.5 gives a schematic overview of a network for input vectors with 40 attributes, 10 hidden layers, 1 output layer and a scalar output. The final output of this network is then a prediction for the target.

An obstacle to avoid is the risk of overfitting. This happens when the error on the training set is so small that the NN loses all generality. It can only reproduce the points of the training set with a very high accuracy. But it should be noted that if the number of attributes in the input vectors is much smaller than the total number of training samples, then there is only little or no chance of overfitting[83].

3.3.2 Pre- and postprocessing

The MATLAB Neural Network Toolbox is equipped with some automated preprocessing functions[84]. The two functions described below are used by the *nftool* function by default:

mapminmax - This function maps the values of the input and target vector attributes to an image in the range $[-1, 1]$. In this way, all attributes have the same weights during the training of the network.

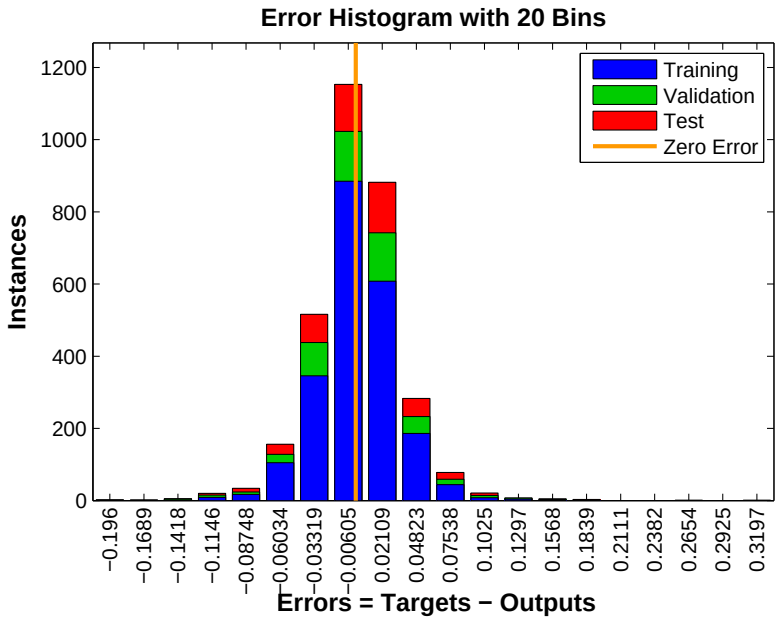


Figure 3.6: Typical error histogram

removeconstantrows - This function removes the attributes that are constant over all samples. If there are many constant features, this function can reduce the runtime of the algorithm significantly.

The postprocessing is in fact just the opposite of the preprocessing. The scaling of the ranges is undone and the constant attributes are added again.

3.3.3 Outcome of the network

Once the NN is created, every possible input vector has a unique output value. Because the target values are known for training, validation and test data, the error on the prediction can be defined as the difference between the target and output values. Figure 3.6 shows a typical error histogram for which the standard deviation (STD) can be calculated for further analysis.

Another useful statistical tool is the regression of the outputs on the targets. Figure 3.7 shows these regressions for the three different subsets and the total set. It is needless to say that the regression of the training set generally will be closer to 1 than the regression of the validation and test set.

Notice that Figure 3.6 and 3.7 are generated with data of which the target values were known in advance. Now that the network is trained it is possible to apply it on input vectors of which the target values are unknown. This returns predictions for the target values and points out the most interesting input vectors for further research.

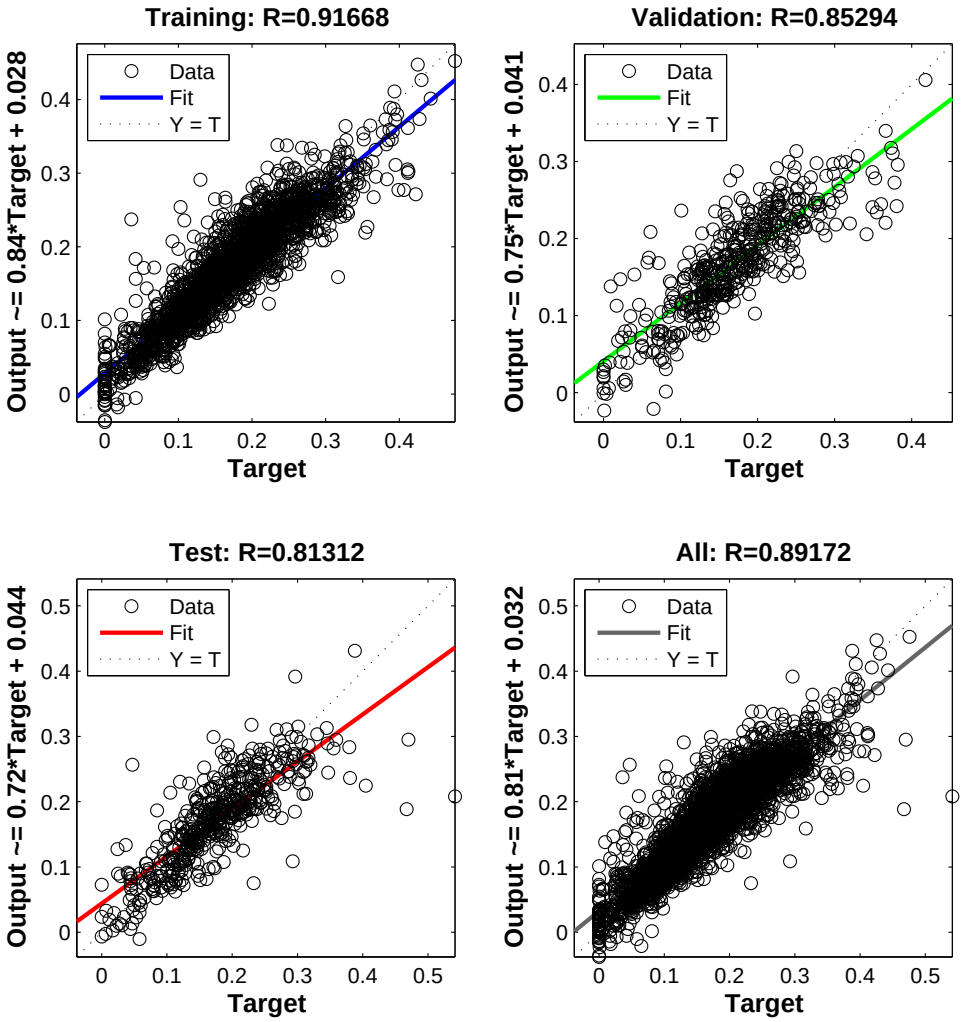


Figure 3.7: Typical regression plots for every data subset apart and for the three data sets combined.

Chapter 4

Calibrating the Neural Network

It has already been stated in Section 1.4 that Dumon[4] computed 4224 QZPs in the Pnma space group. The existence of this completed data set is a substantial assistance to the calibration of the NNs for further research. In Section 4.1 of this chapter, the optimal NN parameters (see Section 3.3.1) for the Pnma set are determined. In Section 4.2 a method is proposed to use the results of the NNs in an efficient way.

4.1 The optimal Neural Network parameters

The NN parameters that are optimized in this section include the relative sizes of the train, validate and test subsets, the numbers of hidden layers and the attributes of the input and target vectors. Section 4.1.1 starts with a basic input vector for testing the three subset sizes and the nature of the target energy. With the results of this first section there will be varied in the number of hidden layers in the NN, see Section 4.1.2. Finally, in Section 4.1.3, the input vectors are supplemented with other elemental information.

4.1.1 Composing a comfortable starting set

The available data is a set of 4224 QZPs in the orthorhombic 1:1:1:4 structure[4] . In the first phase of this analysis, the input vectors are composed by the four atomic numbers Z, the four principal quantum numbers n and the number of electrons in the s, p and d orbitals for each element present in the QZP. These input

Z_{Al}	n_{Al}	s_{Al}	p_{Al}	d_{Al}	Z_{Ba}	n_{Ba}	s_{Ba}	p_{Ba}	d_{Ba}	Z_{Ga}	n_{Ga}	s_{Ga}	p_{Ga}	d_{Ga}	Z_S	n_S	s_S	p_S	d_S
13	3	2	1	0	56	6	2	0	0	31	4	2	1	10	16	3	2	4	0

Z_{Ba}	n_{Ba}	s_{Ba}	p_{Ba}	d_{Ba}	Z_{Al}	n_{Al}	s_{Al}	p_{Al}	d_{Al}	Z_{Ga}	n_{Ga}	s_{Ga}	p_{Ga}	d_{Ga}	Z_S	n_S	s_S	p_S	d_S
56	6	2	0	0	13	3	2	1	0	31	4	2	1	10	16	3	2	4	0

Figure 4.1: Two input vectors for the QZP $BaAlGaS_4$ with the atomic numbers Z, the principal quantum numbers n and the number of valence electrons in s, p and d orbitals as attributes. In the upper input vector the elements are ordered alphabetically, in the lower they are ordered according to the structural formula.

vectors holding 20 attributes will be referred to as the unmodified or unaltered input set. Figure 4.1 gives an example of two possible unaltered input vectors for the QZP $BaAlGaS_4$. For the target energies the two definitions proposed in Section 1.1.2 are possible. The easiest choice is the pure formation energy E_{pure}, calculated after the volume and general relaxations. The alternative is E_{decomp}, for which the decomposition products have to be known. Because the MP[6;7] does not support Po-compounds, these types are omitted for the E_{decomp} calculations. Removing all combinations with Polonium from the Pmna set reduces this set from 4224 to only 3168 QZPs.

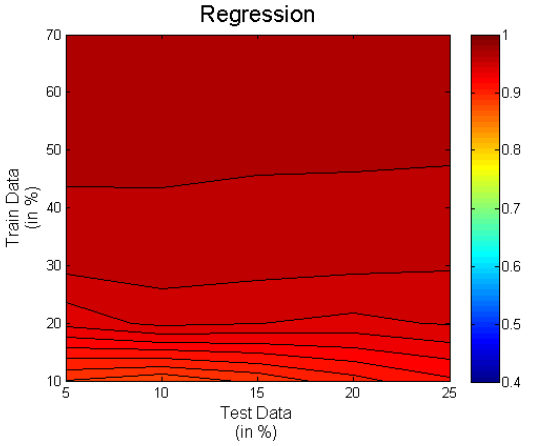


Figure 4.2: Linear regression for the unaltered input set, with E_{pure} as the target energy.

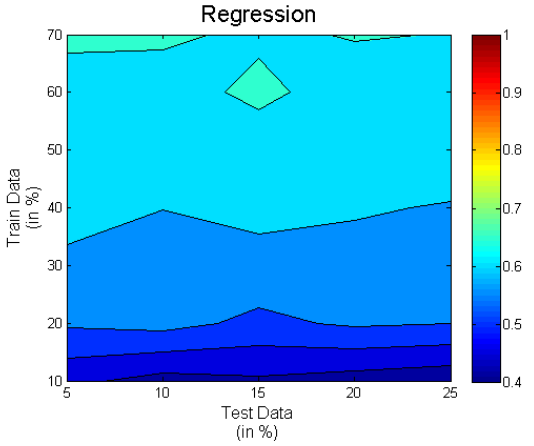


Figure 4.3: Linear regression for the unaltered input set without Polonium compounds and E_{decomp} as target.

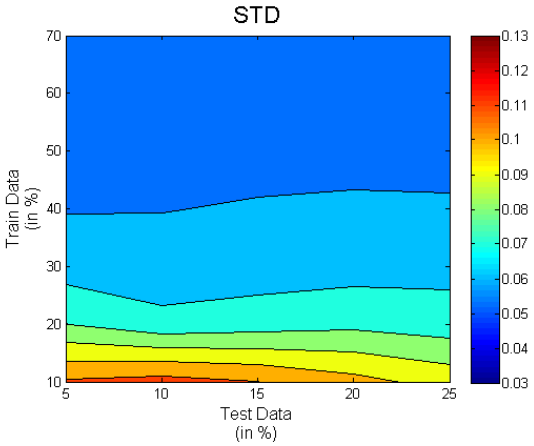


Figure 4.4: Standard deviation for the unaltered input set, with E_{pure} as target energy.

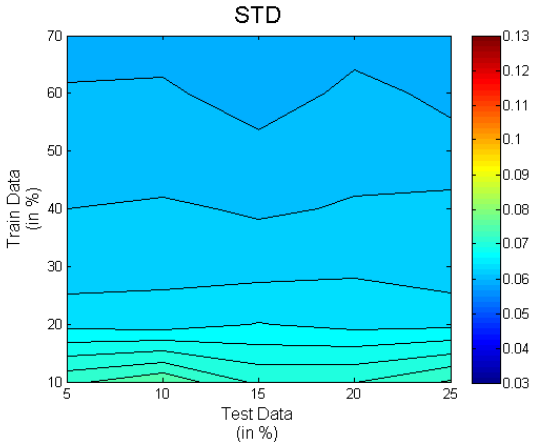


Figure 4.5: Standard deviation for the input set without the Polonium compounds and E_{decomp} as target data.

Bundling the total regressions from similar figures like Figure 3.7 for different sizes of train and test subsets results in colormaps such as Figure 4.2. Each point on this figure is a different ratio of subset sizes and represents the mean of the total regression for three NN simulations. Only the train and test subsets are

depicted on the axes, but the size of the validation subset in each point can easily be calculated as:

$$validation = 100\% - train - test \tag{4.1}$$

Figure 4.2 is generated for an unmodified data set with E_{pure} as target energy. The colormap varies in the vertical direction, which means that the regression depends especially on the training set and the amount of test and validation data is less relevant.

The colormap for the target energy E_{decomp} is given in Figure 4.3. What catches the eye is the worse regression for these E_{decomp} targets. Analysing the standard deviations hints at the same result when comparing E_{pure} and E_{decomp} , see Figure 4.4 and 4.5. Another observation is that for training sets lower than 20% of the total data set, the standard deviations of E_{pure} perform much worse than for E_{decomp}. This effect can be neglected because large training sets perform better anyway and therefore the training set will always have to be maximized.

In this respect one would like to choose for E_{pure} as the target for further NN simulations. Unfortunately, no significant correlation exists between a low E_{pure} and the effective existence of a crystal[4] . E_{decomp}, on the other hand, is an objectively better measure for the stability of a structure. This quantity is much more accurate, but is, as seen above, harder to predict. Summarizing, one can say that a negative E_{pure} is a necessity for the existence of a crystal and a negative E_{decomp}, if all binaries and ternaries are known, a sufficiency . The high number of negative pure formation energies in the Pnma set is the reason to use E_{decomp} as the target function in this report. The task that is left is to find a better input data set to predict E_{decomp}.

Figure 4.1 shows two variations of ordering the attributes in the input vector for the $BaAlGaS_4$ crystal structure. It may be clear that adding morphological via the order of attributes can make a difference in the performance of the NN. The regressions and STDs for the alphabetically ordered input vectors are given in Figure 4.3 and 4.5. Ordering the elements according to the sequence in the structure formula has a drastic effect on the regression and the standard deviation, see respectively Figure 4.6 and 4.7. For the choice of subsets 70% training, 15% validation and 15% test data, a regression coefficient of 0.8932 and a standard deviation of 0.0350 are found as the averages of three simulations. This is in large contrast with the input set where the columns are ordered alphabetically along the elements: 0.6422 and 0.0597 respectively for the same point on the colormaps in Figures 4.3 and 4.5.

4.1.2 Varying the number of nodes

Beside varying the relative size of the data sets, the optimal number of nodes of the NN can be tested. It should be clear that a network of only one node gives bad results. More nodes would increase the accuracy of the NN, but the execution time rises fast. This report tests at what number of nodes this increase of precision converges. Following data set subdivision is used for testing the number of nodes: 70% training, 15% validation and 15% test data and each plot is the mean of 5 runs at each number of nodes. Consider the unaltered set, which has been defined in Section 4.1.1, with the pure formation energy and without taking into account the element order. Figure 4.8 shows that the regression for this set is already converged at 10 nodes. For the correctly sorted data set with E_{decomp} as target, convergence occurs at a higher number of

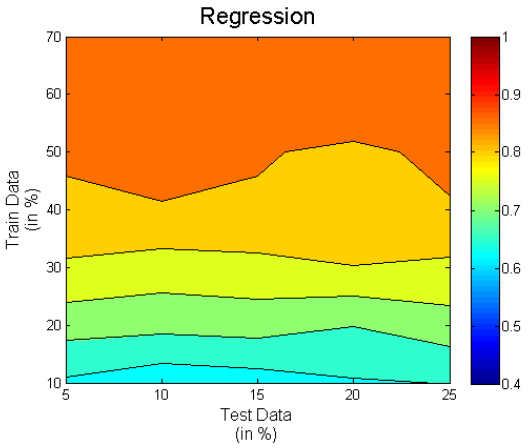


Figure 4.6: Regression for all points of the data set without the Polonium structures, with the attributes in the input vectors ordered as in structure formula and with E_{decomp} as target.

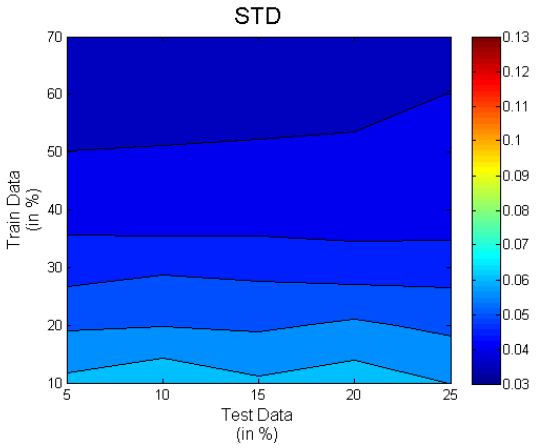


Figure 4.7: Standard deviation for the data set without the Polonium structures, with the attributes in input vectors ordered as in structure formula and with E_{decomp} as target.

nodes. This is due to the fact that E_{decomp} is harder to predict than E_{pure}. In Figure 4.9 it can be seen that the optimal number of nodes for this data set converges around 25 nodes.

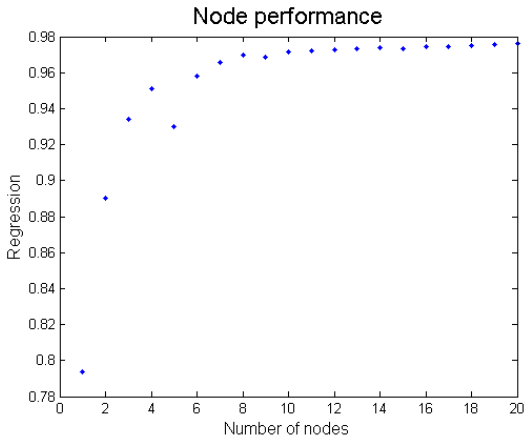


Figure 4.8: Linear regression calculated for NNs from 1 to 20 nodes for the unaltered data set. At 10 nodes the regression seems to be converged.

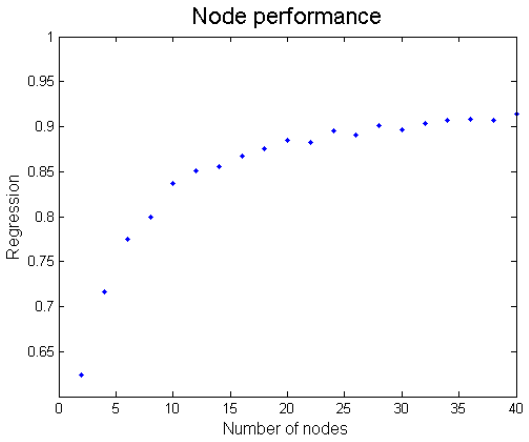


Figure 4.9: Linear regression calculated for NNs from 1 to 40 nodes for the data set with E_{decomp} as target and ordered along the elements in the structure formula. At 25 nodes the regression seems to be converged.

4.1.3 Subsequent modifications

Up to here a sufficient data set was found on which additional modifications can be tested. From here on the geometrically ordered set with E_{decomp} as target will be called the reference set. This reference set will be

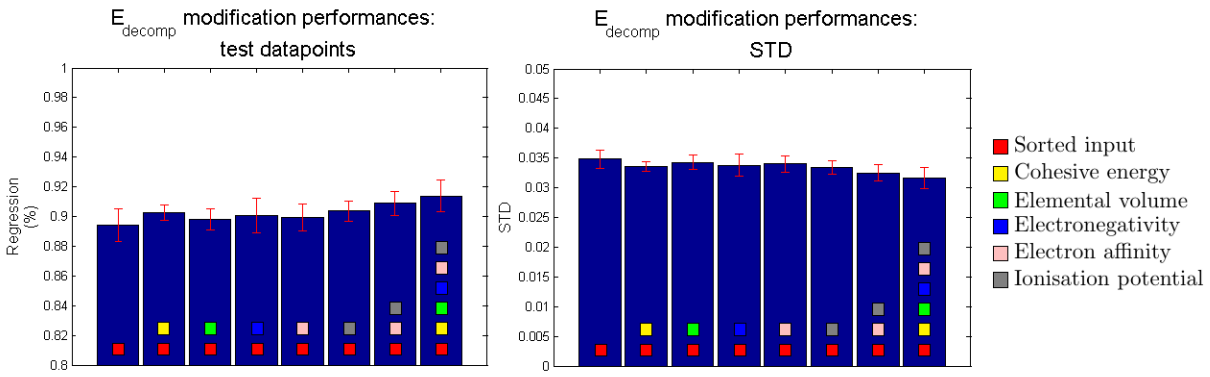


Figure 4.10: Test regression and STD for some additional elemental features to the input vectors of the reference set. Each bar in both diagrams is the mean of five simulations.

supplemented with other elemental data like the cohesive energy, volume of the elemental materials, Pauling electronegativity, electron affinity and ionisation potential. The subdivision for the following tests will again be 70% training, 15% validation and 15% test data. The regressions and STDs of eight combinations of additional elemental features on top of the reference data set can be found in Figure 4.10. The added attributes do not have drastic effects on the performance. Anyway, there will be opted for the best scoring combination of elemental information and that is the full package with 40 attributes per input vector.

4.2 Intelligent screening

Now that a proper NN has been found it is necessary to implement this in the high-throughput cycle, as noticed in Section 2.3.3. Section 4.2.1 explains how this can be achieved. Section 4.2.2 discusses then the results of a simulation of this proposed algorithm on the Pnma space group.

4.2.1 Exploiting the Neural Network

Figure 4.11 depicts the *intelligent screening* algorithm that will be explained here step by step. In the initial stage of the algorithm only a data set of unknown materials is available. These crystal structures are colored white in Figure 4.11 and colored from the moment that their stability is exactly known. From this initial data set only randomly picked materials can be calculated, as the NN can not make predictions yet. After computing a small amount of crystals, the NN can be trained for the first time and the priorities of all uncalculated crystals are updated. This prioritizing is symbolized in the figure as a change of order: the highest prioritized crystal structures are placed just below the earlier calculated materials. From here the iterative stage of the algorithm is entered. The NN updates the priorities each time the number of newly computed crystal structures are sufficiently large. After some iterations the whole dataset is calculated, whereby the most promising crystal structures are intended to emerge first.

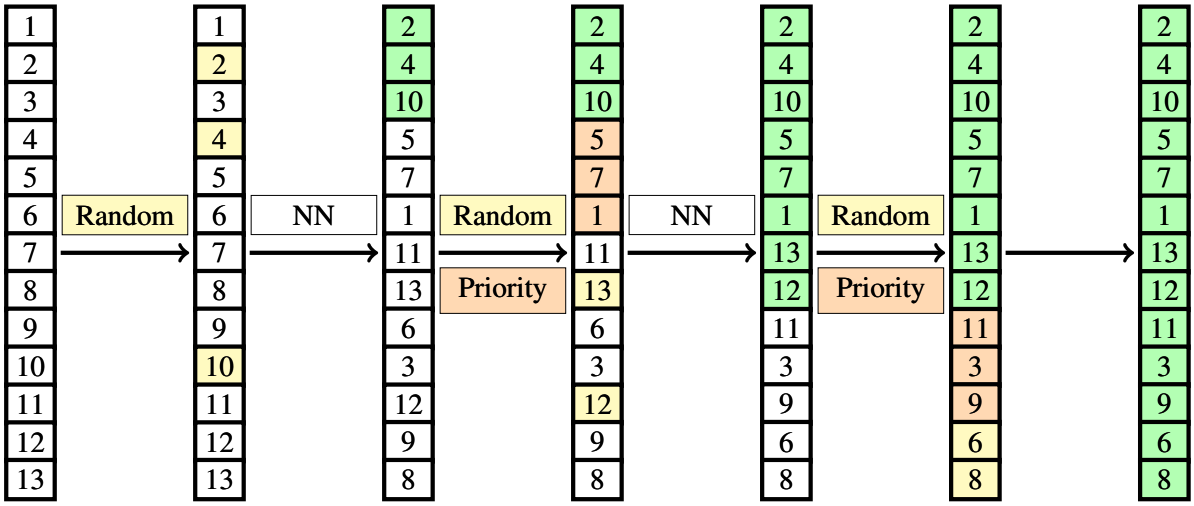


Figure 4.11: This diagram shows the algorithm in a schematic way. Yellow labels are randomly picked materials that are calculated, orange labels are materials that are picked in order of their priority and green labels are materials that were calculated in an earlier stage of the algorithm.

4.2.2 Results of the method

The proposed algorithm is tested with the Pnma set as unknown set. Using this set as a test has two advantages: the time-consuming DFT calculations do not have to be redone and the stable crystals are already identified. It is thus possible to keep track of these stable crystals during the algorithm in order to check the effectiveness of the NN. The materials that are marked as 'stable' are the 198 crystal structures with $E_{decomp} \leq 0.05\,eV$.

For the simulation of the algorithm, iterative steps of 10% of the total data set are taken. In each of these iterative steps half of the to be calculated materials are picked randomly, the other half are the highest prioritized ones. Figure 4.12 shows at what moment during the process the stable crystals emerge. From this figure appears that when only 50% of the whole data set is calculated, almost 90% of all stable Zintls has been found. This is a significant improvement compared to the graph of a brute force approach that only picks random materials.

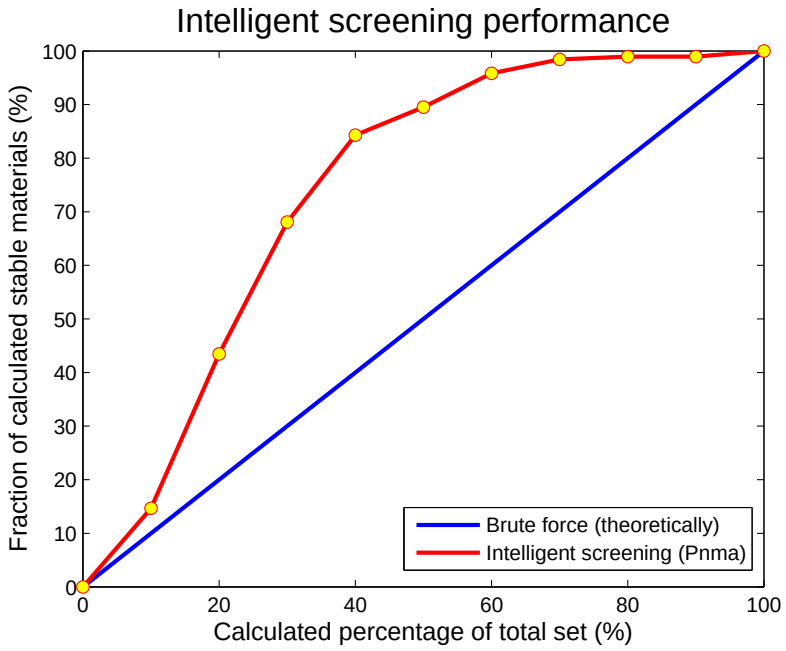


Figure 4.12: The algorithm applied on the Pnma data set. The vertical axis represents the fraction of stable crystals that has already been found, while the horizontal axis is the fraction of the total data set that has been calculated. The yellow dots indicate the NN updates.

Chapter 5

Screening of a straightforward crystal set

This chapter focusses on the NN screening of a set of crystal structures with few atoms in the unit cell. Such simple crystal types have the advantage of being rapidly calculated with DFT. A whole set can be calculated exhaustively and analyzed within the timeframe of this work, as an independent test for the performance of the *intelligent screening* approach. Section 5.1 goes into detail on an experimentally known QZP and presents the preparation of the high-throughput computation on the model of this QZP. Thereafter Section 5.2 discusses not only the properties of the stable crystals, but also the results of a second NN simulation with this new data set.

5.1 Pre-analysis

Li *et al*[85] performed a study on binary, ternary and quaternary tellurides whereby they came across the $K_2BaSnTe_4$ QZP. $K_2BaSnTe_4$ is a semiconductor that has a cubic unit cell and adopts the highly symmetric space group I-43m. This is a body centered cubic lattice with 4-fold rotoinversion axes perpendicular to the face planes, diagonal 3-fold rotation axes and mirror planes diagonally dividing the face planes. Besides, one can notice that the oxidation states add up to zero as is expected for a solid with valence-balanced bonding:

$$2 * \mathrm{K(+I)} + \mathrm{Ba(+II)} + \mathrm{Sn(+IV)} + 4 * \mathrm{Te(-II)} = 0 \tag{5.1}$$

Figure 5.1 shows the crystal as it was originally presented by Li *et al*[85]. From this figure one observes that the $SnTe_4^{-4}$ Zintl anions are located on the corners and at the center of the unit cell. Anions like SnS_4^{-4}, $SnSe_4^{-4}$ and $SnTe_4^{-4}$ are common stable anions that can be seen as basic building blocks for many crystals. The K^{+1} and Ba^{+2} cations are placed on the sides and edges of the unit cell and both occupy the 6b Wyckoff position.

As DFT calculations cannot deal with partial occupancies, the unit cell will be modelled as described in Figure 5.3 and Table 5.1: the K and Ba ions are assigned the 4c and 2a wyckoff positions. This makes a long-range ordered variant of the initial partially disordered crystal. Furthermore, to give more freedom to the crystal structure during the VASP relaxation phase, one can change to the orthorhombic space group I-42m without changing the length of the c-axis. This replaces the diagonal 3-fold rotation axes to 2-fold rotation axes perpendicular on the face planes that lay in the c-direction. After performing these two oper-

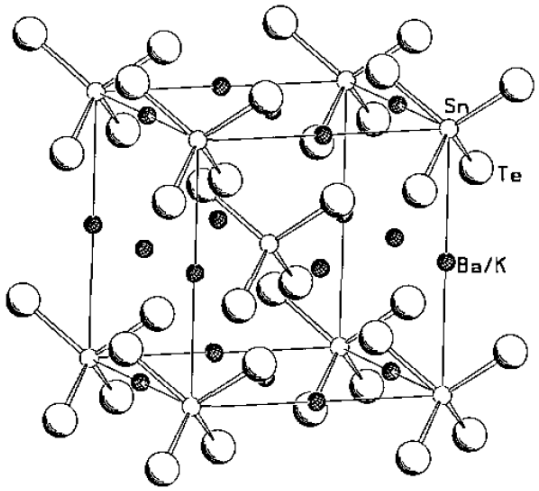


Figure 5.1: The originally published $K_2BaSnTe_4$ crystal structure.[85]

Table 5.1: The Wyckoff positions in the $K_2BaSnTe_4$ crystal structure that adopts the I-42m space group.

Symbol	Coordinates	Wyckoff position
K	$(0.5, 0, 0)$	4c
Ba	$(0, 0, 0)$	2a
Sn	$(0, 0, 0.5)$	2b
Te	$(0.190, 0.810, 0.690)$	8i

ations on the original I-43m crystal structure, the I-42m structure in Figure 5.3 is obtained. From here on only crystal structures in the I-42m space group will be regarded. With the Wyckoff positions[86] given in Table 5.1 one can reconstruct this I-42m unit cell of $K_2BaSnTe_4$ or any other variant crystal that is the result of elemental substitution.

$K_2BaSnTe_4$ is a QZP consisting of two alkali/alkaline-earth metals and two p-block elements. Figure 5.4 demonstrates how the elements will be picked for the creation of a search space of crystals with the same symmetry, hereafter referred to as the "I-42m space" or the "I-42m family". The first two elements in the chemical formula K and Ba will be substituted by elements from the yellow region in Figure 5.4. Sn will be replaced by an element from the light blue 4×3 block and the final element will be picked from S, Se and Te. This partitioning of the periodic system is the same as utilized in Dumon's work[4] for the Pnma family. Using combinatorics, one finds 2016 QZP candidates to be calculated. This set is small enough to fully calculate within the timeframe of this project and thus it will serve as another test family for the *intelligent screening* algorithm.

In Section 2.1.3 it was stated that the cutoff energy for the basis wave functions and the number of points in the k-point grid are crystal dependent and therefore have to be determined for each unique problem. Convergence tests show that a cutoff energy of $400\,eV$ is sufficient for the I-42m family, which was also the

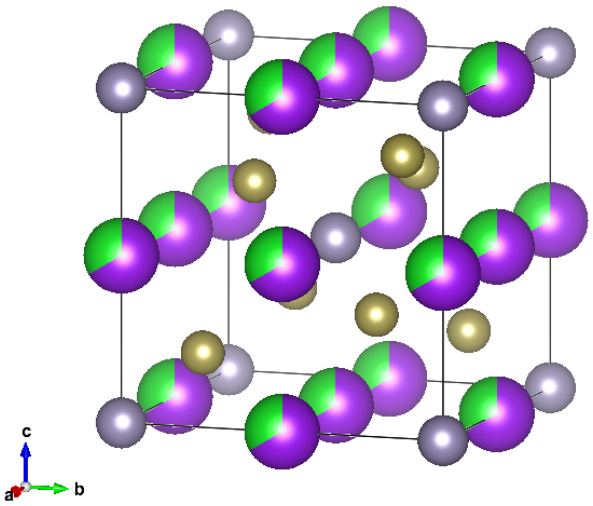


Figure 5.2: The $K_2BaSnTe_4$ crystal structure in the I-43m space group. Purple, green, gray and brown represent respectively K, Ba, Sn and Te.

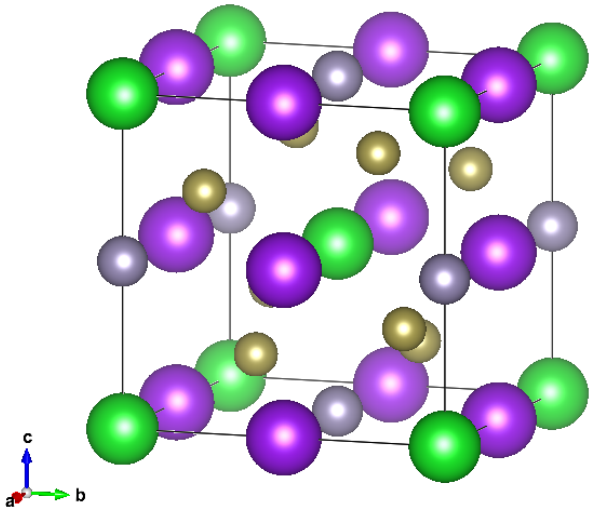


Figure 5.3: The $K_2BaSnTe_4$ crystal structure in the I-42m space group without partially occupancies. Purple, green, gray and brown represent respectively K, Ba, Sn and Te.

case in Dumon's work[4]. Because the k-point mesh has to reflect the cubic shape of the unit cell, tests are performed on $3 \times 3 \times 3$, $5 \times 5 \times 5$, $7 \times 7 \times 7$, $9 \times 9 \times 9$ and $11 \times 11 \times 11$ grids. The $5 \times 5 \times 5$ mesh results already in a converged total energy and will thus be used for the computation of the I-42m family. Another VASP setting that needs some attention is the relaxation of the cell parameters. The experimentally known $K_2BaSnTe_4$ is part of the cubic I-43m space group and would only need a volume relaxation at constant shape to stay in this symmetry. However, there was opted for the orthorhombic I-42m space group and thus the c lattice vector will need to be relaxed in order to give the shape of the initial cubic cell the freedom to change.

In Section 1.1.2 a stable crystal was defined as a crystal for which $E_{decomp} \leq 0$. One has to keep in mind that DFT calculations are not exact and that some stable crystal structures will be overlooked if one clings too strictly to the aforementioned selection criterion. After consulting similar cases in literature[14;87] , a tolerance of 50 meV per atom on the DFT results is selected.

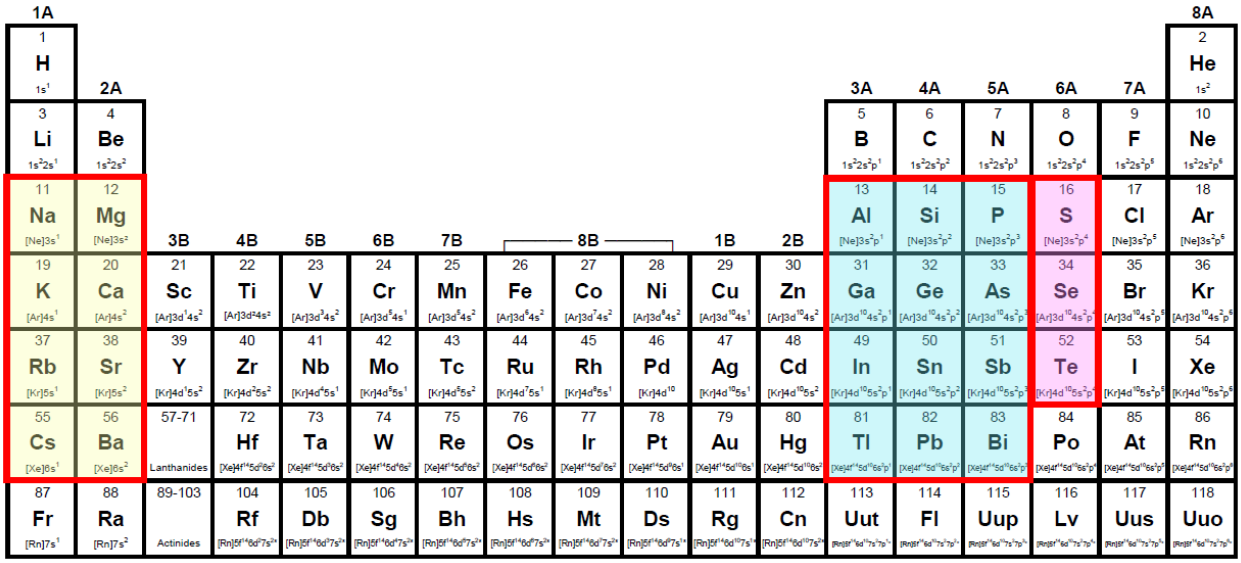


Figure 5.4: A periodic table showing the regions from which the elements are chosen. The alkali and alkaline-earth metals will be represented by two elements from the yellow region. One p-block element comes from the light blue region and the other from the magenta colored elements.

5.2 Results

Section 5.2.1 discusses a simulation of the *intelligent screening* algorithm on the calculated I-42m family. In Section 5.2.2 the total I-42m family is then divided into several subfamilies, according to groups of the elements in the QZP. Each of these subcategories has specific structural, DOS and band properties that will be discussed in this section.

5.2.1 Intelligent screening

The simulation of the *intelligent screening* algorithm for the I-42m family is is performed similarly to to the simulation in Section 4.2.2. Iterative steps of 10% of the total data set are taken and in each of these iterative steps half of the to be calculated materials are picked randomly, the other half are the highest prioritized ones.

Notice again that this simulation is executed after the whole set is calculated. In this way it is possible to track the stable materials during the simulation as done earlier for the Pnma simulation. Figure 5.5 combines both simulations of Pnma and I-42m, and proves that *intelligent screening* gives an equal performance for these two space groups. At 40% of the dataset, again 90% of all 117 stable materials have been found. This is an encouraging perspective for the more challenging QZP family in Chapter 6.

5.2.2 Crystal properties

The creation of the I-42m family did not impose very strict demands for the groups on the periodic table. Group I and II can fill the first two places of the QZP's chemical formula, group III, IV, V are able to take the third spot and the fourth is reserved for group VI. This leaves some variation in the possible stoichiometry

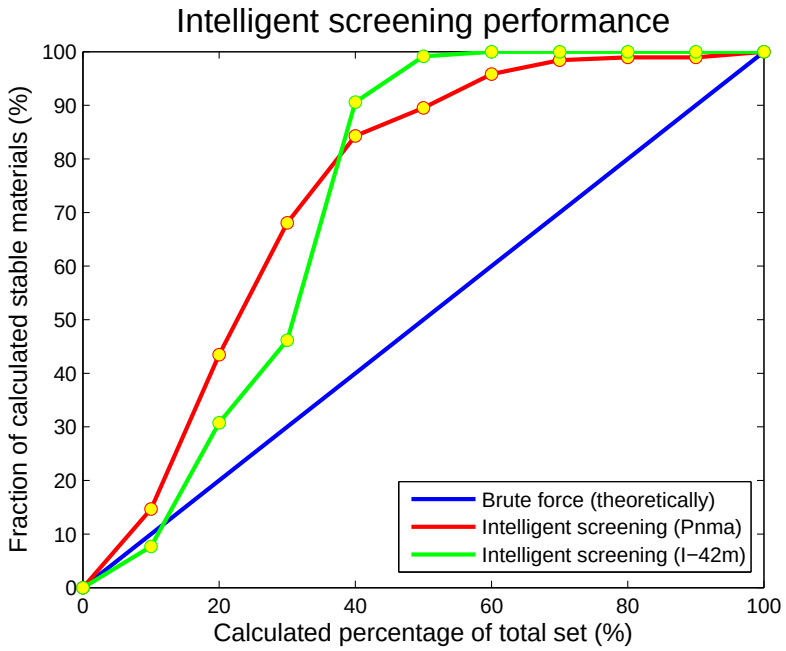


Figure 5.5: The *intelligent screening* algorithm applied on the I-42m data set. The equivalent graph for the Pnma data set from Figure 4.12 is added to compare. The vertical axis represents the fraction of stable crystals that has already been found, while the horizontal axis is the fraction of the total data set that has been calculated. The yellow dots indicate the NN updates.

and causes different subfamilies of the I-42m family. The subsequent analysis of the crystal properties will be based on these subfamilies. As an example for the notation of the newly defined subfamilies, notice that $K_2BaSnTe_4$ is part of the I-II-IV-VI subfamily.

For the 117 crystal structures with $E_{decomp} \leq 50\,meV/atom$, exactly seven subfamilies were found. Four of these have more than five members and will be discussed here on the basis of unit cell shape, electrical characteristics, DOS and band structures. The crystallographic details of the three smaller groups can be found in Table A.1. In the following enumeration the four largest subfamilies are presented with, in the title, their name, the number of crystal structures in this set with $E_{decomp} \leq 0.05\,eV/atom$ and the decomposition energy of the most stable compound.

I-I-IV-VI | 11 members | -0.0419 eV/atom - $K_2RbSiSe_4$ has the lowest $E_{decomp} = -0.0419\,eV/atom$ of this subset. This is a metallic subfamily and the c/a ratio is between 0.9 and 1.10. For more specific information of each crystal one is referred to Table A.2.

I-II-IV-VI | 21 members | -0.1249 eV/atom - With $E_{decomp} = -0.1249\,eV/atom$ $Rb_2MgSiSe_4$ has the lowest decomposition energy of the I-II-IV-VI subfamily and of the space group I-42m in its entirety. All 21 members of this subfamily are semiconductors with a bandgap in the order of $1.5\,eV$. During the general geometry optimization phase the cubic shape of the template crystal shifted to a tetragonal structure for which

the c/a ratios are between 0.79 and 0.86. This results in a flatter structure than the original cubic unit cell and the unit cell of the other subfamilies. For example, one can compare the shapes of the unit cells in Figure 5.6 and 5.7. All 21 stable crystal structures have Mg as the group II element. Information of each particular crystal structure in this set can be found in Table A.3.

I-I-V-VI | 46 members | -0.0290 eV/atom - With 46 out of 129 stable crystals this is the largest subfamily. Rb_2CsPSe_4 is the most promising member with $E_{decomp} = -0.0290\,eV/atom$. All these semiconducting crystals have a cubic-like unit cell for which 80% of the c/a ratios fluctuate between 0.90 and 1.10. The band gaps range between 0.05 eV and 1.93 eV. Table A.4 gives further detailed information on this subfamily.

II-I-IV-VI | 29 members | 0.0072 eV/atom - $Ba_2RbSiTe_4$ has the lowest $E_{decomp} = 0.0072\,eV/atom$ of this crystal collection consisting of 29 members. As was the case with I-I-IV-VI, this is a metallic subfamily. All crystals in this set have a c/a ratio between 0.90 and 1.10, which results in a more or less cubic unit cell. Opposite to the previous subfamily, none of the group II elements is ever Mg. Ca, Sr and Ba occur in the same amount as group II elements. One can observe this in Table A.5.

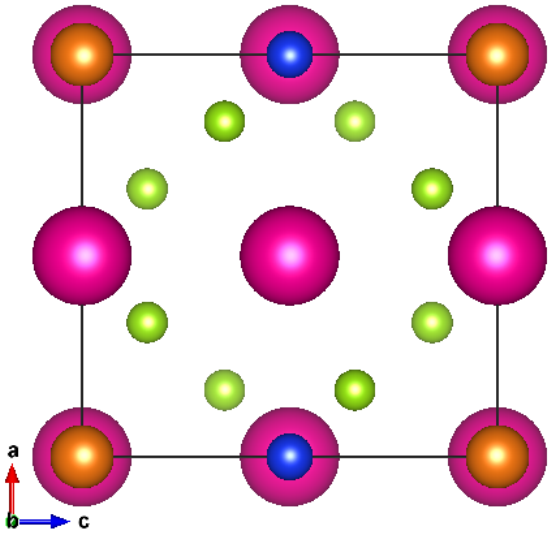


Figure 5.6: The crystal structure of $K_2RbSiSe_4$ (I-I-IV-VI), seen along the b-direction. K, Rb, Si, Se are colored respectively magenta, orange, blue and green.

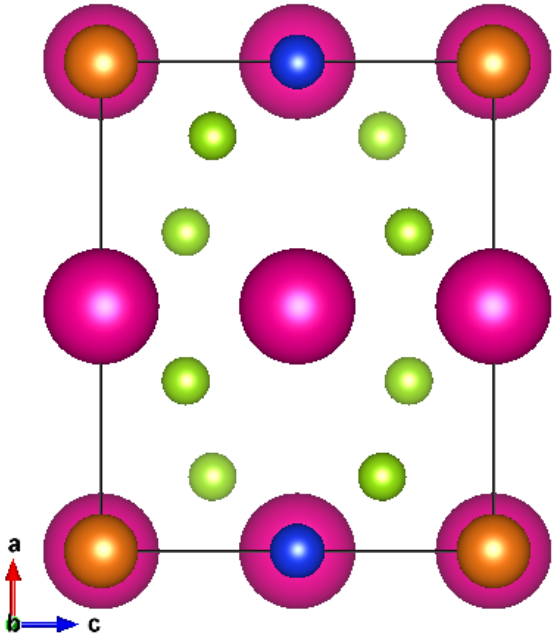


Figure 5.7: The crystal structure of $Rb_2MgSiSe_4$ (I-II-IV-VI), seen along the b-direction. Rb, Mg, Si, Se are colored respectively magenta, orange, blue and green.

A notable trend in all table sof Appendix A is that in none of all stable crystals the group VI element S was found. This implies that these compounds have a much worse stability and the S-ions are more favorable to bond in other decomposition products than Se and Te. This will be discussed in more detail in Chapter 7, parallel to the results of other QZP families.

The first subfamily in the enumeration differs just one group-substitution from the three other large subfamilies. The best way to analyze the properties of the other three families is therefor to compare each of them with the I-I-IV-VI subfamily. This is how the DOS-plots are matched below: the DOS-plot of $K_2RbSiSe_4$ (Figure 5.8) is held against members from other subfamilies with just one differing element. Figure 5.8 shows accessible states just above the Fermi level and an isolated peak between 3 eV and 4 eV that can hold 4 electrons. The DOS of all other members of the I-I-IV-VI set looks similar.

Figure 5.9 gives the DOS-plot of $K_2MgSiSe_4$, member of the I-II-IV-VI set. A rigid-band interpretation applies: the extra valence electron in Mg compared to Rb fills the empty states just above E_F in Figure 5.8. As a result, the DOS in Figure 5.9 looks very similar, but all states are shifted to the left with respect to Figure 5.8, as the Fermi energy E_F lies now at the top of a band. Consequently, this subfamily is semiconducting. Again, all members of this set have a similar DOS.

Figure 5.10 shows the DOS of $Ba_2RbSiTe_4$ (II-I-IV-VI). When K from $K_2RbSiSe_4$ is replaced by Ba, there are two more electrons per formula unit instead of just one, due to the index 2 in the structure formula. This means that the first states across the gap in $K_2RbSiSe_4$ start to be filled. E_F moves considerably to the right – or equivalently: the DOS picture shifts considerably to the left. The result is again metallic, as not all of the states in the first peak across the gap can be occupied.

The third substitution that can be made is the Si-P substitution, which results in K_2RbPSe_4 and its DOS-plot in Figure 5.11. When replacing Si by P, one electron per formula unit is added. This resembles to the $K_2MgSiSe_4$ case, and indeed, Figure 5.11 and Figure 5.9 are very similar. The band that was partially occupied in Figure 5.8 gets now completely occupied, and the crystal is a semiconductor.

Figure 5.12-5.15 give the band structures of the same materials that were discussed with respect to the DOS-plots in Figure 5.8-5.11. The two subfamilies that were found to have the semiconducting property can now be submitted to a more accurate prospection. Figure 5.13 and 5.15 show that I-II-IV-VI and I-I-V-VI are both families of direct semiconductors. Comparison of Figure 5.12 and 5.14 confirms the rigid band interpretation that the isolated peak just below the Fermi level of the II-I-IV-VI subfamily is the same as the isolated peak of the I-I-IV-VI subfamily.

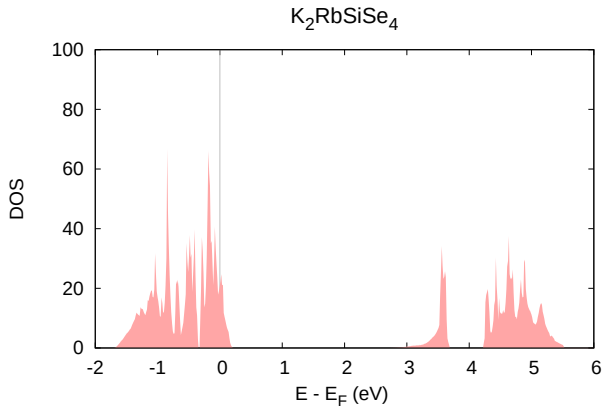


Figure 5.8: DOS of $K_2RbSiSe_4$ (I-I-IV-VI).

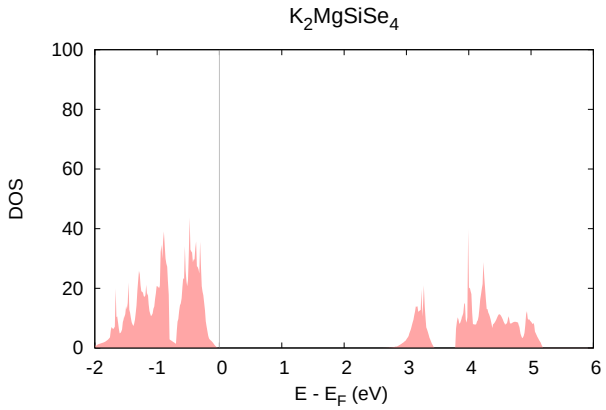


Figure 5.9: DOS of $K_2MgSiSe_4$ (I-II-IV-VI).

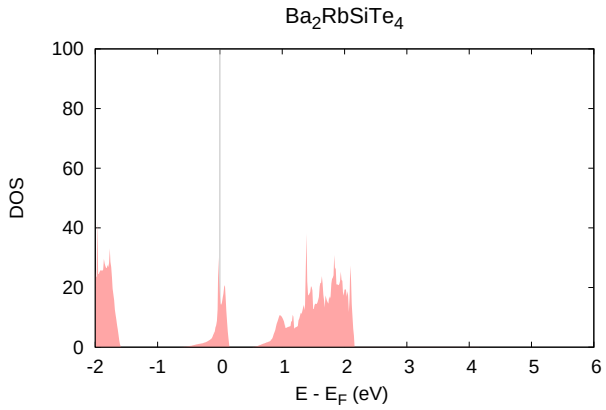


Figure 5.10: DOS of $Ba_2RbSiTe_4$ (II-I-IV-VI).

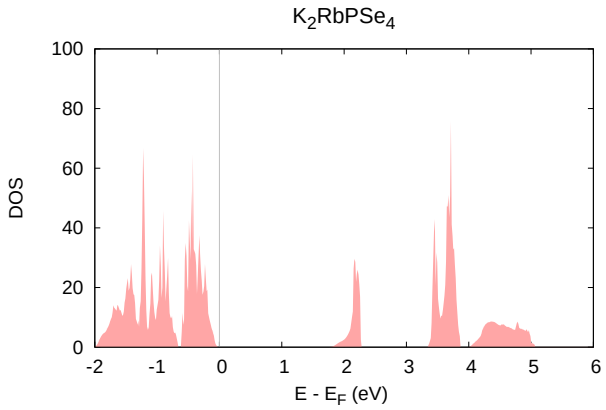


Figure 5.11: DOS of K_2RbPSe_4 (I-I-V-VI).

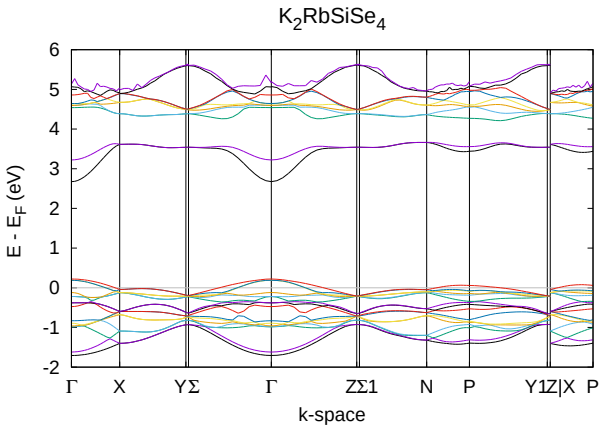


Figure 5.12: BANDS of $K_2RbSiSe_4$ (I-I-IV-VI).

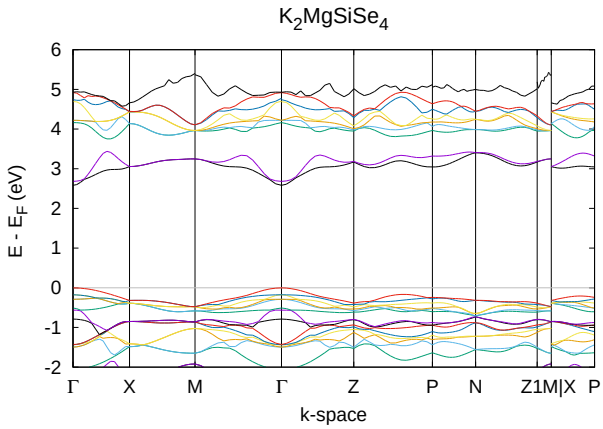


Figure 5.13: BANDS of $K_2MgSiSe_4$ (I-II-IV-VI).

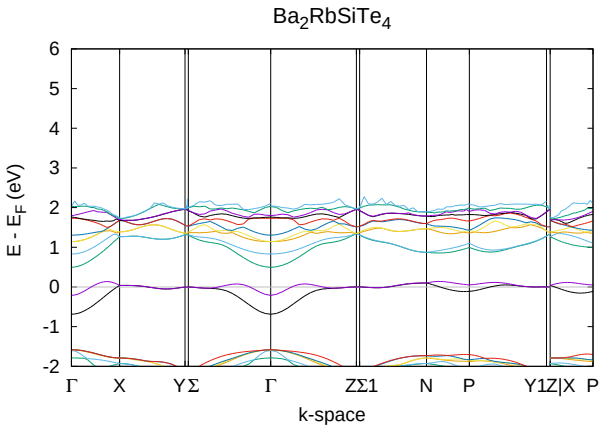


Figure 5.14: BANDS of $Ba_2RbSiTe_4$ (II-I-IV-VI).

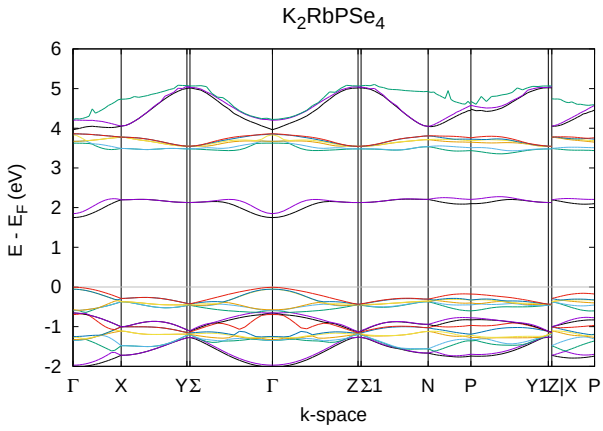


Figure 5.15: BANDS of K_2RbPSe_4 (I-I-V-VI).

Chapter 6

Screening of a challenging crystal set

After the screening of a set of straightforward crystal structures, one is able to aim at a more challenging family of QZPs. This chapter elaborates on the high-throughput computation of such set, but this time not all crystal structures will be calculated. The *intelligent screening* algorithm will predict the most stable crystal structures so that one can safely end the high-throughput procedure before all set members have been evaluated. Section 6.1 introduces two experimentally known QZPs in the same structure and presents the creation of an associated family. Finally, Section 6.2 discusses the performance of the *intelligent screening* algorithm and the properties of the newly found stable crystal structures.

6.1 Pre-analysis

Somer *et al* presented four isotypic crystal structures $K_6AlSb_4Na_3$ [88], $Cs_6AlSb_4K_3$ [88], $K_6GaSb_4Na_3$ [89] and $Cs_6GaSb_4K_3$ [90] in the hexagonal $P6_3/mmc$ space group. This structure is depicted in Figure 6.1 for $K_6AlSb_4Na_3$ as it was originally published. The structure is characterized by an array of columns of condensed $(K_6Na_{6/2})^{9+}$ icosahedra, centered by Sb2 in the figure. These cationic structures are connected via trigonal-planar anions $(AlSb_3)^{6-}$. The unit cell vectors a, b and c as defined in the figure do not differ much from each other: b/a = 1 exactly and c/a = 1.035. The Wyckoff positions of the $K_6AlSb_4Na_3$ crystal structure are given in Table 6.1. With this information one can extend the two experimentally known structures to a large family of isotypic crystals.

The structure formula consists of two alkali/earth alkaline metals and two p-block elements. This time there is no group VI element in the experimental examples, in contrast to the earlier space groups Pnma and I-42m.

Table 6.1: The Wyckoff positions in the $K_6AlSb_4Na_3$ crystal structure.

Symbol	Coordinates	Wyckoff position
K	$(0.208, 0.415, 0.542)$	12k
Al	$(0.333, 0.667, 0.250)$	2c
Sb1	$(0.481, 0.518, 0.250)$	6h
Sb2	$(0, 0, 0)$	2a
Na	$(0.116, 0.232, 0.250)$	6h

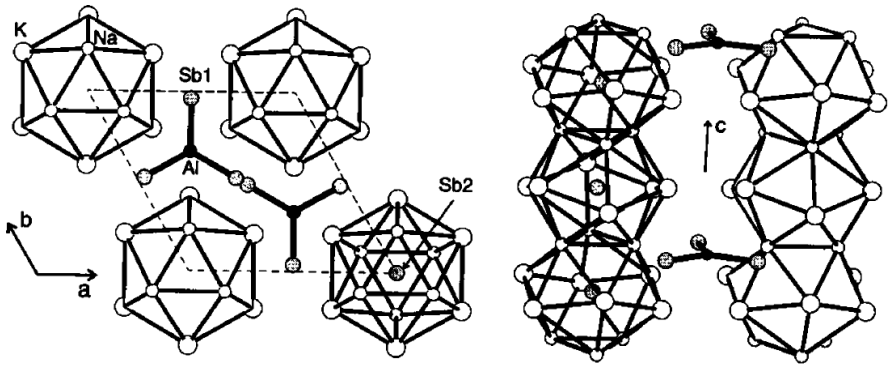


Figure 6.1: The $K_6AlSb_4Na_3$ crystal structure.[88]

This raises the question how the combinatorial substitutions have to be performed, as group VI was consistently treated differently from the other p-block elements in the earlier two QZP families. For the creation of the new symmetry family, hereafter called the "$P6_3/mmc$ family", each group on the periodic table will be a separate pool from which the elements will be picked. This approach results in 6 small pools consisting of 3 or 4 elements as indicated in Figure 6.2. These groups will then be combined with the oxidation numbers in mind: the oxidation state of each constructed QZP candidate will be held zero at the creation of the crystal structure. This was not the case in the earlier two high-throughput experiments, where elemental combinations were made without checking the oxidation state. Afterwards the weighted sum of the oxidation states of the most stable QZPs turned out to be zero for the Pnma and I-42m families. This method explicitly defines six possible stoichiometric subfamilies on beforehand, which are given in Table 6.2 with their specific number of members. It may be clear from this table that the total number of crystals in the $P6_3/mmc$ family will be 1696.

Table 6.2: The subfamilies of the $P6_3/mmc$ space group for which the weighted sum of oxidation states is zero. The order of the groups in the subfamilies is the same as the elements in the structural formula $K_6AlSb_4Na_3$.

Subfamily	Combinations	Number of members
I-(III ∪ V)-V-I	$4*7*4*3$	336
I-(IV ∪ VI)-IV-II	$4*6*4*4$	384
I-IV-VI-II	$4*4*3*4$	192
II-(I ∪ III)-IV-I	$4*7*4*4$	448
II-V-V-I	$4*3*4*4$	192
II-VI-IV-II	$4*4*3*3$	144

Similar to Chapter 5, convergence tests are needed to decide the appropriate settings for cutoff energy and k-point grid. Again, these tests suggest 400 eV as cutoff energy. In principle one must take into account the hexagonal shape of the unit cell when looking into the k-point grids, but because of the small differences between the a, b and c unit vectors a cubic k-point grid will be sufficient. The tested k-point grids are $3\times3\times3$, $5\times5\times5$, $7\times7\times7$, $9\times9\times9$ and $11\times11\times11$, from which the $5\times5\times5$ grid came out as sufficiently converged for further calculations. Again VASP will be allowed to vary both shape and unit cell vectors during the

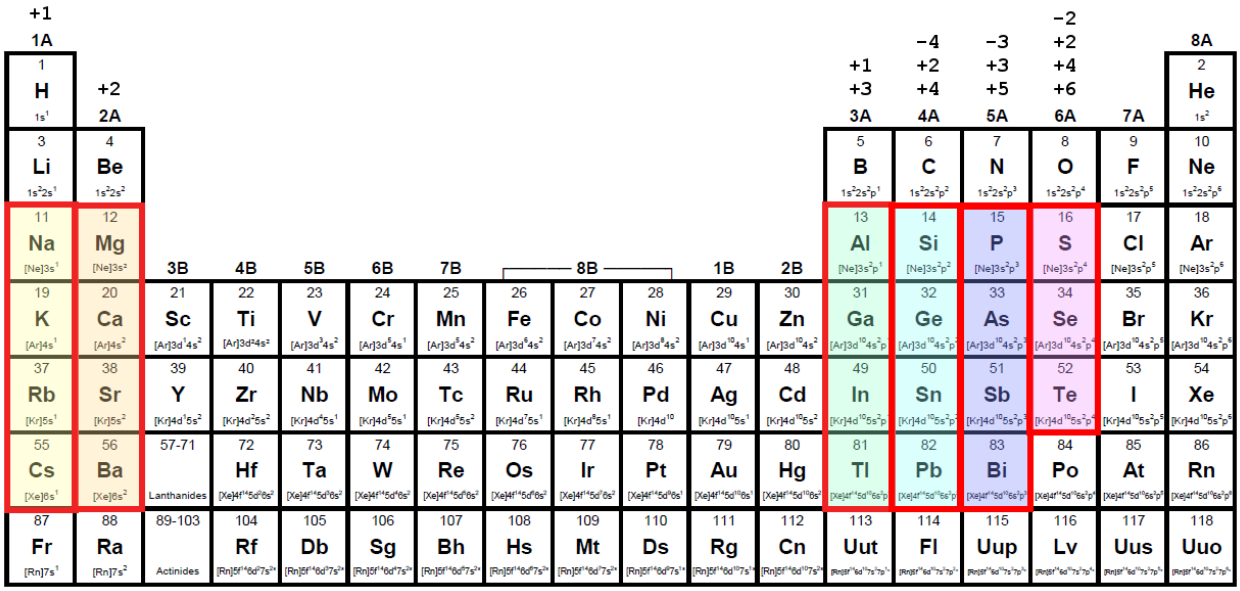


Figure 6.2: The six sets from which the elements are picked for the creation of the $P6_3/mmc$ QZP family. The suitable elements in group I, II, III, IV, V and VI are respectively colored yellow, orange, green, lightblue, violet and magenta. The possible oxidation states for each of the colored groups is added above.

geometry optimization cycle.

6.2 Results

After performing two simulations for testing purpose, it is time to let the *intelligent screening* algorithm prove its merrit. Section 6.2.1 goes into detail on the first real *intelligent screening* execution. After that, Section 6.2.2 elaborates on the crystallographic properties of the obtained stable materials.

6.2.1 Intelligent screening

Interesting experimentally known QZPs have been found, a whole new crystal family has been created and the convergence tests have been performed. So everything is ready for the first real run of the *intelligent screening* algorithm. Initially the priorities of all 1696 crystal structures are set equal so that the Queue Manager picks them randomly for calculation. Figure 6.3 then depicts the progress of the newly found stable QZPs during the high-throughput process. After calculating the total energy of 99 crystal structures, the NN prediction is performed for the first time and the priorities of all uncalculated materials are updated. This results in a steepened graph after this first update in Figure 6.3. At 30.0%, which is the fourth update of the priorities, the discovery rate starts decreasing again. The last three updates of the algorithm are performed at 34.6%, 41.6% and 46.3%. Because no new stables emerge between 34.6% and 46.3%, one can safely decide to stop the experiment at this point. The properties of the 122 crystal structures with $E_{decomp} \leq 0.05$ are discussed in the next section.

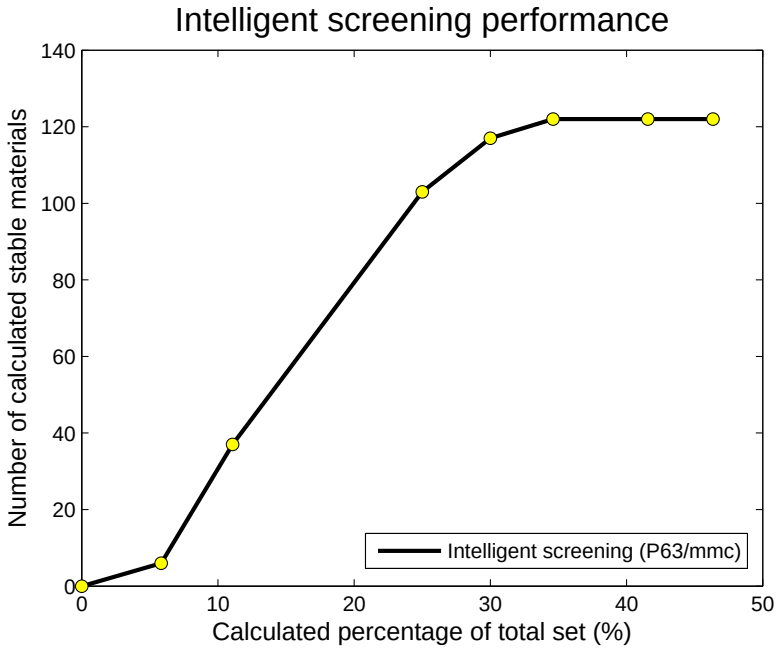


Figure 6.3: This figure shows how many stable crystal structures are found at each stage of the high-throughput calculations of the $P6_3/mmc$ set. The vertical axis represents the total number of stable crystals that has already been found, while the horizontal axis is the fraction of the total data set that has been calculated. The yellow dots indicate the NN updates. Calculations are stopped at 46.3% of the total set.

6.2.2 Crystal properties

From the subfamilies listed in Table 6.2 only I-III-V-I and I-V-V-I crystals occur as stable crystals. The I-III-V-I subfamily has 111 members, I-V-V-I only 11. Given below is a brief overview of the characteristics of both groups. For crystallographic details of all stable materials, one is referred to Appendix B.

I-III-V-I | 111 members | -0.0568 eV/atom -This subfamily contains 111 members that have a decomposition energy below $0.05\,meV$ and even 45 of these energies are below zero. As a result I-III-V-I is by far the most stable subfamily of the $P6_3/mmc$ family. With $E_{decomp} = -0.0568\,eV/atom$, $Cs_6InSb_4Rb_3$ is the most promising crystal structure. The four experimentally known crystals, marked blue in Table B.1, are placed much lower in ranking. Most of the I-III-V-I materials are narrow gap semiconductors with a band gap smaller than $1\,eV$, except for some metallic exceptions that have no band gap. During the relaxation cycle the b/a ratios remained constant for all crystal structures, while the c/a ratios differed slightly in both directions from the initial 1.035.

I-V-V-I | 11 members | 0.0095 eV/atom - This subfamily of 11 members is purely metallic. $Cs_6BiSb_4Rb_3$ has with $E_{decomp} = 0.0095\,eV/atom$ the lowest decomposition energy. Observing the c/a ratios in Table B.2, one can deduce that the unit cells are slightly compressed in the c-direction compared to experimentally known crystals. The b/a ratios remained constant just as in the I-III-V-I subfamily.

Figure 6.4 and 6.5 show the DOS-plots for the two most stable crystal structures of the I-III-V-I family, respectively $Cs_6InSb_4Rb_3$ and $Cs_6AlBi_4K_3$. These density profiles (and the profiles of the rest of this subfamily) have a lot of common characteristics, including a forbidden zone above the Fermi level E_F and a peak around $1\,eV$.

$Cs_6BiSb_4Rb_3$ and $Cs_6BiSb_4K_3$ are the two most stable materials in the I-V-V-I family. From their DOS-plots in Figure 6.6 and 6.7 it is clear that I-V-V-I is a family of semiconductors. When comparing with the I-III-V-I set one sees that there are two extra valence electrons per formula unit. These electrons fill the empty states at $1\,eV$ in Figure 6.4 and 6.5, which causes a shift of the Fermi energy E_F. The states in Figure 6.6 and 6.7 are indeed shifted to the left with respect to Figure 6.4 and 6.5.

The band structures of these four materials are given in the same order in Figure 6.8 - 6.13. From Figure 6.8 and 6.9 it is clear that $Cs_6InSb_4Rb_3$ and $Cs_6AlBi_4K_3$ are direct semiconductors and that the narrow band gap is due to just one state that takes distance from the rest of the conduction band at the Γ point. Figure 6.10 and 6.11 show indeed that the I-III-V-I has some accessible states just above E_F and prove thus their metallic character.

One can ask why some of the members of the first subfamily have the metallic property too. The DOS-plot of the most stable metallic I-III-V-I member in Figure 6.12 and its band structure in Figure 6.13 answer this question. One sees on these figures that the non-existence of the band gap is not due to a peak at the Fermi level but to an extremely long and thin tail from the peak at $1\,eV$. The tail reaches below the Fermi level and overlaps with the valence band in the Γ point. This density profile and band structure can thus be classified with the other DOS- and band-plots of the I-III-V-I subfamily.

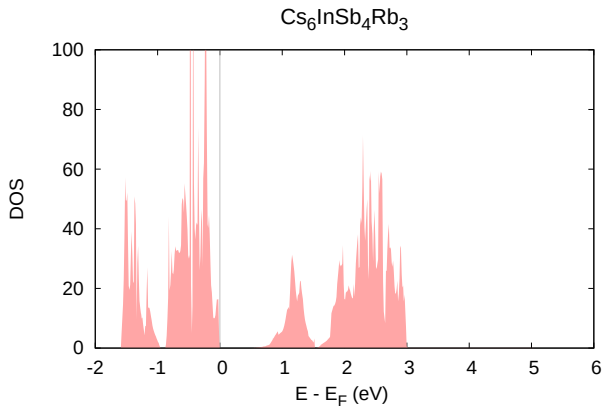


Figure 6.4: DOS of $Cs_6InSb_4Rb_3$ (I-III-V-I).

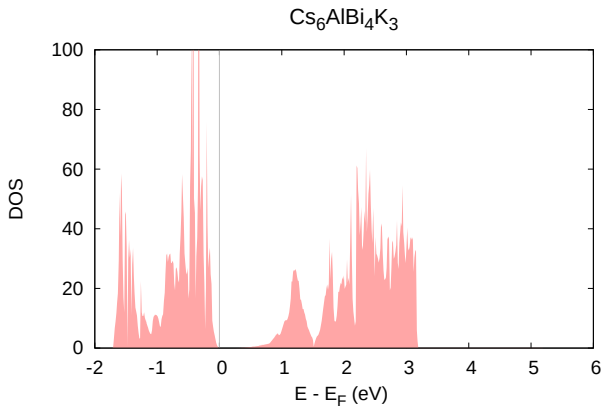


Figure 6.5: DOS of $Cs_6AlBi_4K_3$ (I-III-V-I).

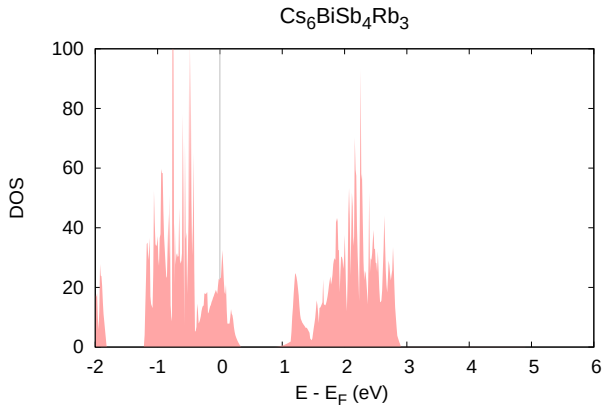


Figure 6.6: DOS of $Cs_6BiSb_4Rb_3$ (I-V-V-I).

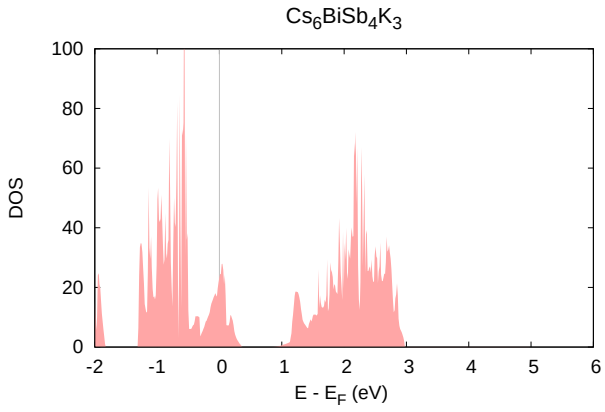


Figure 6.7: DOS of $Cs_6BiSb_4K_3$ (I-V-V-I).

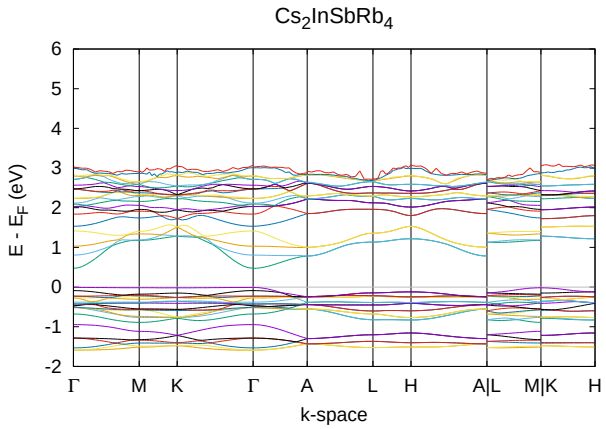


Figure 6.8: BANDS of $Cs_6InSb_4Rb_3$ (I-III-V-I).

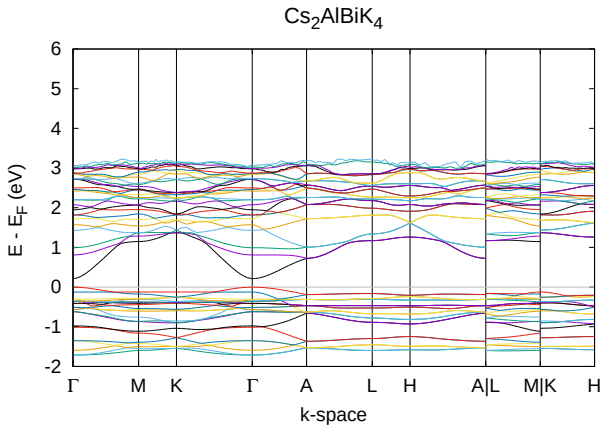


Figure 6.9: BANDS of $Cs_6AlBi_4K_3$ (I-III-V-I).

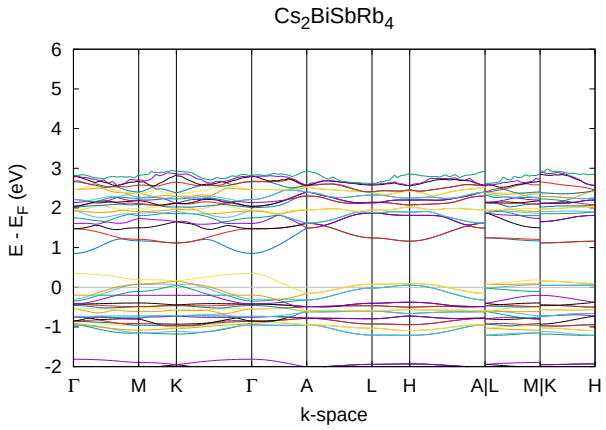


Figure 6.10: BANDS of $Cs_6BiSb_4Rb_3$ (I-V-V-I).

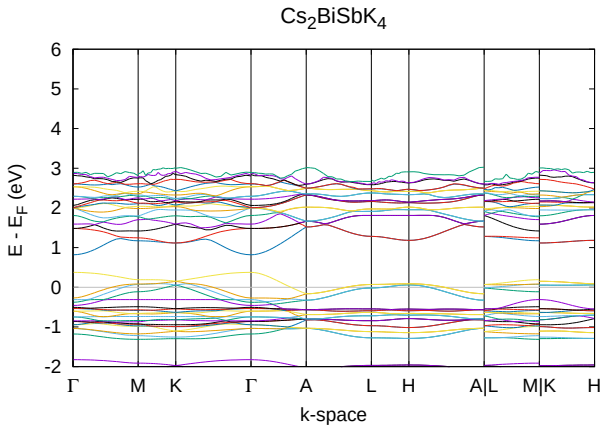


Figure 6.11: BANDS of $Cs_6BiSb_4K_3$ (I-V-V-I).

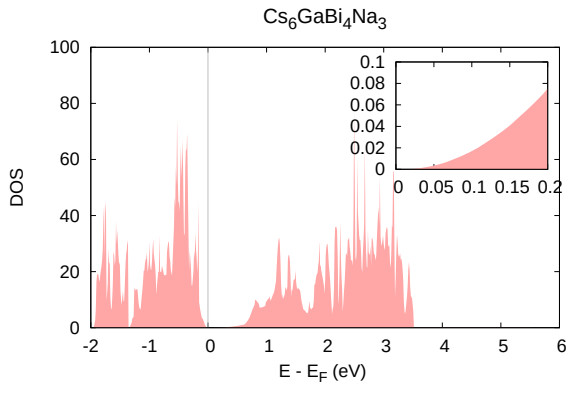


Figure 6.12: DOS of $Cs_6GaBi_4Na_3$ (I-III-V-I).

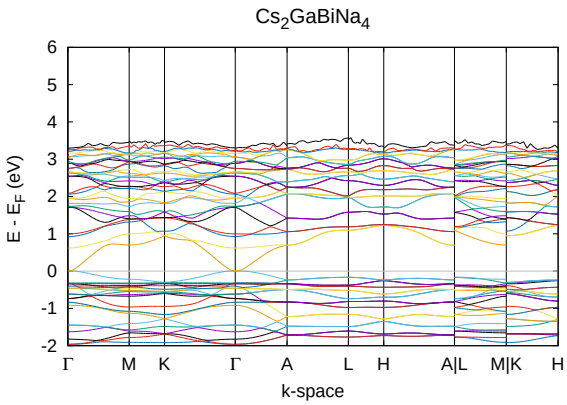


Figure 6.13: BANDS of $Cs_6GaBi_4Na_3$ (I-III-V-I).

Chapter 7

Global results

"How much is left to discover?", this question was the motivation to start this multiannual project. Although it is impossible to answer this question exactly, DFT calculation of large sets of systematically generated crystal structures can unveil the direction into which the answer lies. Dumon[4] initiated this project with the brute force calculation of 4224 QZPs in the Pnma space group, from which 198 materials where found with a decomposition energy below 0.05 eV. The main goal of this thesis was to replace the brute force approach with a more efficient method. With such method one could extract the most promising crystal structures with less than 50% of the computation time required for the brute force approach. This summarizing chapter recapitulates the gathered information in this work and gives an outlook for further research.

The benefits of intelligent screening

In this work an algorithm based on Neural Networks was introduced and tested on three sets of crystal structures, varying in symmetry, number of atoms in the primitive cell and size of the sets. There is no doubt that the *intelligent screening* algorithm that was proposed in Chapter 4 is a fair alternative for the brute force approach.

Figure 7.1 demonstrates that the *intelligent screening* method is able to reduce the execution time to extract all stable crystal structures by a factor 2. It should be noted that the $P6_3/mmc$ graph is treated as if all stables are found: it is hypothesized that the 122 materials with $E_{decomp} \leq 0.05\,eV$ are 100% of the stables in this space group. In this way the graph can be compared with the other graphs of the Pnma and the I-42m families in the same figure. Nevertheless, one is never certain that all stables have been found until all crystal structures have been computed. This is clearly illustrated by the Pnma graph in Figure 7.1, for which the last three stable crystal structures emerged in the final 30% of the simulation.

Figure 7.1 shows that at the first update of the Neural Networks, the predictive capacity rises to a much higher slope than the brute force method. For the $P6_3/mmc$ subfamily this application was already done at 5.8% of the total set instead of the 10% for the earlier two simulations. This reduction of the initial training step of the *intelligent screening* algorithm did not have adverse consequences. It may be recommended to reduce this initial step even more in future applications.

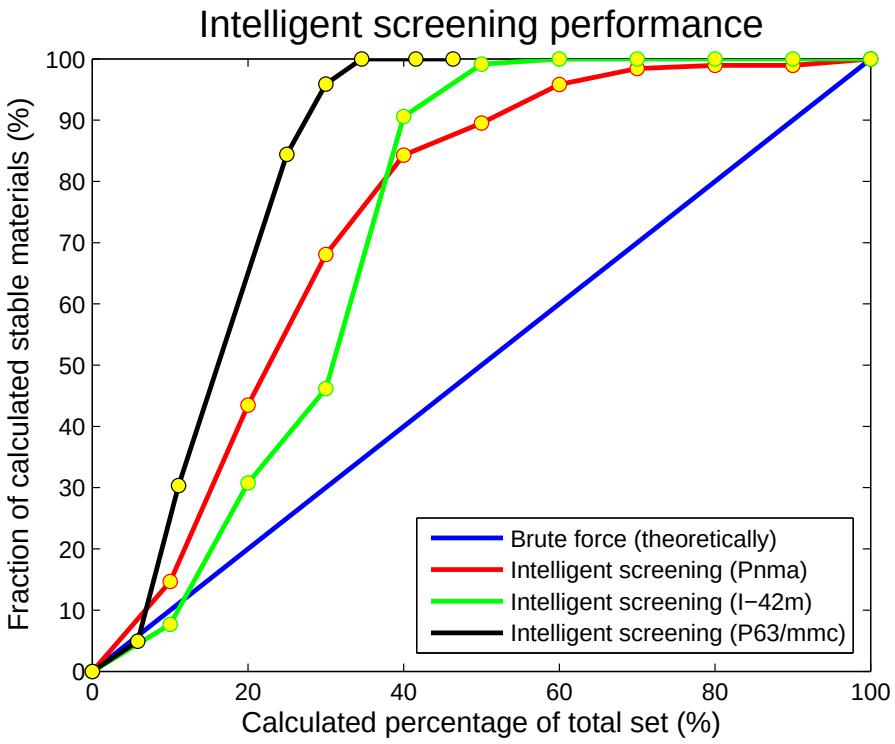


Figure 7.1: The *intelligent screening* algorithm applied on the three crystal families. The vertical axis represents the fraction of stable crystals that has already been found, while the horizontal axis is the fraction of the total data set that has been calculated. The yellow dots indicate the NN updates.

The descriptive attributes that are used to feed the Neural Networks in this project are for each of the four participating elements: the atomic number Z , the principal quantum number n, the number of electrons in the valence s orbital, the number of electrons in the valence p orbital, the number of electrons in the valence d orbital, the cohesive energy, the elemental volume, the electronegativity, the electron affinity and the ionisation potential. The group and row numbers on the periodic table and the atomic radii may be alternative valuable information that can be added in further studies.

Three families of crystal structures

Figure 7.2 combines the results of the three QZP families. Each tile on the subfigures represents an element on the periodic table with an associated average decomposition energy. This average is taken from all calculated compounds that hold this specific element in one QZP family. It has to be stressed that these figures do not depict the pure formation energy as in Dumon's overview[4] that was discussed in Chapter 1. Besides, it has to be mentioned that the contour plot of $P6_3/mmc$ can be seen as a lower limit of the real averages. Since only the best predicted materials with low E_{decomp} are calculated for this family, the averages will rise when completing this material set. From the three contour plots in Figure 7.2 and the lists of all stable crystals, one can draw the global conclusions that are discussed below.

The elements associated with the lowest average decomposition energies are more frequently present in the most stable compounds:

- The figure suggests that the preferred cations in the Pnma family are Cs, Rb, K. Further investigation shows that 85% of the compounds with $E_{decomp} \leq 0.05\,eV$ contain one of these three elements.
- For the I-42m family, Se and Te seem to be in favor for group VI and Si, Ge and Sn for group IV. This is also substantiated by the 117 stable I-42m materials: all of them contain Si or Te and only 44% does not contain Si, Ge or Sn.
- The contour plot of $P6_3/mmc$ clearly reveals three zones of favored elements: group I, group III and group VI. This is consistent with the only two subfamilies that were found: I-III-V-I and I-V-V-I.

Another observation is that the original crystal structures from which the families are constructed are partially reflected in the contour plots. $CsPbPSe_4$ contains with Cs the most favored element of the Pnma family and $K_2BaSnTe_4$ has only Ba that does not lay in a distinct favored zone. As a third example, the four experimentally known $P6_3/mmc$ compounds $K_6AlSb_4Na_3$, $Cs_6AlSb_4K_3$, $K_6GaSb_4Na_3$ and $Cs_6GaSb_4K_3$ only contain elements that are suggested by the $P6_3/mmc$ contour plot. Figure 7.2 shows that it is no coincidence that precisely these compounds are already experimentally known.

For the Pnma and $P6_3/mmc$ families and to a lesser extent for the I-42m family, large cations seem to be preferred. This may be the result of their diffuse outer shell, which makes them more electropositive than smaller cations. Larger cations also create larger spaces between them, which can easier be filled by the anionic structures.

Computational resources

The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government - department EWI. Access was granted to the tier-1 supercomputer facilities (Muk) for 1000 node days. All other calculations were performed on the tier-2 cluster Delcatty. The total number of consumed node days that were used in this project are given in Table 7.1. Each node day is a run of one day (real time) on all cores of one node of a cluster, whereby Muk and Delcatty both contain 16 cores per node. Although only 785 $P6_3/mmc$ crystal structures were computed, this family took 100 node days more calculation time than the 2016 crystal structures of the I-42m family in total.

Other interesting families

Over the past two years already three QZP families were the subject of research: Pnma in the 1:1:1:4 stoichiometry, I-42m in the 1:1:2:4 stoichiometry and $P6_3/mmc$ in the 1:3:4:6 stoichiometry. Of course there are still many combinations of space group and stoichiometry left to be studied. Table 7.2 gives an overview of the most interesting families for further research, based on the ICSD.

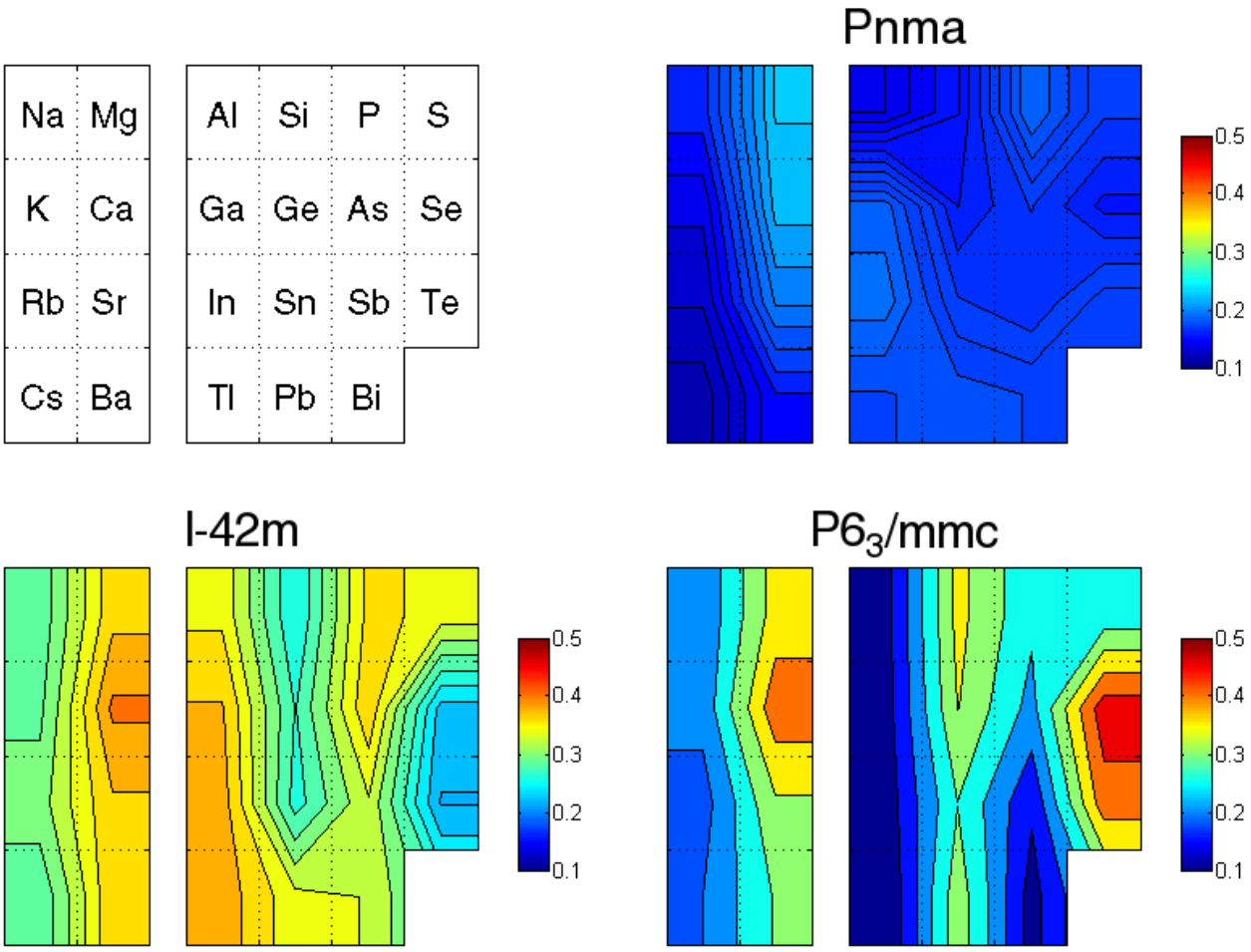


Figure 7.2: These contour plots give an overview of the decomposition energies of the compounds in the three space groups. Each tile represents an element on the periodic table and depicts the average decomposition energy of all compounds in the space group holding this element.

The ZintlDB: new materials at your fingertips

Three different families were studied, for which 11 experimental quaternary crystals were known in advance. Subsequently, the stability of 427 yet undiscovered candidate crystals is predicted in this work. These results suggest that the total number of stable quaternaries that could exist is by no means small. An overview of the stable crystals in these three families is given in Table 7.3.

It would be a shame to keep this information inaccessible for experimentalists. The ZintlDB.com provides information on all materials in the three calculated sets for which $E_{decomp} \leq 0.2\,eV$. For each of these materials the lattice constants, formation and decomposition energy, DOS and band structures are given. These data can be accessed with the objective that some of the featured crystals once will be composed and prove their applicability.

Table 7.1: The total number of consumed node days for the I-42m and $P6_3/mmc$ families on the clusters Muk and Delcatty.

Space group	Node days Muk	Node days Delcatty	Node days Total
I-42*m*	124.8	24.7	149.5
$P6_3/mmc$	102.5	147.5	250.0

Table 7.2: Interesting QZP families for further research.

Space group	Stoichiometry	Number of structures in ICSD
$P6_3mc$	1:2:4:4	11
Pnma	1:1:2:5	7
$P2_1/c$	1:1:1:4	7
*R*3*:H*	1:1:2:6	6
Ibam	1:1:2:2	6
Pa-3	1:1:1:4	4

Table 7.3: Overview of the stable crystals in the three studied QZP families. A crystal is defined stable if $E_{decomp} \leq 0.05 eV/atom$.

Space group	Number of experimentally known crystals	Number of newly suggested crystals
Pnma	6	192
I-42*m*	1	117
$P6_3/mmc$	4	118

Bibliography

[1] V. Atuchin, V. Pankevich, O. Parasyuk, N. Pervukhina, L. Pokrovsky, V. Remesnik, V. Uvarov, and V. Pekhnyo, "Structural and optical properties of noncentrosymmetric quaternary crystal AgCd2GaS4," *Journal of Crystal Growth*, vol. 292, no. 2, pp. 494–499, 2006. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0022024806003733 1

[2] T. Watanabe, H. Yanagi, T. Kamiya, Y. Kamihara, H. Hiramatsu, M. Hirano, and H. Hosono, "Nickel-Based Oxyphosphide Superconductor with a Layered Crystal Structure, LaNiOP," *Inorganic Chemistry*, vol. 46, no. 19, pp. 7719–7721, 2007, pMID: 17705374. [Online]. Available: http://dx.doi.org/10.1021/ic701200e 1

[3] L. Sun, B. Liu, J. Wang, J. Wang, Y. Zhou, and Z. Hu, "Y4Si2O7N2 : A New Oxynitride with Low Thermal Conductivity," *Journal of the American Ceramic Society*, vol. 95, no. 10, pp. 3278–3284, 2012. [Online]. Available: http://doi.wiley.com/10.1111/j.1551-2916.2012.05389.x 1

[4] K. Dumon, "Boosting the discovery rate of quaternary Zintl phases," 2015. 1, 1.4, 1.7, 1.4, 1.8, 1.9, 2.1.3, 2.3, 4, 4.1.1, 4.1.1, 5.1, 7, 7

[5] C. Kittel, *Introduction to Solid State Physics*, 6th ed. New York: John Wiley & Sons, Inc., 1986. 1.1.1

[6] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. Persson, "The Materials Project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1, p. 011002, 2013. [Online]. Available: http://link.aip.org/link/AMPADS/v1/i1/p011002/s1&Agg=doi 1.1.2, 1.3.2, 2.3.2, 2.3.3, 4.1.1

[7] S. P. Ong, L. Wang, B. Kang, and G. Ceder, "Li-Fe-P-O2 Phase Diagram from First Principles Calculations," *Chemistry of Materials*, vol. 20, no. 5, pp. 1798–1807, Mar. 2008. 1.1.2, 4.1.1

[8] G. Bergerhoff, I. D. Brown, F. H. Allen, and others, "Crystallographic databases," *International Union of Crystallography, Chester*, pp. 77–95, 1987. 1.2, 2.3.3

[9] FIZ Karlsruhe ICSD, "ICSD entries," https://icsd.fiz-karlsruhe.de, 2016. 1.2

[10] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD)," *JOM*, vol. 65, no. 11, pp. 1501–1509, 2013. [Online]. Available: http://dx.doi.org/10.1007/s11837-013-0755-4 1.2

[11] S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, and A. Le Bail, "Crystallography Open Database – an open-access collection of crystal structures," *Journal of Applied Crystallography*, vol. 42, no. 4, pp. 726–729, Aug 2009. [Online]. Available: http://dx.doi.org/10.1107/S0021889809016690 1.2

[12] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moeck, R. T. Downs, and A. Le Bail, "Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration," *Nucleic Acids Research*, vol. 40, no. D1, pp. D420–D427, 2012. [Online]. Available: http://nar.oxfordjournals.org/content/40/D1/D420.abstract 1.2

[13] S. C. Glotzer, "Challenges and Opportunities in," *Arbor Ciencia Pensamiento Y Cultura*, vol. 9, no. 10, pp. 730–734, 2014. [Online]. Available: http://dx.doi.org/10.1038/nnano.2014.225 1.2, 1.2

[14] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, "The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies," *Nature Publishing Group*, vol. 1, no. October, p. 15010, 2015. [Online]. Available: http://dx.doi.org/10.1038/npjcompumats.2015.10 1.4, 5.1

[15] Z. H. Liu, M. Zhang, W. Q. Wang, W. H. Wang, J. L. Chen, G. H. Wu, F. B. Meng, H. Y. Liu, B. D. Liu, J. P. Qu, and Y. X. Li, "Magnetic properties and martensitic transformation in quaternary Heusler alloy of NiMnFeGa," *Journal of Applied Physics*, vol. 92, no. 9, pp. 5006–5010, 2002. [Online]. Available: http://scitation.aip.org/content/aip/journal/jap/92/9/10.1063/1.1511293 1.3

[16] I. Galanakis, "Appearance of half-metallicity in the quaternary heusler alloys," *Journal of Physics: Condensed Matter*, vol. 16, no. 18, p. 3089, 2004. [Online]. Available: http://stacks.iop.org/0953-8984/16/i=18/a=010 1.3

[17] S. Wurmehl, G. H. Fecher, K. Kroth, F. Kronast, H. A. Dürr, Y. Takeda, Y. Saitoh, K. Kobayashi, H.-J. Lin, G. Schönhense, and C. Felser, "Electronic structure and spectroscopy of the quaternary heusler alloy co 2 cr 1- x fe x al," *Journal of Physics D: Applied Physics*, vol. 39, no. 5, p. 803, 2006. [Online]. Available: http://stacks.iop.org/0022-3727/39/i=5/a=S06 1.3

[18] Z. Liu, L. Zheng, L. Sun, Y. Qian, J. Wang, and M. Li, "(Cr2/3Ti1/3)3AlC2 and (Cr5/8Ti3/8)4AlC3: New MAX-phase Compounds in Ti–Cr–Al–C System," *Journal of the American Ceramic Society*, vol. 97, no. 1, pp. 67–69, 2014. [Online]. Available: http://dx.doi.org/10.1111/jace.12731 1.3

[19] Z. Liu, E. Wu, J. Wang, Y. Qian, H. Xiang, X. Li, Q. Jin, G. Sun, X. Chen, J. Wang, and M. Li, "Crystal structure and formation mechanism of (Cr2/3Ti1/3)3AlC2 MAX phase," *Acta Materialia*, vol. 73, pp. 186 – 193, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1359645414002456 1.3

[20] B. Anasori, M. Dahlqvist, J. Halim, E. J. Moon, J. Lu, B. C. Hosler, C. N., S. J. May, L. Hultman, P. Eklund, J. Rosén, and M. W. Barsoum, "Experimental and theoretical characterization of ordered MAX phases Mo2TiAlC2 and Mo2Ti2AlC3," *Journal of Applied Physics*, vol. 118, no. 9, 2015. [Online]. Available: http://scitation.aip.org/content/aip/journal/jap/118/9/10.1063/1.4929640 1.3

[21] D. Horlait, S. Grasso, A. Chroneos, and W. E. Lee, "Attempts to synthesise quaternary MAX phases (Zr,M)2AlC and Zr2(Al,A)C as a way to approach Zr2 AlC," *Materials Research Letters*, vol. 3831, no. March, pp. 1–8, 2016. [Online]. Available: http://www.tandfonline.com/doi/full/10.1080/21663831.2016.1143053 1.3

[22] E. Zintl and W. Dullenkopf, *Zeitschrift für Physikalische Chemie, Abteilung B*, vol. B, no. 16, p. 195, 1932. 1.3.1

[23] E. Zintl and G. Brauer, *Zeitschrift für Physikalische Chemie, Abteilung B*, vol. B, no. 20, p. 245, 1933. 1.3.1

[24] E. Zintl and A. Woltersdorf, "Gitterstruktur von LiAl. (18. Mitteilung über Metalle und Legierungen.)," *Zeitschrift für Elektrochemie und angewandte physikalische Chemie*, vol. 41, no. 12, pp. 876–879, 1935. [Online]. Available: http://dx.doi.org/10.1002/bbpc.19350411212 1.3.1

[25] U. Mizutani, *Hume-Rothery Rules for Structurally Complex Alloy Phases*. CRC Press, 2010. [Online]. Available: https://books.google.be/books?id=DBmO34STnZwC 1.3.1

[26] S. C. Sevov, "Zintl Phases," *Intermetallic Compounds: Vol. 3, Principles and Practice*, vol. 3, pp. 113—-132, 2002. 1.3.1, 1.3.1

[27] R. Ferro and A. Saccone, *Intermetallic Chemistry*, ser. Pergamon materials series. Elsevier, 2008. [Online]. Available: https://books.google.be/books?id=I7hAibM9BgoC 1.3.1

[28] R. Nesper and M.-p.-i. Festkrperforschung, "mm pmm , mmma , mma NB mmmBm mlmalmmmam," vol. 38, 1990. 1.3.1, 1.3.2

[29] E. Zintl, "Intermetallische Verbindungen," *Angew. Chem.*, vol. 52, pp. 1–6, 1939. [Online]. Available: http://doi.wiley.com/10.1002/ange.19390520102 1.3.1

[30] D. A. V. F. L. Santamaría-Pérez, "The Zintl –Klemm Concept Applied to Cations in Oxides . II . The Structures of Silicates Periodic Table of Elements," *Group*, no. September, pp. 121–177, 2005. 1.3.1

[31] G. Derrien, M. Tillard, A. Manteghetti, and C. Belin, "Phosphorus Oligomerization in Zintl Phases: Synthesis, Crystal Structure, and Bonding Analysis of Mixed Alkali and Alkaline-Earth Metal Phosphides," *Zeitschrift für anorganische und allgemeine Chemie*, vol. 629, no. 9, pp. 1601–1609, 2003. [Online]. Available: http://dx.doi.org/10.1002/zaac.200300086 1.3.2

[32] J. Thijssen, *Computational Physics*. Cambridge University Press, 1999. 2.1.1

[33] C. Pisani, R. Dovesi, and C. Roetti, *Hartree-Fock Ab Initio Treatment of Crystalline Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1988. 2.1.2

[34] P. Hohenberg and W. Kohn, "Inhomogeneous Electron Gas," *Physical Review B*, vol. 136, no. 3B, pp. B864–B871, 1964. 2.1.2

[35] S. Cottenier, *Density Functional Theory and the Family of (L)APW-methods: a step-by-step introduction*, 2013-2002 (2nd edition). 2.1.2, 2.1.2, 2.1

[36] W. Kohn and L. J. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects," *Physical Review*, vol. 140, no. 4A, pp. A1133–A1138, 1965. 2.1.2, 2.1.2

[37] J. P. Perdew and K. Schmidt, "Jacob's ladder of density functional approximations for the exchange-correlation energy," *AIP Conference Proceedings*, vol. 577, no. Density Functional Theory and Its Application to Materials, pp. 1–20, 2001. [Online]. Available: http://link.aip.org/link/?APC/577/1/1&Agg=doi 2.1.2

[38] J. P. Perdew, "Accurate density functional for the energy: Real-space cutoff of the gradient expansion for the exchange hole," *Phys. Rev. Lett.*, vol. 55, pp. 1665–1668, Oct 1985. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.55.1665 2.1.2

[39] S.-K. Ma and K. A. Brueckner, "Correlation energy of an electron gas with a slowly varying high density," *Phys. Rev.*, vol. 165, pp. 18–31, Jan 1968. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRev.165.18 2.1.2

[40] D. C. Langreth and M. J. Mehl, "Beyond the local-density approximation in calculations of ground-state electronic properties," *Phys. Rev. B*, vol. 28, pp. 1809–1834, Aug 1983. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevB.28.1809 2.1.2

[41] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized Gradient Approximation Made Simple," *Physical Review Letters*, vol. 77, no. 18, pp. 3865–3868, 1996. 2.1.2

[42] ——, "Generalized Gradient Approximation Made Simple- ERRATA," *Physical Review Letters*, vol. 77, no. 18, pp. 3865–3868, 1996. 2.1.2

[43] J. Kitchin, "Modeling materials using density functional theory," pp. 1–298, 2012. 2.1.3, 2.1.3, 2.1.3

[44] D. Sholl and J. Steckel, *Density Functional Theory: A Practical Introduction*. Wiley, 2011. [Online]. Available: https://books.google.be/books?id=_f994dmAdv0C 2.1.3

[45] D. J. Chadi and M. L. Cohen, "Special points in the brillouin zone," *Phys. Rev. B*, vol. 8, pp. 5747–5753, Dec 1973. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevB.8.5747 2.1.3

[46] H. J. Monkhorst and J. D. Pack, "Special points for brillouin-zone integrations," *Phys. Rev. B*, vol. 13, pp. 5188–5192, Jun 1976. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevB.13.5188 2.1.3

[47] P. E. Blöchl, "Projector augmented-wave method," *Phys. Rev. B*, vol. 50, pp. 17 953–17 979, Dec 1994. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevB.50.17953 2.1.3

[48] G. Kresse and D. Joubert, "From ultrasoft pseudopotentials to the projector augmented-wave method," *Phys. Rev. B*, vol. 59, pp. 1758–1775, Jan 1999. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevB.59.1758 2.1.3

[49] G. Kresse and M. Marsman, "Vienna Ab initio Simulation Package VASP the GUIDE," 2016. [Online]. Available: http://cms.mpi.univie.ac.at/vasp/vasp.pdf 2.2, 2.2.1, 2.2.2

[50] F. Birch, "Finite Elastic Strain of Cubic Crystals," *Physical Review*, vol. 71, pp. 809–824, Jun. 1947. 2.2.2

[51] European Grid Infrastructure, "Glossary v2," https://wiki.egi.eu/wiki/Glossary_V2, 2013. 2.3.1

[52] W. F. Maier, K. Stowe, and S. Sieg, "Combinatorial and high-throughput materials science," *Angewandte Chemie - International Edition*, vol. 46, no. 32, pp. 6016–6067, 2007. 2.3.1

[53] R. Hoogenboom, M. A. R. Meier, and U. S. Schubert, "Combinatorial Methods , Automated Synthesis and High-Throughput Screening in Polymer Research : Past and Present," *Macromol. Rapid Commun.*, vol. 24, pp. 15–32, 2003. 2.3.2

[54] S. Curtarolo, D. Morgan, and G. Ceder, "Accuracy of ab initio methods in predicting the crystal structures of metals: A review of 80 binary alloys," *Calphad: Computer Coupling of Phase Diagrams and Thermochemistry*, vol. 29, no. 3, pp. 163–211, 2005. 2.3.2

[55] T. F. T. Cerqueira, S. Pailhes, R. Debord, V. M. Giordano, R. Viennois, J. Shi, S. Botti, and M. A. L. Marques, "Prediction and synthesis of a non-Zintl silicon clathrate," *Chemistry of Materials*, p. acs.chemmater.6b00392, 2016. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/acs.chemmater.6b00392 2.3.2

[56] A. Jain, G. Hautier, C. J. Moore, S. Ping Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, "A high-throughput infrastructure for density functional theory calculations," *Computational Materials Science*, vol. 50, no. 8, pp. 2295–2310, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.commatsci.2011.02.023 2.3.3, 2.2

[57] The phpMyAdmin devel team, "phpMyAdmin's documentation," http://docs.phpmyadmin.net/, 2016. 2.3.3

[58] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design." *Nature materials*, vol. 12, no. 3, pp. 191–201, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23422720 2.3.3

[59] N. Williams, S. Zander, and G. Armitage, "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 5, p. 5, 2006. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1163593.1163596 3.1, 3.1.1

[60] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," *Applied and Environmental Microbiology*, vol. 73, no. 16, pp. 5261–5267, 2007. 3.1.1

[61] C.-C. C. Chih-Wei Hsu and C.-J. Lin, "A Practical Guide to Support Vector Classification," *BJU international*, vol. 101, no. 1, pp. 1396–400, 2008. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf 3.1.1

[62] S. ND4J, "Introduction to deep neural networks," http://deeplearning4j.org/neuralnet-overview.html, 2016. 3.1.1

[63] G. Dahl, A.-R. Mohamed, and G. Hinton, "Phone recognition with the mean-covariance restricted Boltzmann machine," *Advances in neural information*, pp. 469–477, 2010. [Online]. Available: http://machinelearning.wustl.edu/mlpapers/paper{_}files/NIPS2010{_}0160.pdf$\delimiter"026E30F$nhttp://papers.nips.cc/paper/4169-phone-recognition-with-the-mean-covariance-restricted-boltzmann-machine 3.1.1

[64] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012. 3.1.1

[65] L. Deng, M. Seltzer, D. Yu, A. Acero, A.-R. Mohamed, and G. Hinton, "Binary Coding of Speech Spectrograms Using a Deep Auto-encoder," *Interspeech*, no. September, pp. 1692–1695, 2010. 3.1.1

[66] J. Weng, N. Ahuja, and T. Huang, "Learning recognition and segmentation of 3-D objects from 2-D images," *1993 (4th) International Conference on Computer Vision*, 1993. 3.1.1

[67] E. M. Azoff, *Neural Network Time Series Forecasting of Financial Markets*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1994. 3.1.1

[68] G. J. I., S. D. J., D. T. A., and W. L. K., "Real-time camera-based face detection using a modified LAMSTAR neural network system," *Proceedings of the SPIE the International Society for Optical Engineering*, pp. 35–46, 2003. 3.1.1

[69] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2012. 3.1.1

[70] J. Weston, S. Bengio, and N. Usunier, "Large scale image annotation: Learning to rank with joint word-image embeddings," *Machine Learning*, vol. 81, no. 1, pp. 21–35, 2010. 3.1.1

[71] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, "Combinatorial screening for new materials in unconstrained composition space with machine learning," *Physical Review B - Condensed Matter and Materials Physics*, vol. 89, no. 9, pp. 1–7, 2014. 3.1.2

[72] ——, "Combinatorial screening for new materials in unconstrained composition space with machine learning - Supplemantary Info," 2014. [Online]. Available: http://journals.aps.org/prb/abstract/10.1103/PhysRevB.89.094104#supplemental 3.1.2

[73] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A New classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619–1630, 2006. 3.1.2

[74] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," *Physical Review Letters*, vol. 98, no. 14, pp. 1–4, 2007. 3.1.2

[75] R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, and M. Parrinello, "Graphite-diamond phase coexistence study employing a neural-network mapping of the ab initio potential energy surface," *Physical Review B - Condensed Matter and Materials Physics*, vol. 81, no. 10, pp. 18–21, 2010. 3.1.2

[76] H. Eshet, R. Z. Khaliullin, T. D. Kühne, J. Behler, and M. Parrinello, "Ab initio quality neural-network potential for sodium," *Physical Review B - Condensed Matter and Materials Physics*, vol. 81, no. 18, pp. 1–8, 2010. 3.1.2

[77] H. Eshet, R. Z. Khaliullin, T. D. Kühne, J. Behler, and M. Parrinello, "Microscopic origins of the anomalous melting behavior of sodium under high pressure," *Physical Review Letters*, vol. 108, no. 11, pp. 1–4, 2012. 3.1.2

[78] N. Artrith and A. Urban, "An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for TiO2," *Computational Materials Science*, vol. 114, pp. 135–150, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.commatsci.2015.11.047 3.1.2

[79] H. T. L. Y. S. Abu-Mostafa, M. Magdon-Ismail, *Learning From Data*. AMLBook, 2012, california Institute of Technology. 3.2, 3.2

[80] H. Demuth, "Neural Network Toolbox," *Networks*, vol. 24, no. 1, pp. 1–8, 2002. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.123.6691{&}rep=rep1{&}type=pdf 3.3

[81] MATLAB, *version 8.0.0.783 (R2012b)*. Natick, Massachusetts: The MathWorks Inc., 2012. 3.3, 3.5

[82] The MathWorks Inc, "Fit data with a neural network," http://mathworks.com/help/nnet/gs/fit-data-with-a-neural-network.html, 2016. 3.3.1

[83] ——, "Improve neural network generalization and avoid overfitting," http://mathworks.com/help/nnet/ug/improve-neural-network-generalization-and-avoid-overfitting.html, 2016. 3.3.1

[84] ——, "Choose neural network input-output processing functions," http://mathworks.com/help/nnet/ug/choose-neural-network-input-output-processing-functions.html, 2016. 3.3.2

[85] J. Li, H.-Y. Guo, D. M. Proserpio, and A. Sironi, "Exploring Tellurides: Synthesis and Characterization of New Binary, Ternary and Quaternary Compounds," *J. Solid State Chem.*, p. 247, 1995. [Online]. Available: http://chem.rutgers.edu/~jingli/papers/Exploring%20Tellurides%20Synthesis%20and%20Characterization%20of%20New%20B.pdf 5.1, 5.1, 5.1

[86] M. I. Aroyo, J. M. Perez-Mato, D. Orobengoa, E. Tasci, G. De La Flor, and A. Kirov, "Crystallography online: Bilbao crystallographic server," *Bulgarian Chemical Communications*, vol. 43, no. 2, pp. 183–197, 2011. 5.1

[87] S. Bhattacharya and G. K. H. Madsen, "High-throughput exploration of alloying as design strategy for thermoelectrics," *Physical Review B - Condensed Matter and Materials Physics*, vol. 92, no. 8, 2015. 5.1

[88] K. P. M. Somer, W. Carrillo-Cabrera, E.-M. Peters and H. G. von Schnering, "Crystal structure of sodium potassium antimonide triantimonido- aluminate, Na3K6Sb(AlSb3)," *Zeitschrift für Kristallographie*, vol. 210, p. 527, 1995. 6.1, 6.1

[89] ——, "Crystal structure of sodium potassium antimonide triantimonido- gallate,K6Na3Sb[GaSb3]," *Zeitschrift für Kristallographie*, vol. 210, p. 143, 1995. 6.1

[90] K. P. M. Somer and H. G. von Schnering, "Crystal structure of potassium caesium antimonide triantimonidogallate, K3Cs6Sb(GaSb3)," *Zeitschrift für Kristallographie*, vol. 195, pp. 101–102, 1991. 6.1

Appendix A

Crystallographic information of the I-42m family

This appendix contains the results of all I-42m crystals with $E_{decomp} \leq 50\,meV$. The trends between all subfamilies are discussed in detail in Chapter 5.

Table A.1: Crystallographic information of the 3 smaller subfamilies.

Subfamily	Name	E_{decomp} $(eV/atom)$	c/a	E_{gap} (eV)
I-II-III-VI	$Cs_2MgAlTe_4$	0.0307	0.838	-
I-II-III-VI	$Cs_2MgInSe_4$	0.0332	0.852	-
I-II-III-VI	$Cs_2MgAlSe_4$	0.0336	0.830	-
I-II-V-VI	$Rb_2MgSbTe_4$	0.0455	0.832	-
I-II-V-VI	$Cs_2MgSbTe_4$	0.0456	0.818	-
II-I-III-VI	$Ba_2CsAlSe_4$	0.0000	0.915	2.20
II-I-III-VI	$Ba_2RbAlSe_4$	0.0029	1.049	2.25
II-I-III-VI	$Mg_2CsAlSe_4$	0.0097	1.214	1.08
II-I-III-VI	$Mg_2RbAlSe_4$	0.0342	1.218	1.28
II-I-III-VI	$Mg_2RbAlTe_4$	0.0397	1.241	1.14

Table A.2: Crystallographic information of the I-I-IV-VI subfamily.

I-I-IV-VI	E_{decomp} $(eV/atom)$	c/a	E_{gap} (eV)
$K_2RbSiSe_4$	-0.0419	1.033	-
$Na_2RbSiSe_4$	0.0117	1.116	-
Na_2KSiSe_4	0.0141	1.086	-
$K_2NaSiSe_4$	0.0177	0.926	-
Rb_2KSiSe_4	0.0198	0.969	-
$Rb_2NaSiSe_4$	0.0308	0.902	-
$Na_2CsGeSe_4$	0.0312	1.126	-

$K_2CsGeSe_4$	0.0393	1.059	-
$Rb_2CsGeSe_4$	0.0423	1.032	-
$Cs_2NaGeSe_4$	0.0431	0.894	-
$Cs_2RbGeSe_4$	0.0492	1.101	-

Table A.3: Crystallographic information of the I-II-IV-VI subfamily.

I-II-IV-VI	E_{decomp} $(eV/atom)$	c/a	E_{gap} (eV)
$Rb_2MgSiSe_4$	-0.1249	0.802	2.60
$K_2MgSiSe_4$	-0.1144	0.820	2.57
$Cs_2MgGeSe_4$	-0.0912	0.799	2.04
$Cs_2MgSnSe_4$	-0.0710	0.810	1.82
$Rb_2MgSiTe_4$	-0.0287	0.809	1.43
$Rb_2MgGeTe_4$	-0.0009	0.814	0.85
$Rb_2MgSnTe_4$	-0.0007	0.824	1.03
$K_2MgSiTe_4$	0.0030	0.826	1.24
$Cs_2MgGeTe_4$	0.0058	0.800	0.96
$Cs_2MgSnTe_4$	0.0061	0.810	1.13
$Cs_2MgSiTe_4$	0.0223	0.795	1.56
$Na_2MgSiSe_4$	0.0227	0.862	2.03
$K_2MgGeTe_4$	0.0317	0.830	0.68
$K_2MgSnTe_4$	0.0328	0.838	0.90
$K_2MgSnSe_4$	0.0366	0.841	1.83
$Na_2MgSiTe_4$	0.0368	0.857	0.73
$Na_2MgGeSe_4$	0.0373	0.866	1.44
$K_2MgGeSe_4$	0.0380	0.828	1.88
$Rb_2MgSnSe_4$	0.0454	0.824	1.85
$Rb_2MgGeSe_4$	0.0489	0.810	1.99
$Cs_2MgSiSe_4$	0.0496	0.791	2.62

Table A.4: Crystallographic information of the I-I-V-VI subfamily.

I-I-V-VI	E_{decomp} $(eV/atom)$	c/a	E_{gap} (eV)
Rb_2CsPSe_4	-0.0290	1.042	1.79
Na_2RbPSe_4	-0.0182	1.140	1.44
$K_2NaSbSe_4$	0.0000	1.074	1.28
$Rb_2NaSbSe_4$	0.0003	0.890	1.46
$K_2NaAsSe_4$	0.0010	0.911	0.97
$Na_2CsAsSe_4$	0.0015	1.160	1.06
Na_2KPSe_4	0.0024	1.107	1.30

$Na_2RbSbSe_4$	0.0028	1.123	1.19
$Cs_2RbAsSe_4$	0.0051	0.961	1.41
Na_2KAsSe_4	0.0068	1.102	0.71
$Cs_2NaAsSe_4$	0.0069	0.863	1.36
$Rb_2CsAsSe_4$	0.0078	1.041	1.34
$Rb_2CsSbTe_4$	0.0080	1.038	0.65
$Cs_2RbSbTe_4$	0.0098	0.963	0.75
$Cs_2NaSbSe_4$	0.0100	0.869	1.61
$Na_2RbAsSe_4$	0.0113	1.132	0.86
K_2NaPSe_4	0.0123	0.909	1.57
Rb_2NaPSe_4	0.0128	0.882	1.75
Na_2KSbSe_4	0.0137	1.099	1.05
Cs_2RbPSe_4	0.0163	0.961	1.81
Na_2CsPSe_4	0.0164	1.170	1.61
$K_2RbSbSe_4$	0.0166	1.034	1.44
K_2RbPSe_4	0.0178	1.038	1.74
$Na_2CsSbSe_4$	0.0179	1.147	1.35
Rb_2KPSe_4	0.0186	0.964	1.81
$Rb_2NaAsSe_4$	0.0193	0.885	1.18
$K_2RbAsSe_4$	0.0213	1.037	1.17
$K_2CsSbSe_4$	0.0219	1.071	1.51
Cs_2KAsSe_4	0.0221	0.929	1.40
$Rb_2CsSbSe_4$	0.0228	1.039	1.51
$Cs_2NaSbTe_4$	0.0229	0.864	0.67
$K_2CsAsSe_4$	0.0233	1.076	1.23
$Rb_2NaSbTe_4$	0.0239	0.885	0.43
$Cs_2RbSbSe_4$	0.0267	0.963	1.55
Cs_2NaPSe_4	0.0282	0.860	1.93
$K_2CsSbTe_4$	0.0293	1.069	0.51
$K_2RbSbTe_4$	0.0313	1.033	0.42
Rb_2KSbSe_4	0.0327	0.966	1.50
$Na_2RbSbTe_4$	0.0345	1.126	0.06
Rb_2KAsSe_4	0.0394	0.965	1.24
Cs_2KSbSe_4	0.0405	0.932	1.55
$Na_2CsSbTe_4$	0.0415	1.144	0.27
Cs_2KSbTe_4	0.0433	0.932	0.71
K_2CsPSe_4	0.0433	1.079	1.80
Rb_2KSbTe_4	0.0451	1.078	0.52
$K_2NaSbTe_4$	0.0464	0.906	0.21

Table A.5: Crystallographic information of the II-I-IV-VI subfamily.

II-I-IV-VI	E_{decomp} $(eV/atom)$	c/a	E_{gap} (eV)
$Ba_2RbSiTe_4$	0.0072	1.041	-
$Ca_2RbSiTe_4$	0.0153	1.092	-
$Ba_2NaSiTe_4$	0.0161	0.977	-
$Sr_2RbSiTe_4$	0.0180	1.065	-
$Ba_2RbSnTe_4$	0.0219	1.036	-
$Ba_2NaGeSe_4$	0.0230	0.983	-
$Ba_2NaSiSe_4$	0.0233	0.972	-
$Sr_2NaSiSe_4$	0.0244	1.001	-
$Ba_2NaSnSe_4$	0.0262	0.984	-
$Ba_2NaSnTe_4$	0.0293	0.982	-
$Ba_2RbGeTe_4$	0.0325	1.040	-
$Ba_2NaGeTe_4$	0.0343	0.980	-
$Ba_2CsSnTe_4$	0.0347	1.051	-
$Sr_2RbSnTe_4$	0.0371	1.054	-
$Sr_2NaGeSe_4$	0.0372	1.003	-
$Ca_2RbSnTe_4$	0.0392	1.077	-
$Sr_2NaSiTe_4$	0.0402	0.996	-
Ba_2KSiTe_4	0.0408	1.022	-
Ba_2KSnSe_4	0.0430	1.023	-
$Ca_2RbGeTe_4$	0.0430	1.087	-
$Ba_2CsGeTe_4$	0.0439	1.060	-
Sr_2KSiTe_4	0.0443	1.048	-
$Sr_2RbGeTe_4$	0.0443	1.061	-
Ca_2KSiTe_4	0.0451	0.966	-
$Ca_2NaSiSe_4$	0.0457	1.031	-
$Ca_2CsSnTe_4$	0.0459	1.087	-
Ba_2KSnTe_4	0.0459	1.023	-
$Sr_2CsSnTe_4$	0.0460	1.067	-
$Ca_2CsGeTe_4$	0.0492	0.972	-

Appendix B

Crystallographic information of the $P6_3/mmc$ family

This appendix contains the results of all $P6_3/mmc$ crystals with $E_{decomp} \leq 50\,meV$. The trends between all subfamilies are discussed in detail in Chapter 6.

Table B.1: Crystallographic information of the I-III-V-I subfamily. The blue colored crystal structures are experimentally known.

I-III-V-I	E_{decomp} $(eV/atom)$	c/a	E_{gap} (eV)
$Cs_6InSb_4Rb_3$	-0.0568	1.054	0.47
$Cs_6AlBi_4K_3$	-0.0564	1.045	0.20
$Cs_6InSb_4K_3$	-0.0558	1.050	0.65
$Cs_6AlSb_4Rb_3$	-0.0546	1.051	0.50
$Cs_6AlSb_4Na_3$	-0.0545	1.033	0.42
$Cs_6AlBi_4Rb_3$	-0.0530	1.048	0.10
$Cs_6GaBi_4K_3$	-0.0435	1.044	0.19
$Cs_6AlBi_4Na_3$	-0.0433	1.029	0.01
$Cs_6GaBi_4Rb_3$	-0.0408	1.049	0.10
$Cs_6InBi_4Rb_3$	-0.0367	1.051	0.09
$Cs_6TlBi_4Rb_3$	-0.0362	1.053	0.03
$Cs_6TlSb_4Rb_3$	-0.0355	1.055	0.37
$Cs_6InBi_4K_3$	-0.0352	1.047	0.25
$Cs_6TlBi_4K_3$	-0.0338	1.049	0.19
$Cs_6GaBi_4Na_3$	-0.0306	1.025	-
$Rb_6AlAs_4Na_3$	-0.0300	1.050	0.35
$Rb_6AlBi_4Na_3$	-0.0279	1.026	0.02
$Cs_6TlAs_4K_3$	-0.0276	1.067	0.48
$Cs_6InSb_4Na_3$	-0.0267	1.027	0.43
$Cs_6TlSb_4K_3$	-0.0266	1.051	0.54

$K_6AlBi_4Na_3$	-0.0201	1.024	0.10
$Rb_6InSb_4K_3$	-0.0199	1.046	0.25
$Cs_6TlAs_4Rb_3$	-0.0195	1.071	0.29
$Rb_6AlBi_4K_3$	-0.0193	1.040	-
$Rb_6InAs_4K_3$	-0.0183	1.060	0.11
$Rb_6InAs_4Na_3$	-0.0157	1.048	0.41
$Rb_6GaBi_4Na_3$	-0.0155	1.023	0.02
$Rb_6InSb_4Na_3$	-0.0153	1.024	0.57
$Rb_6InBi_4Na_3$	-0.0132	1.020	0.04
$Rb_6GaSb_4Na_3$	-0.0128	1.027	0.56
$Rb_6GaAs_4Na_3$	-0.0121	1.045	0.37
$Cs_6InBi_4Na_3$	-0.0120	1.024	-
$Rb_6InSb_4Cs_3$	-0.0106	1.069	-
$Cs_6AlAs_4K_3$	-0.0093	1.057	0.52
$Rb_6GaAs_4K_3$	-0.0083	1.054	0.17
$Cs_6TlBi_4Na_3$	-0.0081	1.019	-
$Rb_6InBi_4K_3$	-0.0075	1.043	-
$Rb_6GaBi_4K_3$	-0.0075	1.041	-
$K_6GaBi_4Na_3$	-0.0072	1.024	0.06
$Rb_6GaSb_4K_3$	-0.0064	1.044	0.30
$K_6InAs_4Rb_3$	-0.0055	1.072	-
$Rb_6AlSb_4Na_3$	-0.0047	1.031	0.56
$K_6InSb_4Rb_3$	-0.0044	1.058	-
$Cs_6TlSb_4Na_3$	-0.0035	1.023	0.44
$Cs_6AlAs_4Rb_3$	-0.0014	1.062	0.38
$Cs_6GaSb_4Rb_3$	0.0005	1.051	0.48
$Cs_6GaAs_4K_3$	0.0006	1.056	0.52
$Rb_6AlAs_4K_3$	0.0012	1.055	0.22
$K_6AlBi_4Rb_3$	0.0012	1.046	-
$Rb_6AlSb_4K_3$	0.0017	1.043	0.35
$Rb_6TlBi_4Cs_3$	0.0026	1.064	-
$Cs_6InAs_4K_3$	0.0029	1.065	0.57
$Rb_6TlBi_4K_3$	0.0034	1.045	-
$Cs_6InAs_4Rb_3$	0.0046	1.069	0.38
$K_6InBi_4Rb_3$	0.0055	1.053	-
$Rb_6InBi_4Cs_3$	0.0061	1.063	-
$Cs_6AlSb_4K_3$	0.0067	1.046	0.63
$Rb_6TlSb_4Cs_3$	0.0068	1.072	-
$Rb_6TlAs_4K_3$	0.0068	1.062	0.04
$Cs_6GaP_4Na_3$	0.0071	1.054	0.59

$Cs_6GaAs_4Rb_3$	0.0077	1.062	0.39
$Cs_6TlAs_4Na_3$	0.0082	1.047	0.38
$Rb_6GaBi_4Cs_3$	0.0084	1.058	-
$Rb_6AlP_4Na_3$	0.0085	1.058	0.65
$K_6InBi_4Na_3$	0.0106	1.024	0.02
$Rb_6TlBi_4Na_3$	0.0116	1.018	-
$Rb_6TlSb_4Na_3$	0.0117	1.022	0.53
$K_6GaBi_4Rb_3$	0.0120	1.049	-
$K_6GaAs_4Rb_3$	0.0128	1.065	-
$K_6TlBi_4Rb_3$	0.0128	1.058	-
$Cs_6GaSb_4Na_3$	0.0143	1.028	0.42
$K_6TlBi_4Cs_3$	0.0147	1.073	-
$Cs_6GaP_4K_3$	0.0151	1.059	0.77
$K_6GaP_4Na_3$	0.0155	1.048	0.86
$K_6TlAs_4Rb_3$	0.0157	1.075	-
$Cs_6AlAs_4Na_3$	0.0157	1.057	0.32
$K_6GaSb_4Rb_3$	0.0166	1.054	-
$Rb_6TlSb_4K_3$	0.0166	1.048	0.16
$Cs_6GaSb_4K_3$	0.0181	1.046	0.61
$K_6AlSb_4Na_3$	0.0184	1.027	0.71
$K_6GaAs_4Na_3$	0.0187	1.041	0.52
$Rb_6TlAs_4Cs_3$	0.0193	1.079	-
$K_6TlAs_4Cs_3$	0.0208	1.080	-
$Cs_6InP_4K_3$	0.0232	1.071	0.82
$K_6TlBi_4Na_3$	0.0236	1.025	-
$K_6AlAs_4Rb_3$	0.0237	1.063	-
$Rb_6TlAs_4Na_3$	0.0243	1.044	0.41
$K_6AlAs_4Na_3$	0.0251	1.045	0.53
$K_6AlP_4Na_3$	0.0254	1.052	0.87
$K_6AlSb_4Rb_3$	0.0254	1.051	0.04
$K_6GaSb_4Na_3$	0.0260	1.026	0.67
$Cs_6GaAs_4Na_3$	0.0266	1.046	0.33
$Cs_6InP_4Rb_3$	0.0276	1.076	0.61
$Cs_6AlP_4K_3$	0.0278	1.061	0.75
$K_6TlSb_4Rb_3$	0.0286	1.062	-
$K_6InSb_4Na_3$	0.0311	1.026	0.59
$K_6TlAs_4Na_3$	0.0320	1.042	0.39
$Rb_6AlP_4K_3$	0.0360	1.061	0.47
$Rb_6GaP_4Na_3$	0.0365	1.052	0.69
$Rb_6InP_4Na_3$	0.0383	1.060	0.69

$Rb_6InP_4K_3$	0.0396	1.066	0.34
$K_6TlSb_4Na_3$	0.0397	1.027	0.48
$K_6InAs_4Na_3$	0.0405	1.043	0.43
$Cs_6InP_4Na_3$	0.0418	1.081	0.69
$Cs_6InAs_4Na_3$	0.0419	1.053	0.39
$K_6InP_4Na_3$	0.0423	1.051	0.74
$Cs_6AlP_4Rb_3$	0.0443	1.067	0.60
$Cs_6AlP_4Na_3$	0.0465	1.065	0.57
$Cs_6TlP_4K_3$	0.0471	1.073	0.71
$Rb_6InAs_4Cs_3$	0.0476	1.077	-
$Cs_6GaP_4Rb_3$	0.0486	1.066	0.62

Table B.2: Crystallographic information of the I-V-V-I subfamily.

I-III-V-I	E_{decomp} $(eV/atom)$	c/a	E_{gap} (eV)
$Cs_6BiSb_4Rb_3$	0.0095	1.028	-
$Cs_6BiSb_4K_3$	0.0108	1.022	-
$Cs_6BiAs_4K_3$	0.0228	1.034	-
$Cs_6BiAs_4Rb_3$	0.0249	1.041	-
$Cs_6SbAs_4K_3$	0.0257	1.032	-
$Cs_6AsSb_4Rb_3$	0.0291	1.026	-
$Cs_6BiSb_4Na_3$	0.0363	0.985	-
$Cs_6BiAs_4Na_3$	0.0383	0.998	-
$Cs_6AsP_4K_3$	0.0395	1.020	-
$Cs_6SbAs_4Na_3$	0.0420	0.997	-
$Rb_6SbBi_4K_3$	0.0428	1.030	-
$Rb_6BiSb_4K_3$	0.0428	1.023	-