Materials discovery using chemical heuristics and high-throughput calculations

Daniel William Davies

A thesis submitted for the degree of Doctor of Philosophy

University of Bath
Department of Chemistry

September 2018

Contents

List of Figures

Abbreviations

AI Artificial intelligence

CBM Conduction band minimum

CMR Computational materials repository

CV Cross-validation

DFT Density functional theory

EA Electron affinity

GBR Gradient boosting regression

GGA Generalised gradient approximation

HHI$_R$ Herfindahl-Hirschman index for element resources

HPC High performance computing

ICSD Inorganic crystal structure database

IP Ionisation potential

KS Kohn-Sham

LDA Local density approximation

ML Machine learning

MM Molecular Mechanics

MP Materials Project

PAW Projector augmented wave

PES Potential energy surface

QM Quantum mechanics

RMSE Root-mean-squared-error

SCF Self-consistent field

SSE Solid state energy

VASP Vienna *ab-initio* simulation package

VBM Valence band maximum

XC Exchange-correlation

Declaration of work done in conjunction with others

Introductory subsection **1.4.2 Machine learning** is based on a review article of which I am a co-author: K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Machine learning for molecular and materials science*, Nature, **559**, 547-555 (2018).

Three published papers constitute the main results presented in Chapters 4 – 6 and my personal contribution to each one is described before each publication title page. For ease of identification, a coloured bar (▭) is at the top of each publication page. The supplementary information for each publication can be found in the Appendix.

Acknowledgements

My primary supervisor, Prof. Aron Walsh, has been the best mentor I could have hoped for at the beginning of my postgraduate journey. His approach to research and academia in general is uplifting and refreshing, and his seemingly bottomless supply of enthusiasm is never tiring and always contagious. The level of trust he places in his students can at times be a little terrifying, but always seems to pay off. Above all, I have appreciated having a supervisor who is always willing to champion me and my work to visitors, audiences, colleagues and students.

The rest of the Walsh Group members, both past and present, are some of the kindest and cleverest people I have ever had the pleasure to work with. The group has always been an open and friendly forum in which to share ideas and collaborate. In particular, I am thankful to Dr Keith Butler for day-to-day help and advice as well as Dr Jonathan Skelton for assistance on a wide range of topics, their patience and experience has been invaluable. I have also had many helpful discussions with my co-supervisor, Dr Ben Morgan, who has always been prepared to chat in detail about my work and provide a fresh perspective.

It has been a privilege to be a part of the Centre for Sustainable Chemical Technologies and I have appreciated all the work put in by the directors and support staff to make it a success. I am proud to have been a PhD student in its Centre for Doctoral Training and would, I am sure, have been quite lost at times without the support and comic relief provided by the other students, particularly those in my cohort.

There are many other people who have made the past four years enjoyable and fulfilling, including other computational groups in the Chemistry Department at Bath and collaborators from other institutions. I have been lucky to be able to attend more than my fair share of conferences and workshops at which I have met even more fantastic people. I will always be very thankful that Aron encouraged me to attend so many meetings and for the opportunity to visit so many new places in the process.

The calculations carried out for the work in this thesis were possible thanks to two HPC clusters: The UK national system, Archer, and the University of Bath system, Balena. Access to Archer was provided *via* the UK Materials Chemistry Consortium and Balena is maintained by the University of Bath Computing Services.

Lastly, and perhaps most importantly, I am very grateful to my family. My parents and brother have always given me unconditional support and encouragement towards all my endeavours; my partner, Sophie, always manages to convince me of the value of my work when I often cannot see it myself; and my extended family always take a genuine interest in my activities and have also provided 26 years worth of guidance and support. They are all a constant source of inspiration to me.

Abstract

Advances in computational power, first-principles techniques and data-driven methods mean that we live in a world where computational materials design is fast becoming a reality. So far, the composition space for new materials has barely been explored and there is no established protocol for systematically screening such a space. This is a grand challenge that can be approached in many ways, and the work in this thesis explores one such avenue.

Herein, tools are presented for quantifying the search space for inorganic materials. By restricting the search to stoichiometric compounds, and by limiting the stoichiometry of each element to a maximum value, the space becomes finite for binary, ternary and quaternary element combinations. Hierarchical workflows are then used to target specific materials properties such as thermodynamic stability and bandgap. The workflows consist of modular screening steps that are also developed within this thesis, and are based on a mixture of heuristic chemical rules and data-driven approaches. The classic chemical heuristics used include electronegativity and oxidation state, along with more recently developed metrics such as the solid state energy scale. The data-driven screening steps include a machine learning model that predicts bandgap from chemical composition, a probabilistic model to predict likely oxidation state combinations, and a previously reported ionic substitution model that assigns crystal structures to compositions. Finally, these workflows are applied to the search spaces of metal chalcohalides and quaternary metal oxides, with top candidates identified and further characterised using high-throughput first-principles calculations.

Part I

Introduction

Chapter 1

Computational Materials Design

1.1 The role of the computer

1.1.1 General impact

The modern computer is the perfect example of a technology whose overwhelming and comprehensive impact on society could never have been predicted. Retrospectively, this is hardly surprising given how rapidly they have changed since their invention, up to the present day. The first computers, capable of executing calculations and storing information to memory, consisted of thousands of vacuum tubes and occupied large rooms. The fact that we now carry in our pockets and on our wrists far more powerful versions of these early machines is indicative of how our society has been built around them. Personal computers and the world wide web have been the key stepping stones towards our digital, interconnected world, and it is now impossible to imagine how life would be if this technology had not been developed.

The fundamental way in which we use computers has also changed. Born out of a need to crunch numbers – to tabulate a national census or crack a code – computers have long served as "dumb calculators", albeit with exponentially increasing performance (Figure 1.1). They have been used for an ever-growing number of applications but have always, behind the graphical interfaces and other forms of interactivity, been passively crunching numbers according to predefined algorithms. More recently, they have begun to take on a more active role thanks to the advent of artificial intelligence (AI). The exact scope of AI is disputed but can broadly be described as computers using information from their

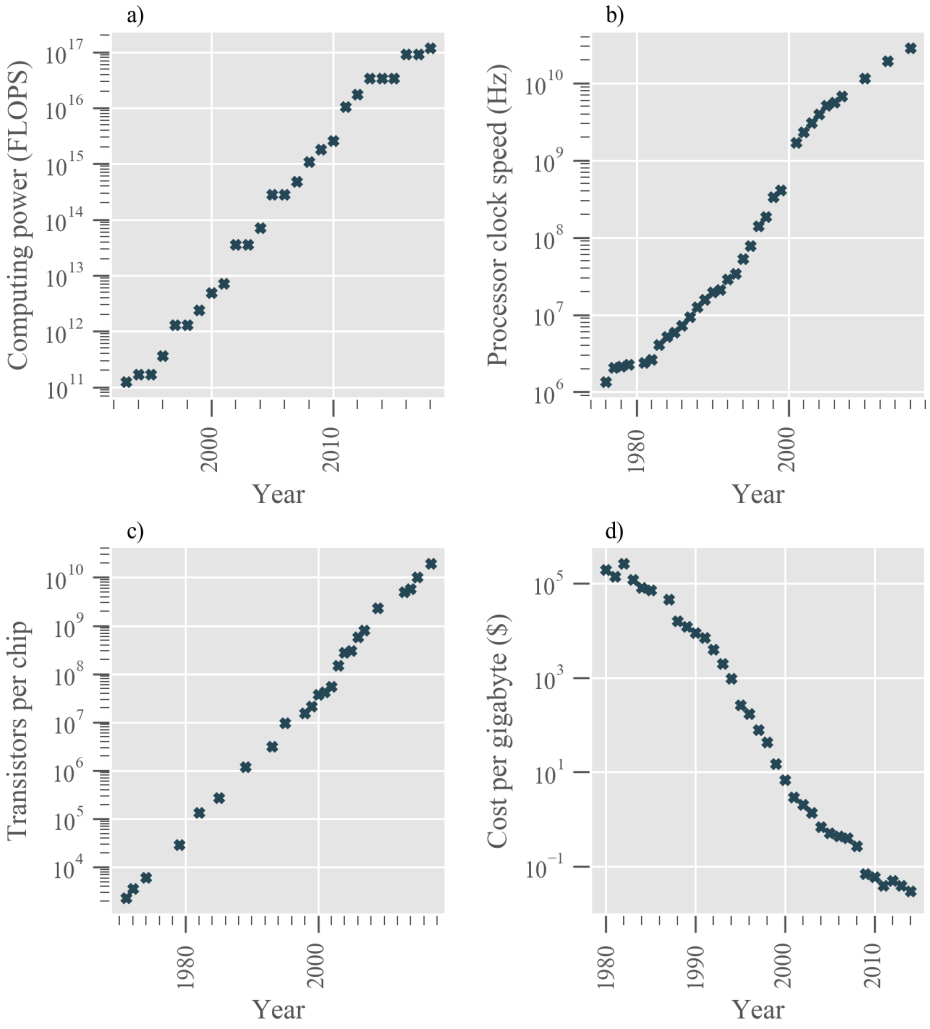

Figure 1.1: Selected computer performance benchmarks over time: a) Number of floating-point operations carried out per second (FLOPS) by the largest supercomputer,[1] b) clock speed of the fastest microprocessors,[2,3] c) largest number of transistors fit into one microprocessor (chip),[4] d) the average cost of one gigabyte of hard disk memory.[5]

environment in order to achieve a goal. This can be seen through the development of language translation, self-driving cars, smart assistants and other emerging technologies. It is possible that we are now at another point in history, similar to the times of the first computers, where the scale of the future impact of AI cannot yet be fully understood.

1.1.2 The four paradigms of science

Computers have played a crucial role in the evolution of science as a whole; in much the same way that they are fully integrated into society, they are deeply embedded into the

way in which scientific progress is made. Furthermore, as is the case in wider society, the relentless increase in computational power along with the advent of AI means that the number of applications is expanding.

An increasing proportion of experimental equipment is now computer controlled and nearly all analysis is carried out using computers. Moreover, calculations (simulations) are now an integral part of the scientific method. Historically, science had always been a purely empirical practice until general laws and models began to be formulated. Examples of the *second paradigm* of science include Newton's laws of motion and the laws of thermodynamics. Once models started to become too complex to be solved analytically, the advent of computers gave rise to the *third paradigm* of science, whereby real-world phenomena could be modelled using the mathematical formulas of the second paradigm.[6]

More recently, calculations are being carried out as part of the third paradigm in such high volumes that they are producing vast amounts of data. Extracting knowledge from large datasets is the realm of informatics and the general area of data-driven discovery has been called the *fourth paradigm* of science (Figure 1.2). In much the same way that the third paradigm uses the laws and models of the second paradigm, the fourth paradigm involves the use of data produced by the third paradigm to make new discoveries. This is done by extracting information using a broad range of techniques, from creating simple plots, to using statistical algorithms that operate in high-dimensional space to reveal hidden trends.

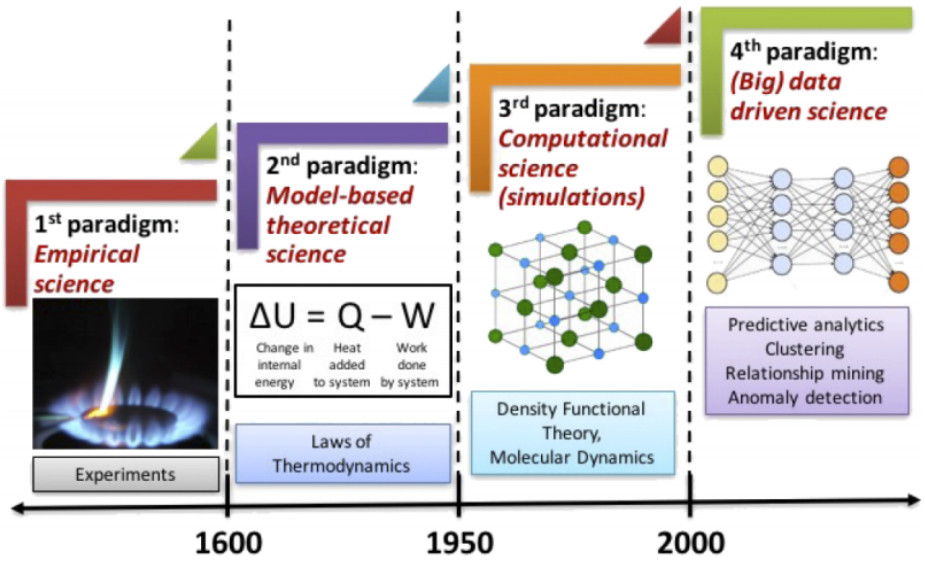

Figure 1.2: The four paradigms of science. (Reproduced from Reference 6 with permission.)

1.2 Modelling materials

1.2.1 Milestones for development

For chemistry, the theory to which computers are applied is quantum theory, thus allowing properties to be calculated from first principles. In the 1920s, the now famous work of Bohr, Planck, De Broglie and Heisenberg had already formalised the idea that the physics at work on the atomic scale could not be described by the equations of classical mechanics. Instead, the wave nature of subatomic particles was embraced and the famous wave equation conceived by Schrödinger was adopted.[7] Solving this equation for a given quantum system yields all the observable quantities *via* its electronic structure. In 1927, quantum theory was applied to the hydrogen molecule, giving rise to the field of quantum chemistry. Even for this simple system, no analytical solution to the wave equation was possible, meaning approximations were needed.[8] Paul Dirac, joint winner of the 1933 Nobel prize in physics along with Schrödinger, summarised:

> *"The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are [...] completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be solved. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation."*[9]

We still cannot exactly solve the equations that arise from quantum theory nearly a century later. However, the development of computational chemistry in general has had a strong focus on making good approximations, and this has facilitated the successful application of computers to innumerable chemical problems of increasing complexity from the 1950s up to the present day.

As was the case for all applications of computers to scientific problems, the initial breakthrough for enabling first-principles calculations was the invention of efficient computers. In 1954, Boys and Price reported first-principles calculations carried out on the systems S, S^-, Cl,and Cl^- using the EDSAC (electronic delay storage automatic calculator) at the University of Cambridge (Figure 1.3).[10] Their comment on the use of this early computer set the tone for the computational chemistry that would follow:

"The present calculation would have been just practicable without these methods, but their use saved considerable labour and they are of considerable importance, since other calculations which would otherwise be prohibitively laborious will be quite feasible by their use."

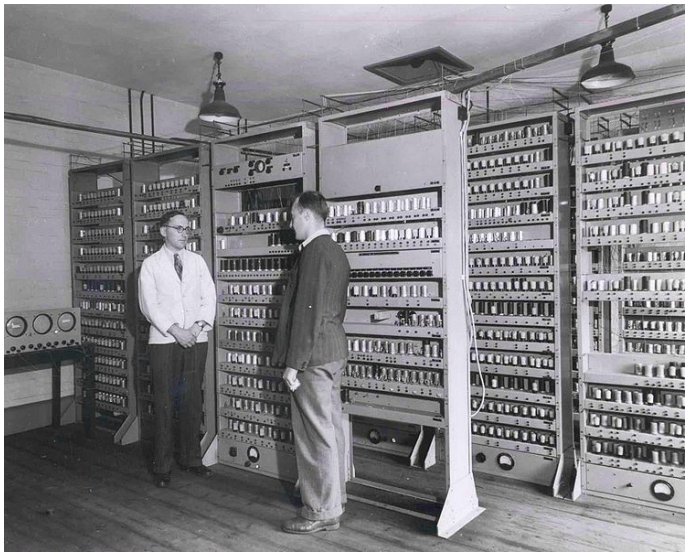

Figure 1.3: The electronic delay storage automatic calculator (EDSAC) at the University of Cambridge featuring Maurice Wilkes (left), head of the Mathematical Laboratory, and Bill Renwick (right), chief engineer of the EDSAC. (Copyright Computer Laboratory, University of Cambridge. Reproduced with permission.)

At the same time, similar progress was being made in the US and the first *ab initio* (from first principles) calculations on diatomic molecules were performed in 1956 at the Massachusetts Institute of Technology.[11]

During the 1960s computers became more user friendly, meaning it was no longer only specialist engineers that could operate them. Many research groups were suddenly able to perform the calculations necessary to apply quantum theory to real chemical problems and to real materials. The desire to exchange software freely and easily in a pre-internet age gave rise to the Quantum Chemistry Program Exchange service. This service was used heavily by theoretical chemists, including students, as well as experimentalists, and assisted in the proliferation of computational chemistry as a mainstream research activity.[12] By the 1970s, John Pople had created the `Gaussian 70` code that could predict the behaviour of molecules of modest size.[13] Also in the 1970s, other efficient computer programs such as `ATMOL`, `IBMOL`, and `POLYAYTOM`, became popular for carrying out first-principles calculations of molecular orbitals, and these software packages led to an increase in applications of computers to chemical problems.

By far the most significant theoretical development for increasing the power of electronic structure calculations came in 1965 in the form of density functional theory (DFT).[14] By making use of the overall electron density as opposed to considering the interactions between individual electrons, the complexity of the subatomic picture considered for a given system is reduced enormously. DFT could provide good enough approximations to the properties of molecules and solids of a much more practical size and continues to strike a harmonious balance between accuracy and computational cost. It has been the workhorse theory for carrying out electronic structure calculations for several decades.[15]

Various methods beyond DFT have also been developed over the past few decades. Hybrid DFT is an important extension to DFT, in which some of the subatomic complexity that is removed by DFT in the first place, is reintroduced. This improves some of the results obtained using standard DFT considerably, depending on the system and properties being calculated, but often requires orders of magnitude more computational power. Modern computing capabilities have made these calculations practically affordable and they have now become fairly routine. Some other approaches that are more expensive than DFT are are under intense development. These include many body perturbation theory in the GW approximation,[16] which is better suited to describing excited state properties due to the fact that interactions between electrons are treated explicitly, as well as time-dependent DFT (TDDFT),[17] in which a time dependent external potential is introduced, such as an electric field.

1.2.2 Calculating properties

Equipped with approaches to simulate fundamental physical laws on computers, we can calculate technologically relevant materials properties. Some calculable properties relate directly to total energies of systems, for example electrode potentials of battery materials can be determined from total energy differences.[18] Additionally, calculations of energy differences between polymorphs of crystalline materials are routinely carried out to establish the most stable structure. Furthermore, mechanical properties including elastic constants as well as tensile and shear strength can also be calculated, and are indispensable in the modelling of high-strength alloys, among other applications.[19,20]

Electronic structure calculations provide a direct link between fundamental physics and optoelectronic properties, and there are many groups applying such techniques to determine light-absorption properties, charge carrier mobilities, ionisation potentials and more. Results from such calculations can have a direct impact on the improvement of elec-

tronic devices, including light-emitting diodes, solar cells and computer components.[21] The plethora of materials properties that can be calculated computationally using first-principles techniques are too numerous to list here, but include spectroscopic and dielectric properties, magnetic properties, surface and interface effects, vibrational and thermal properties, and catalytic activity. There are many reviews that can provide a full overview of current capabilities, including References 22 and 23.

Besides the advances in quantum mechanics methods, modelling based on solving equations of classical mechanics has also been very successful. The use of computers means that for systems of many thousands or millions of atoms we can solve for sets of interatomic potentials. Early examples included modelling hundreds of hard spheres in order to simulate liquid-solid phase transitions.[24] Nowadays, simulations of billions of atoms are possible using interatomic potentials, which in principle represent materials on the micron length scale.[25]

A multitude of important materials properties can be calculated using classical models including mechanical properties such as yield strength and elastic moduli,[18,26] as well as heat and ion conduction.[27,28] While it has only become possible more recently to model non-perfect representations of materials using first-principles techniques, classical mechanics has been used to model features of real materials for many decades. Features such as defects, surfaces and interfaces can be modelled,[29–31] as well as chemical processes that occur at those features.[32] Indeed, classical approaches continue to prove themselves as extremely useful techniques, wherever quantum effects are negligible.

Finally, the use of quantum and classical mechanics is not always mutually exclusive, and increasingly multi-scale modelling is becoming popular. Multi-scale modelling involves a small region of interest being modelled at the quantum level, while the surrounding atoms in the material are modelled classically. In reality, this description is an oversimplification and the material is usually divided into more sub-levels of sophistication, as exemplified in Figure 1.4. Mechanical properties of high strength steels and catalytic metal surface reactions are some of the example areas that benefit from multi-scale techniques.[33,34]

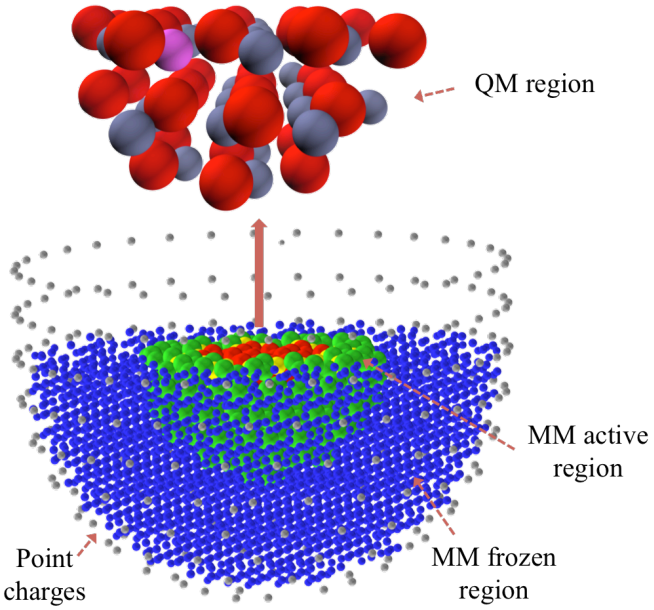

Figure 1.4: Schematic of a multi-scale modelling simulation in which the inner region is treated with quantum mechanics (QM) and surrounding layers give the impression of a true material: A layer of free particles treated with molecular mechanics (MM), then by a layer of fixed particles, and finally point charges.

1.2.3 Current capabilities

1.2.3.1 Calculation reliability

Once DFT and extensions thereof had proven their value, the development of implementations that can deliver the most accurate results in reasonable time frames became an important area of research. We now have a plethora of techniques to obtain approximate solutions using the DFT method at our fingertips, and these are implemented in many different electronic structure codes.[35] Trust in the results produced by these codes is also growing within the wider scientific community, as they are increasingly reproducible and in close agreement with experiment.[36]

There are still some areas where DFT falls short,[37] which is one driver for the continued development of other methods. In general, however, key electronic, optical and mechanical properties of molecules and materials can now be calculated to a high degree of accuracy. Crucially, this is increasingly being done *before* synthesis, such that computational approaches are now being used predictively, as opposed to explaining experimental observations *post hoc*.

1.2.3.2 Large scale calculations

It is not only the factors illustrated in Figure 1.1 that have improved computational capabilities. Modern high-performance computing (HPC) architectures allow for a high level of parallelism with fast communication between individual nodes that are each performing independent calculations. Thus, if an electronic structure code is configured for parallelism, first-principles calculations on very large systems are now achievable compared to the simple atomic or diatomic calculations of the 1950s. Examples include the modelling of over 13,000 atoms in protein nanofibrils[38] and the first-principles molecular dynamics simulation of over 30,000 Si atoms,[39] both using 1024 compute nodes.

On the other hand, it is also possible to perform electronic structure calculations on many individual systems in high throughput. Recently published packages that wrap around electronic structure codes to enable the automation of calculations can considerably reduce the researcher time and effort taken to perform up to tens of thousands of calculations with a consistent set of input parameters,[40–43] which is an important factor for the direct comparison of results. Within the field of inorganic materials alone there are many examples of high-throughput approaches successfully being used to assess stability,[44] compare optical[45–48] and thermal properties,[49,50] calculate dielectric[51,52] and piezoelectric[53] properties, and assess suitability for specific applications such as battery cathode materials.[54]

1.2.4 Outlook

We have reached the point where highly accurate property calculations can be carried out quickly from first principles *in silico*. Thousands of hypothetical compounds can be examined before they undergo the often time-intensive process of being synthesised in the laboratory. This clearly has the potential to impact the discovery of new materials in various ways, and takes us one step close to the goal of being able to truly *design* materials, whereby the paradigm of *input: composition and structure, output: properties* is reversed. We will now explore the role that computers specifically have in the discovery of new materials.

1.3 Materials discovery

1.3.1 The demand for new inorganic materials

The critical role that new materials play in technological advancement cannot be overstated. These advancements have had such vital societal impacts that periods of history are named after the key material of the time. The Bronze Age saw the first examples of deliberate alloying to improve materials properties, and most recently silicon chips replaced vacuum tubes in the Silicon Age. It was recently estimated that of all progress made in computation over the last 40 years, two thirds could be directly attributed to contributions from materials innovation.[55] The same study also found that the relative contribution of materials innovation to overall technological development has been increasing, decade on decade.

It is worth noting that in this context, the *materials* of interest are those which enable a specific function *via* their mechanical, electrical, magnetic or optical properties, or some combination thereof. These properties can depend on many variables but it is well established that they are intimately related to chemical composition and crystal structure. It is often new compounds, therefore, that form the basis for new materials discovery. For example, efficient blue LEDs were fabricated two years after the first growth of single crystalline GaN[56,57] and previously unexplored crystal phases have enabled recent advances in thermoelectric materials.[58] The work in this thesis specifically deals with the application area of solar energy conversion and finding new semiconducting materials for this purpose.

1.3.2 The Edisonian approach

Although the need for new materials is well established, there is typically a lag of around two decades between the invention of a material and its widespread commercial uptake in new technologies, as shown in Figure 1.5. Given that new materials may hold the key to tackling some of the greatest challenges of our time such as climate change and human welfare, there is an immediate need to accelerate all stages of the materials delivery process, from initial discovery to widespread use. The first of these steps – initial discovery – has traditionally taken place *via* a trial-and-error (or Edisonian) method. This approach is intrinsically slow, and heavily reliant on serendipitous discovery. As such, many of the materials that are ubiquitous today were made by accident in the pursuit of some com-

pletely unrelated goal. One example is the non-stick coating Teflon®, which was found on (and subsequently hard to remove from) the inside of a gas cylinder in a laboratory developing fluorinated refrigerant molecules.[59] More systematic approaches are needed in order to accelerate materials discovery. Two critical tools that can be incorporated into such approaches are chemical heuristics and materials informatics.

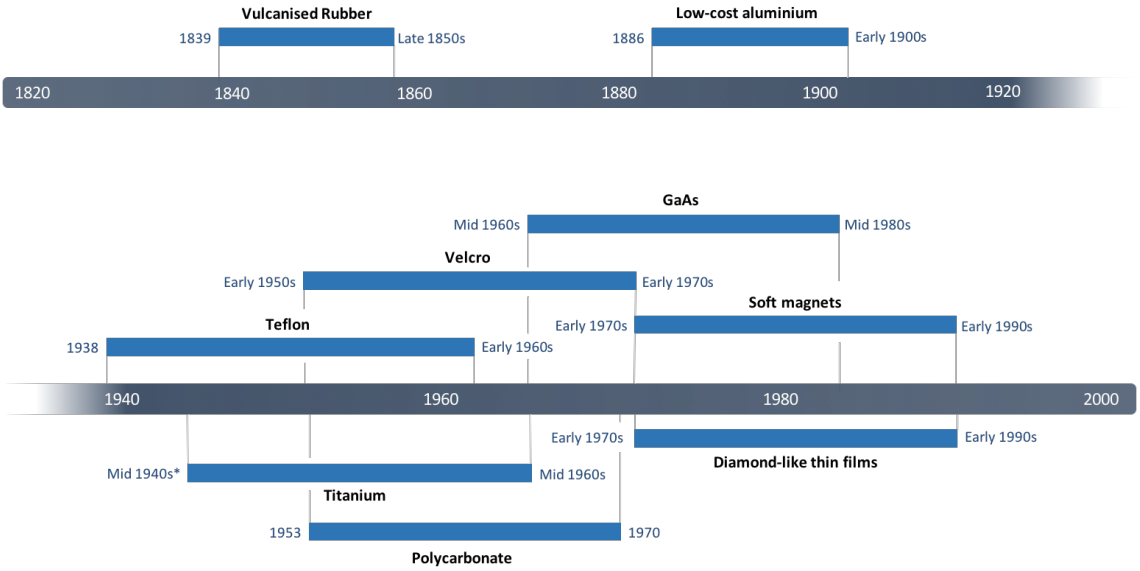

Figure 1.5: The time between the invention and widespread commercialisation of selected materials. * Refers to low-cost production of pure Ti for use in aerospace applications. (Data from Reference 60).

1.3.3 Chemical heuristics

The trial and error discovery process, however inefficient, has not only provided us with all the materials we use today, but crucially it has resulted in a wealth of knowledge that has built up over time. This knowledge can guide further materials design in the form of heuristic rules. For example, radius ratio rules[61] have long been used to predict the propensity of a ternary composition to form the perovskite structure. These rules are still useful today and have recently been extended for applications to hybrid organic-inorganic perovskites.[62,63] There are many other examples that relate various compositional and structural features to particular properties. In the following, only those that are of relevance to this work are outlined.

Of particular importance are rules that can be used to estimate electronic properties based on chemical composition alone, i.e. without crystal structure. Specifically, the prediction of the positions of the valence band maximum (VBM) and the conduction band minimum

(CBM) of semiconducting materials, on an energy scale relative to the vacuum level, which correspond to the highest filled electronic states and the lowest unfilled electronic states, respectively. The difference between these two values – the bandgap – is also highly important for solar energy applications, as this dictates the wavelength of light that will be absorbed by the material.

One such rule comes from the electronegativity scale of elements. Alongside the widely used Pauling electronegativity scale based on bond dissociation energies,[64] Robert Mulliken provided an absolute scale of electronegativity,[65] defined as:

$$M = \frac{|I + E_{EA}|}{2} \tag{1.1}$$

with I the first ionisation energy and E_{EA} the electron affinity. In 1974, Nethercot outlined[66] how this can be extended to compounds by taking the geometric mean of the Mulliken potentials (electronegativities) of all the constituent elements:

$$M_{compound} = (M_A M_B ... M_n)^{1/n} \tag{1.2}$$

This physically represents the mid-gap energy between the VBM and CBM in a solid. Nethercot demonstrated the surprisingly accurate predictive power of this method by testing it on a variety materials with widely accepted bandgaps and ionisation potentials. Since then it has been used to produce results in good quantitative agreement with first-principles methods.[67] Pauling electronegativities can be substituted into Equation 1.2 if a scaling coefficient is used.

Another useful resource is the work of Harrison[68] in which he outlines several techniques for estimating the properties of solids based on electronic structure. He arrives at an equation for determining the bandgap of binary semiconductors based on tabulated s- and p-state eigenvalues of the constituent atoms, determined from approximate electronic structure calculations:

$$E_g \approx 3.60(V_2^2 + V_3^2)^{1/2}(1 - \alpha_m) \tag{1.3}$$

where V_2, V_3 and α_m are terms related to the covalent, polar and metallic bonding energies respectively and are simply calculated from bonding distances and the s- and p-state

eigenvalues as tabulated in the same book.

Combining the Nethercot and Harrison methods gives access to approximate CBM and VBM positions relative to the vacuum level. A different approach to estimate these values directly was recently proposed by Pelatt *et al.*,[69] and is called the solid state energy (SSE) scale. For this scale, the ionisation potentials and electron affinities of 69 binary semiconductors containing 40 different elements were collected. The SSE scale is obtained by assessing an average EA for a cation (empty electronic states) or an average IP for an anion (filled electronic states) for each element by using data from compounds having that specific element as a constituent. Thus, this method provides estimates of the absolute VBM and CBM positions directly and the the bandgap can be estimated simply:

$$Eg \approx SSE^{cation} - SSE^{anion} \tag{1.4}$$

Although derived in a different manner to Mulliken and Pauling electronegativities, it is clear from Figure 1.6 that the SSE scale fits into this family of predictive tools as the periodic trends of electronegativity are captured. The estimations of energy levels that can be made using the above methods are summarised in Figure 1.7.

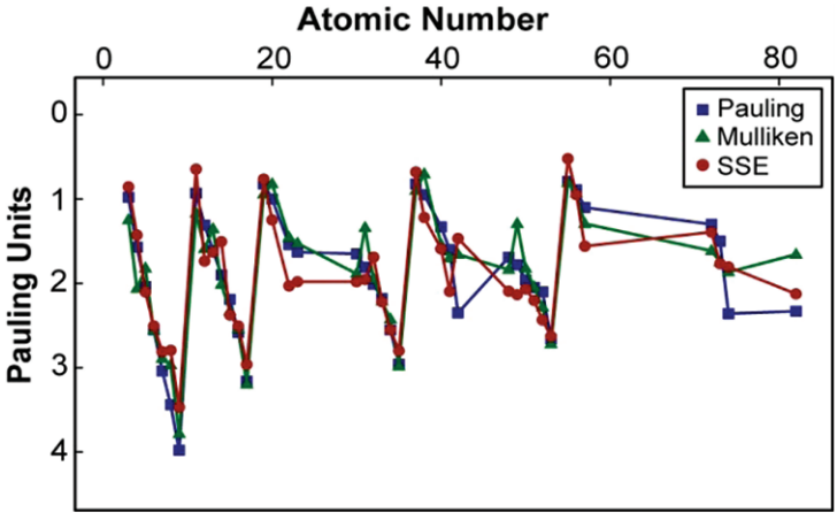

Figure 1.6: Comparison of the SSE scale with Pauling and Mulliken electronegativity scales. Reproduced with permission from Reference 69.

Finally, another key consideration for materials design is environmental sustainability and there are various metrics that can be accounted for such as the crustal abundance, cost and toxicity of the constituent elements. Other heuristics include the Herfindahl-Hirschman Index for element resources (HHI_R) which has recently been developed in the

context of thermoelectric applications.[70] This index includes factors such as the geopolitical influence over the supply and price of elements, thereby giving a more complete picture of sustainability than crustal abundance alone.

All of the heuristics outlined above can be used to link chemical composition to a specific property of interest *via* trivially simple calculations.

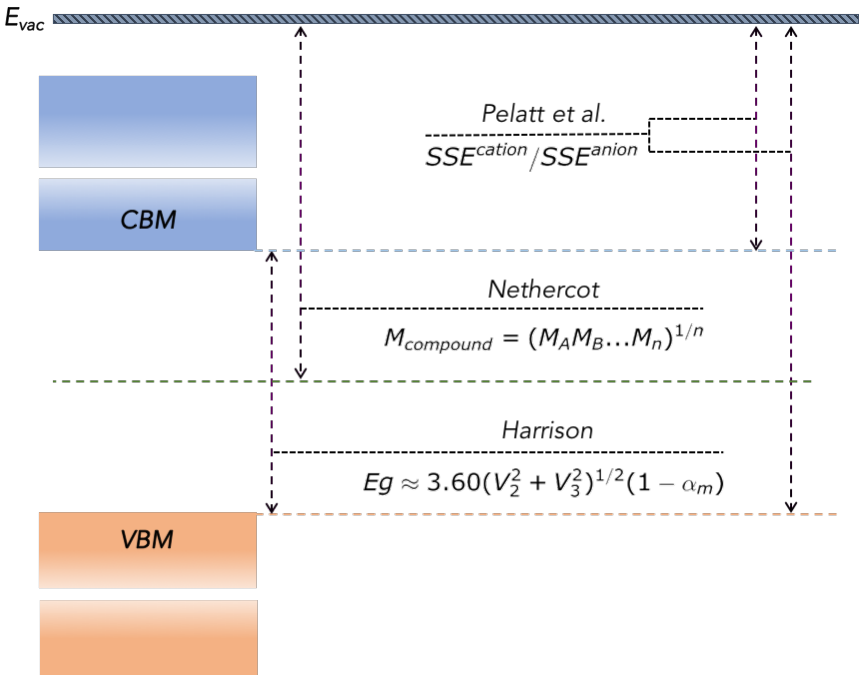

Figure 1.7: Schematic summary of the approaches of Harrison,[68] Nethercot[66] and Pelatt *et al.*[69] for estimating the bandgap, mid-gap energy and conduction/valence band extrema in solids, respectively. These calculations can be done with knowledge of chemical composition only, i.e. without explicitly considering structure.

1.4 Materials informatics: the fourth paradigm

1.4.1 Materials data

Carrying out high-throughput calculations on real and hypothetical materials produces large quantities of potentially useful data. An ultimate goal for computational materials design is to be able to use this data to confidently predict properties of new materials. This information can be applied in a similar way in which chemical knowledge is employed in the form of heuristic rules, avoiding the need to carry out many first-principles calculations.

The availability of open-access databases that cover the calculated properties of known and hypothetical materials has been pivotal for enabling data-driven materials discovery. Historically, outputs from calculations were available only *via* individual journal articles or at best *via* printed data tables. Since then, as fast internet access has developed, more and more on-line resources have emerged. Nowadays, there is a strong move towards open science and open data, whereby enough information should be supplied alongside published work to make the results entirely repeatable. With this trend as a backdrop, the number of useful databases is growing (Table 1.1).

The utility of databases of calculated materials properties has become apparent in recent years. They have been used to aid with further computational studies,[72,73] establish reference points against which to compare new results,[74–76] rapidly verify that new compounds have not been previously reported,[77] estimate thermodynamic stability by comparing total energies of competing phases,[78–84] and perform further screening for specific properties.[85,86] Furthermore, partly thanks to the increased reliability of calculated properties, there are many examples of the experimental community using these datasets in order to validate results,[87,88] establish decomposition pathways,[81,89] and to explain experimental observations in terms of electronic structure.[90]

Finally, in order for databases to be truly useful, they must not only contain a large number of compounds and associated properties, but must be well organised and easy to access:

Organisation: The rapid rise of big data has meant that effective organisation in databases is a challenge faced by many fields of research and industries. One issue for first-principles calculations is that there is no universally agreed data format. In fact the situation is quite the opposite, with every electronic structure code producing output files with a wide variety of different schema. This has been a particular challenge for projects such as NoMaD (Table 1.1), which contains input and output files from 24 different electronic structure codes. The NoMaD database currently contains $>$ 40 million entries for bulk crystal structures that have associated properties, which equates to $>$ 800 million individual data points. These data points are classified using 2,360 code-specific metadata labels, which makes parsing the data into a homogeneous format a daunting task, and only once this has been carried out could the dataset be stored in a code-agnostic format. This exercise has led to suggestions of a common data format,[91] but competing schema are being developed simultaneously.[92]

Table 1.1: Publicly accessible structure and calculated property databases for molecules and solids. Reproduced from Reference 71.

Name	Description	URL
AFLOWLIB	Structure and property repository from high-throughput ab initio calculations of inorganic materials	http://aflowlib.org
Citrination	Computed and experimental properties of materials	http://citrination.com
Computational Materials Repository	Infrastructure to enable collection, storage, retrieval and analysis of data from electronic-structure codes	http://cmr.fysik.dtu/dk
GDB	Databases of hypothetical small organic molecules	http://gdb.unibe.ch/downloads
Harvard Clean Energy Project	Computed properties of candidate organic solar absorber materials	http://cepdb.molecularspace.org
Materials Project	Computed properties of known and hypothetical materials carried out using a standard calculation scheme	http://materialsproject.org
NoMaD	Input and output files from calculations using a wide variety of electronic structure codes	http://nomad-repository.eu
Open Quantum Materials Database (OQMD)	Computed properties of mostly hypothetical structures carried out using a standard calculation scheme	http://oqmd.org
NREL Materials Database	Computed properties of materials for renewable-energy applications	http://materials.nrel.gov
TEDesignLab	Experimental and computed properties to aid the design of new thermoelectric materials	http://tedesignlab.org
ZINC	Commercially available organic molecules in 2D and 3D formats	http://zinc15.docking.org

Easy access: Most databases are easily accessible *via* web browser interfaces that display the materials properties of interest and provide a direct download to the crystal structure in a common format, e.g. CIF file. In order to access large chunks of data simultaneously, some databases (e.g. The OQMD) can be downloaded in their entirety. A more flexible way to expose data is *via* an application program interface (API) over the web. Representational state transfer (RESTful) APIs are common protocols for interacting with data resources and ensure that the data being accessed is up-to-date, that only the needed portion of data is downloaded, and can provide the database owner with analytics such as access statistics. AFLOWLIB and the Materials Project were among the first databases of calculated materials properties to use RESTful API.[93,94] The power of exposing data *via* an API over the web lies in the ability to interact with databases from within programs written in common languages such as Python: up-to-date data can be accessed as part of automated routines and workflows.

1.4.2 Machine learning

Mirroring the situation in wider society, computers are beginning to take on new roles within materials design thanks to the field of AI. In particular, machine learning (ML) – a subfield of AI concerned with the automated building of statistical algorithms whose performance improves with training – is being applied in a wide variety ways.

In essence, ML approaches learn rules that underlie a given dataset by assessing a portion of that data and building a model to make predictions. The training of a ML model can either involve supervised learning, where the goal of the algorithm is to derive a function that predicts a particular output value given a set of input values, or unsupervised learning, where no output values are targeted and the goal is to identify trends in the dataset. There is also semi-supervised learning where a limited number of output values are available. A wide range of model types (learners) exist, and the choice of learner depends on whether output values are continuous or discrete, whether regression or classification is the goal, as well as the size and diversity of the training dataset. Once a learner is chosen, trial models are evaluated in the training phase and the best one is selected. The key test for the accuracy of a ML model is successful application to unseen data. Withholding some data to test the accuracy of the model is a common way to assess the accuracy of the model after the training phase.

Application areas often involve complex problems that are ill-suited to traditional algorithmic approaches. As such, they can benefit most from the analytical models that lie

at the heart of ML approaches. The application of ML to the following areas have the potential to directly accelerate the materials discovery process.

Guiding chemical synthesis: The number of possible transformations at each step of a synthetic route can range from around 80 to several thousand.[95] Additionally, competing objective functions (such as cost, purity, time and toxicity) make synthetic chemistry a challenging field for algorithmic approaches. There have been examples of ML algorithms successfully learning contextual rules from literature examples of chemical syntheses. In one case, it was shown that computers can be more efficient than humans at extracting these rules,[96] and in another, it was shown that trained chemists could not distinguish between the syntheses proposed by human experts and those proposed by the resulting ML models.[95] Other ML approaches have been used to predict reaction conditions,[97] and the propensity of a given molecule to crystallise.[98] In both cases, each model was able to make such accurate predictions because they were trained using both positive an negative results, i.e. molecules that do not crystallise as well as those that do, and results from failed reaction attempts, as well as those that succeeded.

Enhancing theoretical chemistry: The considerable effort that is devoted to finding approximate solutions to the Schrödinger equation has already been mentioned. Improvements to DFT approaches are made by developing new approximations to the exchange-correlation functional that describes non-classical interactions between electrons, and whose exact form is unknown. Learning accurate exchange-correlation functionals from structure-property databases is a new area of research that has already produced some promising results.[99,100] A further example where ML approaches are embedded in the DFT approach itself involves bypassing the expensive equations that link the electron density to the system energy (and all other properties), and instead learning this mapping directly from training systems.[101]

Targeting discovery of new compounds: Of particular relevance to this work is the use of ML approaches to perform target compound searches. Within materials chemistry, the number of examples of ML approaches to design new compounds has risen since 2010.[102,103] These have included using composition-based descriptors to predict the likelihood that a given composition will adopt a particular crystal structure,[104,105] linking electronic band structure features to performance as a photocathode material,[106] and using inexpensive energy estimations to predict DFT total energies (and therefore thermodynamic stabilities) of over 2 million hypothetical crystal structures.[107]

1.5 Beyond existing materials

Heuristic rules are one set of tools for making predictions about the properties of hypothetical compounds. Large datasets of calculated properties of existing and hypothetical compounds are being generated by high-throughput calculations, and these facilitate a new set of tools, which are data-driven approaches like ML. There is no single approach for exploring the composition space of hypothetical compounds in a systematic way in order to apply these tools. Current approaches for methodical generation of hypothetical compounds usually fall into one of two broad categories: 1) search by analogy and 2) crystal structure prediction.

1.5.1 Search by analogy

When searching for new compounds by analogy, the prototype crystal structure is kept fixed and different possible combinations of elements are substituted onto the lattice sites. The substitutions can be a simple isovalent substitution, e.g. the replacement of Zn(II) with Cd(II), or can be extended to aliovalent cross-substitution (also termed cation mutation), e.g. the replacement of two Ga(III) with Zn(II) and Ge(IV). The process of cross-substitution while keeping the valence electron : atom ratio constant was demonstrated by Goodman in the 1950s,[108] who predicted a series of new semiconducting compounds. A systematic implementation of this concept was applied by Pamplin in the 1960s[109] to predicting tetrahedral semiconductors in an exhaustive manner. He derived formulae based on heuristic rules of elemental valency to tabulate a large range of known compounds as well as many previously unknown ternary and quaternary structures.

Aliovalent cross-substitution is used as a strategy to tune properties of semiconductors due to the enhanced chemical and structural freedom that comes from going from two components to three or four. For example, the bandgap of chalcogenides can be sequentially decreased going from binary $II-VI$ to ternary $I-III-VI_2$ to quaternary $I_2-II-IV-VI_4$ structures.[110] Similarly, cross-substitution has been used to design quaternary nitrides based on a parent GaN structure.[78] The resulting compounds were found to be lattice matched to GaN/ZnO and have VBM energies as low as that of ZnO and CBM energies as high as that of GaN, thereby combining two desirable properties of both materials in a single material. The concept of valence has also been successfully applied to predict new intermetallic compounds. For example, Gautier *et al.* investigated 400 “missing” precious metal-containing ternary compounds and predicted that 54 of them should be

stable semiconductors or semimetals, according to first-principles calculations.[111] Similar approaches have also been used for the successful identification of transparent conducting materials.[112]

Hautier *et al.* have devised a data-driven approach to assign likely structures to chemical compositions based on the likelihood of a particular ionic species substituting for another in the same structure type.[113] Underlying this process is the principle of search by analogy, but rather than fixing the crystal structure and making many substitutions of different compositions, the composition is kept fixed and substitutions are systematically trialled on many known crystal structure type.

Historically, elements that have been considered 'similar' (close in size and charge) have been chosen for site substitution, largely based on intuition. In this approach, the similarity factor is quantified based on data-mined information from a database of known structures and the probability of ion substitution carries a numerical value. This model can therefore be used to produce a list of structures that a combination of elements is likely to adopt, ranked in order of probability. Using structures within the inorganic crystal structure database (ICSD), the Hautier *et al.* showed *via* leave-some-out cross-validation that a suitable probability threshold can be chosen in order for the technique to be used predictively. This approach is used multiple times for the work in this thesis and is described more fully in Chapter 3.

1.5.2 Crystal structure prediction

In crystal structure prediction, the element composition is kept fixed and various arrangements of those elements in space are considered. Usually, the configuration with the lowest energy is the target of the search as this corresponds to the most thermodynamically stable. However, there is now a growing appreciation of the fact that a large number of technologically relevant materials are metastable, i.e. do not comprise the ground state atomic configuration.[114,115] It is currently not clear how high in energy above the ground state a compound can be and still be considered a viable synthesis candidate. The fact remains that to ignore all atomic configurations other than the ground state is to ignore potentially useful materials.

The first step of most first-principles calculations workflows is to take a solved crystal structure, usually from experiment, and relax it locally (minimise the forces acting upon each atom) to a given level of theory. Predicting ground-state structures from chemical

compositions alone is much more challenging as global rather than local optimisation is needed. For a system of N atoms per unit cell, there are $3N - 3$ degrees of freedom associated with the atomic positions, plus a further 6 associated with the lattice parameters. Even for small systems ($N \approx 10 - 20$) the number of possible atomic configurations becomes too large for exhaustive sampling and stochastic approaches to explore the relevant areas of configurational space are used in practice, often in conjunction with local optimisation. These approaches include genetic algorithms, evolutionary algorithms, Monte-Carlo sampling, particle-swarm and minima-hopping methods.[22,116]

Several global optimisation codes have been developed[117–120] and a great deal of effort has gone into optimising the algorithms that are implemented within them. However, they rely on a large number of individual DFT calculations being carried out and, although DFT is an efficient use of computer resources for quantum mechanical calculations, this amounts to an intrinsically expensive exercise.[121]

1.5.3 Other search methods

Some novel approaches to generating hypothetical compounds have been reported that are distinct from either of the above categories. For example, Dyer *et al.* have shown how 2D layers of known crystal structures can be combined together to suggest entirely new compounds that exhibit sensible chemical environments.[122] For a particular set of layers, the many thousands of resulting permutations can then be ranked in energy cheaply using classical mechanics, and subsequently using first-principles calculations. This approach was exemplified by the identification and experimental realisation of a new mixed-metal oxide with 148 atoms per unit cell.

Another example is an exercise carried out by Friedrichs *et al.*, who enumerated the number of distinct network topologies for crystalline solids in which the coordination number of the atoms is four.[123] Their results are therefore directly applicable to the design of novel zeolites, silicates, carbon networks, as well as a wealth of tetrahedral inorganic compounds. They find that, even for this relatively simple network type, the number of different possibilities approaches 1,000 when three different atoms are considered. In a similar vein but for structures of more complex formula, the `GRINSP` code allows for the exploration of 3- to 6-connected networks, which can be used to represent hypothetical compounds with corner sharing polyhedra.[124]

The above approaches share a common theme of combinatorially exploring a search space,

within some particular constraints. While the concept of enumerating possible compounds in some way is not new – as we have seen, Pamplin had begun thinking along these lines in the 1960s – the increase in modern computer power now means that larger and larger search spaces can be considered.

1.5.4 The combinatorial perspective

Databases such as the ICSD ($\sim$ 181,000 entries) and the Materials Project ($\sim$ 84,000 entries) can give a rough indication of how many inorganic materials have been experimentally reported. Given that roughly half of the Materials Project entries can be associated with an entry in the ICSD, and that around 80% of the ICSD is formed of non-duplicate entries, the true answer should lie somewhere within these two extremes: 42,500 - 145,000. Given also that $\sim$ 40% of the structures in the ICSD feature partial site occupancy and have therefore not been submitted to the Materials Project, the answer probably lies towards the upper end of that estimate. This raises an interesting question: What proportion of the total composition space for possible new materials does the number of known materials represent? To begin to answer this, a 'bottom-up' approach of combining elements systematically could be adopted. We are restricted to the periodic table of elements as our building blocks for new materials, but they can clearly be combined in a number of different ways to form chemical compounds.

There exists no established protocol to explore the vast composition space for new inorganic materials. While this is a broad and multi-faceted challenge, the focus of this thesis is to explore one particular way in which the search space could be constructed, and develop tools that can be used to filter through such a space sequentially, in order to discover materials with target properties at reasonable computational costs.

In the following chapters, tools are presented for enumerating the search space for stoichiometric compounds. The search space becomes finite for binary, ternary, quaternary etc. element combinations under certain stoichiometric restrictions. Having constructed a compositional search space, the next task is to search through it for target materials.

Initially, we take a more detailed look at oxidation states and establish which combinations of oxidation states are likely to be exhibited in new compounds based on existing materials data. Next, two closely-related screening studies are presented, in which hierarchical workflows are used to target stable materials with bandgaps that would be useful for solar applications. The workflows consist of steps where simple heuristic rules are used

to filter compositions at very low computational cost. In addition, data-driven techniques, including the results from the oxidation states study along with another ML model, are used to screen for candidates with target properties. The data-mined ionic substitution model (Hautier *et al.*[113]) is used to assign crystal structure and in one screening study this approach is compared to a global optimisation algorithm for crystal structure prediction. Finally, high-throughput first-principles calculations are carried out to predict accurate properties of leading candidates. Overall, the aim is to develop systematic approaches to designing new inorganic materials that are computationally affordable and fit in to the new era of data-driven science.

Bibliography

[1] *Top500 The List*, https://www.top500.org/statistics/perfdevel/-[Accessed:02-06-2018].

[2] E. R. Berndt and N. J. Rappaport, *Am. Econ. Rev.*, 2001, **91**, 268–273.

[3] Semiconductor Industry Association, *ITRS 2002 Update*, Semiconductor industry association technical report, 2002.

[4] K. Rupp, *42 years of microprocessor trend data*, 2018, https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/-[Accessed:02-06-2018].

[5] M. Komorowski, *A history of storage cost*, 2014, http://www.mkomo.com/cost-per-gigabyte-update/-[Accessed:03-06-2018].

[6] A. Agrawal and A. Choudhary, *APL Mater.*, 2016, **4**, 053208.

[7] E. Schrödinger, *Phys. Rev.*, 1926, **28**, 1049–1070.

[8] W. Heitler and F. London, *Zeitschrift für Phys.*, 1927, **44**, 455–472.

[9] P. A. M. Dirac, *Proc. R. Soc. London A Math. Phys. Eng. Sci.*, 1929, **123**, 714–733.

[10] S. F. Boys and V. E. Price, *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, 1954, **246**, 451–462.

[11] J. D. Bolcer and R. B. Hermann, in *Rev. Comput. Chem.*, ed. K. B. Lipkowitz and D. B. Boyd, VCH Publishers Inc., 2007, vol. 5, ch. 1, pp. 1–63.

[12] D. B. Boyd, *ACS Symp. Ser.*, 2013, **1122**, 221–273.

[13] J. A. Pople, *Angew. Chemie Int. Ed.*, 1999, **38**, 1894–1902.

[14] P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864.

[15] R. Haunschild, A. Barth and W. Marx, *J. Cheminform.*, 2016, **8**, 52.

[16] M. S. Hybertsen and S. G. Louie, *Phys. Rev. Lett.*, 1985, **55**, 1418–1421.

[17] E. Runge and E. Gross, *Phys. Rev. Lett.*, 1984, **52**, 997–1000.

[18] Y. Han and J. Elliott, *Comput. Mater. Sci.*, 2007, **39**, 315–323.

[19] H. Ikehata *et al.*, *Phys. Rev. B*, 2004, **70**, 174113.

[20] M. Černý, J. Pokluda, M. Šob, M. Friák and P. Šandera, *Phys. Rev. B*, 2003, **67**, 035116.

[21] P. J. Hasnip *et al.*, *Philos. Trans. A. Math. Phys. Eng. Sci.*, 2014, **372**, 20130270.

[22] K. T. Butler, J. M. Frost, J. M. Skelton, K. L. Svane and A. Walsh, *Chem. Soc. Rev.*, 2016, **45**, 6138–6146.

[23] J. Hafner, C. Wolverton and G. Ceder, *MRS Bull.*, 2011, **31**, 659–668.

[24] B. J. Alder and T. E. Wainwright, *J. Chem. Phys.*, 1957, **27**, 1208–1209.

[25] K. Kadau, T. C. Germann and P. S. Lomdahl, *Int. J. Mod. Phys. C*, 2006, **17**, 1755–1761.

[26] J. S. Kallman *et al.*, *Phys. Rev. B*, 1993, **47**, 7705–7709.

[27] W. G. Hoover and C. G. Hoover, *Condens. Matter Phys.*, 2005, **8**, 247–260.

[28] B. J. Morgan and P. A. Madden, *Phys. Rev. B*, 2014, **89**, 054304.

[29] E. Pearson, T. Takai, T. Halicioglu and W. A. Tiller, *J. Cryst. Growth*, 1984, **70**, 33–40.

[30] M. S. Daw and M. I. Baskes, *Phys. Rev. B*, 1984, **29**, 6443.

[31] G. W. Watson, E. T. Kelsey, N. H. de Leeuw, D. J. Harris and S. C. Parker, *J. Chem. Soc. Faraday Trans.*, 1996, **92**, 433–438.

[32] T. X. T. Sayle, S. C. Parker and C. R. A. Catlow, *Surf. Sci.*, 1994, **316**, 329–336.

[33] S. Hao, W. K. Liu, B. Moran, F. Vernerey and G. B. Olson, *Comput. Methods Appl. Mech. Eng.*, 2004, **193**, 1865–1908.

[34] M. Salciccioli, M. Stamatakis, S. Caratzoulas and D. Vlachos, *Chem. Eng. Sci.*, 2011, **66**, 4319–4355.

[35] N. Mardirossian and M. Head-Gordon, *Mol. Phys.*, 2017, **115**, 2315–2372.

[36] K. Lejaeghere *et al.*, *Science*, 2016, **351**, 3000.

[37] A. J. Cohen, P. Mori-Sánchez and W. Yang, *Chem. Rev.*, 2012, **112**, 289–320.

[38] K. A. Wilkinson, N. D. M. Hine and C.-K. Skylaris, *J. Chem. Theory Comput.*, 2014, **10**, 4782–4794.

[39] M. Arita, D. R. Bowler and T. Miyazaki, *J. Chem. Theory Comput.*, 2014, **10**, 5419–5425.

[40] W. Setyawan and S. Curtarolo, *Comput. Mater. Sci.*, 2010, **49**, 299–312.

[41] S. P. Ong *et al.*, *Comput. Mater. Sci.*, 2013, **68**, 314–319.

[42] A. Jain *et al.*, *Concurr. Comput.*, 2015, **27**, 5037–5059.

[43] K. Mathew *et al.*, *Comput. Mater. Sci.*, 2017, **139**, 140–152.

[44] G. Hautier, S. P. Ong, A. Jain, C. J. Moore and G. Ceder, *Phys. Rev. B*, 2012, **85**, 155208.

[45] W. Setyawan, R. M. Gaume, S. Lam, R. S. Feigelson and S. Curtarolo, *ACS Comb. Sci.*, 2011, **13**, 382–390.

[46] L. Yu and A. Zunger, *Phys. Rev. Lett.*, 2012, **108**, 068701.

[47] Y. Wu, P. Lazic, G. Hautier, K. Persson and G. Ceder, *Energy Environ. Sci.*, 2013, **6**, 157–168.

[48] I. E. Castelli *et al.*, *Adv. Energy Mater.*, 2015, **5**, 1400915.

[49] C. Toher *et al.*, *Phys. Rev. B*, 2014, **90**, 174107.

[50] W. Chen *et al.*, *J. Mater. Chem. C*, 2016, **4**, 4414–4426.

[51] K. Yim *et al.*, *NPG Asia Mater.*, 2015, **7**, e190.

[52] I. Petousis *et al.*, *Sci. Data*, 2017, **4**, 160134.

[53] R. Armiento, B. Kozinsky, G. Hautier, M. Fornari and G. Ceder, *Phys. Rev. B - Condens. Matter Mater. Phys.*, 2014, **89**, 134103.

[54] G. Hautier *et al.*, *Chem. Mater.*, 2011, **23**, 3495–3508.

[55] C. L. Magee, *Complexity*, 2012, **18**, 10–25.

[56] H. P. Maruska and J. J. Tietjen, *Appl. Phys. Lett.*, 1969, **15**, 327–329.

[57] J. Pankove, E. Miller, D. Richman and J. Berkeyheiser, *J. Lumin.*, 1971, **4**, 63–66.

[58] G. J. Snyder and E. S. Toberer, *Nat. Mater.*, 2008, **7**, 105–114.

[59] R. J. Plunkett, in *High Perform. Polym. Their Orig. Dev.*, ed. G. S. Kirshenbaum, Springer Netherlands, Dordrecht, 1986, pp. 261–266.

[60] T. Eagar, *Technol. Rev.*, 1995, **98**, 42–49.

[61] V. M. Goldschmidt, *J. Chem. Soc.*, 1937, 655–673.

[62] G. Kieslich, S. Sun and T. Cheetham, *Chem. Sci.*, 2015, **6**, 3430–3433.

[63] W. Travis, E. N. K. Glover, H. Bronstein, D. O. Scanlon and R. G. Palgrave, *Chem. Sci.*, 2016, **7**, 4548–4556.

[64] L. Pauling, *J. Am. Chem. Soc.*, 1932, **54**, 3570–3582.

[65] R. S. Mulliken, *J. Chem. Phys.*, 1934, **2**, 782–793.

[66] A. H. Nethercot, *Phys. Rev. Lett.*, 1974, **33**, 1088–1091.

[67] M. A. Butler and D. S. Ginley, *J. Electrochem. Soc.*, 1978, **125**, 228–232.

[68] W. A. Harrison, *Electronic Structure and the Properties of Solids*, Dover Publications Inc., New York, 1980.

[69] B. D. Pelatt, R. Ravichandran, J. F. Wager and D. a. Keszler, *J. Am. Chem. Soc.*, 2011, **133**, 16852–16860.

[70] M. W. Gaultois *et al.*, *Chem. Mater.*, 2013, **25**, 2911–2920.

[71] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.

[72] G. Purja Pun, K. Darling, L. Kecskes and Y. Mishin, *Acta Mater.*, 2015, **100**, 377–391.

[73] M. A. Gibson and C. A. Schuh, *Scr. Mater.*, 2016, **113**, 55–58.

[74] G. Miletić and A. Drašner, *J. Alloys Compd.*, 2015, **622**, 1041–1048.

[75] R. Sarmiento-Pérez, S. Botti and M. A. L. Marques, *J. Chem. Theory Comput.*, 2015, **11**, 3844–3850.

[76] T. S. Jauho, *Phys. Rev. B*, 2015, **92**, 115140.

[77] I. Valencia-Jaime *et al.*, *J. Alloys Compd.*, 2016, **655**, 147–154.

[78] Z.-H. Cai *et al.*, *Chem. Mater.*, 2015, **27**, 7757–7764.

[79] I. E. Castelli *et al.*, *Energy Environ. Sci.*, 2012, **5**, 15729–15735.

[80] I. E. Castelli *et al.*, *Energy Environ. Sci.*, 2012, **5**, 5814–5819.

[81] T. Krishnamoorthy *et al.*, *J. Mater. Chem. A*, 2015, **3**, 23829–23832.

[82] R. Sarmiento-Pérez *et al.*, *J. Chem. Phys.*, 2015, **142**, 024710.

[83] K. Choudhary *et al.*, *Comput. Mater. Sci.*, 2016, **113**, 80–87.

[84] M. Pandey, A. Vojvodic, K. S. Thygesen and K. W. Jacobsen, *J. Phys. Chem. Lett.*, 2015, **6**, 1577–1585.

[85] A. Seko *et al.*, *Phys. Rev. Lett.*, 2015, **115**, 205901.

[86] T. Tada, S. Takemoto, S. Matsuishi and H. Hosono, *Inorg. Chem.*, 2014, **53**, 10347–10358.

[87] M. J. Young *et al.*, *J. Electrochem. Soc.*, 2015, **162**, A2753–A2761.

[88] J. C. Weber *et al.*, *Nanotechnology*, 2014, **25**, 415502.

[89] A. J. Martinolich and J. R. Neilson, *J. Am. Chem. Soc.*, 2014, **136**, 15654–15659.

[90] M. Fondell, T. J. Jacobsson, M. Boman and T. Edvinsson, *J. Mater. Chem. A*, 2014, **2**, 3352–3363.

[91] L. M. Ghiringhelli *et al.*, *npj Comput. Mater.*, 2017, **3**, 46.

[92] Citrine Informatics, *PIF Documentation*, 2018, `http://citrineinformatics.github.io/pif-documentation/schema{_}definition/index.html{%}5C-[Accessed:03-06-2018]`.

[93] R. H. Taylor *et al.*, *Comput. Mater. Sci.*, 2014, **93**, 178–192.

[94] S. P. Ong *et al.*, *Comput. Mater. Sci.*, 2015, **97**, 209–215.

[95] M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.

[96] T. Klucznik *et al.*, *Chem*, 2018, **4**, 522–532.

[97] P. Raccuglia *et al.*, *Nature*, 2016, **533**, 73–76.

[98] J. G. P. Wicker *et al.*, *CrystEngComm*, 2015, **17**, 1927–1934.

[99] J. Wellendorff *et al.*, *Phys. Rev. B*, 2012, **85**, 235149.

[100] N. Mardirossian and M. Head-Gordon, *J. Chem. Phys.*, 2016, **144**, 214110.

[101] F. Brockherde *et al.*, *Nat. Commun.*, 2017, **8**, 872.

[102] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran and R. Ramprasad, *Sci. Rep.*, 2013, **3**, 2810.

[103] G. Hautier, C. C. Fischer, A. Jain, T. Mueller and G. Ceder, *Chem. Mater.*, 2010, **22**, 3762–3767.

[104] A. O. Oliynyk *et al.*, *Chem. Mater.*, 2016, **28**, 7324–7331.

[105] F. Legrain, J. Carrete, A. van Roekeghem, G. K. H. Madsen and N. Mingo, *J. Phys. Chem. B*, 2018, **122**, 625–632.

[106] T. Moot *et al.*, *Mater. Discov.*, 2016, **6**, 9–16.

[107] F. A. Faber, A. Lindmaa, O. A. Von Lilienfeld and R. Armiento, *Phys. Rev. Lett.*, 2016, **117**, 135502.

[108] C. H. L. Goodman, *J. Phys. Chem. Solids*, 1958, **6**, 305–314.

[109] B. Pamplin, *J. Phys. Chem. Solids*, 1964, **25**, 675–684.

[110] S. Chen, X. G. Gong, A. Walsh and S.-H. Wei, *Phys. Rev. B*, 2009, **79**, 165211.

[111] R. Gautier *et al.*, *Nat. Chem.*, 2015, **7**, 308–316.

[112] X. Zhang, L. Zhang, J. D. Perkins and A. Zunger, *Physical Review Letters*, 2015, **176602**, 1–6.

[113] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain and G. Ceder, *Inorg. Chem.*, 2011, **50**, 656–663.

[114] W. Sun *et al.*, *Sci. Adv.*, 2016, **2**, e1600225.

[115] M. Aykol, S. S. Dwaraknath, W. Sun and K. A. Persson, *Sci. Adv.*, 2018, **4**, eaaq0148.

[116] A. R. Oganov, A. O. Lyakhov and M. Valle, *Acc. Chem. Res.*, 2011, **44**, 227–237.

[117] D. Lonie and E. Zurek, *Comput. Phys. Commun.*, 2011, **182**, 372–387.

[118] A. R. Oganov *et al.*, *Nature*, 2009, **457**, 1–13.

[119] Y. Wang, J. Lv, L. Zhu and Y. Ma, *Comput. Phys. Commun.*, 2012, **183**, 2063–2070.

[120] D. E. E. Deacon-Smith, D. O. Scanlon, C. R. A. Catlow, A. A. Sokol and S. M. Woodley, *Adv. Mater.*, 2014, **26**, 7252–6.

[121] A. R. Oganov, Y. Ma, A. O. Lyakhov, M. Valle and C. Gatti, *Rev. Mineral. Geochemistry*, 2010, **71**, 271–298.

[122] M. S. Dyer *et al.*, *Science*, 2013, **340**, 847–52.

[123] O. D. Friedrichs, A. W. Dress, D. H. Huson, J. Klinowski and A. L. Mackay, *Nature*, 1999, **400**, 644–647.

[124] A. Le Bail, *J. Appl. Crystallogr.*, 2005, **38**, 389–393.

Part II

Theory and Methods

Chapter 2

First-principles Calculations

2.1 The Schrödinger equation

First-principles (*ab-initio*) methods involve using fundamental physics in order to describe chemical systems. The many-body time-independent Schrödinger equation is the starting point for accessing ground state electronic properties:

$$E\Psi = \hat{H}\Psi \tag{2.1}$$

With E the energy of an n-particle system, Ψ the many-body wavefunction and $\hat{H}$ the Hamiltonian operator.[1] The Hamiltonian operator can be constructed from the different types of interaction that occur within molecules and solids:

$$\hat{H} = T_{nuc} + T_e + U_{ee} + U_{ne} + U_{nn} \tag{2.2}$$

Where the kinetic energy terms T_{nuc} and T_e are for nuclei and electrons, respectively, and U_{ee}, U_{ne} and U_{nn} are potential energy terms for electron-electron, nucleus-electron and nucleus-nucleus interactions, respectively. For a hydrogen atom, the U_{nn} and U_{ee} terms can be removed and the Schrödinger equation can be solved analytically. For systems with more than one electron the equation cannot be solved. While electrostatic forces acting on electrons can be accounted for, we lack the necessary mathematical tools to solve the equations that come from the quantum mechanical interactions. This is an example of a *many-body problem* and approximations must be made in order to carry out practical calculations.

The first step is to introduce the Born-Oppenheimer approximation, which recognises that nuclei are much more massive than electrons so can be considered stationary for the time scales on which electrons move. Now that the wavefunction can be separated into a nuclear and electronic component, the electronic wavefunction is solved for a fixed set of nuclear positions and only electronic terms need to be explicitly considered for the Hamiltonian:

$$\hat{H}_e = T_e + U_{ee} + U_{ne} \tag{2.3}$$

This can be written more fully as:

$$\hat{H}_e = -\sum_i \frac{\hbar^2}{2m_i}\nabla_i^2 + \sum_{i\neq j}\frac{e^2}{r_{ij}} - \sum_i\sum_I \frac{e^2 Z_I}{r_{iI}} \tag{2.4}$$

were i and I denote electrons and nuclei respectively, m is mass, e is the charge of an electron, r is distance, Z is charge and ∇^2 is the Laplacian operator ($\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$). From now on, we will only consider the electronic terms and $\hat{H}_e$ will be written as $\hat{H}$. The potential energy term for electron-electron interaction (U_{ee}) is still problematic for a many-body system for the reasons stated above, so further simplification is needed.

2.2 The Hartree–Fock method

In the Hartree–Fock method,[2,3] two key assumptions are made. The first is that the n-electron wavefunction can be approximated as a set of one-electron wavefunctions. The Hamiltonian is a sum of one-electron Hamiltonians:

$$\hat{H} = \sum_i \hat{h}_i \tag{2.5}$$

where the electron-electron interaction term has been dropped from $\hat{h}_i$:

$$\hat{h}_i = -\frac{\hbar^2}{2m_i}\nabla_i^2 - \sum_I \frac{e^2 Z_I}{r_{iI}} \tag{2.6}$$

This is known as the *independent electron approximation*. The Schrödinger equation can then be evaluated for each one-electron wavefunction ψ_i to get the eigenvalue ϵ_i, and the wavefunction for the system is given by their product:

$$\begin{aligned} h_i\psi_i &= \epsilon_i\psi_i \\ \Psi &= \psi_1\psi_2...\psi_i \end{aligned} \tag{2.7}$$

The second assumption reintroduces some electrostatic forces due to other electrons while maintaining the simplicity of a one-electron interpretation. This is achieved using the *mean field approximation*, in which each electron experiences an average field of the other electrons in the system. A *Hartree potential* term ν_i is added to the one-electron Hamiltonian[4] and is calculated as:

$$\begin{aligned} \nu_i &= \sum_j \int \frac{\rho_j}{r_{ij}} dr' \\ \hat{h}_i &= -\frac{\hbar^2}{2m_i}\nabla_i^2 - \sum_I \frac{e^2 Z_I}{r_{iI}} + \nu_i \end{aligned} \tag{2.8}$$

where ρ_j is the electron density and is given by:

$$\rho_j = |\psi_j|^2 \tag{2.9}$$

Hence, in order to solve the eigenvalue problem for one electron, it is a requirement to know the electron density to construct the Hamiltonian. In an apparent contradiction, however, the electron density itself is calculated using the one-electron wavefunction. In practice, the solution is an iterative process, whereby a trial set of one-electron wavefunctions ψ_i are used to construct a corresponding set of $\hat{h}_i$, which are then used to generate new ψ_i via the Schrödinger equation. The electron density ρ_i can then be recalculated and the process repeats. A solution is reached when the density generated by the wavefunctions is equal (within some practical tolerance) to the density they produce. This is the self-consistent field (SCF) method and relies on the variational principle, which states that the true ground state energy of the system is always less than or equal to the value produced by any trial Hamiltonian.[5]

In order to form an acceptable set of wavefunctions, further constraints must be applied that come from the fundamental physics of fermions. Fermions are characterised as having half-integer spin and obeying the Pauli exclusion principle, which means that no two electrons with the same spin can occupy the same quantum state. The total wavefunction of the system Ψ must therefore be anti-symmetric with respect to electron exchange – the sign of the wavefunction is reversed when the position x of two electrons is switched. The Pauli exclusion principle is enforced by writing the total wavefunction as a matrix

determinant, known as a Slater determinant:[6]

$$\Psi = \frac{1}{\sqrt{n!}} \begin{vmatrix} \psi_1(x_1) & \psi_2(x_1) & ... & \psi_n(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & ... & \psi_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(x_n) & \psi_2(x_n) & ... & \psi_n(x_n) \end{vmatrix} \tag{2.10}$$

The Slater determinant for an n-electron system must also satisfy the requirement of normalisation; as the square of the wavefunction represents the electron density, the probability of finding the n electrons anywhere in space must be exactly unity:

$$\int ... \int |\Psi(x_1, x_2, ..., x_n)|^2 dx_1 dx_2 ... dx_n = 1 \tag{2.11}$$

As well as normality, the restriction of orthogonality may also be enforced, which means that there is no net overlap for any two wavefunctions in a given system:

$$\int \psi_i \psi_j dr = 0 \text{ (if } i \neq j) \tag{2.12}$$

Normality and orthogonality can be described together as a single requirement: orthonormality.

By approximating the exact n-electron wavefunction using Slater determinants, the Slater determinant that gives the lowest energy is selected by individually varying one-electron wavefunctions.

Finally, the Hartree–Fock approach naturally introduces an additional potential term U_{ij}^X:

$$U_{ij}^X = -\sum_j \psi_j(r) \int \frac{\psi^*(r')\psi(r')}{|r - r'|} dr' \tag{2.13}$$

This is the *exchange* potential and represents a quantum effect felt between electrons of like spin, at positions r and r'. It acts as a negative correction to the self interaction error that arises from calculation of the Hartree potential. The unphysical self interaction term occurs due to electrons experiencing their own field, i.e. when $i = j$ in the top part of Equation 2.8. A corresponding effect between electrons of different spins is still not accounted for, and this so-called *correlation* energy must be incorporated by taking a different approach to approximating solutions to the Schrödinger equation.

2.3 Density functional theory

2.3.1 Principles of DFT

The problem with the Hartree–Fock technique is that it neglects electron correlation. The way in which each electron contributes to the field that is felt by other electrons is not, in reality, independent. A method of accounting for this correlation but simultaneously avoiding the many-body problem is needed. The breakthrough came in the form of a proof by Hohenberg and Kohn in 1964,[7] which showed that for an interacting electron gas in an external potential, knowledge of the electron density is all that is required to access the electronic ground state properties. The existing suspicions that moving to a density-based scheme was possible, as exemplified in the work of Thomas and Fermi in the 1920s,[8] were confirmed. The two Hohenberg–Kohn theorems provide the basis for this conclusion and, briefly stated, are:

1. The external potential, and hence the total energy, is a unique functional of the ground state electron density.
2. It is possible to obtain the ground state energy of a system (but no lower) by varying the electron density.

In essence, this shows that moving to a density-based description does not result in the loss of any information.

While the proof by Hohenberg and Kohn meant that in principle the ground state energy should exist as a universal functional of electron density $E = F[\rho(r)]$, the exact form of the functional was not known, and is still not known. In 1965, Kohn and Sham outlined an implementation method that allowed for practical DFT calculations.[9] Their method essentially involves mapping the n-electron interacting system onto n one-electron non-interacting systems and partitioning contributions to the total energy into those that can be solved exactly for this for the fictitious non-interacting system, and those that cannot.

In the non-interacting system, Kohn–Sham (KS) orbitals ϕ_i are considered, and the sum of the squares of ϕ_i returns the electron density of the real system:

$$\rho(r) = \sum_i |\phi_i(r)|^2 \tag{2.14}$$

Going back to the Hartree–Fock approach, the energies can be summarised as kinetic,

external, Hartree and exchange energies, respectively:

$$E = E_k + E_{ext} + E_H + E_X \tag{2.15}$$

These terms can be redistributed according to interacting and non-interacting contributions:

$$\begin{aligned} E_k &= E_k^{non} + E_k^{int} \\ E_H + E_X &\rightarrow E_H + E_X + E_C^{int} \\ E_{XC} &= E_X + E_C^{int} + E_k^{int} = E_X + E_C \end{aligned} \tag{2.16}$$

In the above, the kinetic energy term E_k is first separated into interacting and non-interacting contributions, while a new contribution from electron correlation E_C^{int} that is neglected in the Hartree–Fock approach is introduced. The interacting terms are grouped together as the exchange-correlation (XC) energy E_{XC}, which consists of the contributions from electron exchange and electron correlation, including interacting (correlated) kinetic effects. Within the DFT framework the energy is dealt with as:

$$E[\rho(r)] = E_k^{non}[\phi(r)] + E_{ext}[\rho(r)] + E_H[\rho(r)] + E_{XC}[\rho(r)] \tag{2.17}$$

where the external term E_{ext} is the effect felt by the electrons due to the nuclei. It is possible to calculate the first three terms exactly, and the final term is unknown and subject to approximations. Although the kinetic energy term is strictly calculated using KS orbitals, not electron density, the two quantities are interrelated via Equation 2.14. The derivatives of the above energy functionals with respect to electron density enable the calculation of the corresponding KS Hamiltonian, which consists of the non-interacting kinetic energy part, and an effective potential ν_{eff} that includes the three potential terms. This is implemented in the form of the Schrödinger equation, which can then be solved self-consistently:

$$\begin{aligned} \hat{h}_{KS} &= -\frac{\hbar^2}{2m_i}\nabla^2 + \nu_{eff}(r) \\ \nu_{eff}(r) &= \nu_{ext}(r) + \int \frac{\rho(r')}{|r - r'|} dr' + \nu_{XC}(r) \\ \hat{h}_{KS}\phi_i(r) &= \epsilon_i \phi_i(r) \end{aligned} \tag{2.18}$$

For a given system, the set of KS equations is solved simultaneously. In the self-consistent cycle, the electron density is calculated from the KS orbitals and the Hamiltonian is constructed using the electron density. This is then used in Equation 2.18 to generate new KS

orbitals and electron density. Convergence is reached when the difference in the old and new energies differs by less than some threshold.

While the same restrictions of orthonormality are imposed on the KS orbitals as in the Hartree–Fock approach, the KS orbitals are constructed from the electron density so do not have the same interpretation as the one-electron wavefunctions in the Hartree–Fock theory, the product of which reproduces the many electron wavefunction.

In summary, while Hartree–Fock theory is approximate, but can be solved exactly, DFT is formally exact, but practical solutions require approximation of the E_{XC}. DFT provides a computationally efficient alternative to the Hartree–Fock approach that is more accurate in practice. The increase in speed is due to the many-body problem being recast as n one-body problems where integration over the electron density in three dimensions is required. The increase in accuracy is due to the inclusion of electron correlation, however this is only included as an approximation along with electron exchange in the form of E_{XC}. Given that an exact XC energy would in principle lead to all many-body effects being included within the DFT approach, the search for accurate and efficient XC functionals has been an area of intense focus for many decades.

2.3.2 Exchange-correlation functionals

The accurate estimation of E_{XC} is crucial for successful DFT calculations. Although this term only constitutes around 10% of the total energy, its calculation provides details about chemical bonding and determines key properties such as bandgap. In principle, the exchange part could be calculated exactly as:

$$E_X = -\sum_{ij}^{n} \int\int \frac{\phi_i^*(r)\phi_j^*(r')\phi_i(r')\phi_j(r)}{|r-r'|} dr dr' \tag{2.19}$$

However, this requires reverting back to the Hartree–Fock framework. Instead, various ways to approximate the XC functional have been developed over the years.

Local density approximation (LDA): The simplest XC functionals are based purely on the density of electrons, $\rho(r)$ at a given point r. It is possible to calculate an exact XC functional for a homogeneous electron gas, in which electrons are evenly distributed with a uniform external potential to maintain charge neutrality. In LDA it is assumed that the XC energy for an electron at point r, ϵ_{XC} is the same as the XC energy for an electron in a

homogeneous electron gas of the same density ϵ_{XC}^{homo}:

$$E_{XC}^{LDA}[\rho(r)] = \int \rho(r)\epsilon_{XC}^{homo}[\rho(r)]dr \tag{2.20}$$

In general, the E_X contribution is overestimated in LDA and E_C is underestimated, leading to a cancellation of errors and reasonable performance. One major problem is that binding energies between atoms are usually overestimated leading to incorrect ground state geometries. The approximation really only works well for systems with slowly varying electron density such as metals. In the majority of chemical systems, including most ionic and covalent solids, the requirement of a slowly varying electron density is simply not met and higher levels of accuracy are required.

Generalised gradient approximation (GGA) : In GGA, not only is $\rho(r)$ taken into account, but also the density gradient, $\nabla\rho(r)$:

$$E_{XC}^{GGA}[\rho(r)] = \int \rho(r)\epsilon_{XC}^{GGA}[\rho(r), \nabla\rho(r)]dr \tag{2.21}$$

In practice, $E_{XC}^{GGA}[\rho(r)]$ is fitted to satisfy various physical constraints, and is expressed as the LDA form with an enhancement factor $F(s)$ to modify the energy directly:

$$E_{XC}^{GGA}[\rho(r), s] = \int \epsilon_{XC}^{LDA}[\rho(r)]\rho(r)F(s)dr \tag{2.22}$$

The value of s is calculated from $\rho(r)$ and $\nabla\rho(r)$ directly:

$$s = C\frac{|\nabla\rho(r)|}{\rho^{4/3}(r)} \tag{2.23}$$

where C is a constant. Typical values of $F(s)$ vary from 1.0 to 1.6. There have been many GGA functionals proposed over the years and one popular choice is the Perdew-Burke-Ernzerhof (PBE) functional,[10] which is transferable between many different types of system. In periodic solids, the PBE functional is found to overestimate bond lengths (conversely to LDA which tends to over-bind the atoms).[11] A solution to this problem comes in the form of the PBEsol functional,[12] which accounts for the increase in exchange energies in systems with consistent density gradients (such as solids). It is an empirically modified version of the PBE functional and (unless otherwise stated) is used for the work in this thesis for geometry optimisation, as it generally arrives at reasonable geometries at a computational cost amenable to high-throughput DFT.

Hybrid functionals: In hybrid XC functionals, a portion of the exact exchange from

Hartree–Fock is incorporated into E_{XC} in order to improve calculation accuracy:

$$E_{XC}^{hybrid} = E_{XC}^{GGA} + \alpha(E_X^{HF} - E_X^{GGA}) \tag{2.24}$$

The value of α is arbitrary, although it is usually around 0.25 as this tends to give the most accurate properties with respect to experiment across a range of materials.

As the Hartree–Fock exchange contribution requires evaluation of the exchange integrals, hybrid functionals are more computationally expensive than pure DFT methods. In particular, the exchange energy converges slowly over long distances. In the HSE06 method,[13,14] only short range exchange energies are calculated with the Hartree–Fock approach, while the GGA functional is used over the whole calculation region. This is an example of a *screened* hybrid XC functional, in which another parameter must be set: the range over which the Hartree–Fock exchange energy is calculated. In HSE06, the range separation parameter is set to a value which is optimal for predicting bandgaps of semiconductors, when α is 0.25.

While hybrid methods provide greater accuracy than standard DFT, it is not currently feasible for them to be implemented in very high-throughput calculations due to their computational cost. It is also important to note that for a wide range of properties such as bond lengths and various mechanical properties, there is often little or no improvement on values calculated using a good GGA functional.

2.4 Basis sets

The energy in DFT is a function of electron density which, in turn, is constructed from KS orbitals. These orbitals must be represented by a set of mathematical functions - a basis set. This is also the case for any approach where many-body properties are built using single particle functions, including in the Hartree–Fock approach. The basis set used can either be localised, such that functions are well-fitted to orbitals around individual atoms, or formed of plane waves, which span the whole space equally and are non-local.

In calculations of solids, it is common to assume the atoms are arranged in a perfect repeating pattern which allows for the imposition of periodic boundary conditions. Under these conditions, the simulation box (often one unit cell of a crystal structure) is surrounded by translational images of the same box in three dimensions. As such, periodic functions are best suited to act as a basis set, as opposed to localised, atom-centred or-

bitals which would be best suited to the simulation of an isolated molecule.

In Blöch's theorem, the wavefunction of an electron in a periodic potential can be separated into two parts:

$$\phi_{n,k} = u_n(r)e^{ikr} \tag{2.25}$$

where $u_n(r)$ is a function with the periodicity of the lattice, while e^{ikr} describes a plane wave similar to that of a propagating free electron. Used together in this way, an electron occupying an energy level, often referred to as a band n, with a wave vector k is represented. Relevant entities such as potentials and electron density can now be expressed in terms of a repeating periodic pattern.

The lattice periodic part $u_n(r)$ can be expressed as a sum of plane waves such that the whole wave function can be expressed as:

$$\phi_{n,k} = \sum_G C_{n,k+G} e^{i(k+G)r} \tag{2.26}$$

Where the sum is over reciprocal lattice vectors G. All the plane waves used have the periodicity of the lattice and in theory the sum is infinite. In practice, a plane wave cutoff is specified, which is the maximum frequency of wave to be used in the summation to construct the electronic wavefunctions. This restricts the wave vectors to a maximum radius in reciprocal space and is usually expressed as an energy:

$$\frac{\hbar^2}{2m}|k+G|^2 \leq E_{cut} \tag{2.27}$$

Higher values of E_{cut} give better accuracy in general, although the coefficients $C_{n,k+G}$ become smaller as G increases.

The wave vector k is also an index to each wave function. In reciprocal space (or k-space) only the values of k in a single unit cell need to be sampled. The full symmetry of the reciprocal lattice is contained within the *first Brillouin zone*, so in theory the wavefunctions need to be derived at all values of k within this region. While it is impossible to integrate over the entire Brillouin zone, wavefunctions vary slowly with k, so in practice a grid of k-points of a certain density is sampled and the integration is replaced with a weighted sum.

A key advantage of the use of plane waves is that only one parameter E_{cut} needs to be tuned to increase accuracy. A disadvantage is that very high values of E_{cut} are needed to represent core electron wavefunctions, which is computationally expensive. The reason for this

is that it is necessary to represent a spatially rapidly varying function, so a high resolution of basis plane waves is needed to capture this. While many *all-electron* codes do exist,[15,16] it is rare for them to follow this approach and they tend to use localised basis sets. An alternative is to use pseudopotentials to represent nuclei and core electrons together. The philosophy here is that core electrons do not contribute very much to chemical bonding and materials properties so can be included along with the nuclei. Pseudopotentials are designed to reproduce the effect of a nucleus screened by core electrons, but such that the wavefunctions do not vary as quickly near the nucleus, allowing for the use of a plane waves and a low E_{cut}. Furthermore, it is then only necessary to consider valence electrons explicitly, reducing the number of electrons that need to be treated by the Schrödinger equation.

The Vienna Ab Initio Simulation Package (VASP) code[17,18] is used for all first principles calculations in this work and employs projector augmented wave (PAW) pseudopotentials as proposed by Blöchl,[19] in which site-local functions (projectors) are added to the plane wave basis. Crucially, this method still allows for all-electron energies and electron densities to be obtained from the resulting pseudo-wavefunction.

2.5 Calculating properties

Having introduced the theoretical background to DFT in the previous section, this section will describe how various properties of interest can be obtained using this approach.

2.5.1 Geometry optimisation

It is essential to ensure that the crystal structure being simulated is in its ground state geometry. Many properties of interest are sensitive to ionic coordinates and lattice parameters so misleading results would be obtained should an off-equilibrium structure be imposed. The exact atomic positions and lattice parameters for a relaxed structure will vary between XC functionals, so geometries from experiment or other calculations cannot be taken at face value. Furthermore, within the Born-Oppenheimer approximation, finite temperature effects that would influence experimental crystal structure are automatically neglected in standard DFT. Instead, a starting structure is fed into a geometry optimisation algorithm in which the net forces acting on the atoms are minimised. This is done iteratively, and constitutes an outer loop (Figure 2.1) such that atomic moves take

place after each electronic SCF cycle.

In the GGA and in Hybrid DFT, the first derivatives of the potential energy surface, which are the forces on the atoms, are necessarily calculated. The forces are then used within minimisation algorithms to search for the lowest energy structure. The simplest algorithm to find the minimum energy structure is the steepest descent algorithm, in which choice of direction in which to move on the potential energy surface (PES) is always the steepest descent direction (calculated from the gradient at the present point). This method can result in many steps being taken to approach the minimum, many of which are not directed towards the minimum. In conjugate gradient methods, the energy and gradient at previous points, not just the current point, are taken into account in order to choose a more optimal minimisation direction. More complex algorithms that provide faster minimisation such as quasi-Newton methods are also used in practice.

The geometry is judged to be converged when forces acting on the atoms are minimised to below a certain threshold. It is common to relax structures at GGA level, even if Hybrid DFT will subsequently be used to calculate properties. This is largely due to the computational cost of performing many ionic steps at Hybrid DFT level. For this work, structures are relaxed using the PBEsol functional with a force threshold set to 0.01 eVÅ^{-1} unless otherwise stated.

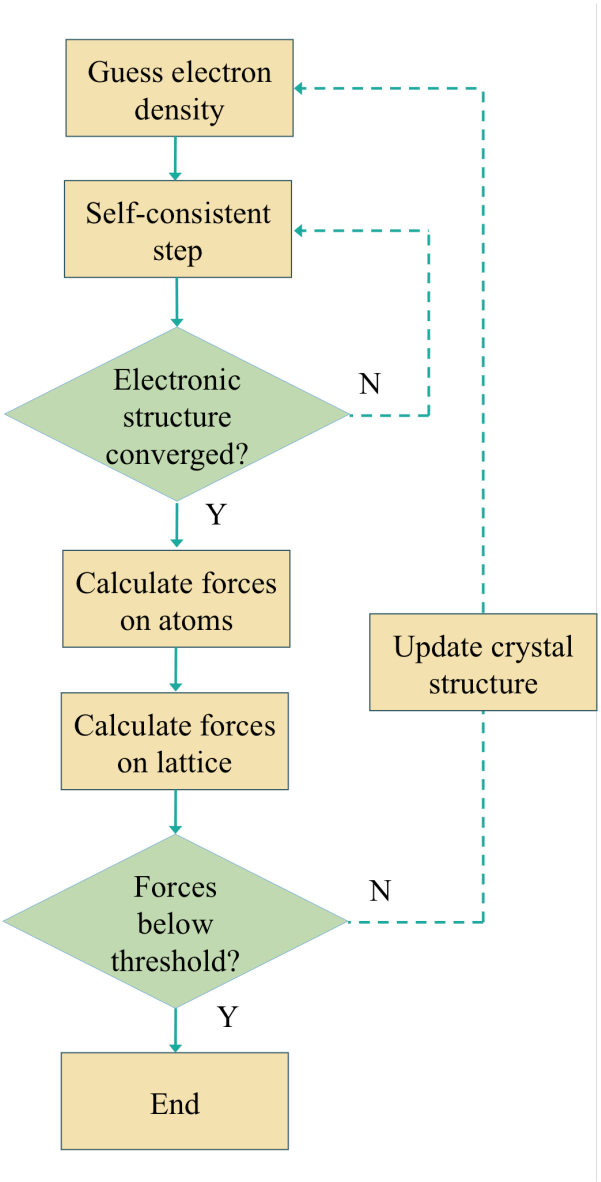

Figure 2.1: Iterative workflow for geometry optimisation using density functional theory (DFT).

2.5.2 Bandgap calculation

Building on the concepts introduced by Blöch's theorem, it is common to construct band structure plots of energy vs. wave vector k for each of the n bands, as shown schematically in Figure 2.2. Often a certain path through reciprocal space is chosen, and values of k along straight lines connecting symmetry points form the x-axis. Band structure plots relate directly to many important electronic structure properties. For example, in insulators and semiconductors, the plots feature a bandgap between the highest filled and lowest empty states. If the lowest energy state above the gap (conduction band minimum – CBM) has the same value of k as the highest energy state below the gap (valence band maximum – VBM), the bandgap is said to be direct. If the values of k differ, the bandgap is indirect.

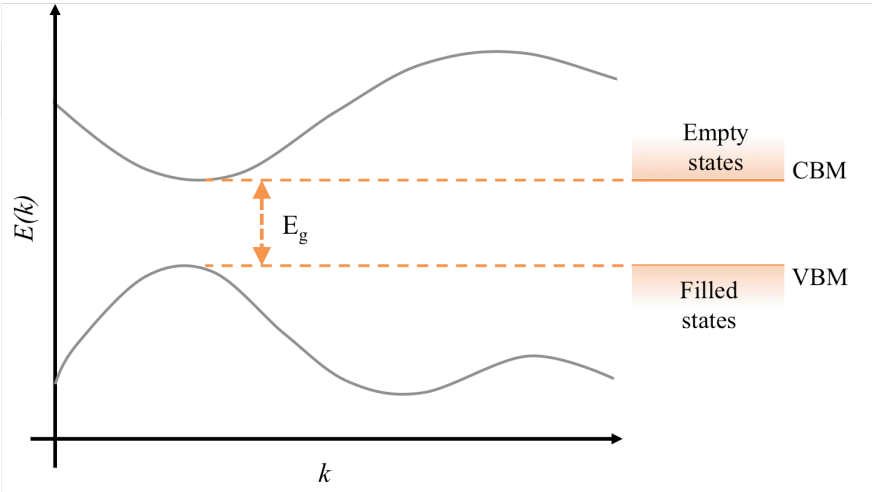

Figure 2.2: Schematic of a band structure (band dispersion) diagram. Energy E is plotted as a function of wave vector k and the bandgap E_g between the valence band maximum (VBM) and conduction band minimum (CBM) is direct in this case. Simple diagrams that are agnostic of k (right hand side) can be used to compare bandgaps of different materials.

One of the main shortcomings of DFT approaches is the ability to predict accurate bandgaps. Although the band dispersion (band width) is often calculated accurately, standard DFT (LDA and GGA) tends to consistently underestimate bandgaps in semiconductors and insulators. This is an issue in the context of screening studies as the bandgap is often – as in the work in this thesis – a key property of interest. The reason for this underestimation is that the self interaction energy from the Hartree potential is not totally cancelled out, as it is by the exchange energy in the Hartree–Fock scheme. The self-interaction causes an increase in energy of the occupied states and results in GGA bandgaps being underestimated, often severely.[20] Conversely, in Hartree–Fock theory the total lack of any treatment of electron correlation means bandgaps tend to be overestimated. Thus

by incorporating a portion of the exact Hartree–Fock exchange (Equation 2.24), hybrid XC functionals can provide much more accurate bandgaps. The increase in accuracy is essentially a cancellation of errors and depends strongly on the amount of exact exchange included in the calculation (α in Equation 2.24). The HSE06 XC functional is used in this work to calculate bandgaps of candidate materials.

2.5.3 Carrier effective mass

The motion of an electron in a periodic potential is often very different to the motion of an electron in a vacuum (free electron). The effective mass approximation assumes that the response of the electron in the potential is the same as the response of a free electron with a renormalised (effective) mass. Electrons are therefore assigned an effective mass m^* that is usually quoted in units of the rest mass of an electron m_e. With a newly defined mass, an electron becomes a quasiparticle. Another important quasiparticle in semiconductor physics is the positively charged electron hole, which arises due to the aggregate motion of electrons in the valence band. The effective mass is inversely proportional to conductivity, so smaller values are desirable for efficient semiconductors.

The carrier effective mass is related to the dispersion of the band and often the parabolic approximation is used, in which a quadratic least-squares fit is made to the CBM and VBM to obtain the electron and hole effective masses, respectively. The curvature of the band can then be used to calculate the effective mass directly:

$$m^* = \left(\frac{\partial^2 E}{\partial k^2} \frac{1}{\hbar^2} \right)^{-1} \tag{2.28}$$

The parabolic approximation is not always valid, particularly at high carrier concentrations where the bands are often not as parabolic as at the extrema. Furthermore when the curvature approaches 0, the effective mass approaches infinity. For m^* calculations in this thesis, k-points that are $k_B T$ above the CBM and below the VBM at standard temperature ($\sim$ 25 meV) are included in the least-squares fit.

2.5.4 Absolute electron energies

While the relative position of electronic bands determines the energies of optical transitions *via* the bandgap, the absolute positions of the bands relative to the vacuum level

are needed to determine other important properties. For example, absolute electron energies in two semiconductors determine whether electrons will flow easily between their conduction bands when in contact.[21] Absolute electron energies also determine whether a semiconductor is able to drive the water splitting reaction and produce O_2 and H_2 gases.[22]

The absolute energy of an electron is not an intrinsic bulk property and can only be specified relative to some other state.[23] As we are concerned with the addition and removal of electrons to and from the material, the key properties of interest are the electron affinity and ionisation energy. For completeness, it should be noted that in solids the carrier concentration gives rise to one further quantity, the Fermi energy, which is affected by the level of doping in a material. Absolute electron energies are influenced by two factors; a bulk binding energy and a dipole that originates from a redistribution of charges at the surface.

One approach to calculating absolute electron energies wile taking surface effects into account within the DFT framework is to build a slab model of the material of interest. The material is represented as semi-infinite, repeating in two dimensions with a surface to vacuum in the third. Periodic boundary conditions are still formally employed in 3D, but with a vacuum region separating several repeating units of the compound, constituting a slab. The planar average of the electrostatic potential calculated in the vacuum region is then used to align the electron energies. The planar average[24] as a function of the z coordinate (normal to the surface) of the 3D electrostatic potential V is calculated by taking averages of the planar surface area slices S:

$$\bar{V}(z) = \frac{1}{S}\int_S V(x,y,z)dxdy \tag{2.29}$$

The macroscopic average can then be calculated using the unit cell length a in the z direction:

$$\bar{\bar{V}}(z) = \frac{1}{a}\int_{z-\frac{a}{2}}^{z+\frac{a}{2}} \bar{V}(z')dz' \tag{2.30}$$

The difference between the macroscopic average of the electrostatic potential in the slab region normal to the surface and the electrostatic potential in the vacuum region gives the surface dipole D_s, as show schematically in Figure 2.3. While D_s is in fact an energy difference, it is conventional for this quantity to be referred to as the surface dipole, which forms due to electron leakage into the vacuum at any surface, even a formally non-polar surface. The ionisation energy is then calculated as $D_s - \epsilon_{vbm}$, where ϵ_{vbm} is the VBM as calculated in a standard bulk calculation. Similarly, the CBM from a bulk calculation ϵ_{cbm} can be used to calculate electron affinity. In this work, the relevant output files from

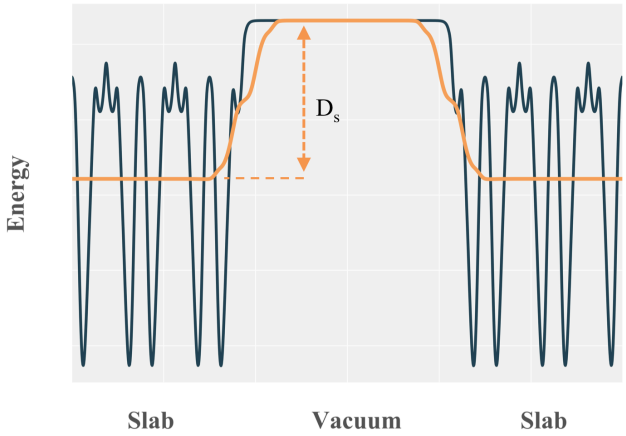

Figure 2.3: Schematic illustrating how the surface dipole is obtained from a slab calculation The macroscopic average (orange line) of the planar average of the electrostatic potential (blue line) is plotted normal to the surface and into the vacuum region.

the VASP code that store local potential information are analysed using the `MacroDensity` Python library.[25]

Practically, choosing how to cleave the bulk material to expose a surface is an important problem, as there are usually multiple low-energy surfaces, each of which can result in different energies. If a particular surface is known to be preferred experimentally, or is of particular relevance for a specific application, the choice becomes easier. For the calculation of electron energies of hypothetical materials in this work, the lowest index surface that is Type I according to Tasker's categorisation[26] – where there is no net dipole perpendicular to the surface – is selected. All bulk structures are relaxed as described above before generating a surface, and none of the surfaces are relaxed further. Finally, it is important to ensure that both the slab and vacuum layers are thick enough so that no surface effects are felt by the opposite surface, hence convergence of electron energies with respect to these parameters must be established.

2.5.5 Optical absorption

The dielectric tensor, calculated within the PAW methodology as described elsewhere,[27] consists of a real and imaginary part $\epsilon = \epsilon_r + i\epsilon_i$. The complex modulus of the dielectric

tensor $|\underline{\epsilon}|$ is then used to calculate the extinction coefficient κ:

$$\begin{aligned} |\underline{\epsilon}| &= \sqrt{\epsilon_r^2 + \epsilon_i^2} \\ \kappa &= \sqrt{\frac{|\underline{\epsilon}| - \epsilon_r}{2}} \end{aligned} \tag{2.31}$$

The extinction coefficient can then be used to calculate the absorption coefficient α at different wavelengths of electromagnetic radiation λ:

$$\alpha = \frac{4\pi}{\lambda}\kappa \tag{2.32}$$

For the absorption coefficients reported in this work, ϵ is averaged over the three Cartesian coordinates to give a powder average, assuming random orientation of crystals in a sample with respect to the direction of incident light.

2.5.6 Dynamic stability

When the geometries of crystal structures are relaxed by minimising net forces on the ions, there is a no guarantee that the optimisation algorithm has found a local minimum. Local maxima, or more commonly saddle points, on the PES also lead to an absence of net-forces on the ions and in these instances, given a small perturbation, the structure would relax into a lower energy structure. Finite displacement calculations are carried out to obtain vibrational (phonon) frequencies which gives an insight into the nature of the PES.

Within the harmonic approximation, the phonon frequencies and atomic displacement patterns of a crystal structure are determined from the second-order force-constant matrices $\Phi_{\alpha\beta}(jl, j'l')$:

$$\Phi_{\alpha\beta}(jl, j'l') = -\frac{\Delta F_\alpha(jl)}{\Delta r_\beta(j'l')} \tag{2.33}$$

where the force F on atom j is induced by the displacement of atom j' from position r. The indices l and l' refer to the unit cells of the atoms, and α and β label the Cartesian directions x, y and z. The matrix is built up by performing small displacements of atoms along symmetry-inequivalent directions and is known as the *finite displacements* method. In practice, one static DFT calculation is carried out for each of the displacements, and these can be done independently.

The force-constant matrix can then be transformed to the dynamical matrix for a given

phonon wave vector q. The dynamical matrix captures the wavelength and propagation of the atomic-displacement wave and its diagonalisation leads to a set of phonon frequencies $\omega(q, s)$ and atomic displacement patterns $W(q, s)$. The dynamical matrices are given by:

$$D_{\alpha\beta}(j, j', q) = \frac{1}{\sqrt{m_j m_{j'}}} \sum_{l'} \Phi_{\alpha\beta}(j0, j'l') e^{iq(r(j'l')-r(j0))} \tag{2.34}$$

in which m_j are the atomic masses and where each of the j atoms in the first unit cell are considered ($l = 0$).

The energy of a phonon mode is given by:

$$E = \frac{1}{2}\omega^2 Q^2 \tag{2.35}$$

where Q is the normal mode coordinate (amplitude of the oscillation). If $E < 0$ when $Q \neq 0$, it implies that the structure is at a local maximum or saddle point on the PES. Since Q cannot be less than 0, ω^2 must be less than 0, hence ω drops out as a complex number and such modes are termed *imaginary modes*. For the work in this thesis, net forces on atoms are reduced to below 0.005 eVÅ^{-1} before carrying out finite displacement calculations. Force-constant matrices and dynamical matrices are constructed using the `Phonopy` Python library.[28,29]

Bibliography

[1] E. Schrödinger, *Phys. Rev.*, 1926, **28**, 1049–1070.

[2] D. R. Hartree and W. Hartree, *Proc. R. Soc. London A Math. Phys. Eng. Sci.*, 1935, **150**, 9–33.

[3] J. C. Slater, *Phys. Rev.*, 1951, **81**, 385–390.

[4] D. R. Hartree, *Math. Proc. Cambridge Philos. Soc.*, 1928, **24**, 89–312.

[5] V. Fock, *Zeitschrift fur Phys.*, 1930, **61**, 126–148.

[6] J. C. Slater, *Phys. Rev.*, 1929, **34**, 1293–1322.

[7] P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864–B871.

[8] L. H. Thomas, *Math. Proc. Cambridge Philos. Soc.*, 1927, **23**, 542.

[9] W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, A1133.

[10] J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

[11] W. Koch and M. Holthausen, *A Chemist's Guide to Density Functional Theory*, Wiley-VCH, Weinheim, Germany, 2nd edn., 2001, p. 313.

[12] J. P. Perdew *et al.*, *Phys. Rev. Lett.*, 2008, **100**, 136406.

[13] J. Heyd, G. Scuseria and M. Ernzerhof, *J. Chem. Phys.*, 2003, **118**, 8207–8215.

[14] A. V. Krukau, O. A. Vydrov, A. F. Izmaylov and G. E. Scuseria, *J. Chem. Phys.*, 2006, **125**, 224106.

[15] V. Blum *et al.*, *Comput. Phys. Commun.*, 2009, **180**, 2175–2196.

[16] R. Dovesi *et al.*, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2018, **8**, e1360.

[17] G. Kresse and J. Furthmüller, *Comput. Mater. Sci.*, 1996, **6**, 15–50.

[18] G. Kresse and J. Furthmüller, *Phys. Rev. B*, 1996, **54**, 11169–11186.

[19] P. E. Blöchl, *Phys. Rev. B*, 1994, **50**, 17953–17979.

[20] J. M. Crowley, J. Tahir-Kheli and W. A. Goddard, *J. Phys. Chem. Lett.*, 2016, **7**, 1198–1203.

[21] M. Gratzel, *Prog. Photovoltaics Res. Appl.*, 2000, **8**, 171–185.

[22] A. Walsh and K. T. Butler, *Acc. Chem. Res.*, 2014, **47**, 364–372.

[23] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, Cambridge University Press, Cambridge, 2004.

[24] M. Peressi, N. Binggeli and a. Baldereschi, *J. Phys. D. Appl. Phys.*, 1998, **31**, 1273–1299.

[25] K. T. Butler, *MacroDensity*, `https://github.com/WMD-group/MacroDensity/-[Accessed:01-05-2017]`.

[26] P. Tasker, *J. Phys. C Solid State Phys.*, 1979, **12**, 4977.

[27] M. Gajdoš, K. Hummer, G. Kresse, J. Furthmüller and F. Bechstedt, *Phys. Rev. B*, 2006, **73**, 045112.

[28] A. Togo, F. Oba and I. Tanaka, *Phys. Rev. B*, 2008, **78**, 134106.

[29] A. Togo and I. Tanaka, *Scr. Mater.*, 2015, **108**, 1–5.

Chapter 3

Machine Learning

3.1 Gradient boosting regression

3.1.1 Machine learning workflow

A supervised learning approach can be used to build a model that predicts target values based on a set of input values. In this thesis, such an approach is used to predict bandgap values from chemical composition. The steps involved are outlined in Figure 3.1, along with the tools used at each step.

3.1.2 Data acquisition and representation

The Materials Project (MP) database is accessed using the Materials API[1] via the `Pymatgen` Python library.[2] The Computational Materials Repository (CMR) dataset is downloaded in its entirety from `https://cmr.fysik.dtu.dk/mp_gllbsc/mp_gllbsc.html` and accessed using the `Pysqlite` python library. The compounds in the CMR dataset are a subset of the compounds in the MP database but with additional information on bandgaps calculated with different XC functionals. Each entry in the CMR dataset has an associated `mp-id` which uniquely identifies the compound in the MP database.

The `Matminer` Python library[3] is used to represent compositions as vectors. This process is called featurisation or vectorisation. The ML-ready data takes the form of a dataframe of n columns and i rows, where each of $n-1$ columns is a feature of the composition* and

*Features used for the specific application in Chapter 7 are detailed in the same chapter.

the last column is the target property. Each compound – or more generally each *sample* – is represented as one of i rows.

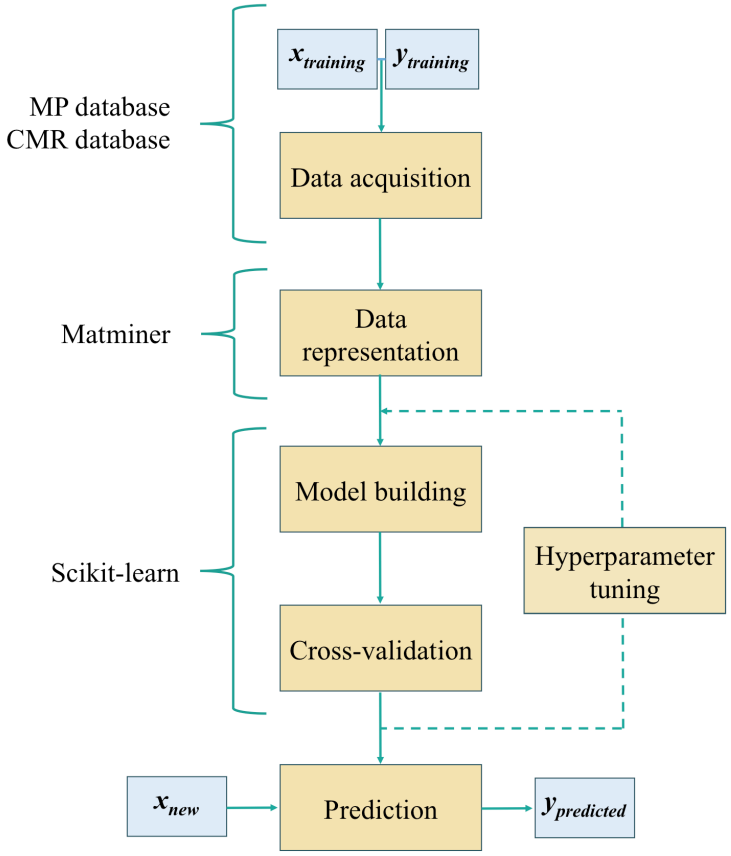

Figure 3.1: General workflow for supervised machine learning. MP and CMR refer to Materials Project and Computational Materials Repository, respectively.

3.1.3 Decision trees

Gradient boosting is an ensemble method, used to improve the performance of individual weak learners. The weak learner most frequently used is the decision tree where the goal is to create a model that predicts the value of a target variable by learning simple decision rules from features of a dataset. It is possible to apply decision trees to classification and regression problems, so the target variable can be discrete or continuous. When the application is regression, the corresponding ensemble model is referred to as gradient boosting regression (GBR). Individual trees are constructed using the classification and regression trees (CART) algorithm, outlined below.[4]

For training vectors x_i and a label vector (target) y, the space is partitioned into two parts,

such that samples with the same labels are grouped together. Let Q represent the data at node j, for a candidate split θ, consisting of a feature f and some threshold t_j, the data is partitioned into $Q_{left}(\theta)$ and $Q_{right}(\theta)$:

$$\begin{aligned} Q_{left}(\theta) &= (x, y)|x_f \leq t_j \\ Q_{right}(\theta) &= Q \setminus Q_{left}(\theta) \end{aligned} \tag{3.1}$$

The impurity at j is calculated using an impurity function $H()$:

$$C(Q, \theta) = \frac{n_{left}}{N_j} H(Q_{left}(\theta)) + \frac{n_{right}}{N_j} H(Q_{right}(\theta)) \tag{3.2}$$

where n_{left} and n_{right} are the number of samples in each partition out of the total number of samples a the node N_j.

The threshold and possible split points are evaluated and chosen in a greedy manner, such that parameters are always chosen that minimise the impurity:

$$\theta^* = \underset{\theta}{\mathit{argmin}}\, C(Q, \theta) \tag{3.3}$$

This process is repeated recursively for subsets $Q_{left}(\theta^*)$ and $Q_{right}(\theta^*)$ until some stopping criteria is fulfilled: Either the predefined maximum tree depth is reached, $N_j = 1$, or $N_j \leq$ some predefined value (minimum samples at a leaf node).

Finally, for regression the criteria to be minimised is the mean squared error, which is generally calculated as:

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.4}$$

for a vector $\hat{y}$ of n predictions, where y is the vector of observed values.

For decision trees, the mean values at terminal nodes are used:

$$c_j = \frac{1}{N_j} \sum_{i \in N_j} y_i \tag{3.5}$$

The equation for the impurity function then becomes:

$$H(X_j) = \frac{1}{N_j} \sum_{i \in N_j} (y_i - c_j)^2 \tag{3.6}$$

where X_j is the training data at node j.

3.1.4 Boosting

In gradient boosting, the individual decision trees are considered in an additive way. The overall model is given by:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{3.7}$$

where $h_m(x)$ are the weak learners (decision trees) and γ_m are the step lengths. In practice, the step lengths are usually fixed and called the learning rate. Decision trees of a fixed size are added to the model sequentially, in a forward stagewise fashion:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{3.8}$$

A fixed number of decision trees to be added to the model is set at the outset.

At each addition, the decision tree is chosen to minimise a loss function L given the current model fit $F_{m-1}(x_i)$:

$$F_m(x) = F_{m-1}(x) + \underset{h}{\mathit{argmin}} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h(x)) \tag{3.9}$$

This minimisation problem is solved numerically via steepest descent. The negative gradient of the differentiable loss function evaluated at the current model F_{m-1} gives the steepest descent direction:

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \nabla_F L(y_i, F_{m-1}(x_i)) \tag{3.10}$$

Finally, the loss function itself is the squared error loss function:

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \nabla_F (y_i - F_{m-1}(x_i))^2 \tag{3.11}$$

and evaluating the negative gradient of this simply gives the residuals of the model multiplied by 2:

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} 2(y_i - F_{m-1}(x_i)) \tag{3.12}$$

In this way, each consecutive decision tree is trained on the residuals of the current model. If a decision tree were able to completely correct for the residuals, the ensemble would give

predictions without errors. In practice, this is never the case and the process is carried out iteratively, as above. For the work in this thesis, the Python library `Scikit-learn`[5] is used to construct GBR models.

3.1.5 Cross-validation

Total error in ML approaches comes from a combination of bias, variance and irreducible errors. Shallow decision trees such as those used in GBR are prone to high bias (error from erroneous assumptions about the training data). Gradient boosting reduces bias of individual trees, but runs the risk of increasing the variance (error from sensitivity to noise in the training data, or overfitting). Upon changing some parameter such as the number of decision trees, it is crucial to check how the model performs on unseen data, even if the fit to the training data appears to be improving.

Each time a GBR model is built, 10-fold cross validation (CV) is performed. The model is initially trained on 90% of the data, then tested on the remaining 10%. The predicted bandgap values are compared to the ground truth by calculating the root-mean-squared-error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.13}$$

where n is the number of entries in the test set, y_i is the true bandgap value of the i^{th} entry and $\hat{y}_i$ is the bandgap value predicted by the model.

Each 10% portion of the data is used as the test set in turn, and the mean RMSE of all splits is calculated. CV is carried out for each set of hyperparameters that is trialled for the model.

3.1.6 Hyperparameter tuning

The hyperparameters of a GBR model are both tree-specific parameters and boosting parameters that are set before the training process begins. The key tree-specific parameters are:

1. Minimum number of samples needed to split a node
2. Minimum number of samples allowed at a leaf (terminal) node

3. Maximum depth (size) of trees
4. Maximum number of features each tree can use

While optimum values for all of these are problem-specific and cannot be predicted *a priori*, it is generally true that parameter 3 must be kept small in for GBR to minimise variance. Parameter 4 introduces a degree of diversity into the ensemble which helps to further reduce overfitting.

The key boosting specific parameters are:

1. Learning rate
2. Number of trees (boosting stages)
3. Fraction of samples in the dataset to fit each tree

Given that learning rate is essentially the step size in the minimisation, this is decreased as the number of trees is increased to give a more accurate model. A common approach is to keep the number of trees low and learning rate high initially, while optimum values are found for tree-specific parameters at low computational cost. This is done in the following order, prioritised by the impact each parameter tends to have on overall performance:

1. Perform a grid search for the optimum values of maximum tree size and minimum number of samples to split a node.
2. Fix the maximum tree size and perform a grid search on the number of samples to split a node and the minimum samples at a leaf node.
3. Fix all other parameters and tune the maximum features each tree can consider.

Subsequently, the fraction of samples in the dataset to fit to each tree is tuned, followed by increasing the number of trees and decreasing the learning rate until no further improvement in RMSE from CV is seen.

3.1.7 Feature importance

It is possible to determine the relative importance of each feature by extracting some basic information about the structure of the final model. The exact method used is the mean

decrease impurity method,[4] whereby importance of a node j is calculated as:

$$mi_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \tag{3.14}$$

where w_j is the weighted number of samples that traverse node j, C_j is the impurity in this node, and *left(j)* and *right(j)* are the children nodes. Feature importance for an individual tree can then be calculated as:

$$fi_i = \frac{\sum_{j:m_j} mi_j}{\sum_{j:m_{tot}} mi_j} \tag{3.15}$$

where m_j are the nodes that split on feature i, and m_{tot} are all the nodes in the tree. Values of fi_i are then averaged across all trees.

3.2 Structure substitution algorithm

The ionic substitution method for predicting likely crystal structures from composition was introduced in Chapter 1. The model quantifies how likely it is for a set of species to substitute onto the sites in a known crystal structure and is used heavily throughout the work in this thesis, so an overview of how it is built and used is given here. A full description of the methodology can be found in the original paper by Hautier *et al.* (Reference 6). For the work in this thesis, a pre-trained model as implemented in `Pymatgen`[2] is used, i.e. the values of λ (see below) are already determined.

3.2.1 Model structure

They key quantity of interest is the probability p of two compounds $\boldsymbol{X}$ and $\boldsymbol{X'}$ existing in the same crystal structure. This is can be expressed in terms of the constituent species:

$$p(\mathbf{X}, \mathbf{X'}) = p(X_1, X_2, ..., X_n | X'_1, X'_2, ..., X'_n) \tag{3.16}$$

where X_j and X'_j are ions present at the position j in the crystal structure common to the two compounds. For example, $p(Ni^{2+}, Li^{+}, P^{5+}, O^{2-} | Fe^{2+}, Li^{+}, P^{5+}, O^{2-})$ represents how likely Fe^{2+} is to be substituted by Ni^{2+} in a lithium transition metal phosphate.

The value of $p(\mathbf{X}, \mathbf{X'})$ cannot be calculated directly and must be approximated. This is done using feature functions, which only return values of 1 or 0. A weighted sum of feature

functions is used to approximate $p(\mathbf{X}, \mathbf{X'})$:

$$p(\mathbf{X}, \mathbf{X'}) \approx \frac{e^{\sum_i \lambda_i f_i(\mathbf{X}, \mathbf{X'})}}{Z} \tag{3.17}$$

in which λ_i is the weight given to the feature function f_i and Z is a partition function ensuring normalisation of the probability function. The feature functions themselves are simple binary feature functions of the form:

$$f^{a,b}(\mathbf{X}, \mathbf{X'}) = \begin{cases} 1 & X = a \text{ and } X' = b \\ 0 & \text{else} \end{cases} \tag{3.18}$$

such that each pair of ions a and b is assigned a feature function with a corresponding weight $\lambda^{a,b}$ indicating how likely ions a and b are to substitute for one another. The weights also satisfy the constraint $\lambda^{a,b} = \lambda^{b,a}$. An example of a feature function is that which relates to the Ca^{2+} – Ba^{2+} substitution:

$$f^{Ca^{2+},Ba^{2+}}(\mathbf{X}, \mathbf{X'}) = \begin{cases} 1 & X = Ca^{2+} \text{ and } X' = Ba^{2+} \\ 0 & \text{else} \end{cases} \tag{3.19}$$

which will have an associated weight $\lambda^{Ca^{2+},Ba^{2+}}$ whose magnitude indicates how likely this particular substitution is.

3.2.2 Training the model

The weights $\lambda^{a,b}$ are the key to the accuracy of the model and are as yet unknown. These are determined from the information present in the ICSD. A structure comparison algorithm is used to find similar structures,[7] then specific instances of $(\mathbf{X}, \mathbf{X'})$ can be found. For example, $BaTiO_3$ and $CaTiO_3$ both form cubic perovskite structures with Ca^{2+} and Ba^{2+} on the same site. An entire crystal structure database D will lead to m individual assignments $(\mathbf{x}, \mathbf{x'})$ of this type:

$$D = \{(\mathbf{x}, \mathbf{x'})_1, (\mathbf{x}, \mathbf{x'})_2, ..., (\mathbf{x}, \mathbf{x'})_m\} \tag{3.20}$$

A maximum-likelihood approach is used, in which the set of weights maximising the likelihood of observing the training data is selected. The log-likelihood l can be represented as:

$$l(D, \boldsymbol{\lambda}) = \sum_{t=1}^{m} \log p((\mathbf{x}, \mathbf{x'})_t | \boldsymbol{\lambda}) \tag{3.21}$$

where the vector $\boldsymbol{\lambda}$ represents the set of weights. This can be evaluated, according to the approximation in Equation 3.17, as:

$$l(D, \boldsymbol{\lambda}) = \sum_{t=1}^{m}[\sum_{i} \lambda_i f_i((\mathbf{X}, \mathbf{X'})_t) - \log Z(\boldsymbol{\lambda})] \tag{3.22}$$

The feature weights are chosen by solving:

$$\boldsymbol{\lambda} = \underset{\boldsymbol{\lambda}}{argmax}\, l(D, \boldsymbol{\lambda}) \tag{3.23}$$

3.2.3 Implementation

In order to predict a likely structure formed by a set of candidate species a, b, c the following procedure is followed:

1. For compound i in the database containing the species x_i^1, x_i^2, x_i^3, the probability of forming a new compound by substituting the candidate species into compound i ($p(a, b, c|x_i^1, x_i^2, x_i^3)$) is determined using Equation 3.17.
2. If the probability is above a given threshold and the resulting structure is charge neutral, the new compound is added to the list of possible structures.
3. The next compound $i + 1$ in the database is considered an the process repeats.

In this work, a threshold of of 10^{-5} (log probability threshold of -5) is used unless otherwise stated. This value is found in the original paper to maximise the true positive rate and minimise the false positive rate during cross-validation. Different values are required for different chemistries due to the fact that the original database the model is trained on is biased towards certain compounds (e.g. towards oxides). Where different threshold values are used, they are empirically chosen as the minimum value required to return some suggested crystal structures for the majority of compositions in a given search.

Bibliography

[1] S. P. Ong *et al.*, *Comput. Mater. Sci.*, 2015, **97**, 209–215.

[2] S. P. Ong *et al.*, *Comput. Mater. Sci.*, 2013, **68**, 314–319.

[3] L. Ward *et al.*, *Comput. Mater. Sci.*, 2018, **152**, 60–69.

[4] C. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, *Classification and Regression Trees*, Taylor & Francis Group, Boca Raton, 1984, vol. 1.

[5] F. Pedregosa *et al.*, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

[6] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain and G. Ceder, *Inorg. Chem.*, 2011, **50**, 656–663.

[7] R. Hundt, J. C. Schön and M. Jansen, *J. Appl. Crystallogr.*, 2006, **39**, 6–16.

Part III

Results

Chapter 4

The Inorganic Composition Space

4.1 Introduction

This chapter deals primarily with the use of chemical heuristics to build up a search space of stoichiometric inorganic compositions. Although this necessarily results in the exclusion of all hypothetical compounds that are not exactly stoichiometric, it has been shown that stoichiometric compounds can be used as a starting point for further computational or experimental study, during which stoichiometry can be fine-tuned to target specific properties.[1] For example, the low energy structure $Y_2Ba_2Ca_4Fe_{7.5}Cu_{0.5}O_{21}$ was identified computationally, leading to experimental identification of $Y_{2.24}Ba_{2.28}Ca_{3.48}Fe_{7.44}Cu_{0.56}O_{21}$, which has the necessary properties to function as a solid oxide fuel cell cathode.[2] Furthermore, a primary aim in this chapter is to gain some insight into what proportion of the search space has already been explored. There is no shortage of stoichiometric compounds with useful properties and the list is still growing,[3] so it stands to reason that if enough of the composition space is unknown then it is likely that there are many more waiting to be discovered.

The approaches in the following chapters of this thesis exhibit a consistent screening methodology; first filtering by composition with no structural consideration, and subsequently assigning structure to top candidate compositions. This can be thought of visually as traversing the x-axis of Figure 4.1, then moving in the y direction to consider crystal structures of chosen compositions. Composition-property and structure-property rela-

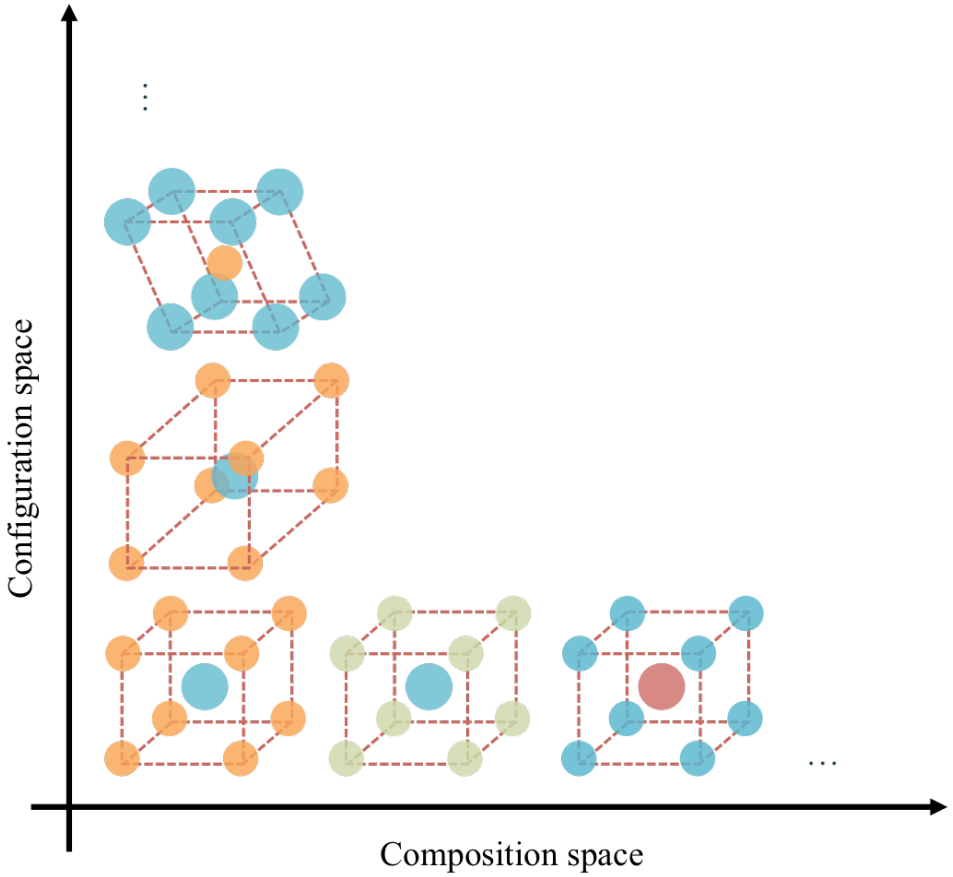

Figure 4.1: Schematic representation of the search space for inorganic compounds. The elemental composition is varied as the x-axis is traversed, while the arrangement of atoms in space is varied along the y-axis.

tionships both play an important role, so screening based on composition alone is only expected to predict properties with a modest level of accuracy. The ability to screen such a high volume of hypothetical compositions should mitigate against this drawback.

On a practical note, another aim is that the composition space can be constructed efficiently on a desktop computer. The overall materials design approach in this thesis consists of hierarchical screening workflows, whereby the initial screening steps are computationally cheap enough to cope with a large number of compositions. Therefore, it would be useful if the initial step of constructing a search space, subject to some chosen constraints, was as computationally efficient as possible.

Finally, this chapter also presents two short examples of using the composition space in a real search for target materials. In the first, the SSE scale (as described in Chapter 1) is used to target novel chalcohalide materials for photoelectrochemcial water splitting applications and this is followed up in detail in Chapter 6. The second example involves the

application of the Goldschmidt radius ratio rules[4] in order to enumerate the number of possible perovskite structures.

4.2 Statement of authorship

The following paper entitled *Combinatorial Screening of All Stoichiometric Inorganic Materials* reports on original research I conducted during the period of my Higher Degree by Research candidature.

Personal contributions: *Formulation of ideas (60%):* I have been heavily involved with all decisive stages of development of the project with guidance from Dr Keith Butler. *Design of methodology (70%):* Dr Adam Jackson, Dr Keith Butler and I contributed equally to the primary code base of the `smact` package. Subsequently, I performed additional refinement and testing of the code in order to improve its performance, along with Dr Jonathan Skelton. *Experimental work (60%):* I carried out the work on enumerating element combinations, as well as the photoelectrode search (Results subsections 2 and 4) and assisted Dr Keith Butler with the perovskite search (Results subsection 5). *Presentation of data in journal format (60%):* The first draft of the manuscript was written with equal contribution from Dr Adam Jackson, Dr Keith Butler and me, with input from the other co-authors. The finalised manuscript was prepared for submission by Prof. Aron Walsh and me.

4.3 Access statement

4.4 Publication 1

Computational Screening of All Stoichiometric Inorganic Materials

Daniel W. Davies,[†1] Keith T. Butler,[†1] Adam J. Jackson,[†1] Andrew Morris,[1] Jarvist M. Frost,[1] Jonathan M. Skelton,[1] Aron Walsh[1,2,3]

[†]DWD, KTB, and AJJ contributed equally to this manuscript.

1. Centre for Sustainable Chemical Technologies and Department of Chemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK

2. Global E^3 Institute and Department of Materials Science and Engineering, Yonsei University, Seoul 120-749, Korea

3. Department of Materials, Imperial College London, Exhibition Road, London SW7 2AZ, UK

4.4.1 Abstract

Forming a four component compound from the first 103 elements of the periodic table results in more than 10^{12} combinations. Such a materials space is intractable to high-throughput experiment or first-principles computation. We introduce a framework to address this problem and quantify how many materials can exist. We apply principles of valency and electronegativity to filter chemically implausible compositions, which reduces the inorganic quaternary space to 10^{10} combinations. We demonstrate that estimates of bandgaps and absolute electron energies can be made simply based on the chemical composition and apply this to search for new semiconducting materials to support the photoelectrochemical splitting of water. The applicability to crystal structure prediction by analogy with known compounds is shown, including exploration of the phase space for ternary combinations that form a perovskite lattice. Computer screening reproduces known perovskite materials and predicts the feasibility of thousands more. Due to the simplicity of the approach, large-scale searches can be performed on a single workstation.

4.4.2 Introduction

There are currently over 184,000 entries in the inorganic crystal structure database (ICSD) based on 9,141 structure types.[5] 66,814 of these materials have also been subject to quantum mechanical calculations, with the basic electronic structure and thermodynamic information included in the Materials Project[6] (powered by the pymatgen infrastructure[7]).

The configurational phase space for *new* materials is immense, and a blind exploration of the periodic table is a daunting task. Fortunately, over a century of research in the physical sciences has provided us with myriad rules for assessing the feasibility of a given stoichiometry and the likelihood of particular crystal arrangements. Examples of chemical phenomenology include the radius ratio rules[4] and Pettifor maps[8] for structure prediction, as well as electronegativity and chemical hardness for predicting reactivity.[9] Pauling's rules[10] provide predictive power for ionic or heteropolar crystals. A wealth of knowledge exists for understanding the physical properties of tetrahedral semiconductors.[11] Recent examples of searches for new materials that draw from existing chemical knowledge include 18-electron ABX compounds,[12] hyperferroelectric superlattices,[13] and organic-inorganic perovskites.[14,15]

The reliability and predictive power of atomistic materials simulations is increasing.[16,17] Many approximations are being removed as high-performance supercomputers reach petaflop scale. This includes more accurate quantum mechanical treatment of electron-electron interactions in the solid state,[18] as well as more realistic models of chemical disorder.[19] However, owing to the computational cost, high-throughput screening with first-principles techniques is usually limited to hundreds or thousands of materials — a small fraction of the overall phase space.

We report a procedure to navigate the materials landscape with low computational effort, which can be achieved using simple chemical descriptors. We first explore the magnitude of the task at hand, by enumerating combinations of elements and ions for binary, ternary and quaternary compositions. We demonstrate that chemical constraints can narrow the search space drastically. Examples of how deeper insights can be gained are illustrated for electronic (photoelectrodes for water splitting) and structured (perovskite type) materials. The procedure can be used to comfortably explore the vast compositional space or as the first step in a multi-stage high-throughput screening process. Instead of being a roadblock to achieving new functionality, the combinatorial explosion for multi-component compounds provides fertile ground for innovative materials discovery.

4.4.3 Results

4.4.3.1 Elemental combinations

To begin, one can map chemical space by enumerating the ways in which the constituent elements of the periodic table can combine. If we restrict ourselves to the first 103 elements (to the end of the actinide series), the combinations (i.e. C_n^{103}) for two, three and four components are 5,253, 176,851 and 4,421,275, respectively. For five components, the combinations exceed 87 million.

Physically, the situation is more complex. Elements can combine in different ratios leading to variation in material stoichiometry, e.g. the binary combinations AB, AB_2, A_2B_3, A_3B_4. Given elements may also adopt multiple oxidation states, each with a unique chemical behaviour, e.g. Sn(II)O, Sn(IV)O_2 and Sn(II)Sn(IV)O_3. For our enumeration of feasible compounds, we next consider the accessible oxidation states of each element in stoichiometry up to quaternary $A_wB_xC_yD_z$, where the integers $w, x, y, z \leq 8$. This definition includes, for example, common ternary pyrochlore oxides ($A_2B_2O_7$) and quaternary double perovskites (A_2BCO_6). Using the most common oxidation states extends the first 103 elements of the periodic table to 403 unique ions.

The number of combinations is now drastically increased, as shown in Table 4.1, with four component candidate materials exceeding 10^{12}. In order to reduce this composition space we can introduce selection rules (filters) from chemical theory.

We note that the estimations discussed here represent a lower limit on the number of accessible materials. We consider regular inorganic compounds and exclude, for example, non-stoichiometry, organic systems, hybrid organic-inorganic materials, electrides, and intermetallics, where additional considerations are required to predict viability.[20–22]

4.4.3.2 Chemical filters

Rule 1: Charge neutrality. Ions tend to combine into charge neutral aggregates. The same thinking applies to both ionic solids and more covalently bonded semiconductors. Any periodic solid must be charge neutral otherwise it would have an infinite electrostatic potential. Balancing of oxidation states and fulfilment of the valence octet rule are equivalent, e.g. III–V semiconductors such as GaAs can be represented as $Ga^{3+}As^{3-}$. Our implementation is inspired by the work of Pamplin[23] and Goodman[24] on the subject of

Table 4.1: Estimates for the number of possible inorganic materials allowing for variable oxidation state and stoichiometry with the constraints of charge neutrality (q) and electronegativity balance (χ).

Type	Constraint	Number
A_wB_x		3,483,129
A_wB_x	q	58,614
A_wB_x	$q + \chi$	14,721
$A_wB_xC_y$		4,753,229,039
$A_wB_xC_y$	q	174,081,685
$A_wB_xC_y$	$q + \chi$	32,157,899
$A_wB_xC_yD_z$		4,139,315,402,300
$A_wB_xC_yD_z$	q	267,381,955,246
$A_wB_xC_yD_z$	$q + \chi$	32,381,953,858

multi-component semiconductors.

A charge neutrality constraint significantly reduces the total number of candidate materials. The rule states that the formal charges (q) of the components sum to zero, i.e.

$$wq_A + xq_B + yq_C + zq_D = 0. \tag{4.1}$$

Charge neutrality contracts the compositional space by at least an order of magnitude for binaries, ternaries and quaternaries (Table 4.1).

Rule 2: Electronegativity balance. Further to assuming that all charge-neutral combinations of oxidation states are accessible, we can implement a second constraint based on the electronegativity of the component elements. The empirical electronegativity (χ) scale represents the 'attraction' of a particular atom for electrons. For a stable compound the relation $\chi^{\text{cation}} < \chi^{\text{anion}}$ should be obeyed, i.e. the most electronegative element carries the most negative charge. Here we employ the Pauling electronegativity scale, which reduces the allowed compositions by a factor of between 4 – 10 for the different numbers of components (Table 4.1).

It is also instructive to consider existing materials databases (the ICSD and Materials Project). For binary compounds we find fewer combinations from `smact` than the ICSD (Figure 4.2), which can largely be attributed to our exclusion of intermetallics and polymorphs. In the Materials Project multiple entries for a single composition are removed and the number of compositions are in close agreement. For ternaries and quaternaries the compositions passing both charge and electronegativity tests continues to rise exponentially, while the number in existing databases remains relatively constant. The increased complexity of ternary and quaternary systems means that their synthesis, charac-

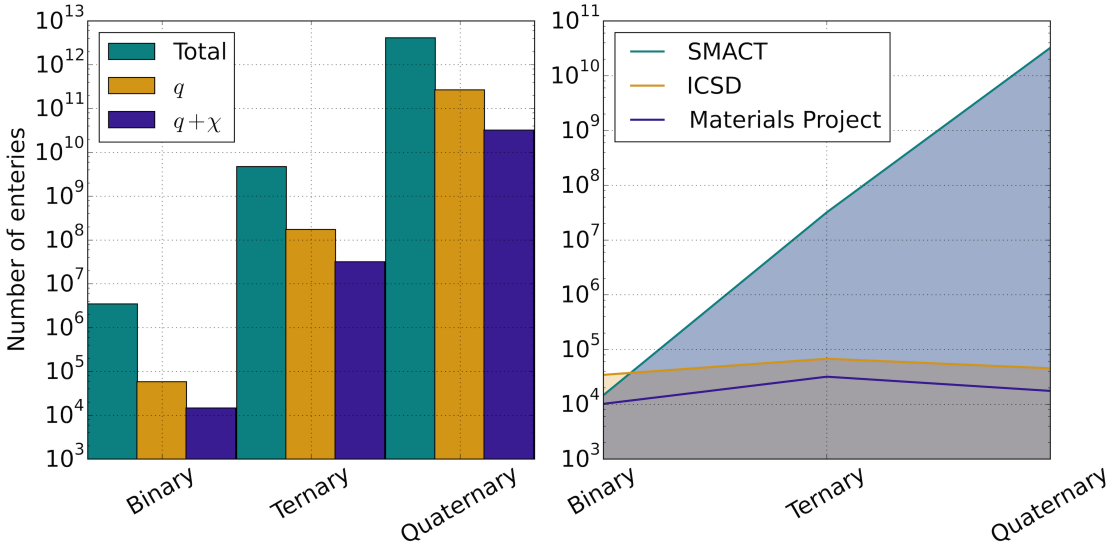

Figure 4.2: (left) Narrowing of compositional space for inorganic materials by imposing chemical constraints of charge (q) and electronegativity (χ). (right) Comparison of the accessible materials predicted by `smact` and those reported in the ICSD[5] and the Materials Project[25] database.

terisation and reporting are more challenging than for binary systems. Nevertheless, the large differences between numbers of potential and reported materials suggests that wide areas of unexplored compositional space may contain stable and useful materials.

The numbers reported in this section are vast, and performing quantitative screening for application using modern electronic structure techniques is unimaginable. Exploration of the hitherto neglected compositional space will require further guidelines. In the following sections we demonstrate how additional descriptors can be applied to identify materials for specific applications.

4.4.3.3 Compositional descriptors

Several useful properties can be estimated based on knowledge of the chemical composition alone, and here we explore the application of some of these approaches.

Descriptor 1: Electronic chemical potential. The concept of atomic electronegativity has been successfully extended to solids, where the geometric mean becomes the single-value descriptor, i.e.

$$\chi^{\text{solid}} = \sqrt[w+x+y+z]{\chi_{\text{A}}^{w}\chi_{\text{B}}^{x}\chi_{\text{C}}^{y}\chi_{\text{D}}^{z}} \tag{4.2}$$

This descriptor represents a mid-gap energy between the filled (valence band) and empty (conduction band) electronic states. This corresponds to the electronic chemical potential (Fermi level) at 0 K.[26] Butler and Ginley[27] found a linear correlation between the solid electronegativity and the electrochemical flat-band potentials for a range of semiconductors. This was subsequently extended to a wider data set including metal oxides, chalcogenides and halides.[28] The method provides a rapid procedure for the estimation of absolute electron energies for multi-component materials. It is now commonly employed in the computational screening of new materials for electrochemical applications.[29–32]

Descriptor 2: Electronic structure. Many tight-binding model Hamiltonians exist for semiconductors and dielectrics.[11] One recent approach is based on the atomic solid-state energy (SSE) scale,[33] which provides information on valence and conduction bands based on the frontier orbitals of the constituent ions. While the Mulliken definition of electronegativity is an average of the ionisation potential (IP) and electron affinity (EA) of an atom, the SSE reflects the IP of an anion (filled electronic states) and EA of a cation (empty electronic states). The energies of 40 elements were fitted from a test set of 69 closed-shell binary inorganic semiconductors,[33] which has recently been extended to 94 elements.[34] Based on the tabulated SSE scale, the bandgap (E_g) can be estimated as

$$E_g^{SSE} = \mathrm{SSE}^{\mathrm{cation}} - \mathrm{SSE}^{\mathrm{anion}} \tag{4.3}$$

For multi-component systems the limiting values (cation with highest EA and anion with lowest IP) are used. The SSE has a root-mean-squared-deviation of 0.66 eV against the measured bandgaps of 35 ternary semiconductors (see Table S1). This simple method allows for rapid screening of bandgaps and absolute band edge alignment.

Both methods (Equations 4.2 and 4.3) have been implemented for arbitrary compositions based on tabulated atomic data in the `smact` package. Since no crystal structure information is included at this level, the results are qualitative and the models do not distinguish, for example, between polymorphs.

4.4.3.4 Electronic structure: photoelectrodes

We now use the compositional space and chemical descriptors defined above to search for potential materials for solar fuel generation via photoelectrochemical water splitting.

The properties that are required for viable photoelectrodes include: (i) a bandgap in the visible range of the electromagnetic spectrum in order to absorb a significant fraction of sunlight; (ii) the upper valence and lower conduction bands bridge the water oxidation and reduction potentials, enabling the redox reaction. We set an optimal bandgap range of between 1.5 and 2.5 eV. Whilst the free energy for water dissociation is 1.2 eV, the combination of loss mechanisms found in practical devices may require a bandgap as large as 2.2 eV.[35,36]

Metal oxides combine many attractive properties for water splitting (e.g. stability and cost), but they usually have bandgaps too large to absorb a significant fraction of sunlight. The formation of multi-anion compounds offers a route to modifying the electronic structure. We consider ternary metal chalcohalides (i.e. $A_xB_yC_z$), with $B = [O, S, Se, Te]$ and $C = [F, Cl, Br, I]$. We restrict the A cations to those with an SSE higher than the water reduction potential (approx. -4.5 V relative to the vacuum at pH = 0). The conditions of charge neutrality and electronegativity are used perform an initial screening that yields 52,094 combinations. With the additional bandgap criterion, the combinations are reduced to 7,676, while the pool of cations is reduced from 25 to 7 with $A = [B, Ti, V, Zn, Ga, Cd, Sn]$. We further rule out any boron containing combinations at this stage, as these are known to form discrete molecular units (e.g. BClSe).

Finally, we screen compositions based on the environmental sustainability of the elements. We employ the Herfindahl–Hirschman Index for elemental reserves (HHI_R), which has been developed in the context of thermoelectric applications.[37] This index includes factors such as geopolitical influence over materials supply and price. The HHI_R for a given composition can be obtained as the weighted average over the constituent elements. At this stage, since stoichiometry is variable, we consider the mean HHI_R for each $A_1B_1C_1$ combination.

The band edge positions of the 20 candidates with the smallest HHI_R values are presented in Figure 4.3. The HHI_R has the effect of eliminating all combinations containing Ga, Te and Br*. There are no entries in the ICSD for the majority of candidates that we identified; however, reports can be found for $Cd_2O_6I_2$, Sn_2SI_2 and $Zn_6S_5Cl_2$.[38–40] Both $Cd_2O_6I_2$ and Sn_2SI_2 feature in the Materials Project and have bandgaps of 3.3 and 1.6 eV, respectively, calculated within density functional theory (DFT). These compare to the SSE bandgaps of 2.5 and 2.0 eV. The third compound, $Zn_6S_5Cl_2$ is reported to have an optical gap of 2.7 eV,[40] which compares to the SSE bandgap of 2.4 eV.

*Though relatively abundant, the majority of the world's Br is produced from the Dead Sea, making it geopolitically sensitive as reflected in a high HHI_R.

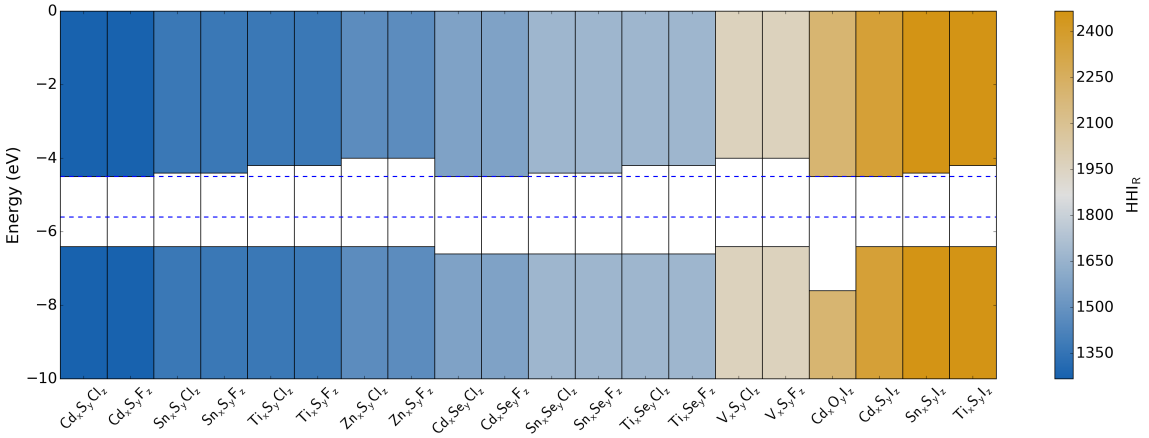

Figure 4.3: Calculated band edge positions, relative to the vacuum level, of 20 promising element combinations for water-splitting applications based on the solid-state energies (SSE) of the constituent elements. Blue dashed lines indicate the water reduction (above) and oxidation (below) potentials with respect to vacuum.

Only one oxygen containing compound survived the bandgap screening criterion: the values for metal oxyhalides are generally too large. For O_yI_z, the iodide forms the upper valence band (low binding energy of I 5p), while for other halides it is the oxide (O 2p). However, the sensitivity of the oxide ion to its crystal environment is well documented,[31,41] and consequently its SSE carries the greatest uncertainty.[33] This is one aspect where knowledge of the local structure (electrostatic potential) could significantly improve the accuracy of the results.

We must connect composition to crystal structure in order to make more accurate property predictions. Global optimisation of crystal structures from first-principles is a formidable task; although, great progress is being made in this area.[42] We instead adopt an approach based on analogy with known structures through chemical substitutions as developed by Hautier *et al.*[43] It employs data-mined probability functions, as implemented in the Materials Project.

To demonstrate the translation from composition to material, we have performed the crystal structure mining for the four combinations with the lowest HHI_R. The 88 predicted structures were then subjected to a full DFT lattice optimisation procedure and ranked by total internal energy. Finally, accurate bandgaps are predicted for the lowest energy structures using hybrid DFT (HSE06 electron exchange and correlation[44,45]). The compound $Sn_5S_4Cl_2$ has an indirect bandgap of 1.6 eV and a direct gap of 1.8 eV, which lies within the target range. The bandgaps of the other three lowest energy compounds are calculated to be between 3.0 and 3.4 eV. Full details of the workflow (Figure S1) and bandgaps (Table S2) can be found in the supplementary information.

The newly identified compound, $Sn(II)_5S_4Cl_2$, adopts a structure formed of two distinct Sn centred polyhedra: a distorted octahedron with equatorial S and apical Cl ions, and a distorted tetrahedron with 4 S ions and a sterochemically active Sn lone pair (Figure S2). The polyhedra form interlocking chains in three dimensions. The electronic density of states reveals an upper valence band comprised of hybridised Sn s – Cl p orbitals; such Sn s based valence bands are considered promising indicators for hole mobility.[46] The lower conduction band is comprised mainly of overlapping Sn p orbitals. The chemical structure and bonding characteristics suggest that this material should have favourable carrier transport, crucial for optoelectronic applications.

4.4.3.5 Crystal structure: perovskites

One of the most successful approaches to discover new materials is structural analogy. The concept is to take a crystal structure with a known chemistry and to replace elements within the structure to tune the properties. In the most simple case, this involves direct isovalent substitution, e.g. Zn(II)S $\longrightarrow$ Cd(II)S. Structural analogy can be extended to aliovalent cross-substitution (also termed cation mutation), e.g. Zn(II)S $\longrightarrow$ Cu(I)Ga(III)S_2. A systematic methodology was outlined more than 40 years ago in a paper by Pamplin for enumerating charge-neutral tetrahedral semiconductors.[23]

The challenge of going beyond tetrahedral semiconductors is predicting crystal structure. The radius of ions within a lattice has a long history as a geometric descriptor of structural stability, with a key example being the application of radius ratio rules by Goldschmidt[47] to predict the propensity of an ABC_3 combination to form the perovskite structure:

$$t = \frac{r_A + r_C}{\sqrt{2}(r_B + r_C)} \tag{4.4}$$

where t is the tolerance factor and r is the ionic radius. Values of $t > 1$ imply a relatively large A site favouring a hexagonal structure, $0.9 < t < 1$ predicts a cubic structure, whereas $0.7 < t < 0.9$ means the A site is small, preferring an orthorhombic structure. For $t < 0.7$ other (non-perovskite) structures are predicted. These rules have recently been extended to describe structure-property relationships in hybrid organic-inorganic perovskites.[14,15]

In this section, we apply our screening procedure to include knowledge of the crystal structure and estimate the size of the perovskite materials space. We start by enumerating the elemental combinations. We then reduce the set by requiring an octahedral coordination

environment for the B site — as contained in the Shannon dataset[48] – and require a combination of oxidation states that are charge neutral. This list is then assessed in terms of t as defined by the Shannon ionic radii.[48]

We consider single-anion compositions based on C=[O,S,Se,F,Cl,Br,I]. The charge neutrality and octahedral B-site constraints reduce the 176,851 elemental combinations to 41,725. The tolerance factor constraint, $0.7 < t < 1.0$, further reduces this to 26,567. For potential applications in the energy sector, we can consider candidates with HHI_R smaller than that of CdTe (a commercial thin-film photovoltaic material) resulting in a final population of 13,415.

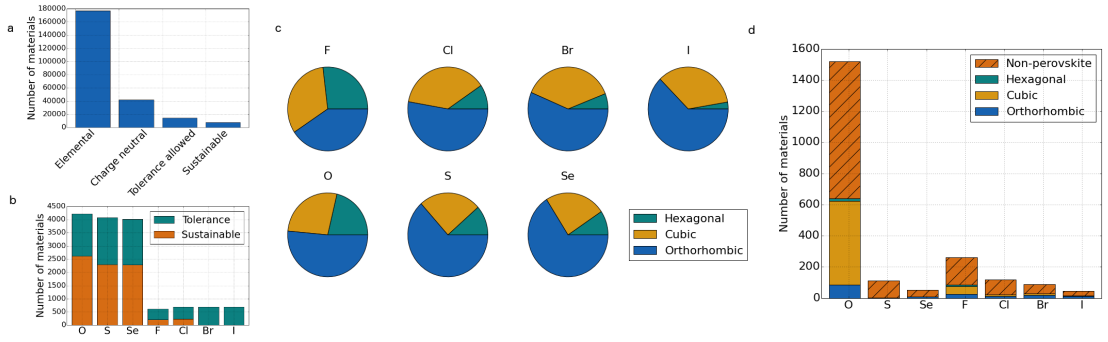

Figure 4.4: Counting experiments with perovskites. (a) Combinations found at each stage of the screening procedure. (b) Perovskite compounds with a HHI_R lower than CdTe for each anion. (c) The distribution of hexagonal, cubic and orthorhombic perovskite structures predicted based on the Goldschmidt tolerance factor and Shannon radii of the ions. (d) ABC_3 combinations found in the Materials Project database, sorted into structure type based on spacegroup (here orthorhombic and lower symmetry perovskites are grouped together).

For each anion, an orthorhombic perovskite structure is the most common prediction, with hexagonal most rarely predicted (Figure 4.4). The fraction of cubic perovskite structures remains roughly constant within the respective halide and oxide/chalcogenide series; however, it is more dominant for the halides. The presence of Br or I makes a material less sustainable (higher HHI_R); otherwise, there is little to differentiate the anions.

There are far more oxide and chalcogenide perovskites predicted than halides. The higher anion charge allows for three distinct cation combinations (I-V, II-VI, III-III) in comparison to the (I-II) halides. In addition, a greater radius compatibility is found for the group VI anions. We find that the number of plausible perovskite structures increases with the anion radius; however, the lower crustal abundance for heavier elements reduces the number that meet the sustainability criterion.

A search of the Materials Project over the same anion space reveals 920 materials, a small

fraction of those predicted from `smact` (26,567). The search includes all standard perovskite space groups.[49] For oxide perovskites 8.26 % of the number identified from `smact` are found in the Materials Project, for sulfides this falls to 0.45 % and selenides to 0.12 %. To some extent, the greater number of oxide perovskites discovered reflects the greater research activity in this field; however, synthesis of chalogenide perovskites has been reported[50–52] and there is interest in these materials for technological applications.[53,54] Of the ABC_3 materials reported in the Materials Project, 48 % of oxides, 35 % of sulfides and 20 % of selenides are in perovskite space groups.

Why are there so few chalcogenide perovskites? The tolerance factor arguments that work well for metal oxides may not hold for chalcogenide perovskites. Oxygen forms more ionic compounds due to a higher electronegativity and lower polarisability than S, Se and Te. When covalent bonding becomes prevalent it is known to result in deviations from tolerance factor behaviour. An example is the case of $NaSbO_3$, where $t = 0.92$ is commensurate with cubic perovskite formation, but which forms the non-perovskite ilmenite structure. Goodenough and Kafalas explained this deviation as a result of strong σ bonding between Sb and O.[55]

This procedure demonstrates the power of searching through materials based on structural analogy. Only a small fraction of possible perovskite materials have been synthesised. While some may not represent thermodynamic ground-states, they could be accessible through kinetic control of crystal growth or templated on a substrate. In particular, there are many interesting chalcogenide perovskites waiting to be discovered. The final pool of 13,415 feasible compositions is within the grasp of explicit computation using quantum mechanical methods, albeit as part of an ambitious project. Indeed, high-throughput screening of 5,400 multi-anion cubic perovskite structures using density functional theory has been reported,[29,56] which itself revealed 32 promising new materials for water splitting applications.

4.4.4 Conclusion

We have demonstrated the utility of chemical theory in quantifying the magnitude of the compositional space for multi-component inorganic materials. Even after the application of chemical filters, the space for four-component materials exceeds 10^{10} combinations. We further estimate the five-component space to exceed 10^{13} combinations. There are many applications where materials with even higher-order compositions have been developed, e.g. in high-temperature superconductors where 6 – 7 component materials

are common. The number of potential materials is not infinite, but it is immense. The scale of the combinatorial explosion emphasises the need for effective materials design procedures that employ existing chemical and physical knowledge in a targeted manner. Stochastic sampling of this chemical space is unlikely to be effective in yielding materials with specific functionality. We have presented a procedure that employs simple descriptors to support materials exploration, discovery and design.

4.4.5 Experimental procedures

4.4.5.1 Code implementation

The Python toolkit developed in this work, `smact`, is available online at `https://github.com/WMD-group/SMACT`. It is free software made available under the Gnu Public License (GPL) version 3.

All the element counts and plots presented in this paper were created with custom codes based on `smact` and written in the Python 2.7 programming language. Elemental data is collated from multiple sources (see Table 4.2), and made algorithmically accessible in a unified object orientated interface. Example routines are provided which check element/oxidation-state combinations against the conditions of charge-neutrality and electronegativity.

Scripts which generate the results and plots reported in this paper are made available with the `smact` codes. A number of tutorials working through the combinatorial explosion are provided at `https://github.com/WMD-group/SMACT_practical`.

The codes, collectively named Semiconducting Materials by Analogy and Chemical Theory, are inspired by the pen-and-paper procedure reported by B. R. Pamplin in 1964.[23]

Author contributions All authors contributed to the development of the `smact` package, while the primary coding was performed by KTB, AJJ and DWD. AW, DWD and KTB wrote the first draft manuscript with input, discussion and analysis from all co-authors.

4.4.6 Acknowledgements

We acknowledge the contributions of T. Gauntlett and J. Evans to the additions of Shannon radii and sustainability data in `smact`.

Funding: This research has been supported by the Royal Society, European Research Council (Grant No. 277757) and the EPSRC (EP/J017361/1, EP/K004956/1, EP/K016288/1, EP/L017792/1, EP/M009580/1, EP/G03768X/1).

Conflicts of interest: The authors declare that they have no competing interests.

Table 4.2: Data sources for `smact`. Where possible, values recommended by National Institute of Standards and Technology (NIST) are used.

Data type	Source
Abundance	Estimated crustal abundance of elements from the CRC Handbook of Physics and Chemistry[57]
Atomic mass	NIST Standard Reference Database 144,[58] where the relative abundance of isotopes is unknown or a range of values is provided, a simple mean was taken
Covalent radius	Scientific paper[59]
Electron affinity	Scientific paper,[60] no default value is used for elements which lack electron affinity data
Eigenvalues	Highest occupied p-state and s-state eigenvalues were tabulated by Harrison[61] from the approximate Hartree-Fock calculations of Herman & Skillman[62]
HHI	Elemental Herfindahl–Hirschman index calculated from geological and geopolitical data[37]
Ionisation potential	NIST Atomic Spectra Database[63]
Pauling electronegativity	Updated values of electronegativity on Pauling's scale were compiled in the CRC Handbook.[57] For elements 95 (Am) and above, Pauling's recommended value of 1.3 is employed.[64] The value for Krypton (3.0) was derived from the bond energy of KrF_2 and reported in a scientific paper[65]
SSE	"Solid-state energy" model of semiconductors and dielectrics[33,34]
SSE (Pauling)	Extended solid-state energy estimates from correlation between known values and Pauling electronegativity[34]

4.5 Remarks

By considering only non-zero oxidation states, the methodology outlined in this chapter automatically and deliberately ignores huge areas of chemical space such as intermetallics. Even so, the approach clearly provides us with a large enough composition space to use in a high-throughput screening process (see Figure 4.2). In addition to the conclusions drawn in the publication, some further comments can be made. Firstly, the `smact` code is able to enumerate the search space for a given set of elements – including the application of the charge neutrality and electronegativity order rules – on a desktop computer, in the times listed in Table 4.3. The quaternary space takes approximately 40 minutes to compute and the quinternary space takes a markedly longer time (over a week).

Table 4.3: Performance of the `smact` code for enumerating the composition space with stoichiometries $v, w, x, y, z \leq 8$. The code was run on an Apple iMac with a 4 GHz Intel Core i7 processor.

Order	Time
A_vB_w	0.1s
$A_vB_wC_x$	9.9s
$A_vB_wC_xD_y$	41m 48s
$A_vB_wC_xD_yE_z$	9d 9h 56m 12s

In earlier versions of `smact`, the same exercise took much longer, with the quaternary count taking well over two weeks. The current speed is only achievable thanks to several improvements made to the code:

1. The code makes use of the multiprocessing library in Python, which allows for approximately linear scaling and an n-fold speed up on an n-core workstation.
2. A caching system is implemented to avoid repeated, unnecessary file reads. Once the data has been read the first time, it is stored and read from memory. This is essential as many millions of `Species` objects are constructed, each with certain attributes.
3. Before an enumeration run, a list of all the different oxidation state combinations for the entire set of elements is constructed. These unique combinations are then used as keys in a lookup table which contains the number of charge neutral combinations that exist of that combination, for the stoichiometry limit in question. This is much faster than calculating the number of combinations possible each time, for each individual set of species.

One extra consideration is that, for a given element, the oxidation states deemed to be "accessible" is open to some interpretation. The inclusion of different oxidation states

will clearly impact on the compositions that are allowed (charge neutral) for a given set of elements. The default list of oxidation states in `smact` was constructed by the code authors and is designed to be as inclusive as possible of oxidation states that are stable under standard conditions. Other sets of oxidation states available in the code include all those which feature in the ICSD, as well only those that feature in the `Pymatgen` code, in order to maximise compatibility.

Choosing which set of oxidation states to include is problem specific and not always easy. As an example, it is widely accepted that in ionic compounds, K only exists as K^+, and that the only negative oxidation state of Cl is Cl^-. However, under extreme pressures, this is not the case, with phases such as K_3Cl, KCl_5 and KCl_7 being found as stable.[66] The topic of accessible oxidation states was of sufficient interest to investigate further and in the next chapter it is approached from a data-driven perspective.

Bibliography

[1] Q. D. Gibson *et al.*, *J. Am. Chem. Soc.*, 2017, **139**, 15568–15571.

[2] M. S. Dyer *et al.*, *Science*, 2013, **340**, 847–52.

[3] Y. Hinuma *et al.*, *Nat. Commun.*, 2016, **7**, 11962.

[4] V. M. Goldschmidt, *Naturwissenschaften*, 1926, **14**, 477–485.

[5] FIZ Karlsruhe, *Inorganic Crystal Structure Database*, `http://icsd.cds.rsc.org/-[Accessed:27-08-2017]`.

[6] *The Materials Project*, `https://materialsproject.org/-[Accessed:01-01-2016]`.

[7] S. P. Ong *et al.*, *Comput. Mater. Sci.*, 2013, **68**, 314–319.

[8] D. G. Pettifor, *J. Phys. C Solid State Phys.*, 1986, **19**, 285–313.

[9] R. G. Pearson, *Inorg. Chem.*, 1988, **27**, 734–740.

[10] L. Pauling, *J. Am. Chem. Soc.*, 1929, **51**, 1010–1026.

[11] J. Phillips, *Bonds and Bands in Semiconductors*, Academic Press, New York, 1973, p. 300.

[12] R. Gautier *et al.*, *Nat. Chem.*, 2015, **7**, 308–316.

[13] K. F. Garrity, K. M. Rabe and D. Vanderbilt, *Phys. Rev. Lett.*, 2014, **112**, 127601.

[14] G. Kieslich, S. Sun and T. Cheetham, *Chem. Sci.*, 2015, **6**, 3430–3433.

[15] W. Travis, E. N. K. Glover, H. Bronstein, D. O. Scanlon and R. G. Palgrave, *Chem. Sci.*, 2016, **7**, 4548–4556.

[16] K. Lejaeghere *et al.*, *Science*, 2016, **351**, 3000.

[17] K. T. Butler, Y. Kumagai, F. Oba and A. Walsh, *J. Mater. Chem. C*, 2016, **4**, 1149–1158.

[18] G. H. Booth, A. Grüneis, G. Kresse and A. Alavi, *Nature*, 2013, **493**, 365–370.

[19] F. Liu *et al.*, *Adv. Energy Mater.*, 2016, **6**, 1502206.

[20] G. Kieslich *et al.*, *Chem. Commun.*, 2015, **51**, 15538–15541.

[21] C. Jiang and B. P. Uberuaga, *Phys. Rev. Lett.*, 2016, **116**, 105501.

[22] C. Caetano, K. T. Butler and A. Walsh, *Phys. Rev. B*, 2016, **93**, 144205.

[23] B. Pamplin, *J. Phys. Chem. Solids*, 1964, **25**, 675–684.

[24] C. H. L. Goodman, *J. Phys. Chem. Solids*, 1958, **6**, 305–314.

[25] `https://www.materialsproject.org/-[Accessed:01.01.16]`.

[26] A. H. Nethercot, *Phys. Rev. Lett.*, 1974, **33**, 1088–1091.

[27] M. A. Butler and D. S. Ginley, *J. Electrochem. Soc.*, 1978, **125**, 228–232.

[28] Y. Xu and M. A. Schoonen, *Am. Mineral.*, 2000, **85**, 543–556.

[29] I. E. Castelli *et al.*, *Energy Environ. Sci.*, 2012, **5**, 9034–9043.

[30] L. A. Burton and A. Walsh, *J. Solid State Chem.*, 2012, **196**, 157–160.

[31] A. Walsh and K. T. Butler, *Acc. Chem. Res.*, 2014, **47**, 364–372.

[32] V. Stevanović, S. Lany, D. S. Ginley, W. Tumas and A. Zunger, *Phys. Chem. Chem. Phys.*, 2014, **16**, 3706–3714.

[33] B. D. Pelatt, R. Ravichandran, J. F. Wager and D. a. Keszler, *J. Am. Chem. Soc.*, 2011, **133**, 16852–16860.

[34] B. D. Pelatt *et al.*, *J. Solid State Chem.*, 2015, **231**, 138–144.

[35] T. Bak, J. Nowotny, M. Rekas and C. Sorrell, *Int. J. Hydrogen Energy*, 2002, **27**, 991–1022.

[36] B. A. Pinaud *et al.*, *Energy Environ. Sci.*, 2013, **6**, 1983–2002.

[37] M. W. Gaultois *et al.*, *Chem. Mater.*, 2013, **25**, 2911–2920.

[38] H. Bach and H. Küppers, *Acta Crystallogr. Sect. B*, 1978, **34**, 263–265.

[39] F. Thévet, Nguyen-Huy-Dung, C. Dagron and J. Flahaut, *J. Solid State Chem.*, 1976, **18**, 175–182.

[40] W.-T. Chen, H.-M. Kuang and H.-L. Chen, *J. Solid State Chem.*, 2010, **183**, 2411–2415.

[41] J. H. Harding and N. C. Pyper, *Philos. Mag. Lett.*, 1995, **71**, 113–121.

[42] A. R. Oganov, A. O. Lyakhov and M. Valle, *Acc. Chem. Res.*, 2011, **44**, 227–237.

[43] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain and G. Ceder, *Inorg. Chem.*, 2011, **50**, 656–663.

[44] J. Heyd and G. Scuseria, *J. Chem. Phys.*, 2004, **121**, 1187–1192.

[45] J. Heyd, G. E. Scuseria and M. Ernzerhof, *J. Chem. Phys.*, 2006, **124**, 219906.

[46] M. J. Wahila *et al.*, *Chem. Mater.*, 2016, **28**, 4706–4713.

[47] V. M. Goldschmidt, *J. Chem. Soc.*, 1937, 655–673.

[48] R. D. Shannon, *Acta Crystallogr. Sect. A*, 1976, **32**, 751–767.

[49] P. M. Woodward, *Acta Crystallogr. Sect. B Struct. Sci.*, 1997, **53**, 44–66.

[50] H. Hahn and U. Mutschke, *Zeitschrift fur Anorg. und Allg. Chemie*, 1957, **288**, 269–278.

[51] R. Lelieveld, D. J. W. IJdo and IUCr, *Acta Crystallogr. Sect. B Struct. Crystallogr. Cryst. Chem.*, 1980, **36**, 2223–2226.

[52] T. Nitta, K. Nagase and S. Hayakawa, *J. Am. Ceram. Soc.*, 1970, **53**, 601–604.

[53] Y.-Y. Sun, M. L. Agiorgousis, P. Zhang and S. Zhang, *Nano Lett.*, 2015, **15**, 581–585.

[54] J. W. Bennett, I. Grinberg and A. M. Rappe, *Phys. Rev. B*, 2009, **79**, 235115.

[55] J. B. Goodenough and J. A. Kafalas, *J. Solid State Chem.*, 1973, **6**, 493–501.

[56] I. E. Castelli, J. M. García-Lastra, F. Hüser, K. S. Thygesen and K. W. Jacobsen, *New J. Phys.*, 2013, **15**, 105026.

[57] W. M. Haynes, *CRC Handbook*, CRC Press, 92nd edn., 2011, p. 2656.

[58] J. S. Coursey, D. J. Schwab and J. J. Tsai, NIST Physical Measurement Laboratory, Standard Reference Database 144, `http://www.nist.gov/pml/data/comp.cfm-[Accessed12.04.16]`.

[59] B. Cordero *et al.*, *Dalt. Trans.*, 2008, 2832–2838.

[60] T. Andersen, H. K. Haugen and H. Hotop, *J. Phys. Chem. Ref. Data*, 1999, **28**, 1511–1533.

[61] W. A. Harrison, *Electronic Structure and the Properties of Solids*, Dover Publications Inc., New York, 1980.

[62] F. Herman and S. Skillman, *Atomic Structure Calculations*, Prentice Hall, New Jersey, 1963.

[63] A. Kramida, Yu. Ralchenko, J. Reader and and NIST ASD Team, NIST Atomic Spectra Database (ver. 5.3), [Online]. Available: `http://physics.nist.gov/asd` [2016, April 12]. National Institute of Standards and Technology, Gaithersburg, MD., 2015.

[64] L. Pauling, *The Nature of the Chemical Bond*, Cornell University Press, Ithaca, 3rd edn., 1960.

[65] L. C. Allen and J. E. Huheey, *J. Inorg. Nucl. Chem.*, 1980, **42**, 1523–1524.

[66] W. Zhang *et al.*, *Sci. Rep.*, 2016, **6**, 26265.

Chapter 5

Probabilistic Oxidation States Model

5.1 Introduction

The method for building up a composition space described in the previous chapter relies on combining species in charge neutral stoichiometries, where a species is an element with an associated oxidation state. Intuitively, it might be expected that some combinations are not likely to coexist in the same material. For example, some transition metals might only adopt high oxidation states in the presence of a sufficiently electronegative anion. There may also be geometric constraints; given that only so many anions can be packed around a certain cation in a crystal lattice, it may be that singly charged anions (e.g. halides) are not present in sufficient quantity for the cation to access the highest oxidation states.

In this chapter, the aim is to construct a model that quantifies the probability of a given combination of species, based on existing data. This model can then be used, along with a probability threshold, to reduce the size of the composition space at low computational cost. In order to assign oxidation states to elements in a large dataset of structures, we carry out bond valence analysis, as described by Brese and O'Keeffe.[1] For each symmetry inequivalent atom i we calculate the sum (BV_{sum}) of all the values of bond valence BV_{ij} between atom i and its surrounding atoms j (see Equations 5.6 and 5.7). A maximum *a posteriori* (MAP) method is then used to predict the oxidation states at each site using the values of BV_{sum}. The posterior probabilities of oxidation states O are calculated as:

$$P(O|BV_{sum}) = P(BV_{sum}|O)P(O) \tag{5.1}$$

where $P(BV_{sum}|O)$ is a Gaussian with a mean μ and standard deviation σ determined by analysis of the ICSD:

$$P(BV_{sum}|O) = \frac{1}{\sigma} e^{-\frac{(BV_{sum}-\mu)^2}{2\sigma^2}} \tag{5.2}$$

and the prior distribution $P(O)$ is simply the frequency of the species in the ICSD. The oxidation states for each site in the structure are then ranked in order of probability and the most likely combination of oxidation states that is charge neutral is selected. The parameters μ, σ and $P(O)$ are part of the implementation built into the `Pymatgen` package, and the analysis of the ICSD to generate them was not repeated as part of this work.

While this method is robust in the majority of cases, it is possible that certain unusual anion oxidation states could lead to incorrect cation assignments. An attempt to add oxidation states to metal peroxide compounds present in the MP database led the algorithm to conclude that it could not assign oxidation states in all cases. This is promising, showing that oxygen was not being erroneously assigned a -2 charge which would cause incorrect oxidation states to be assigned to metals for charge balance. However, a more complete investigation into the robustness of this algorithm towards other unconventional chemistries will constitute important future work.

5.2 Statement of Authorship

The following paper entitled *Materials Discovery by Chemical Analogy: Role of Oxidation States in Structure Prediction* reports on original research I conducted during the period of my Higher Degree by Research candidature.

Personal contributions: *Formulation of ideas (80%):* After the initial conception of the idea of the study by Prof. Aron Walsh, I have made the majority of decisions relating to the development of the project with guidance from Dr Keith Butler. *Design of methodology (90%):* I wrote all of the code enabling this study that has now been added to the `smact` package. *Experimental work (90%):* I carried out the statistical analysis, including building the probabilistic model, and all high-throughput DFT calculations. The DFT total energies of all ICSD compounds according to the AFLOW-ML model were generated and supplied by Dr Olexandr Isayev. *Presentation of data in journal format (80%):* The first drafts of the manuscript were written by me, with input from Dr Keith Butler and Prof. Aron Walsh at each stage of revision. The finalised manuscript was prepared by Prof. Aron Walsh and me. I presented the work at the RSC Faraday Discussion entitled *Methods and applications of crystal structure prediction* (Cambridge, July 2018).

5.3 Access statement

5.4 Publication 2

Materials Discovery by Chemical Analogy: Role of Oxidation States in Structure Prediction

Daniel W. Davies,[1] Keith T. Butler,[1] Olexandr Isayev,[2] Aron Walsh[1,3,4]

1. Centre for Sustainable Chemical Technologies and Department of Chemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK

2. Laboratory of Molecular Modelling, Division of Chemical Biological and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599, USA

3. Global E^3 Institute and Department of Materials Science and Engineering, Yonsei University, Seoul 120-749, Korea

4. Department of Materials, Imperial College London, Exhibition Road, London SW7 2AZ, UK

5.4.1 Abstract

The likelihood of an element to adopt a specific oxidation state in a solid, given a certain set of neighbours, might often be obvious to a trained chemist. However, encoding this information for use in high-throughput searches presents a significant challenge. We carry out a statistical analysis of the occurrence of oxidation states in 16,735 ordered, inorganic compounds and show that a large number of cations are only likely to exhibit certain oxidation states in combination with particular anions. We use this data to build a model that ascribes probabilities to the formation of hypothetical compounds, given the proposed oxidation states of its constituent species. The model is then used as part of a high-throughput materials design process, which significantly narrows down the vast compositional search space for new ternary metal halide compounds. Finally, we employ a machine learning analysis of existing compounds to suggest likely structures for a small subset of the candidate compositions. We predict two new compounds, $MnZnBr_4$ and $YSnF_7$, that are thermodynamically stable according to density functional theory, as well as four compounds, $MnCdBr_4$, $MnRu_2Br_8$, $ScZnF_5$ and $ZnCoBr_4$, which lie within the window of metastability.

5.4.2 Introduction

The idea of ascribing an oxidation state to a metal can be traced back almost 200 years.[2] As the phrase suggests, it was used to describe the amount of oxygen bound to an element that is known to form multiple oxides. Since then, oxidation states have helped in the formation of many fundamental chemistry concepts. For example, a plot of the periodicity of accessible oxidation states (Figure 5.1) by Irving Langmuir was one of the key pieces of evidence that led to the adoption of the octet rule around 100 years ago.[3] The English term itself, "oxidation state" (or equally "oxidation number"), first came into common use in the realm of electrochemistry in the 1930s,[4] and in the 1940s it had gained widespread use[5] to replace the less-than-perfect system of appending *-ous* and *-ic* to the lower and higher oxidation states of metals, respectively. Ferrous became Fe(II), ferric became Fe(III), and transition metals with more than two oxidation states could now be unambiguously described. The term has remained an indispensable heuristic tool in almost all sub-disciplines of the physical sciences. It is integral to the way in which chemists think about the interaction of elements within molecules and solids.

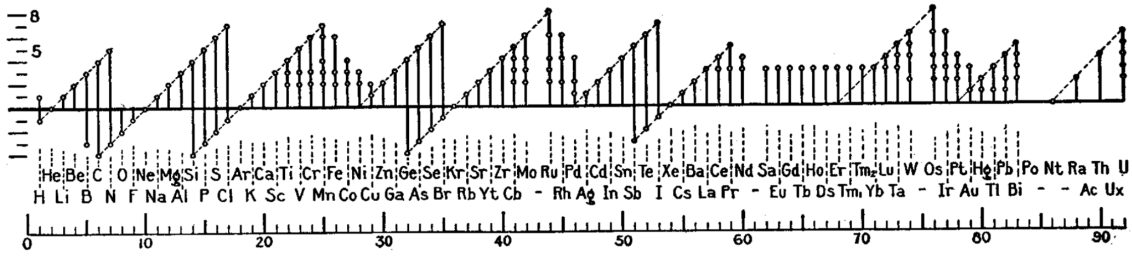

Figure 5.1: Plot of accessible oxidation states reproduced from a 100 year old paper by Irving Langmuir[3] on the octet rule.

Linus Pauling first postulated that oxidation states could be determined by approximating bonds as 100% ionic according to the electronegativities of the elements involved.[6] This simple approach did not initially gain acceptance as the use of Pauling's electronegativity scale[7] resulted in some unusual assignments. Nevertheless, his approach is reflected in the modern definition given by IUPAC: *"An atom's charge after ionic approximation of its heteronuclear bonds"*.[8] In practice, knowledge of the chemical formula is sufficient to assign formal oxidation states in many inorganic compounds; however, there are cases were ambiguities exist (e.g. mixed-valence compounds, electrides, polyanions and polycations). As highlighted in a recent essay by Karen, the "atom's charge", its "heteronuclear bonds" and the "ionic approximation" are all terms that need clarification, and there are choices to be made about how each is defined.[9]

The subtlety of assigning oxidation states is still the subject of many lively discussions in both pedagogical and research contexts.[10–14] For practical purposes, we emphasise the insight of Jansen and Wedig, who point out that:

> *"It is a purely formal concept; nowhere within the definition is it claimed that a particular oxidation state can be associated with a real charge. Nevertheless, the term is certainly useful, since a specific oxidation state can be correlated to real properties."*[15]

It is this correlation to real properties that useful in a materials design context. Oxidation states have had a role to play in materials design for many decades. In the 1950's and 1960's, Goodman and Pamplin were able to systematically and exhaustively design superlattices of multicomponent compounds by substitution of the cations in simple binary semiconductors, while ensuring the octet rule remains satisfied.[16,17] This cation substitution (mutation) concept continues to inspire modern computational work on semiconductor design.[18,19]

Knowledge of accessible oxidation states for each element is advantageous because we can generate many stoichiometric combinations while ensuring that there is overall charge neutrality. For example, the formal oxidation states q of any ternary combination $\mathrm{A}_x\mathrm{B}_y\mathrm{C}_z$ must sum to zero:

$$xq_{\mathrm{A}} + yq_{\mathrm{B}} + zq_{\mathrm{C}} = 0. \tag{5.3}$$

We have previously demonstrated that many chemically plausible formulas can be generated in this way.[20] For example, if the stoichiometry values (x, y and z in the above equation) are limited to integers ≤ 8, the search space for ternary combinations exceeds 1×10^8, and for quaternary combinations it is over 2×10^{11}. The resulting formulas can be fed into a high-throughput screening workflow that uses machine learning structure prediction models to screen for new functional materials.[21]

In this study, we first carry out a statistical analysis on the occurrence of oxidation states in 16,735 stoichiometric, inorganic compounds in order to highlight trends and show that many elements only exhibit certain oxidation states in the presence of particular elements. We then go on to construct a screening model based on this data and apply it to the search space for ternary transition metal halides. The model we propose can be used as a general chemical filter when dealing with large composition search spaces, in order to remove those combinations of elements that are unlikely to form stable compounds. For example, we find that many transition metals are only likely to adopt their highest accessible

oxidation states in the presence of sufficiently electronegative anions.

5.4.3 Results

5.4.3.1 Data curation

We focus on the variation of oxidation states of metals in the presence of common anions. The anions we include are the first four group VI and VII elements in their most common oxidation states, i.e.: O^{2-}, S^{2-}, Se^{2-}, Te^{2-}, F^{-}, Cl^{-}, Br^{-}, I^{-}. These provide a reasonable range of electronegativities (Table 5.1) and as such we do not include group V anions. This also allows us to avoid the metalloids As and Sb. The compounds included in the dataset originate from the Inorganic Crystal Structure Database (ICSD) and are downloaded from the Materials Project (MP)[22] using their API.[23] Full details of how the dataset was refined can be found in the Methods section. In broad terms, all the compounds meet the following criteria (total number of compounds remaining in the dataset shown in brackets):

1. Feature in both the ICSD and MP databases (34,913)
2. Calculated to be less than 100 meV/atom above the thermodynamic convex hull by the MP (30,781)
3. Oxidation states of all elements can be determined automatically using a bond valence analysis algorithm* (24,376)
4. Contain at least one anion (as defined above) and at least one metal (16,735)

Figure 5.2 shows the resulting metals that are included after this refinement has been applied. In total, 16,735 different compounds are included.

5.4.3.2 Occurrence of oxidation states

In the first instance, we examine the occurrence of metal oxidation states as a function of the most electronegative anion present in each compound (see Table 5.1). In each case, we normalise by the total number of compounds containing a given species (metal in a given oxidation state), i.e., we look at how the total number of instances of each species

*This rules out those elements that were not included in the original study which proposed the algorithm[24] as well as intermetallic compounds.

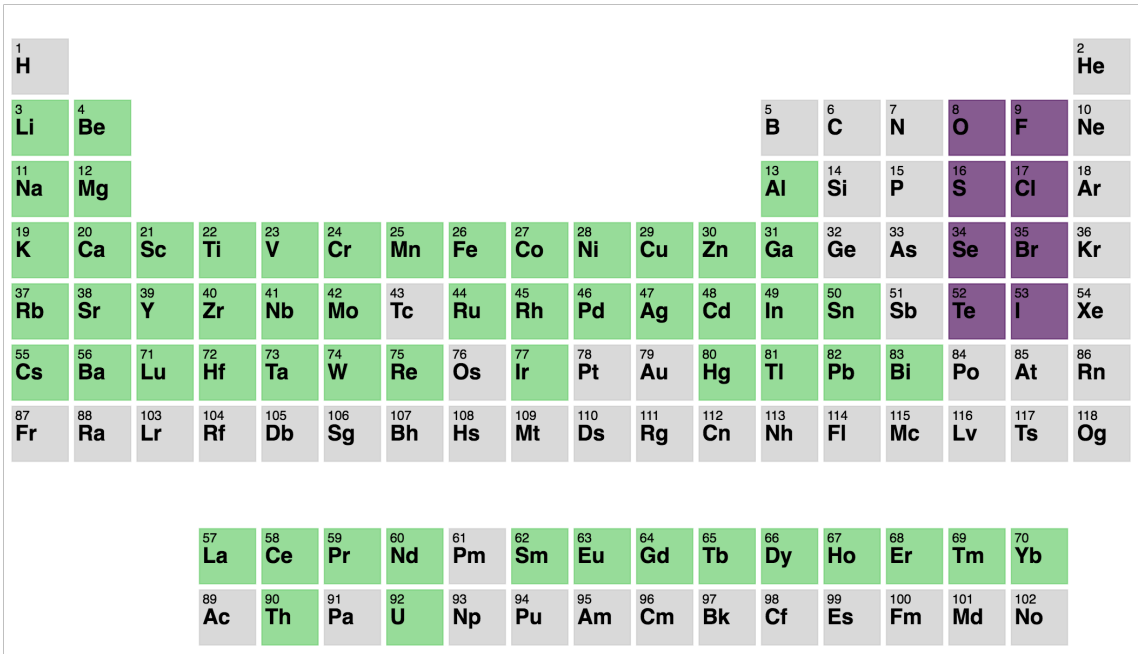

Figure 5.2: Periodic table illustrating the metals included in our statistical analysis (green) and the anions considered (purple).

Table 5.1: Anion electronegativities (χ) and number of compounds in which each anion is the most electronegative element.

Anion	χ	Occurrence
F	3.98	1,759
O	3.44	10,546
Cl	3.16	924
Br	2.96	444
I	2.66	499
S	2.58	1,489
Se	2.55	759
Te	2.10	320

is distributed across the compounds. This is given by the ratio $\frac{N_{SX}}{N_S}$, where N_{SX} is the number of compounds containing the species S where the most electronegative anion is X, and N_S is the total number of compounds containing the species S. These values are shown graphically for all species in the Supplementary Information.

Transition metals have the largest number of accessible oxidation states. Figure 5.3 shows the distribution of some first-row d-block species. Each of these exhibits the same general trend: The likelihood of finding a metal in a higher oxidation state increases when a more electronegative anion is present in the compound (increase in the relative heights of red bars in Figure 5.3). Meanwhile, metals are more likely to exhibit low oxidation states when the most electronegative anion present is of low electronegativity. More specific trends can also be extracted. For example, the higher oxidation states of Mn (Mn^{5+} – Mn^{7+}) are exclusively exhibited in oxides. This is also the case for Cr^{6+}, while Cr^{5+} is limited to oxides and fluorides.

For higher oxidation states, the likelihood of finding the metal with an anion of moderate electronegativity, such as Cl^-, Br^- and I^-, often goes to zero before the likelihood of finding it with an anion of low electronegativity, such as S^{2-}, Se^{2-}, and Te^{2-}. This is a trend that may not necessarily be expected, for example, going from V^{2+} to V^{4+}. It is also important to mention at this stage that oxides nearly always dominate each distribution as 10,546 of the 16,735 compounds contain oxygen. This point is addressed later when using the data predictively, in order to minimise bias.

Figure 5.4 shows a similar trend in the distribution of some second-row d-block species. Again, compounds containing lower electronegativity anions, in the absence of any higher electronegativity anions, are more likely to contain lower oxidation state metal species. For the highest oxidation states of Ru (Ru^{5+} and Ru^{6+}), the presence of F^- or O^{2-} is required.

The distribution of oxidation states is more even across moderate and low electronegativity anions for second-row transition metals compared to the first row. This is consistent with established principles of chemical hardness,[†] as applied to inorganic compounds by Pearson.[26] The order of chemical hardness for the halides is $F^- > Cl^- > Br^- > I^-$ and, in general, the halide anions are harder than the chalcogenide anions, which is consistent with the electronegativity ordering in Table 5.1. For cations, hardness increases with increasing charge. The species in the second row have consistently lower chemical hardness than the corresponding species above in the periodic table with the same oxidation state, so it should be expected that they form more compounds with softer halides.

[†]Chemical hardness is estimated by $\frac{I-A}{2}$ where I is the ionisation potential and A the electron affinity. This represents half the energy gap between the highest occupied orbital and lowest unoccupied orbital. Absolute electronegativity,[25] distinct from Pauling's definition,[7] is defined as $-\frac{I+A}{2}$ and represents the midpoint between the two orbitals.

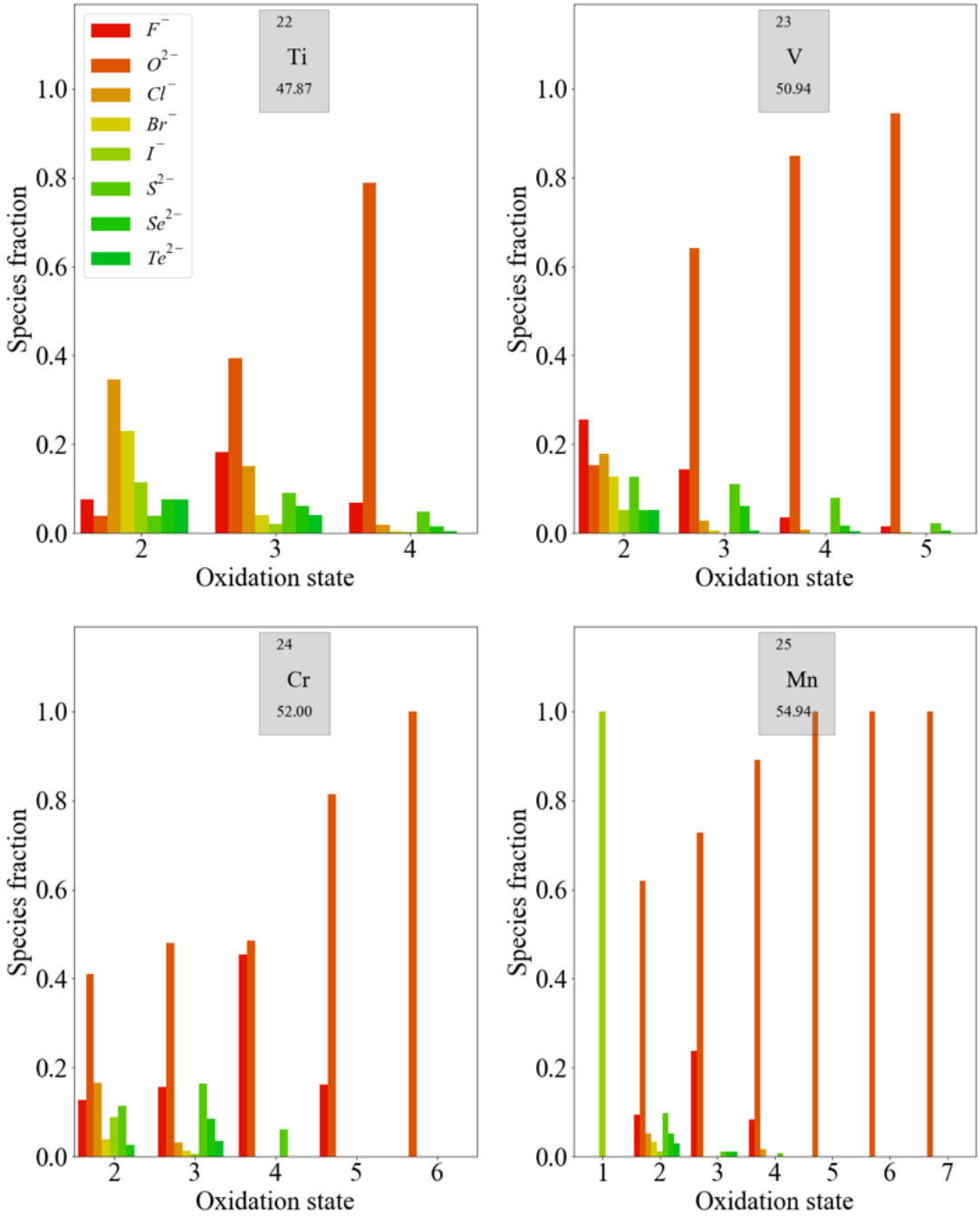

Figure 5.3: Distribution of oxidation states in known inorganic crystals containing some first row transition-metal species. The color scale represents the electronegativity of the most electronegative anion present in the compound from dark red (F, most electronegative) to dark green (Te, least electronegative).

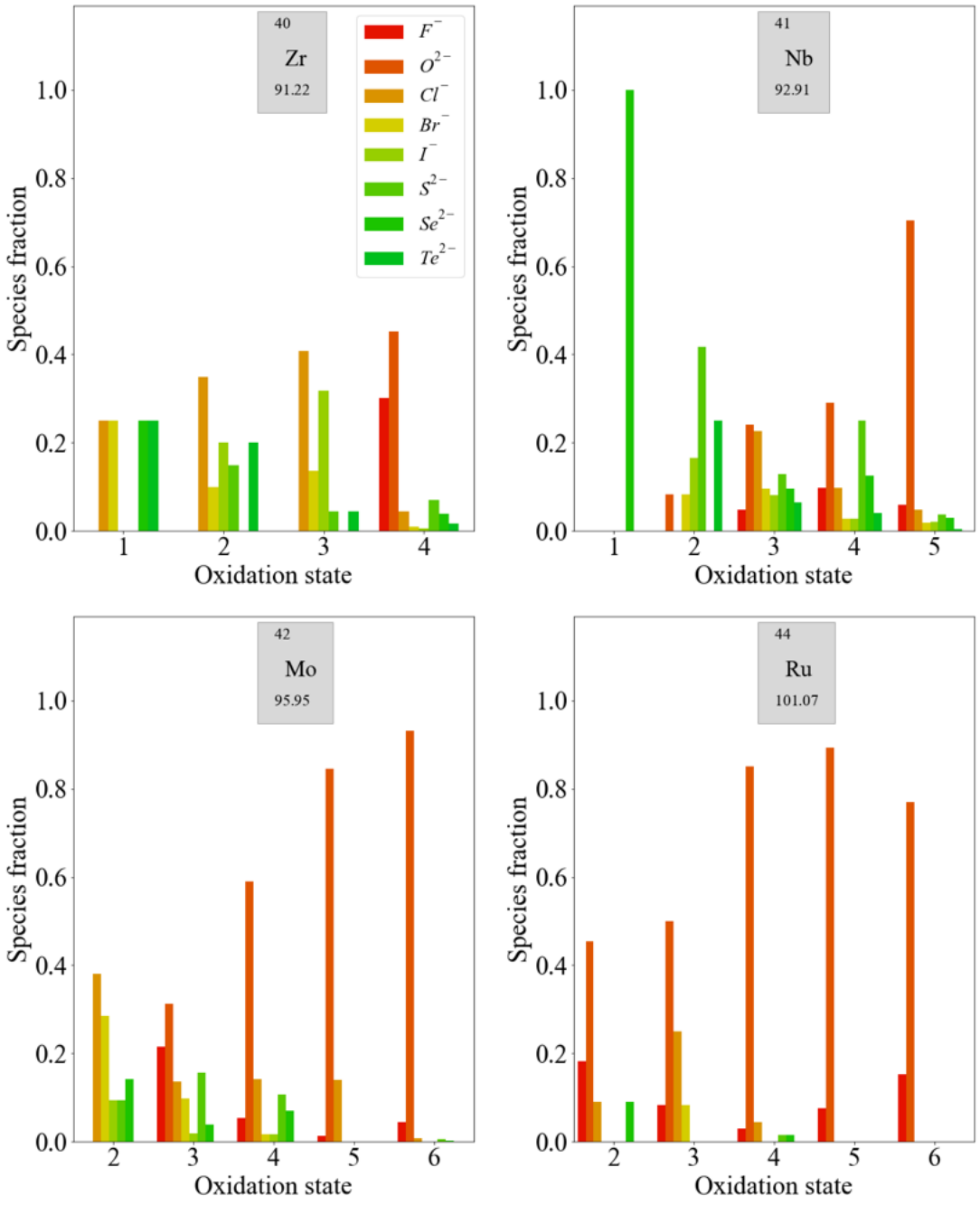

Figure 5.4: Distribution of oxidation states in known inorganic crystals containing some second-row transition-metal species. The color scale represents the electronegativity of the most electronegative anion present in the compound from dark red (F, most electronegative) to dark green (Te, least electronegative).

For p-block metals, the trends are less pronounced. As shown in Figure 5.5, F^- and O^{2-} containing compounds are likely to exhibit the higher oxidation states of the metals. Moving from low to high oxidation states, there is less of a reduction in the fraction of compounds containing lower electronegativity anions compared with the transition metal compounds discussed so far. The reduction in the fraction of compounds containing moderate electronegativity anions is more pronounced for these metals. The general observation from this data that the oxidation states of these metals are more weakly correlated to the electronegativity of the counter-ions than transition metals, is expected based on the fact that transition metals have multiple, readily accessible oxidation states by virtue of their partially occupied d-bands. This is not the case for p-block metals, for which adding or removing electrons results in more significant energy differences.

The third row transition metals and lanthanide series display similar trends to the first and second row transition metal series (see Supplementary Information). For completeness, we note that the alkali and alkali-earth metals only exhibit +1 and +2 oxidation states, respectively, for the vast majority of compounds. Similarly, Sc, Y, Zn and Cd are usually not strictly classified as transition metals as there is a strong energetic preference for them to adopt the oxidation states that lead to empty (Sc^{3+}, Y^{3+}) or filled (Zn^{2+}, Cd^{2+}) valence d-orbitals, not partially filled as the definition dictates. We also note that later d-block metals (Ni, Cu, Pd, Ag) do not exhibit trends as clear as those for the rest of the d-block. This is due to similar effects as above, whereby particular closed (or pseudo-closed) shell configurations are favourable, for example, the d^8 electronic configuration of Ni^{2+} and Pd^{2+}.

The abundance or scarcity of particular species–anion pairings in this dataset may not always reflect what is chemically accessible. Even assuming that the dataset is sufficiently diverse, heightened interest in particular compounds or compound classes can result in their over-representation, which is a general problem in data mining. Nevertheless, we have shown analysis of the dataset both recovers established chemical concepts and provides new insights. We now go on to develop a simple model that can be universally applied based on the dataset as a whole.

5.4.3.3 Probabilistic model of species combinations

There are more compounds where O is the most electronegative anion present than any other anion as shown in Table 5.1. To use the information from our analysis, we must ensure that the occurrence of each anion does not bias the results. To this end, we define

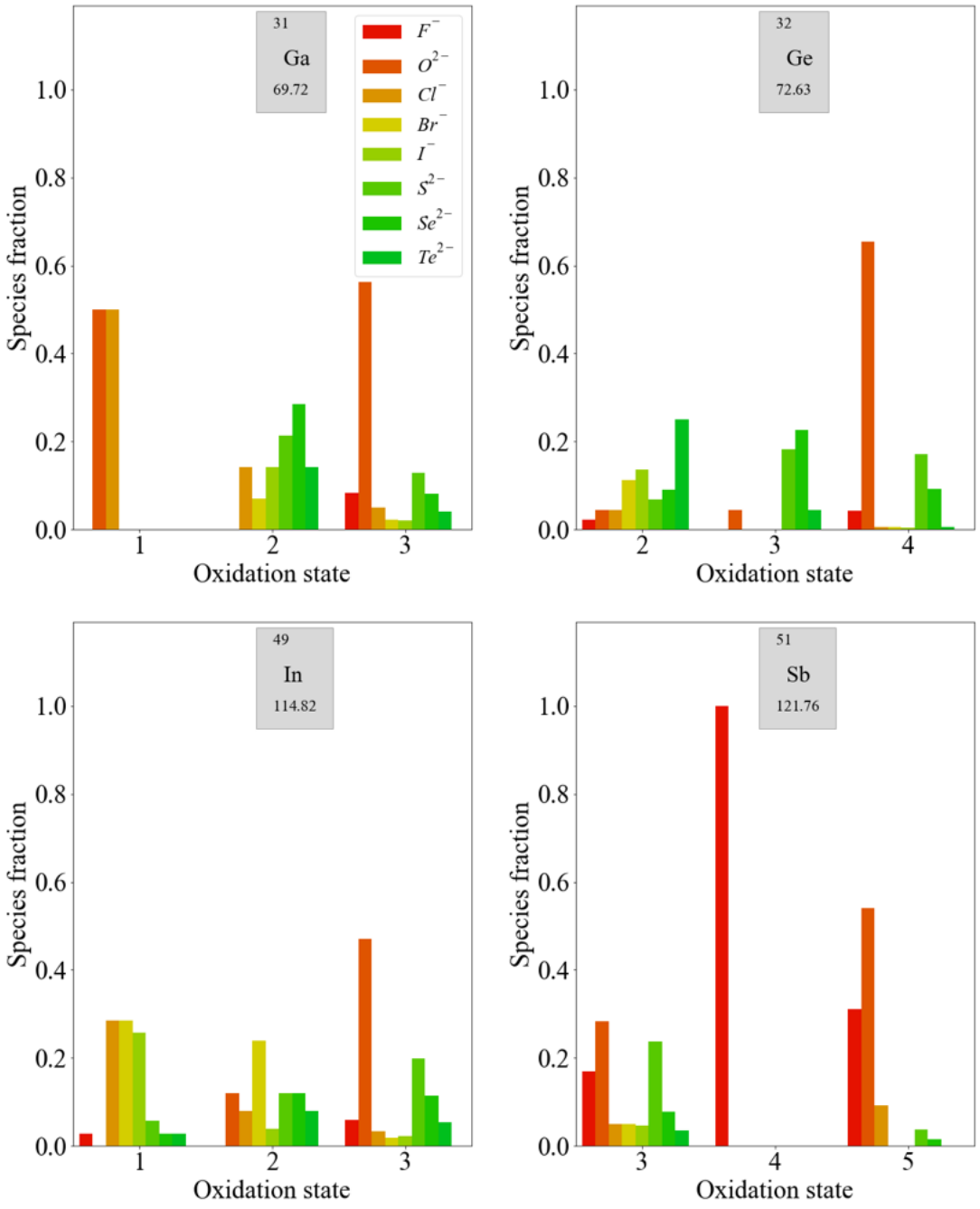

Figure 5.5: Distribution of oxidation states in known inorganic crystals containing some p-block species. The color scale represents the electronegativity of the most electronegative anion present in the compound from dark red (F, most electronegative) to dark green (Te, least electronegative).

the *probability* that a species is present with a given anion as:

$$P_{SA} = \frac{N_{SX}}{N_{MX}} \tag{5.4}$$

where N_{MX} is the total number of compounds containing the metal *element* where X is the most electronegative anion.

We use this formula to construct a lookup table of 1,320 species–anion pair probabilities (P_{SA}). The table contains 411 probabilities that equal 0, and 195 probabilities that equal 1. The former represent all the pairings that do not occur within the dataset and the latter represent pairings whereby for a given anion, the metal only exhibits one particular oxidation state. The P_{SA} values are also presented graphically in the Supplementary Information. We note that this still does not mitigate against limitations that are intrinsic to the dataset. For example, there are over 100 distinct CdI_2 crystal structures in the dataset (owing to the large number of distinct polymorphs) giving rise to an anomalously high probability for the Cd^{2+}–I^- pairing.

An overall *compound probability* can be calculated as the product of the individual P_{SA} values. For example, for a ternary metal halide $A_aB_bX_x$, the compound probability is calculated as:

$$P_{A_aB_bX_x} = P_{AX}P_{BX} = \frac{N_{AX}}{N_{M_AX}} \times \frac{N_{BX}}{N_{M_BX}} \tag{5.5}$$

where M_A and M_B are the metal elements corresponding to species A and B. Stoichiometries are not factored in to the probability calculation, such that $P_{A_aB_bX_x} = P_{ABX}$. This ensures that compounds featuring elements that all have only one oxidation state are assigned a probability of 1.0. For example, Ca^{2+} and Al^{3+} are the only species in the database of Ca and Al, hence $P_{CaAl_2O_4} = 1.0$. The number of compounds in the dataset that have compound probabilities above a given threshold, *t*, is shown in Figure 5.6. The number decreases steadily and linearly, before dropping off more rapidly as the threshold becomes more strict.

5.4.3.4 High-throughput compound design

We now use these compound probabilities to inform a high-throughput design workflow. Specifically, we explore the compositional landscape for ternary transition metal halide compounds. An overview of the workflow is shown in Figure 5.7. The `smact` code[20] is used to generate 54,484 $A_aB_bX_x$ compositions. Of these only 4,276 are in known chemical systems (A-B-X) within the MP database. The compositions are assigned probabilities as

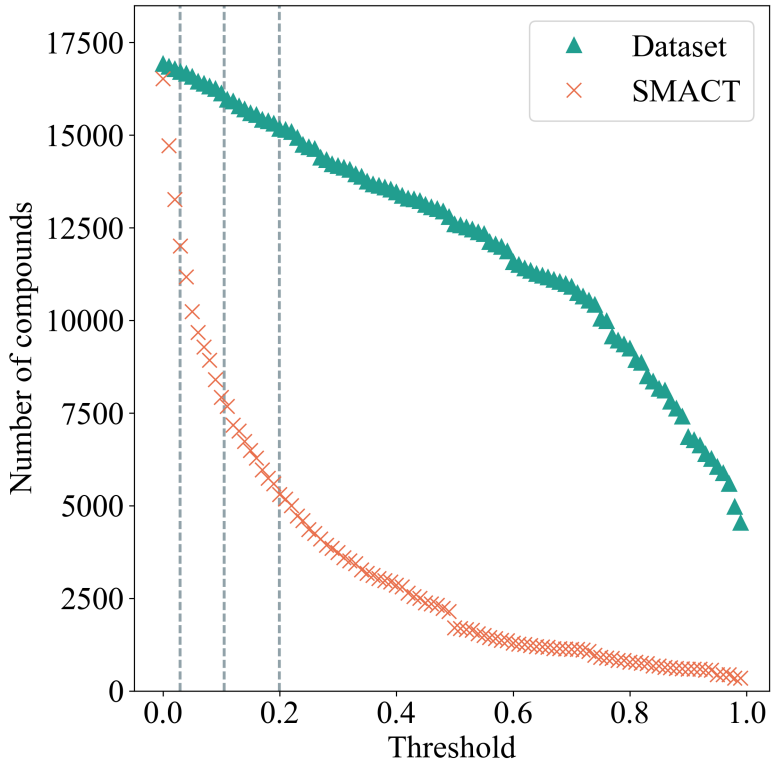

Figure 5.6: Total number of allowed compounds from the entire dataset (green triangles) and of allowed compositions *for ternary metal halides only* generated by `smact` (red crosses) as a function of compound probability threshold, *t*. Dotted vertical lines represent cut-offs that return 99%, 95% and 90% of the original dataset.

per Equation 5.5, and only 18,164 are non-zero, which represents an immediate three-fold reduction in the search space.

The number of compositions produced by `smact` that pass through this probability filter as the threshold, *t*, is increased from zero is also shown in Figure 5.6. Many compositions have low probabilities, hence, contrary to the scenario for the compound dataset, the total number drops off rapidly as the threshold increases. This separation between the two curves would, in principle, allow for a threshold to be chosen that eliminates many suggested structures but is still inclusive of the majority of the structures in the dataset. For example, choosing a threshold that includes 90% of the structures in the dataset results in a further three-fold reduction of the search-space to $<$ 6,000 compositions.

If we set a probability threshold of $t = 1$, there are 346 compositions that pass through the filter and this equates to 88 distinct sets of three species. In order to demonstrate the rest of our workflow, we take 10 of these sets (Table 5.2) to the next step, which is to

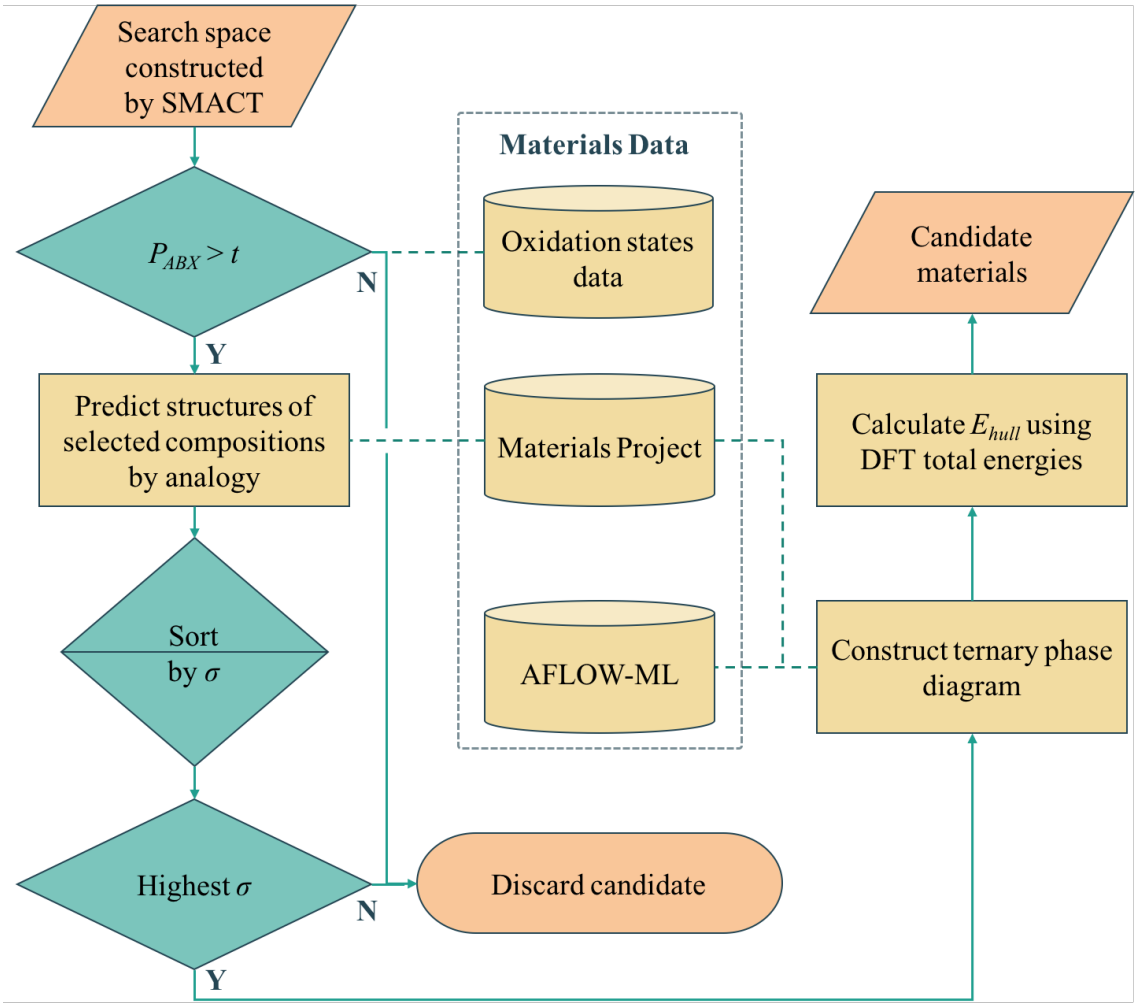

Figure 5.7: Data-driven design workflow used to generate new stable compounds. P_{ABX} is the compound probability from oxidation states analysis, which must be greater than the threshold, t. The structure prediction procedure has a separate threshold, σ. The structure with the highest σ is placed onto a phase diagram constructed using compounds from the MP database, and corresponding energies from the AFLOW-ML approach. Density function theory (DFT) is used to calculate the total energies of competing phases in order to determine the energy above the convex hull, E_{hull}, of the new compound.

assign them to likely crystal structures (first yellow box in Figure 5.7). For this, we adopt the structure substitution algorithm developed by Hautier *et al.*[27] This method also uses a statistical model and relies on a database of known compounds including oxidation state information: A combination of species is substituted onto lattice sites in known structures from the dataset of known materials. Each species substitution is associated with a certain probability, which comes from a model trained on the compounds that already exist in the ICSD. If the overall probability for a given set of substitutions is above a certain threshold, σ, it is added to a list of possible structures. This substitution process is performed, for each set of species, on each known crystal structure in the MP database. The structure with the highest overall probability for each of the set of 10 species is taken forward to the next stage of the workflow (second and third green diamonds in Figure 5.7). These are listed in Table 5.2 along with their parent compounds.

Table 5.2: Energy above the convex hull (E_{hull}) of proposed compounds along with the chemical formula of the parent compound found by the structure predictor algorithm for each composition.

Species set	Formula	Parent formula	E_{hull} (meV/atom)
Co^{2+} Ru^{3+} Br^{-}	$CoRu_2Br_8$	$TiAl_2Cl_8$	287
Mn^{2+} Cd^{2+} Br^{-}	$MnCdBr_4$	$CdCuF_4$	99.5
Mn^{2+} Co^{2+} Br^{-}	$MnCoBr_4$	$CdCuF_4$	130
Mn^{2+} Ru^{3+} Br^{-}	$MnRu_2Br_8$	$TiAl_2Br_8$	73.2
Mn^{2+} Zn^{2+} Br^{-}	**$MnZnBr_4$**	**$GaCuI_4$**	**0**
Sc^{3+} Zn^{2+} F^{-}	$ScZnF_5$	$MnCdF_5$	48.3
Y^{3+} Co^{2+} I^{-}	Y_2CoI_8	$TiAl_2Br_8$	181
Y^{3+} Zr^{4+} F^{-}	**$YZrF_7$**	**$YSnF_7$**	**0**
Zn^{2+} Cd^{2+} Cl^{-}	$ZnCd_2Cl_6$	$ZnPb_2F_6$	132
Zn^{2+} Co^{2+} Br^{-}	$ZnCoBr_4$	$CdCuF_4$	40.4

Each structure is placed on a phase diagram in order to determine likely competing phases, which requires total energies as calculated using DFT. The key quantity of interest is the energy above the convex hull (E_{hull}) that is formed by drawing straight lines between thermodynamically stable phases. It was recently estimated by Sun *et al.* that around half of all known inorganic materials are metastable,[28] so to focus solely on thermodynamically stable compounds would be to potentially overlook kinetically stabilised, useful materials. The likelihood of existence drops off exponentially as E_{hull} increases. The rate of decay depends on the chemistry of the system and we use 100 meV/atom as a guiding principle for the maximum E_{hull}. The set of competing phases on which DFT calculations were performed was determined using a trained machine learning model (AFLOW-ML) in which structures are represented as property labelled fragments.[29]

This stage reveals a key advantage of pursuing only those compositions with higher probabilities based on the oxidation states analysis: the parent binary compounds are well de-

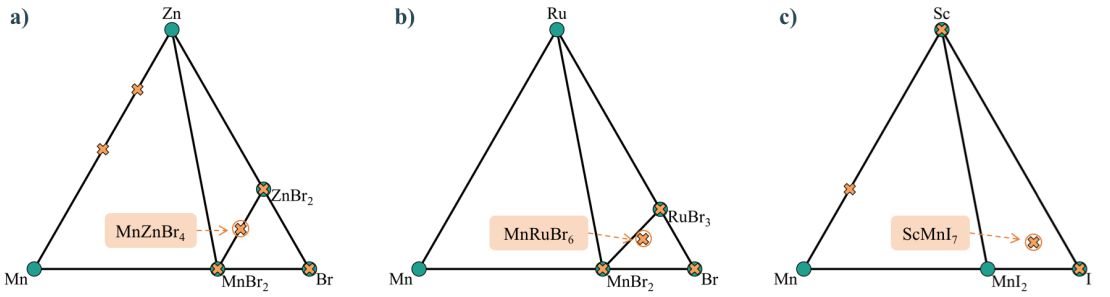

Figure 5.8: Ternary phase diagrams of the hypothetical compositions a) $MnZnBr_4$, b) $MnRuBr_6$ and c) $ScMnI_7$. Stable phases (green circles) are connected to form the convex and unstable or proposed phases (orange crosses) sit above the convex hull.

fined. Competing binary compounds exhibiting the metals in the same oxidation states as in the ternary (or multernary) compound are more likely to be known and amenable to total energy calculations to determine phase stability. Arbitrary combinations of species can result in stoichiometries that require energies of competitive gas or liquid phases which are subject to larger errors in DFT simulations. Figure 5.8 illustrates this point with a comparison between the phase diagram of three proposed ternary compositions. $MnZnBr_4$ has a probability of 1.0 as both $MnBr_2$ and $ZnBr_2$ are known and these decomposition products do not require a change of oxidation state of either metal. The ternary, therefore sits on the tie-line between the two binaries. The proposed compositions $MnRuBr_6$ and $ScMnI_7$, however, both have probabilities of zero, as in each case one or more species–anion pair is not known to occur. These compositions sit in an equilibrium triangle as opposed to on a tie line, and the stability of the proposed compounds now depend on the chemical potential of the anion.

Final E_{hull} values are shown in Table 5.2. Two of the new compounds, $MnZnBr_4$ and $YZrF_7$, are predicted to be thermodynamically stable with respect to competing phases. Of the remaining eight compounds, four sit within the metastability window of $0 < E_{hull} < 100$ meV/atom, while four are unlikely to form stable compounds. The crystal structures of the two compounds identified as stable are shown in Figure 5.9. By comparison with previous work where similar a workflow was employed,[21] this result provisionally indicates that the additional step of considering compound probabilities based on our oxidation states analysis increases the chance of identifying stable compounds.

The main limitation of the procedure outlined here is that it is based on analysis of known materials with extrapolation to new systems. This assumes that the range of structure types and chemistries found in current materials databases provide a complete basis for materials design. While this is a reasonable starting point, advances in materials synthesis

- for example in the area of hybrid organic-inorganic solids - will require adaptations and the development of alternative approaches. We have noted that there are many instances were oxidation states themselves become ill-defined, which often is associated with interesting and important physical behaviour (e.g. superconductivity). Before tackling such challenge cases, we have highlighted[20] that a vast amount of "conventional" materials space remains unexplored.

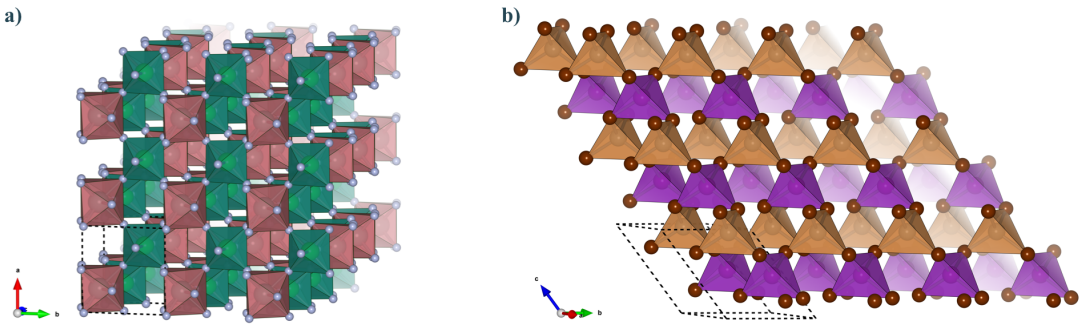

Figure 5.9: Two new stable ternary metal halides predicted using this workflow. a) $YZrF_7$ consists of vertex sharing irregular polyhedra of YF_8 (red) and octahedra of ZrF_6 (green). b) $MnZnBr_4$ consists of vertex sharing $ZnBr_4$ (orange) and $MnBr_4$ (purple) with both metals in a tetrahedral coordination environment.

5.4.4 Conclusion

We have performed a statistical analysis of the occurrence of oxidation states in 16,735 inorganic compounds and shown that qualitative trends in keeping with chemical intuition can be extracted from the data. Many of the highest oxidation states of transition metals are only observed in the presence of the most electronegative anions, O^- and F^-, whilst an absence of these anions are required for many of the lower oxidation states of transition metals. We go on to use the data to construct a model that is applied to inform a high-throughput search for new stable ternary halide materials. The application of the model results in an immediate three-fold reduction in the search space of 54,484 compositions. The search space is reduced to those compositions which are more likely to have known chemically similar compounds as competing phases, such as binary halides, thereby increasing the confidence we have in their calculated stability. Our workflow is able to identify two new stable compounds, $YZrF_7$ and $MnZnBr_4$, using modest computing resources.

5.4.5 Methods

5.4.5.1 Dataset

The MP API[23] is used to download the structures of all the compounds that are associated with at least one ICSD entry and with a calculated energy above the hull of < 100 meV/atom. An attempt is made to add oxidation states to all species in each structure using the `pymatgen`[30] *bond_valence* module (See oxidation state assignment subsection for details). Those compounds for which oxidation states cannot be assigned are discarded. Finally, the dataset is limited to compounds that feature at least one metal element and one of the anions of interest, i.e. [O^{2-}, S^{2-}, Se^{2-}, Te^{2-}, F^{-}, Cl^{-}, Br^{-}, I^{-}].

5.4.5.2 Oxidation state assignment

In order to assign integer numbers of electrons to atoms, the bond order must be determined. This task easily carried out for molecules but not for extended solids. The bond valence (BV) is a quantity similar to bond order that is used instead and, for atoms i and j, is calculated by

$$BV_{ij} = exp(\frac{R^0_{ij} - d_{ij}}{B}) \tag{5.6}$$

where d is the distance between the atoms and B is a parameter usually fixed to 0.37. R^0 is the single bond length between the two atoms, although in practice it is a function of the coordination number and oxidation state of the approximated cation for a given approximated anion and is fitted to a set of structures. In the general implementation by Brese and O'Keeffe[1] it is calculated as

$$R_{ij} = r_i + r_j - \frac{r_i r_j (\sqrt{c_i} - \sqrt{c_j})^2}{c_i r_i + c_j r_j} \tag{5.7}$$

where r and c are parameters related to the size and electronegativity of the atoms, respectively.

We use the maximum a posteriori (MAP) estimation method to determine oxidation states using the BV approach, as implemented within the `pymatgen` code[30] with a maximum nearest-neighbour radius of 4 Å.

5.4.5.3 Compound design

Using as input the metal species for which we have P_{SA} values, we use the `smact` package[20] to generate all charge neutral $A_aB_bX_x$ compositions where A and B are d-block metals, X is one of the first four halides, and the stoichiometries a, b, x are integers ≤ 8. For structure prediction, we use the structure substitution algorithm developed by Hautier *et al.*,[27] as implemented in the `Pymatgen` framework[30] with a probability threshold, σ, of 0.00001. The structure with the highest probability that does not contain more than 40 atoms/unit cell is selected as the candidate compound for a given set of species.

5.4.5.4 Total energy calculations

For calculating E_{hull}, first-principles calculations are carried out using Kohn-Sham DFT with a projector-augmented plane wave basis[31] as implemented in the Vienna Ab-initio Simulation Package (VASP).[32,33] We use the PBEsol exchange-correlation functional[34] and a k-point grid is generated for each calculation with a density of 120 Å^3 in the reciprocal lattice. The kinetic-energy cut-off is set at 600 eV and the forces on each atom minimised to below 0.005 eVÅ^{-1}.

We note that no Hubbard +U parameters have been used in the calculations to correct for the self-interaction error present in the generalised gradient approximation (GGA) for some transition metals.[35,36] The use of GGA+U has been shown the improve stability estimates of ternary oxides,[37] however, in the absence of any reliable U parameters fitted to metal halides, we use GGA for all calculations for consistency.

5.4.6 Data access statement

The `smact` package can be accessed from `https://github.com/WMD-group/SMACT`. Screening results from these calculations may be reproduced using the Python code available on-line from `https://github.com/WMD-group/SMACT/tree/master/examples`. Optimised structures are available on-line from `https://github.com/WMD-group/Crystal_structures/tree/master/TM_halides`. All other data may be obtained from the authors on request.

5.4.7 Acknowledgements

DWD gratefully acknowledges support from the Engineering and Physical Sciences Research Council (EPSRC) via the Centre for Doctoral Training in Sustainable Chemical Technologies (grant no. EP/L016354/1). Calculations were carried out on the Balena HPC cluster at the University of Bath, which is maintained by Bath University Computing Services. AW acknowledges support from the Royal Society and the Leverhulme Trust.

5.5 Remarks

The final model used to calculate compound probabilities is very simple; a product of individual P_{SA} values is used and stoichiometry is ignored. It is likely that a more sophisticated model would give more accurate results. For example, a model similar to that used by Hautier *et al.*,[27] where parameters are introduced to maximise the likelihood of the observed data, could be used. In such a case, the observed data would be the distribution of oxidation states in the ICSD. However, the model presented here is very easily interpreted and still performs well as a composition-based screening step, as evidenced by Figure 5.6. The interpretability of the model is a positive point; as the popularity of using complex ML models in chemistry increases, there is some concern that many of them operate as a "black box" and it is therefore hard to extract physical meaning from them.

For this study, a small set of compounds (Table 5.2) suggested by the model was taken forward to calculate stabilities. Of these, two are predicted to be thermodynamically stable and four are < 100 meV/atom above the convex hull. Now that this model is written into the `smact` code, some important future work will be to ascertain how this result compares to a similar search where the oxidation states model is not used. Similarly, it would be interesting to see whether stable compounds can be identified for species combinations that are predicted to be highly unlikely by the model.

This article was prepared as a Royal Society of Chemistry Faraday Discussion paper. The unique format of the Faraday Discussions involves a short (5 min) presentation of the article followed by a long (25 min) discussion session, which is minuted and later published alongside the article. The discussion of this article (yet to be published) raised some additional interesting points.

Firstly, it was pointed out that one of the compounds found to be stable, $YZrF_7$, has been previously reported and adopts the same crystal structure.[38] In one sense, this is a positive point because the structure was not in the database used build the model, and was subsequently correctly identified by the workflow. It also highlights that it is important to thoroughly check different databases before reporting a compound as "new".

Another important point, linked to the simplistic nature of the model, is that using the product of probabilities (e.g. $P_{CaAl_2O_4} = P_{CaO}P_{AlO}$) only strictly makes sense if the probabilities in question are independent. Indeed, this highlights an intrinsic assumption in the approach, and it is possible that by incorporating other correlations between metal species a more sophisticated model could be built. Another avenue for future work will

be to see if the incorporation of this kind of information can reduce the number of false negative results.

Finally, the use of 100 meV/atom as a limit for the value of E_{hull} should be mentioned, as the concept of metastability complicates the prediction of new structures. As yet, there is no clear way to determine whether a compound is close enough to the convex hull to be feasibly synthesised. What is known is that the likelihood of a compound being feasible drops off exponentially as E_{hull} increases. Sun *et al.* have shown that, based on the available data, the feasibility limit varies by the chemistry of the system.[28] For chlorides and iodides, there are very few compounds above 50 meV/atom and 25 meV/atom respectively, while fluoride compounds with E_{hull} values of 100 meV/atom are known. Some recent work that involves using energies of amorphous states,[39] is one example of studies that are beginning to define the upper limit of metastability more clearly. The role that metastability plays in realising new compounds will most likely be an active area of research for many years to come. One important factor that should also be considered is dynamic stability and this is included in the form of additional DFT calculations in the materials design workflow in the next chapter. The compound probabilities from this model are not used in the next chapter, but are employed in a separate search in Chapter 7.

Bibliography

[1] N. E. Brese and M. O'Keeffe, *Acta Crystallogr. Sect. B*, 1991, **47**, 192–197.

[2] F. Wöhler, *Unorganische Chemie*, Duncker und Humblot, Berlin, 3rd edn., 1835.

[3] I. Langmuir, *J. Am. Chem. Soc.*, 1919, **41**, 868–934.

[4] W. M. Latimer, *The Oxidation States of the Elements and their Potentials in Aqueous Solutions*, Prentice Hall, 1938.

[5] W. P. Jorissen, H. Bassett, A. Damiens, F. Fichter and H. Rémy, *J. Am. Chem. Soc.*, 1941, **63**, 889–897.

[6] L. Pauling, *J. Chem. Soc.*, 1948, **0**, 1461–1467.

[7] L. Pauling, *J. Am. Chem. Soc.*, 1932, **54**, 3570–3582.

[8] IUPAC, *Compendium of Chemical Terminology (the "Gold Book")*, Blackwell Scientific Publications, Oxford, 2nd edn., 1997.

[9] P. Karen, *Angew. Chemie Int. Ed.*, 2015, **54**, 4716–4726.

[10] D. W. Smith, *J. Chem. Educ.*, 2005, **82**, 1202–1204.

[11] G. Parkin, *J. Chem. Educ.*, 2006, **83**, 791–799.

[12] H.-P. Loock, *J. Chem. Educ.*, 2011, **88**, 282–283.

[13] L. Jiang, S. V. Levchenko and A. M. Rappe, *Phys. Rev. Lett.*, 2012, **108**, 166403.

[14] A. Walsh, A. A. Sokol, J. Buckeridge, D. O. Scanlon and C. R. A. Catlow, *J. Phys. Chem. Lett.*, 2017, **8**, 2074–2075.

[15] M. Jansen and U. Wedig, *Angew. Chemie Int. Ed.*, 2008, **47**, 10026–10029.

[16] B. R. Pamplin, *J. Phys. Chem. Solids*, 1964, **25**, 675–684.

[17] C. H. L. Goodman, *J. Phys. Chem. Solids*, 1958, **6**, 305–314.

[18] A. Walsh, S.-H. Wei, S. Chen and X. G. Gong, 2009 34th IEEE Photovolt. Spec. Conf., 2009, pp. 001875–001878.

[19] Z.-H. Cai *et al.*, *Chem. Mater.*, 2015, **27**, 7757–7764.

[20] D. W. Davies *et al.*, *Chem*, 2016, **1**, 617–627.

[21] D. W. Davies *et al.*, *Chem. Sci.*, 2018, **9**, 1022–1030.

[22] A. Jain *et al.*, *APL Mater.*, 2013, **1**, 011002.

[23] S. P. Ong *et al.*, *Comput. Mater. Sci.*, 2015, **97**, 209–215.

[24] M. O'Keeffe and N. E. Brese, *J. Am. Chem. Soc.*, 1991, **113**, 3226–3229.

[25] R. S. Mulliken, *J. Chem. Phys.*, 1934, **2**, 782–793.

[26] R. G. Pearson, *Inorg. Chem.*, 1988, **27**, 734–740.

[27] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain and G. Ceder, *Inorg. Chem.*, 2011, **50**, 656–663.

[28] W. Sun *et al.*, *Sci. Adv.*, 2016, **2**, e1600225.

[29] O. Isayev *et al.*, *Nat. Commun.*, 2017, **8**, 15679.

[30] S. P. Ong *et al.*, *Comput. Mater. Sci.*, 2013, **68**, 314–319.

[31] G. Kresse and D. Joubert, *Phys. Rev. B*, 1999, **59**, 1758–1775.

[32] G. Kresse and J. Furthmüller, *Comput. Mater. Sci.*, 1996, **6**, 15–50.

[33] G. Kresse and J. Furthmüller, *Phys. Rev. B*, 1996, **54**, 11169–11186.

[34] J. P. Perdew *et al.*, *Phys. Rev. Lett.*, 2008, **100**, 136406.

[35] V. I. Anisimov, J. Zaanen and O. K. Andersen, *Phys. Rev. B*, 1991, **44**, 943–954.

[36] V. I. Anisimov, F. Aryasetiawan and A. I. Lichtenstein, *J. Phys. Condens. Matter*, 1997, **9**, 767–808.

[37] G. Hautier, S. P. Ong, A. Jain, C. J. Moore and G. Ceder, *Phys. Rev. B*, 2012, **85**, 155208.

[38] M. Poulain, M. Poulain and J. Lucas, *Mater. Res. Bull.*, 1972, **7**, 319–326.

[39] M. Aykol, S. S. Dwaraknath, W. Sun and K. A. Persson, *Sci. Adv.*, 2018, **4**, eaaq0148.

Chapter 6

Design of Metal Chalcohalide Photoelectrodes

6.1 Introduction

We now have a method for generating a search space of inorganic compositions as well as various tools that can be used to screen through the search space. In this chapter, an example of applying these tools to the ternary chaclohalide space is presented. The work follows on directly from the chalcohalide example introduced briefly in Chapter 4. The SSE scale, HHI_R and structure substitution algorithm are all used.

We also compare the total energies of the structures found by the substitution algorithm to those identified when a global search is carried out using the evolutionary algorithm in the `USPEX` code.[1] The objective of the evolutionary algorithm is to find the global minimum of the energy landscape, using an initial set of candidate structures that *evolve* to produce more promising (lower energy) structures. This is analogous to the Darwinian survival of the fittest concept from evolutionary biology. For inorganic compounds, the `USPEX` code uses four variation operators to produce child structures:

1. Heredity: Planar slabs are cut from two parent structures, then combined.
2. Lattice mutation: Random deformations are applied to the unit cell.
3. Permutation: Atoms of different elements are swapped.
4. Soft mutation: Atoms are moved along the softest mode eigenvector (requires calculation of the dynamical matrix).

A mixture of these techniques is used in a standard search and the other main variables that need to be set are the number of atoms (number of formula units) and the number of evolutions, which are both limited practically by computational resources. In addition, the USPEX code performs DFT local optimisations for each new child structure, as described in Chapter 2. A global search such as this is never guaranteed to find the lowest energy structure for a given conformation. However, it constitutes a much more thorough configurational search than the structure substitution algorithm and is capable of finding novel structure types, whereas the substitution algorithm is not. The low energy structures identified by USPEX are subject to finite displacement calculations to ensure that they are true local minima (and not saddle points) on the potential energy surface.

While a bandgap of appropriate energy is the key criteria for any solar device, there are numerous other important properties that can be calculated from first principles.[2] Further calculations are therefore carried out on leading candidates to investigate optoelectronic properties in more detail including carrier effective masses from electronic band structures, simulated absorption spectra *via* dielectric properties and absolute electron energies from slab calculations. Absolute electron energies are crucial for the application of photoelectrochemical water splitting as the VBM and CBM must bridge the water oxidation and reduction potentials in order to drive the O_2 and H_2 evolution reactions.

6.2 Statement of Authorship

The following paper entitled *Computer-aided Design of Metal Chalcohalide Semiconductors: From Chemical Composition to Crystal Structure* reports on original research I conducted during the period of my Higher Degree by Research candidature.

Personal contributions: *Formulation of ideas (70%):* I have been heavily involved with all decisive stages of development of the project with guidance from Dr Keith Butler on the initial screening procedure and Dr Jonathan Skelton on the dynamic stability and simulated absorption spectra. *Design of methodology (70%):* The SMACT code that was written previously was used to carry out initial screening. I set up the high-throughput DFT workflow used to assess thermodynamic stability. *Experimental work (80%):* I carried out the initial screening based on compositional descriptors, structure prediction using the substitution algorithm and the high throughput DFT calculations to get E_{hull} values. Congwei Xie carried out the global structure searches using USPEX under the supervision of Prof. Artem Oganov. I carried out phonon calculations, investigating imaginary

phonon modes using the `ModeMap` code supplied by Dr Jonathan Skelton, and performed the DFT and hybrid DFT calculations to obtain bandgaps, carrier effective masses, and electron energies. *Presentation of data in journal format (80%):* The first drafts of the manuscript were written by me, with input from Dr Keith Butler and Prof. Aron Walsh at each stage of revision. The finalised manuscript was prepared by Prof. Aron Walsh and me, with input from all co-authors.

6.3 Access statement

Reprinted with permission from D. W. Davies *et al.*, *Chem. Sci.*, 2018, **9**, 1022-1030.

6.4 Publication 3

Computer-aided Design of Metal Chalcohalide Semiconductors: From Chemical Composition to Crystal Structure

Daniel W. Davies,[1] Keith T. Butler,[1] Jonathan M. Skelton,[1] Congwei Xie,[2] Artem R. Oganov,[3,4,5] Aron Walsh[1,6,7]

1. Centre for Sustainable Chemical Technologies and Department of Chemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK
2. Science and Technology on Thermostructural Composite Materials Laboratory, International Center for Materials Discovery, School of Materials Science and Engineering, Northwestern Polytechnical University, Xian, Shaanxi 710072, Peoples Republic of China
3. International Center for Materials Discovery, School of Materials Science and Engineering, Northwestern Polytechnical University, Xian, Shaanxi 710072, Peoples Republic of China
4. Skolkovo Institute of Science and Technology, 3 Nobel Street, Moscow Region 143026, Russia
5. Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region 141700, Russia
6. Global E^3 Institute and Department of Materials Science and Engineering, Yonsei University, Seoul 120-749, Korea
7. Department of Materials, Imperial College London, Exhibition Road, London SW7 2AZ, UK

6.4.1 Abstract

The standard paradigm in computational materials science is INPUT: Structure; OUTPUT: Properties, which has yielded many successes but is ill-suited for exploring large areas of chemical and configurational hyperspace. We report a high-throughput screening procedure that uses compositional descriptors to search for new photoactive semiconducting compounds. We show how feeding high-ranking element combinations to structure prediction algorithms can constitute a pragmatic computer-aided materials design

approach. Techniques based on structural analogy (data mining of known lattice types) and global searches (direct optimisation using evolutionary algorithms) are combined for translating between chemical composition and crystal structure. The properties of four novel chalcohalides ($Sn_5S_4Cl_2$, Sn_4SF_6, $Cd_5S_4Cl_2$ and Cd_4SF_6) are predicted, of which two are calculated to have bandgaps in the visible range of the electromagnetic spectrum.

6.4.2 Introduction

The past decade has seen the emergence of many databases for computed materials properties from quantum mechanical calculations.[3–9] This has made it possible to virtually screen through enormous amounts of data in the search for promising materials for energy applications such as photovoltaics,[10–12] solar fuels,[13–17] and thermoelectrics.[18–20] Furthermore, these databases are facilitating the move towards more predictive materials design using data-mining, machine learning, and other statistical techniques to reveal hitherto undiscovered trends and rules.[21–31] In order to search for Earth-abundant materials for energy applications, it is important to move beyond known materials and extend screening criteria to new compositions and structures.

There are vast areas of unexplored chemical space for inorganic compounds.[32] Such a space is intractable to high-throughput first-principles computation, even with tremendous advances in computing power and algorithms. As such, a different approach is required to efficiently explore the search space – one that is less computationally demanding overall, but sufficiently accurate.

One modern tool that is providing impressive leaps forward in this area is machine learning (ML), a subfield of artificial intelligence that involves statistical algorithms whose performance improves with experience. A growing infrastructure of ML tools has enabled its application to complex problems in many areas of chemistry and materials science.[8,22,23] This includes the development of models that relate system descriptors to desirable properties in order to reveal structure-property relationships,[33] the prediction of the likelihood of a composition to adopt a given crystal structure,[34] and the use of quantum-mechanics results as training data to extrapolate and discover new materials at a fraction of the computational cost.[31,35]

Another approach is to apply a hierarchy of screening steps, based on pre-existing methods, whereby the fact that accuracy is low in initial steps is counteracted by the idea that as the size of the search space that can be screened is so large, the chance of finding a

promising material at the end of the process remains high. Here we present one such workflow incorporating simple chemical descriptors, data mining from public databases, density functional theory (DFT) calculations and global structure searching algorithms (Figure 6.1) to translate from a compositional search space to compounds predicted to have target properties by quantum-mechanical calculations.

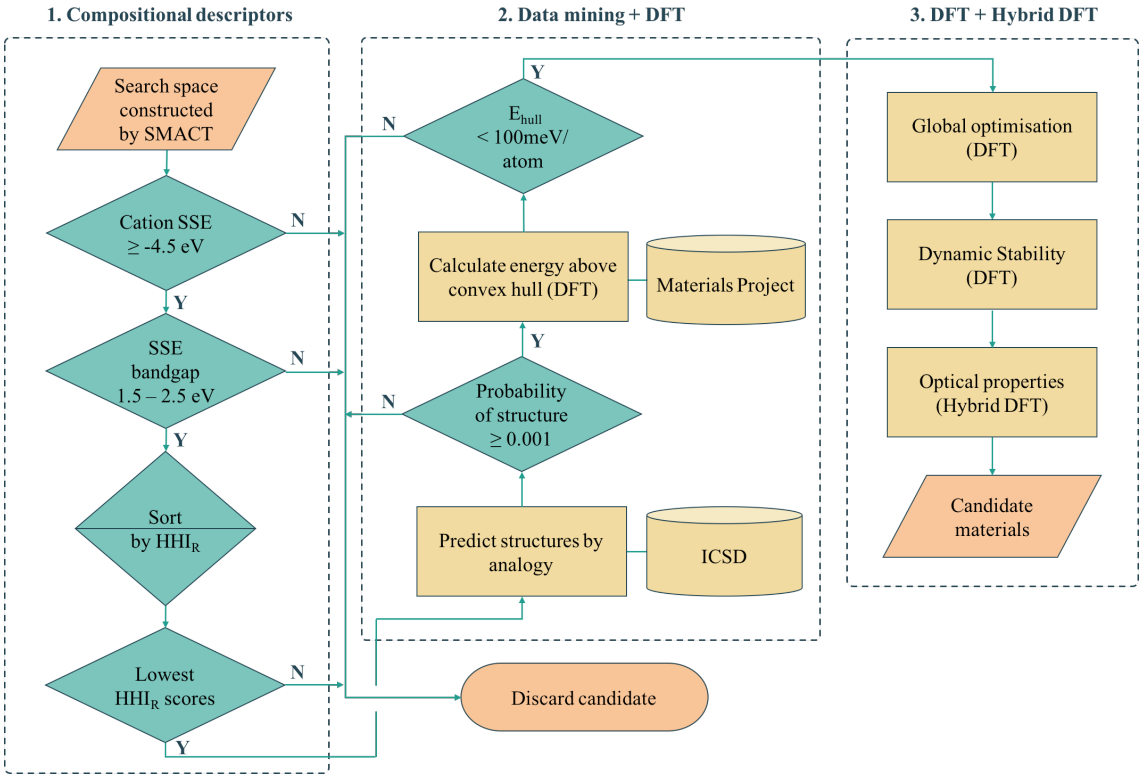

Figure 6.1: Computer-aided-design workflow used for exploring novel photoactive semiconductors. `SMACT` refers to our screening package, SSE refers to the solid-state energy scale, HHI_R refers to the Herfindahl-Hirschman Index for sustainability, while DFT refers to density functional theory.

We employ a multi-stage screening approach in a search for new photoactive semiconductors. While metal oxides combine many attractive properties for energy materials (e.g. chemical stability and low cost), they usually have bandgaps too large to absorb a significant fraction of sunlight. The formation of multi-anion compounds offers a route to modifying the electronic structure, so we consider all ternary metal chalcohalides, (i.e., $A_xB_yC_z$ with B = [O,S,Se,Te] and C = [F,Cl,Br,I]). As a target application, we search for materials for solar fuel generation, specifically for photoelectrochemical water splitting, where a set of well-defined screening criteria enables us to quickly narrow down the search space. Our searching methodology is built on already established and freely available materials design tools (`SMACT`, `Pymatgen` and `USPEX`) and can be adapted to search for different classes of materials, in a wide range of contexts of technological interest.

6.4.3 Results

6.4.3.1 $A_xB_yC_z$ compositional screening

There exist various compositional descriptors that enable the low-cost filtering of chemical space. One such tool is the solid-state energy (SSE) scale,[36] which can be used to estimate the positions of the valence band maxima (VBM) and conduction band minima (CBM) of a semiconductor with respect to the vacuum level using solely the identity of the constituent ions. We employ the SSE scale to carry out our compositional screening (see Computational Methods section for details).

First, the `smact` code[32] is used to narrow down the ternary compound search space of roughly 32 million compositions to the chalcohalide search space of 161,000 compositions. The SSE scale is then used to screen for suitable bandgaps and band-edge positions. The A cations are restricted to those with a SSE higher than the water reduction potential (approximately 4.5 V in relation to the vacuum at pH = 0) and the bandgap window was set to 1.5 – 2.5 eV. The latter criterion is set to a value range higher than the free energy for water dissociation (1.2 eV), in order to compensate for the combination of loss mechanisms found in practical devices that mean a bandgap as large as 2.2 eV could be required.[37,38] This results in in 7,676 candidate $A_xB_yC_z$ compositions with unique x, y, z stoichiometries.

Next, we sort the candidates by the sustainability of their constituent elements based on the Herfindahl–Hirschman Index for elemental reserves (HHI_R).[39] The HHI_R includes factors such as geopolitical influence over materials supply and price, and for a given composition can be obtained as the weighted average over the constituent elements. At this stage, because stoichiometry is variable, we consider the mean value for each $A_xB_yC_z$ chemical system. The six most sustainable chemical systems according to this scale are $Sn_xS_yX_z$, $Cd_xS_yX_z$ and $Ti_xS_yX_z$, where X = [Cl, F]. Of these, the Sn- and Cd-containing compositions are selected and Ti^{3+} compounds are excluded due to the d^1 electronic configuration being linked to fast electron-hole recombination, and, more practically, the well-known challenges for electronic-structure modelling due to the high correlation.[40]

The HHI_R scores of $Zn_xS_yX_z$ and $Cd_xSe_yX_z$ are the next lowest in the ranking, making these the next most sustainable according to this scale. This is because Zn and Se have higher HHI_R scores than Ti and S respectively. These systems could be of interest for future studies in the same spirit, particularly the Zn-containing compositions due to their low toxicity. This rapid screening process based on composition alone constitutes the first

phase of our overall procedure (Part 1 of Figure 6.1).

6.4.3.2 From chemical composition to crystal structure

Although compositional screening is a key initial step in materials exploration, the precision with which physical properties can be predicted from chemical composition alone is limited. In order to move to the next level of accuracy and make quantitative predictions, we must introduce a three-dimensional model of the arrangement of atoms in space. To our knowledge, no compounds of the compositions identified by our screening process have yet been reported, so the crystal structures must be predicted. Crystal structure prediction is a long-standing challenge in materials science,[41] due to the large number of degrees of freedom (lattice vectors and internal coordinates) and poor scaling with increasing system complexity.

We combine two machine learning approaches for generating candidate crystal structures from chemical composition, *viz.* 1. analogy with known crystal structures reported in crystallographic databases, and 2. direct global crystal structure searching. The first approach has a much lower computational cost, exploiting data on existing compounds, and we use this step to assess the metastability of a candidate composition. Those compounds that fall within an acceptable window of metastability are then passed to the second method, which is a more rigorous search of configurational space and allows for new structure types to be adopted.

For crystal structure prediction by analogy, we adopt the structure substitution algorithm developed by Hautier *et al.*,[42] as implemented in the `Pymatgen` framework.[43] In this method, a combination of ions are substituted onto lattice sites in known structures from the Inorganic Crystal Structure Database (ICSD).[44] Each ion substitution is associated with a certain probability, which comes from a statistical model trained on the compounds that already exist in the ICSD. If the overall probability for a given set of substitutions is above a certain threshold, it is added to a list of possible structures. This substitution process is performed on each known crystal structure in the database.

For each of the four compositions, the candidate crystal structures are locally optimized using DFT calculations and the structure with the lowest energy per atom selected. Figure 6.2 illustrates this process for one of the structures suggested by the algorithm for the $Cd_xS_yCl_z$ chemical system. In this case, the structure suggested is based on $Hg_5O_4Cl_2$ due to the high probabilities associated with both Hg^{2+}/Cd^{2+} and O^{2-}/S^{2-} substitutions.

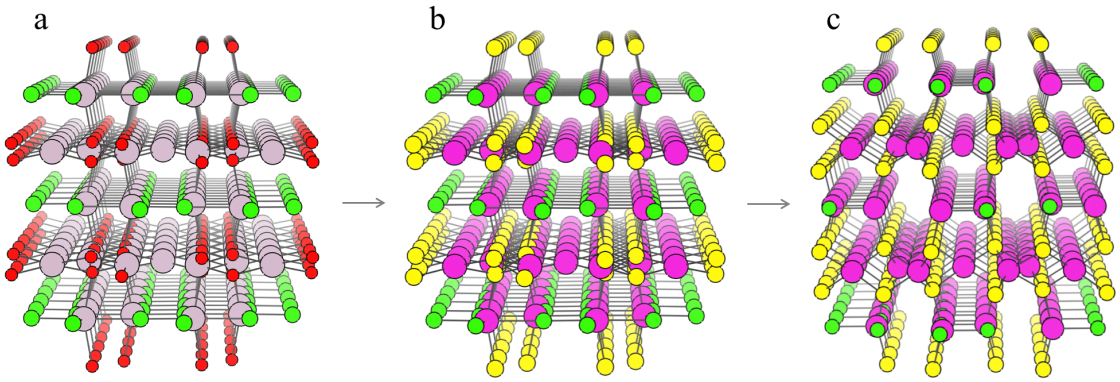

Figure 6.2: Illustration of the process of crystal structure prediction by ion substitution into existing lattice types. The $Hg_5O_4Cl_2$ structure (a) is identified as a candidate structure for the $Cd_xS_yCl_z$ chemical system. The Hg^{2+} (grey balls) and O^{2-} ions (red balls) are replaced by Cd^{2+} (purple balls) and S^{2-} ions (yellow balls), respectively, to produce the $Cd_5S_4Cl_2$ structure (b). Forces on the ions are then minimised using DFT with the PBEsol functional[45] to produce the relaxed structure (c).

Table 6.2 contains the chemical formulae of the four compounds deemed to be the most stable as a result of this process, along with the formulae of their parent structures in the ICSD. We next assess the thermodynamic stability of the candidate materials.

6.4.3.3 Thermodynamic metastability

By calculating the total energies of all the competing phases of a chemical system, one can construct an energy – composition phase diagram and assess the stability of a given compound with respect to polymorphic transformations and phase separation. By creating a bounding surface between the lowest energy phases of each composition, a convex hull is constructed above which metastable compounds fall. A key value of interest for assessing the metastability of a compound is this energy above this convex hull (E_{hull}).

Fortunately, the existence of databases of DFT total energies have all but eliminated the need for carrying out calculations for all phases of a given chemical system. Instead, one can perform calculations on new compounds using identical parameters to those used for the data in a given database, thus allowing for direct comparison of energies. Similarly, one can use the energy values in a database to construct a phase diagram and identify where on the diagram the new phase would appear. In doing so, the set of polymorphs and decomposition products that require explicit calculation can be identified. We note that it is standard to calculate such convex hulls based on internal energies, which neglect finite temperature contributions to the free energy of a compound.

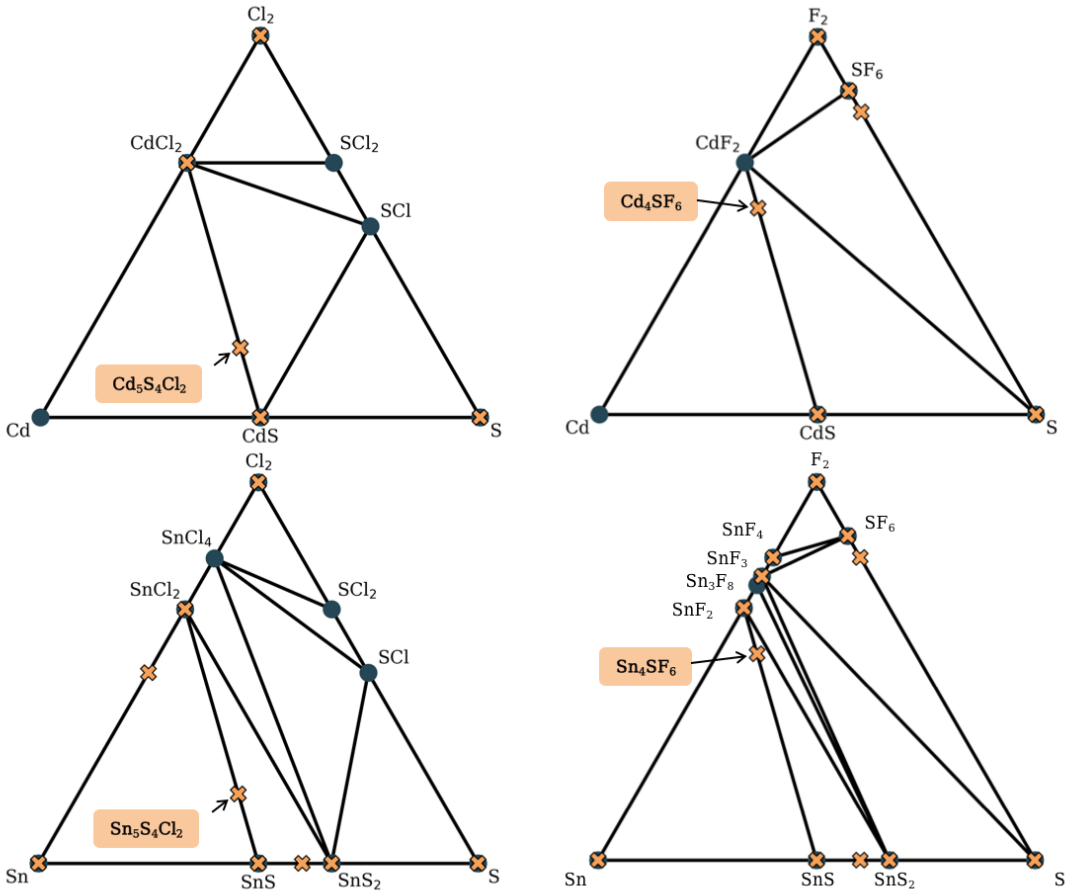

Figure 6.3: Simulated phase diagrams for the Cd−S−Cl_2, Cd−S−F_2, Sn−S−Cl_2 and Sn−S−F_2 chemical systems. Stable phases (circles) are connected by black tie-lines forming the convex hull, and unstable phases (crosses) sit above the hull. Those that are above a stable phase are unstable with respect to polymorphic changes and those above a tie-line are unstable with respect to decomposition into the stable phases at each end. The labels indicate the new phases discovered in this work.

Here, we use the Materials Project database to construct phase diagrams using the `Pymatgen` code,[43] and hence identify decomposition products. As mentioned above, and as depicted in the phase diagrams in Figure 6.3, it is not necessary to consider competing polymorphs as no compounds have yet been reported for these compositions. As can be seen from Table 6.2, all of the values of E_{hull} for the structures predicted by analogy lie between 18 and 97 meV/atom. Hence, all the compounds can formally be described as thermodynamically metastable at 0 K, but does this rule out their existence?

Metastable materials exist and are ubiquitous in both nature and technology. This includes obvious examples such as diamond vs. the lower energy allotrope of carbon, graphite, as well as classes of materials such as zeolites and metal-organic frameworks.[46] It was

recently estimated by Sun *et al.* that around half of all known inorganic materials are metastable.[27] Whether or not the value of E_{hull} is enough to predict the likelihood of successful synthesis of a material is a question that has yet to be answered. In the same work by Sun *et al.*, it was shown that the likelihood of existence drops off exponentially as E_{hull} increases. The exact rate of the drop depends on the chemistry of the system. We use 100 meV/atom as a guiding principle for the maximum E_{hull}, as this criteria covers approximately 90% of compounds in the Materials Project database that represent fully-characterised structures in the ICSD. The four structures found by analogy all fall within this metastability window, so they are all carried forward to the global structure searching stage.

6.4.3.4 Global structure search

The structure from analogy approach provides an attractive route to obtaining sensible crystal structures with reasonable energies, however it does not provide a rigorous route to obtaining the true ground state. Finding the true global ground state structure for a given chemical composition is one of the outstanding problems of theoretical chemistry. Whilst exhaustive searching of parameter space is the only way to find a guaranteed global minimum structure, this approach quickly becomes impractically large for even simple chemical systems. Global searching, based on evolutionary algorithms offer a solution to this problem and have had enormous success in discovering new ground state crystal structures. Here we use `USPEX` to apply an evolutionary algorithm and perform a global structure search.

For each of the four compositions, the global structure search algorithm[1,47] yields a different crystal structure to that found by analogy with known structures (Figure 6.4). For each of the structures generated by the global search, there is no way in which the data-mining algorithm could have arrived at the same result. This is an intrinsic limitation of the data-mining approach, as it relies on a database of known structures and it is therefore incapable of predicting new structure types. Three of the four compounds adopt structure types that have not yet been reported, disregarding those with fractional occupancy on some lattice sites. The remaining compound, $Cd_5S_4Cl_2$, adopts the same structure type as Li_5BiO_5.[48] However, this substitution is rejected by the structure prediction algorithm on the basis that the resulting formula is not charge neutral – the structure we find is partially inverted in terms of anion / cation occupancy.

The values of E_{hull} for the structures predicted by global structure search are also shown

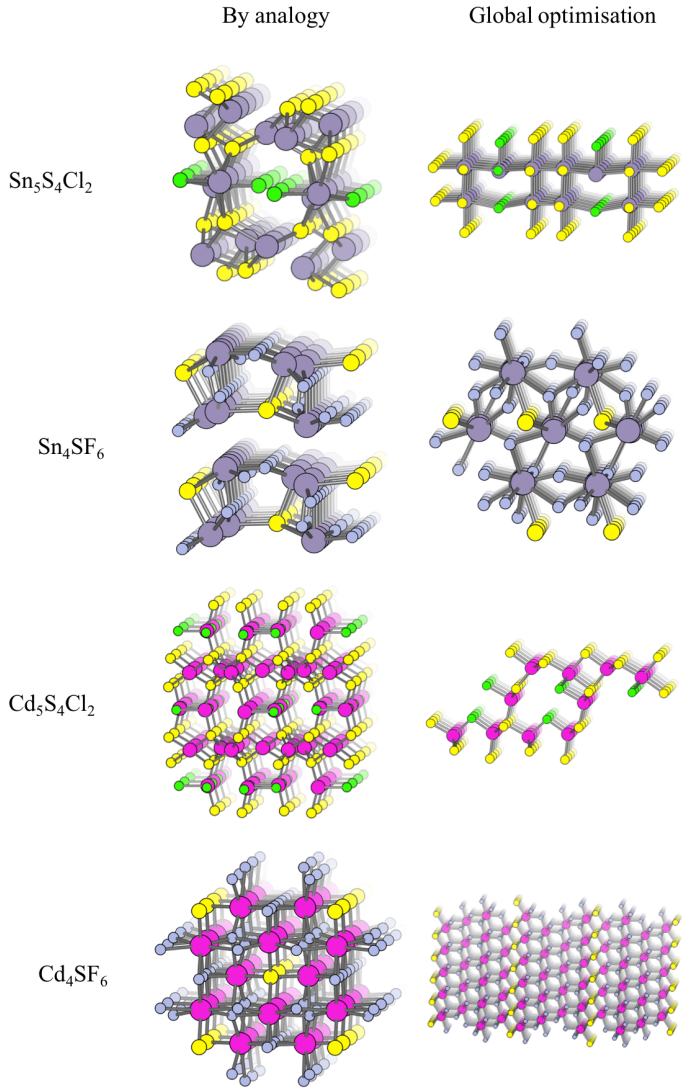

Figure 6.4: Crystal structures of the four candidate compositions as predicted by analogy through data mining of other structures and by a first-principles global structure search algorithm.

in Table 6.2, and are universally lower than those found by analogy. While the structural analogy procedure is limited by the diversity of known structure types, the global structure search approach is restricted only by the structural complexity (number of formula units) included in the search. A holistic assessment of performance in the context of high-throughput screening must however also take into account time and resources: the data-mining algorithm takes only a few minutes to run on a desktop computer, while the global structure searching requires a supercomputing resource where around 10,000 CPU hours were needed for each material.

In addition to thermodynamic stability, another factor that cannot be ignored is dynamic stability, to ensure that the crystal structures are true local minima (and not saddle points) on the potential energy surface. Finite-displacement calculations were carried out to obtain the vibrational (phonon) frequencies of each of the compounds, and no negative-frequency (imaginary) phonon modes were found at the zone centre (Γ point) for any of the structures. Full details of this analysis can be found in the Supplementary Information.

6.4.3.5 Crystal structures and bonding environments

Table 6.1 contains the space groups and lattice parameters of the four minimum energy compounds identified at the end of the screening process.

$Sn_5S_4Cl_2$: Eight Sn(II) atoms per crystallographic unit cell adopt an octahedral environment, forming bilayers of edge-sharing SnS_5Cl polyhedra in the *bc* plane. The polyhedra are vertex sharing at the Cl atoms, and the other two Sn atoms in the unit cell reside in the same plane as the halide ions.

Sn_4SF_6: Sn(II) adopts both 6-and 4-coordinate environments, with space for a lone pair in each. The Sn-centred polyhedra are all vertex sharing and have either 6 F vertices (6-coordinate Sn) or 3 F vertices and 1 S vertex (4-coordinate Sn).

$Cd_5S_4Cl_2$: Two Cd(II) atoms per unit cell locate at the centre of CdS_4 tetrahedra, and seven Cd atoms form the centre of CdS_3Cl tetrahedra. The other two Cd atoms form trigonal bipyramids with 3 S and 2 Cl vertices. All of the polyhedra are vertex sharing bar one of the trigonal bipyramids, which is edge sharing with two of the tetrahedra.

Cd_4SF_6: Eight Cd(II) atoms per unit cell adopt a distorted 8-fold coordination with Cl atoms. The S atom locates in monolayers in the *ab* plane, and the four Cd atoms that are adjacent to these layers are 7-coordinate with 3 neighbouring S and 4 neighbouring F neighbouring atoms. All of the polyhedra in the structure are edge sharing.

Table 6.1: Structural information for the minimum energy compounds.

Compound	Space group	*a* (Å)	*b* (Å)	*c* (Å)	Formula units per cell
$Sn_5S_4Cl_2$	$Pma2$	17.529	5.771	5.817	2
Sn_4SF_6	$R3$	8.615	8.615	9.528	3
$Cd_5S_4Cl_2$	Cm	14.507	4.212	15.631	2
Cd_4SF_6	$R\bar{3}m$	3.832	3.832	37.148	3

Having established promising compositions and their candidate structures, we next go on to perform quantitative analyses of the detailed electronic structure of these materials.

6.4.3.6 Optoelectronic properties

The most critical property for any light-harvesting material, whether for photovoltaic or solar fuel applications, is the bandgap (E_g). Indeed, the screening procedure we have employed thus far relies on making initial estimates of E_g at an early stage, before considering structure or stability. The calculations required to accurately predict bandgaps are significantly more computationally demanding than those which can satisfactorily predict equilibrium geometry.

The first-principles values of E_g are presented in Table 6.2 alongside the bandgaps estimated using the SSE scale. Two of the compounds found by the screening procedure, $Cd_5S_4Cl_2$ and Cd_4SF_6, have bandgaps in the visible range of 2.75 and 2.15 eV, respectively. $Sn_5S_4Cl_2$ has a bandgap of 0.9 eV, which is better suited for solar cell or thermoelectric applications. This is encouraging, given the small set of compounds that have been brought through to this stage of the screening process and the qualitative nature of the SSE metric employed to screen the bandgaps.

Beyond the bandgap, quantum-mechanical calculations can also provide access to optical absorption spectra *via* computation of the complex dielectric function. Figure 6.5 shows the simulated spectra of the four compounds of interest. The Cd compounds display moderate absorption in the visible region, indicating their potential for use as solar fuel or photovoltaic materials. Of the two, Cd_4SF_6 absorbs photons with energy across more of the visible range but quite weakly, suggesting that thicker layers would be needed in a device. Meanwhile, $Cd_5S_4Cl_2$ absorbs more strongly but at a higher energy, so would

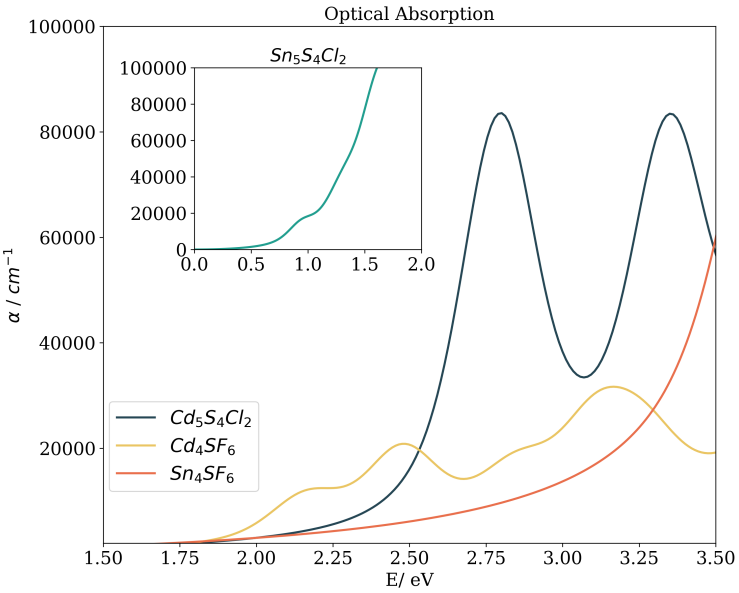

Figure 6.5: Simulated optical absorption spectra of the candidate materials from the complex dielectric function. Calculations were performed within DFT and the non-local HSE06 exchange-correlation functional, using the independent particle approximation (excluding excitonic and phonon-assisted transitions).

be suited to incorporation into a tandem solar cell.

The absolute band edge positions are also calculated using surface (non-polar slab) models of the four materials. The CBM position is the negative of the electron affinity (EA), and as indicated in Table 6.2, the EA values are all < 4.5 eV. This indicates that as well as having promising bandgaps, the two Cd-based compounds have potential for use in photoelectrochemical water splitting applications, with VBM and CBM positions that bridge the water oxidation and reduction potentials, enabling the redox reaction. For Sn_4SF_6, no slab without an overall dipole could be found, so we instead report a likely range for the EA and IP values after applying a dipole correction in the slab calculation (see Computational Methods Section). This material also bridges these energies, but has too wide a band gap, while the other Sn-containing compound, $Sn_5S_4Cl_2$, has an appropriate EA, but too small a bandgap, as has already been discussed. This is summarised in the energy band alignment diagram, Figure 6.6.

Finally, carrier effective mass (m^*) is a quantity that can also provide preliminary insight into the performance of a semiconducting material, with smaller m^* values being more desirable as this quantity is inversely proportional to conductivity. The two Cd-containing compounds have lower m_e^* values than the Sn-containing compounds (Table 6.2). This is a result of the metallic s-states forming the lower conduction band in the former case which

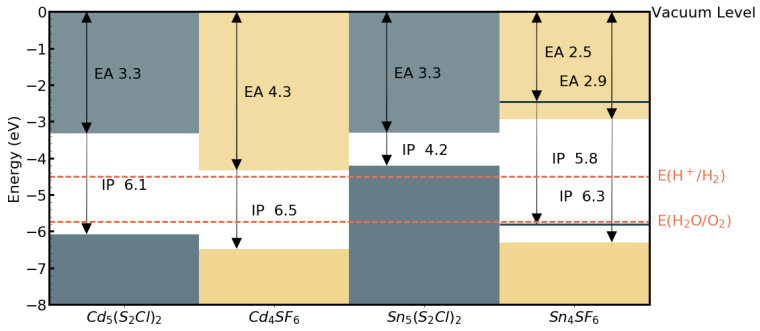

Figure 6.6: Electron affinities (EA) and ionisation potentials (IP) for the candidate materials, from DFT calculations of non-polar crystal terminations. The water redox potentials (dashed orange lines) are also shown. For Sn_4SF_6, a dipole correction was added resulting in lower and upper (blue solid lines) bounds for the IP and EA values.

give higher dispersion than the more directional metallic p-states in the latter (Figure 6.7a and Figure 6.7b). The m_h^* values are in general much higher, with the sulphur and halide p-states dominating the upper valence band. One notable exception is $Sn_5S_4Cl_2$ with a value of 0.40 m_e. This is a result of strong hybridisation between the Sn s and S p orbitals which form a two-dimensional Sn-S network along which carriers can transport without encountering a Cl atom (Figure 6.4). This is possible due to the Sn^{2+} oxidation state, which results in the Sn s orbitals remaining occupied. In the case of Sn_4SF_6, no such Sn-S network exists and S p states dominate the VBM, while F p states also contribute (Figure 6.7c).

The calculated band structure of $Sn_5S_4Cl_2$ reveals the presence of multiple band extrema ("multi-valley"), a sought-after feature in the design of thermoelectric materials.[49] Furthermore, the effective number of extrema is increased by the presence of multiple bands within a few k_BT in energy of each other at the R, T, S and U points in the Brillouin zone (see Supplementary Information Figure S4).

6.4.4 Conclusion

We have introduced a hierarchical screening procedure and used it to search through a large space of over 161,000 compositions to identify promising candidate photoactive semiconductors. Using our approach, which relies on compositional descriptors and exploits existing data, first-principles calculations were carried out on a subset of compounds in order to establish thermodynamic stability, and global structure searching was employed for the most promising candidates. This procedure has enabled us to identify four new chalcohalide compounds, two of which, $Cd_5S_4Cl_2$ and Cd_4SF_6, have bandgaps

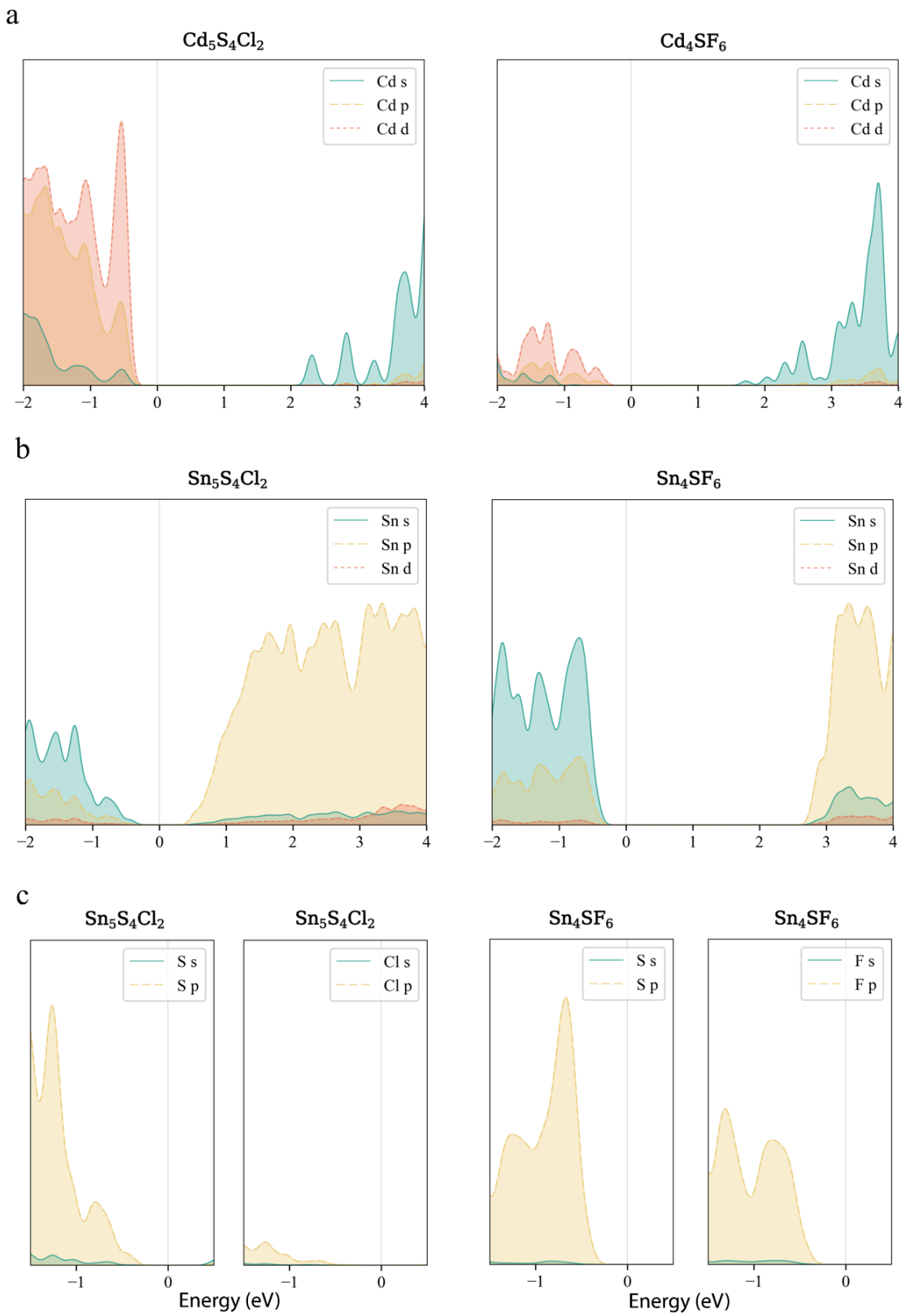

Figure 6.7: Orbital-projected local electronic density of states of $Cd_5S_4Cl_2$, Cd_4SF_6, $Sn_5S_4Cl_2$ and Sn_4SF_6. s- p- and d-orbital contributions from the metal species to the density of states near the band edges for the Cd-containing (a) and Sn-containing (b) compounds. The s- and p-orbital contributions from S and the halide species to the upper valence band for the Sn-containing compounds are also shown (c).

Table 6.2: The parent-structure formulae from the ICSD compounds identified by analogy that led to the lowest energy structures after DFT relaxation are shown along with the energies above the convex hull ($E_{hull}^{analogy}$), the corresponding energies predicted after global structure search (E_{hull}^{global}). The estimated bandgaps from SSEs (E_g^{SSE}) used at the beginning of the workflow, bandgaps (E_g), electron affinities (EA) and ionisation potentials (IP) calculated using a hybrid exchange-correlation functional at the end of the screening workflow, and effective masses for carrier electrons and holes from GGA calculations (m_e^* and m_h^*) are also displayed.

Compound	Parent	$E_{hull}^{analogy}$ (meV/ atom)	E_{hull}^{global} (meV/ atom)	E_g^{SSE} (eV)	E_g (eV)	EA (eV)	IP (eV)	m_e^*	m_h^*
$Sn_5S_4Cl_2$	$Hg_5(O_2Cl)_2$	96.5	61.8	2.0	0.91	3.30	4.21	0.50	0.40
Sn_4SF_6	Hg_4OF_6	51.8	46.7	2.0	3.36	2.45–2.94[†]	5.81–6.30[†]	0.86	2.01
$Cd_5S_4Cl_2$	$Hg_5(O_2Cl)_2$	83.5	50.2	1.9	2.75	3.33	6.08	0.18	2.58
Cd_4SF_6	Cd_4F_6O	18.2	18.0	1.9	2.15	4.33	6.48	0.25	2.00

[†] When only polar surfaces could be found, a dipole correction term was added to the calculation of the surface dipole, which yields upper and lower bounds to the EA and IP values (see Computational Methods Section).

in the visible range and good absorption properties for solar fuel applications. Further detailed investigation into the electronic structure of these materials show that effective electron and hole conduction should be possible. The approach constitutes a computer-aided materials design procedure that employs existing knowledge in a targeted manner in order to traverse the vast chemical hyperspace.

6.4.5 Computational methods

6.4.5.1 Compositional screening

Construction of the search space and subsequent screening based on SSE and HHI_R is carried out with Python 3 on a desktop computer using the `smact` library, which is publicly available online at `https://github.com/WMD-group/SMACT`. First, the compositional search space of ternary chalcohalides is constructed using the `smact` package: The stoichiometry maximum is set to 8 and only those compositions which pass both the charge neutrality and electronegativity balance tests form part of the initial search space. Every possible combination of $A_xB_yC_z$ is generated with B = [O,S,Se,Te] and C = [F,Cl,Br,I]. All known oxidation states of all elements in each combination are considered and charge neutrality is assessed by

$$xq_A + yq_B + zq_C = 0 \tag{6.1}$$

where q is the formal charge associated with each species in the considered oxidation state. Combinations satisfy electronegativity balance when $\chi^{\text{cation}} < \chi^{\text{anion}}$, where χ is the Paul-

ing electronegativity[50] of an element. This ensures the most electronegative elements carry the most negative charge. For full details of this method of search space construction, the reader is referred to Ref. 32.

The SSE scale[36] is used to limit the A cations to those with a SSE higher than the water reduction potential and set the bandgap window was to 1.5 – 2.5 eV. The SSE provides information on valence and conduction bands on the basis of the frontier orbitals of the constituent ions. It reflects ionisation potential of an anion (filled electronic states) and electron affinity of a cation (empty electronic states). The energies of 40 elements were originally fitted from a test set of 69 closed-shell binary inorganic compounds, and now the SSE values for 94 elements are available.[51] The bangap (Eg) can then be estimated from the tabulated SSE values as

$$E_g^{SSE} = \mathrm{SSE}^{\mathrm{cation}} - \mathrm{SSE}^{\mathrm{anion}} \tag{6.2}$$

For multicomponent systems, the limiting SSE values are used.

6.4.5.2 Crystal structure prediction by analogy

We use the structure substitution algorithm developed by Hautier *et al.*,[42] as implemented in the `Pymatgen` framework[43] with a probability threshold of 0.001. For a given composition the procedure is carried out for each common oxidation state of the metal (e.g. for $Sn_xS_yCl_z$ both Sn(II) and Sn(IV) must be considered).

6.4.5.3 Crystal structure prediction by global searching

Global crystal structure searches are carried out for each of the candidate compositions using the same stoichiometries as the lowest energy crystal structures from the prediction by analogy. This step is only carried out if a structure found by analogy falls within the defined "metastability window" of 100 meV/atom. Using the evolutionary structure prediction algorithm `USPEX`,[1,47] we perform global structure searches for the candidate compositions. No constraints are imposed on the shape or volume of the unit cell, but the search is restricted to one (11 atoms/cell) and two (22 atoms/cell) formula units for each of the four compositions. In the evolutionary optimisation procedure, the first generation contains 80 randomly generated structures, and the succeeding generations (each with 60 structures) are produced by random (20%), heredity (50%), permutation (10%),

soft-mutation (10%), and lattice mutation (10%) operations as described elsewhere.[47]

6.4.5.4 First-principles calculations

All first principles calculations are carried out using Kohn-Sham DFT with a projector-augmented plane wave basis[52] as implemented in the Vienna Ab-initio Simulation Package (VASP).[53,54]

Total energies: For calculating E_{hull} we use the PBEsol exchange-correlation functional.[45] A Monkhorst-Pack k-point grid is generated for each calculation with k-point spacing of 0.242 $Å^{-1}$. The kinetic-energy cutoff is set at 520 eV and the force on each atom converged to within 0.005 $eVÅ^{-1}$. The Materials Project API[55] is used to retrieve DFT total energies of known phases for each chemical system. Phase diagrams are constructed to identify decomposition products and the total energies of these products recalculated in the same manner as described above.

Dynamical stabilities: Structures are further relaxed using a kinetic energy cutoff of 700 eV. The normal modes are calculated within the harmonic approximation, using the PHONOPY package[56–58] to construct and evaluate the force constants. The finite displacement method (FDM) approach is used with a step size of 0.01 Å. Each of the unit cells contains N atoms (where N = 22 or 33) so has 6N (132 or 198) possible displacements. The number of unique displacements is reduced to between 11 and 44 depending on the crystal symmetry. For computational efficiency, phonons are considered at the Γ point only.

Optoelectronic properties: Semi-local exchange-correlation treatments such as the PBEsol functional provide an accurate description of crystal structures but tend to underestimate the electronic bandgaps of semiconductors. To overcome this issue, more accurate electronic structure calculations are performed using the hybrid non-local functional HSE06,[59] which includes 25% screened Hartree-Fock exact exchange. Γ-centred homogeneous k-point meshes are used, the spacings of which are determined by the magnitude of the lattice vectors, as per Yu *et al.*[60] and the kinetic energy cutoff is set at 520 eV. For optical absorption calculations, the dielectric tensor is calculated using the VASP code following the Kubo-Greenwood method. This is then used to calculate the absorption *via* the Kramers-Kronig relation.

Absolute electron energies (IP and EA values) are calculated by generating 2D slab models of low Miller index, non-polar surfaces of the crystal structures. Hybrid DFT (HSE06 functional) is used to calculate the surface dipole, D, which is the difference between the average electrostatic potential in the slab and that in the vacuum level. The VBM and CBM positions from the bulk calculations can then be used to calculated the true VBM and CBM positions. These are simply the differences between D and VBM_{bulk}, and D and CBM_{bulk}, respectively. Convergence with respect to slab thickness and vacuum distance was achieved within two repeat layers and 15 Å respectively, in all cases. When no non-polar surfaces could be found for a material, the dipole-dipole interaction correction is added to the potential, as implemented in the VASP code. This leads to an upper and lower limit of the potential in the vacuum level, hence an upper an lower limit to D.

Carrier effective masses are calculated using band structures generated from hybrid DFT (HSE06 functional) calculations. The SeeKpath code[61] is used to generate a suitable path through the Brillouin zone, which is sampled at a resolution of 0.01Å^{-1} between each k-point. In order to calculate effective masses, a parabola is fit to all points from the minimum (maximum) of the CBM (VBM) to the points k_BT higher (lower).

6.4.6 Data access statement

The `smact` package can be accessed from `https://github.com/WMD-group/SMACT`. Screening results from these calculations may be reproduced using the Python code available on-line from `https://github.com/WMD-group/SMACT/tree/master/examples`. Optimised structures are available on-line from `https://github.com/WMD-group/Crystal_structures/Chalcohalides` All other data may be obtained from the authors on request.

6.4.7 Acknowledgements

DWD gratefully acknowledges support from the Engineering and Physical Sciences Research Council (EPSRC) via the Centre for Doctoral Training in Sustainable Chemical Technologies (grant no. EP/L016354/1). JMS. gratefully acknowledges support from the EPSRC (grant nos. EP/K004956/1 and EP/P007821/1). Calculations were carried out on the Balena HPC cluster at the University of Bath, which is maintained by Bath University Computing Services. Some of the calculations were also carried out on the UK national Archer HPC facility, accessed through membership of the UK Materials Chemistry Consortium, which is funded by EPSRC grant no. EP/L000202.

6.5 Remarks

In terms of the materials design workflow itself, the fact that one material out of just four candidate materials (Cd_4SF_6) had a bandgap within the target window, and at the position required to drive the water redox reaction, is positive. The bandgaps of the other materials are too wide or too narrow according to the criteria used for the initial screening, but are all approximately in the correct region for water splitting; the band edges of three materials out of four bridge the water oxidation and reduction potentials. This is an advantage of using the SSE scale, which is derived from experimental ionisation potentials and electron affinities.

Theoretically, the $Cd_5S_4Cl_2$ compound also has a bandgap that could be acceptable for a water splitting material (2.75 eV). The drawback is that bandgaps as wide as this result in a low maximum efficiency, as only a small fraction of the solar spectrum can be absorbed.[62] Another practical issue that is not addressed in the publication is the stability of the materials in water. Unlike photovoltaic cells which need to be stable in air, or can be encapsulated, photoelectrode materials necessarily need to make contact with water. Photocorrosion is a major problem for many candidate water splitting materials and occurs when photogenerated charge carriers cause oxidation or reduction of the material itself.[62] Although there has been some limited investigation into the use of metal chalcohalides for this application,[63] it is not clear how stable they would be, and it is conceivable that the halide ions could be oxidised to halogen gases.

In summary, this chapter has shown one way in which it is possible to screen a large number of hypothetical compositions for stable structures with target properties using a hierarchical screening workflow. Relatively few first-principles calculations (DFT total energies of ~100 crystal structures and hybrid DFT calculations on just four crystal structures) were needed to identify target materials, making the overall process computationally affordable. Although the global structure search identified lower energy configurations for all four leading candidate compositions, and these were ultimately taken forward to calculate their properties, the structure substitution algorithm did provide sensible structures of similar energies. This algorithm is therefore a much cheaper alternative to link compositions to crystal structures and is more amenable to a high-throughput search. In the next chapter, this algorithm will be used again in another materials design workflow, along with a new ML model for compositional screening that is an alternative to using the SSE scale.

Bibliography

[1] A. R. Oganov and C. W. Glass, *J. Chem. Phys.*, 2006, **124**, 244704.

[2] A. M. Ganose, C. N. Savory and D. O. Scanlon, *Chem. Commun.*, 2017, **53**, 20–44.

[3] J. Hill *et al.*, *MRS Bull.*, 2016, **41**, 399–409.

[4] W. Setyawan and S. Curtarolo, *Comput. Mater. Sci.*, 2010, **49**, 299–312.

[5] W. Setyawan, R. M. Gaume, S. Lam, R. S. Feigelson and S. Curtarolo, *ACS Comb. Sci.*, 2011, **13**, 382–390.

[6] D. D. Landis *et al.*, *Comput. Sci. Eng.*, 2012, **14**, 51–57.

[7] A. Jain *et al.*, *APL Mater.*, 2013, **1**, 011002.

[8] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, *Jom*, 2013, **65**, 1501–1509.

[9] *The NoMaD Repository*, `http://nomad-repository.eu/-[Accessed: 02-09-2017]`.

[10] L. Yu, R. S. Kokenyesi, D. A. Keszler and A. Zunger, *Adv. Energy Mater.*, 2012, **3**, 43–38.

[11] T. Krishnamoorthy *et al.*, *J. Mater. Chem. A*, 2015, **3**, 23829–23832.

[12] Y. Hinuma *et al.*, *Nat. Commun.*, 2016, **7**, 11962.

[13] I. E. Castelli *et al.*, *Energy Environ. Sci.*, 2012, **5**, 9034–9043.

[14] I. E. Castelli *et al.*, *Energy Environ. Sci.*, 2012, **5**, 5814–5819.

[15] Y. Wu, P. Lazic, G. Hautier, K. Persson and G. Ceder, *Energy Environ. Sci.*, 2013, **6**, 157–168.

[16] I. E. Castelli *et al.*, *Adv. Energy Mater.*, 2015, **5**, 1400915.

[17] M. Pandey, A. Vojvodic, K. S. Thygesen and K. W. Jacobsen, *J. Phys. Chem. Lett.*, 2015, **6**, 1577–1585.

[18] C. Toher *et al.*, *Phys. Rev. B*, 2014, **90**, 174107.

[19] T. D. Sparks, M. W. Gaultois, A. Oliynyk, J. Brgoch and B. Meredig, *Scr. Mater.*, 2016, **111**, 10–15.

[20] A. Faghaninia *et al.*, *Phys. Chem. Chem. Phys.*, 2017, **19**, 6743–6756.

[21] C. C. Fischer, K. J. Tibbetts, D. Morgan and G. Ceder, *Nat. Mater.*, 2006, **5**, 641–646.

[22] G. Hautier, C. C. Fischer, A. Jain, T. Mueller and G. Ceder, *Chem. Mater.*, 2010, **22**, 3762–3767.

[23] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran and R. Ramprasad, *Sci. Rep.*, 2013, **3**, 2810.

[24] B. Meredig *et al.*, *Phys. Rev. B*, 2014, **89**, 094104.

[25] O. Isayev *et al.*, *Chem. Mater.*, 2015, **27**, 735–743.

[26] P. V. Balachandran, J. Theiler, J. M. Rondinelli, T. Lookman and A. P. Sutton, *Sci. Rep.*, 2015, **5**, 13285.

[27] W. Sun *et al.*, *Sci. Adv.*, 2016, **2**, e1600225.

[28] G. Pilania *et al.*, *Sci. Rep.*, 2016, **6**, 19375.

[29] J. Lee, A. Seko, K. Shitara, K. Nakayama and I. Tanaka, *Phys. Rev. B*, 2016, **93**, 115104.

[30] W. Chen *et al.*, *J. Mater. Chem. C*, 2016, **4**, 4414–4426.

[31] T. Moot *et al.*, *Mater. Discov.*, 2017, **6**, 9–16.

[32] D. W. Davies *et al.*, *Chem*, 2016, **1**, 617–627.

[33] A. Pulido *et al.*, *Nature*, 2017, **543**, 657–664.

[34] A. O. Oliynyk *et al.*, *Chem. Mater.*, 2016, **28**, 7324–7331.

[35] F. A. Faber, A. Lindmaa, O. A. Von Lilienfeld and R. Armiento, *Phys. Rev. Lett.*, 2016, **117**, 135502.

[36] B. D. Pelatt, R. Ravichandran, J. F. Wager and D. a. Keszler, *J. Am. Chem. Soc.*, 2011, **133**, 16852–16860.

[37] T. Bak, J. Nowotny, M. Rekas and C. Sorrell, *Int. J. Hydrogen Energy*, 2002, **27**, 991–1022.

[38] B. A. Pinaud *et al.*, *Energy Environ. Sci.*, 2013, **6**, 1983–2002.

[39] M. W. Gaultois *et al.*, *Chem. Mater.*, 2013, **25**, 2911–2920.

[40] B. J. Morgan and P. A. Madden, *Phys. Rev. B*, 2012, **86**, 035147.

[41] S. M. Woodley and R. Catlow, *Nat. Mater.*, 2008, **7**, 937–946.

[42] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain and G. Ceder, *Inorg. Chem.*, 2011, **50**, 656–663.

[43] S. P. Ong *et al.*, *Comput. Mater. Sci.*, 2013, **68**, 314–319.

[44] FIZ Karlsruhe, *Inorganic Crystal Structure Database*, `http://icsd.cds.rsc.org/-[Accessed:27-08-2017]`.

[45] J. P. Perdew *et al.*, *Phys. Rev. Lett.*, 2008, **100**, 136406.

[46] C. H. Hendon *et al.*, *Chem. Mater.*, 2017, **29**, 3663–3670.

[47] C. W. Glass, A. R. Oganov and N. Hansen, *Comput. Phys. Commun.*, 2006, **175**, 713–720.

[48] C. Greaves and S. Katib, *Mater. Res. Bull.*, 1989, **24**, 973–980.

[49] G. Tan, L.-D. Zhao and M. G. Kanatzidis, *Chem. Rev.*, 2016, **116**, 12123–12149.

[50] L. Pauling, *J. Am. Chem. Soc.*, 1932, **54**, 3570–3582.

[51] B. D. Pelatt *et al.*, *J. Solid State Chem.*, 2015, **231**, 138–144.

[52] G. Kresse and D. Joubert, *Phys. Rev. B*, 1999, **59**, 1758–1775.

[53] G. Kresse and J. Furthmüller, *Comput. Mater. Sci.*, 1996, **6**, 15–50.

[54] G. Kresse and J. Furthmüller, *Phys. Rev. B*, 1996, **54**, 11169–11186.

[55] S. P. Ong *et al.*, *Comput. Mater. Sci.*, 2015, **97**, 209–215.

[56] A. Togo, L. Chaput, I. Tanaka and G. Hug, *Phys. Rev. B*, 2010, **81**, 174301.

[57] J. M. Skelton, S. C. Parker, A. Togo, I. Tanaka and A. Walsh, *Phys. Rev. B*, 2014, **89**, 205203.

[58] A. Togo, L. Chaput and I. Tanaka, *Phys. Rev. B*, 2015, **91**, 094306.

[59] A. V. Krukau, O. A. Vydrov, A. F. Izmaylov and G. E. Scuseria, *J. Chem. Phys.*, 2006, **125**, 224106.

[60] L. Yu and A. Zunger, *Phys. Rev. Lett.*, 2012, **108**, 068701.

[61] Y. Hinuma, G. Pizzi, Y. Kumagai, F. Oba and I. Tanaka, *Comput. Mater. Sci.*, 2017, **128**, 140–184.

[62] C. Jiang, S. J. A. Moniz, A. Wang, T. Zhang and J. Tang, *Chem. Soc. Rev.*, 2017, **46**, 4645–4660.

[63] H. Kunioku, M. Higashi and R. Abe, *Sci. Rep.*, 2016, **6**, 32664.

Chapter 7

Design of Quaternary Oxide Solar Materials

7.1 Introduction

We have seen previously how various heuristic tools can be used to quickly and cheaply estimate bandgaps of materials. In particular the SSE scale, that was used in Chapter 4 to estimate chalcohalide bandgaps, relies on experimental ionisation potential (IP) and electron affinity (EA) values. For this reason, the SSE is a more useful scale for some elements than others. For example it performs particularly poorly for predicting the bandgap of oxide compositions, as the range of IP values for the 56 binary oxides used in the construction of this model is 4.9 eV, and the standard deviation across all values is 1.44 eV. The SSE for O therefore carries the highest uncertainty among all elements.

In this chapter, the aim is to build a ML model that is trained using much more information about the constituent elements of compounds than IP and EA alone and that should be able to provide reasonable predictions of bandgaps for oxide compositions. This will then be used as the first screening step of an overall workflow (Figure 7.1) which brings together the tools developed in previous chapters to search for new oxide materials that have a bandgap suitable for solar applications.

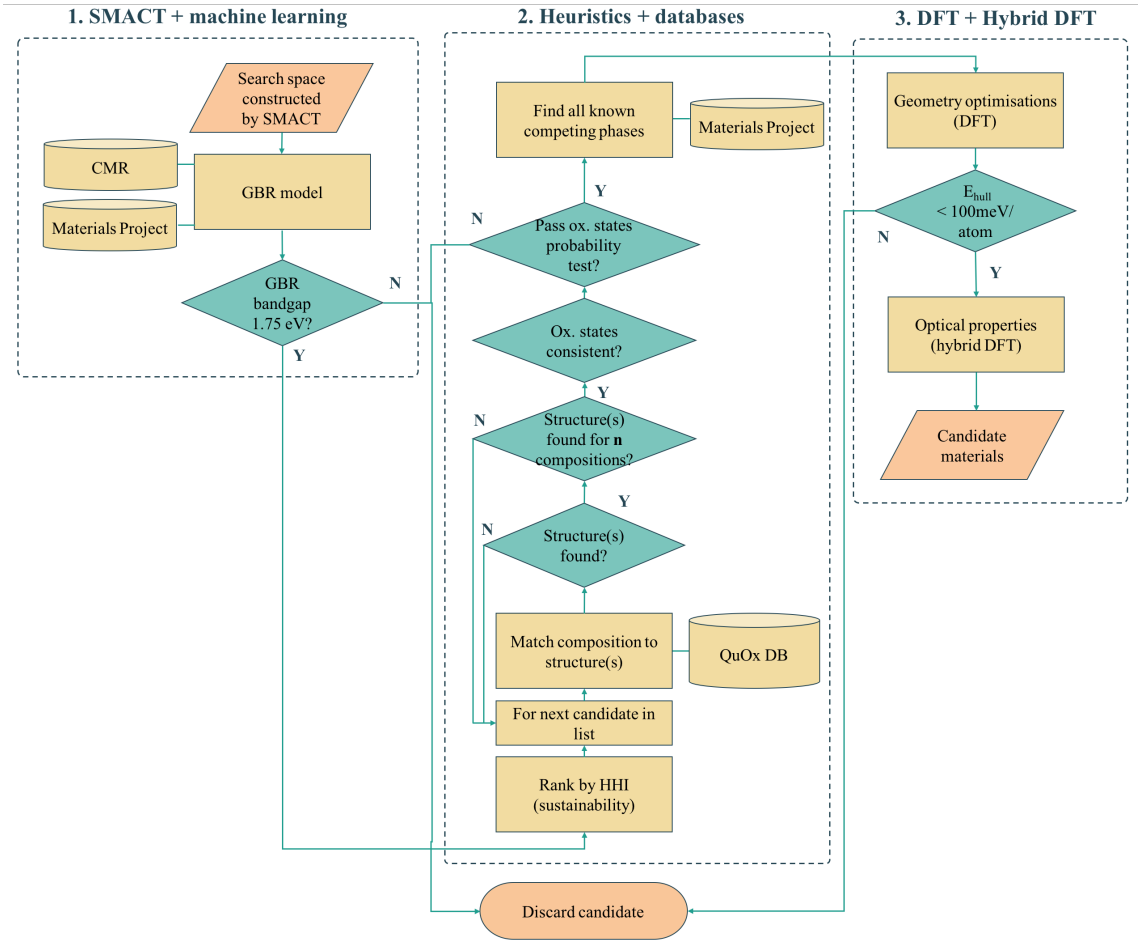

Figure 7.1: Computer-aided design workflow. 1. Data from the computational materials repository (CMR) and Materials Project databases are used to construct a gradient boosting regression (GBR) model, which is used to filter for bandgaps. 2. Compositions are ranked using the Herfindahl Hirschman Index (HHI_R) and matching structures sought from a database of quaternary oxides (QuOx DB). The probabilistic oxidation state model filters out unlikely species combinations and competing phases are found from the MP database. 3. Thermodynamic stability and bandgaps are calculated from first principles using density functional theory (DFT) and hybrid DFT.

7.2 Machine learning model

7.2.1 Representation of training data

The target property for the ML model is the bandgap calculated using the GLLB-sc XC functional,[1] which has been shown to predict bandgap values more accurately than GGA XC functionals.[2] The increase in accuracy is due to the efforts that have been made in the construction of this functional to estimate the *derivative discontinuity*. While it is beyond the scope of this chapter to fully review the bandgap-predicting accuracy of different levels of theory, it should be noted that the Kohn-Sham gap in DFT differs from the "true" gap (i.e. the difference between the ionisation potential and electron affinity) by the derivative discontinuity. The GLLB-sc bandgap, therefore, represents an affordable alternative to higher levels of theory such as GW that in principle can give the true bandgap directly. The bandgap values in the dataset produced by Castelli *et al.* are used as a training set,[3] and are available from the Computational Materials Repository (CMR) database.[4] This set is comprised of 2,289 inorganic materials, 799 of which are oxides (i.e. contain oxygen and at least one other element).

The compositions of the materials are represented using the element properties in the Magpie package.[5] These are the minimum, maximum, range, mean, mode and mean absolute deviation of atomic number, Mendeleev number, atomic mass, melting temperature, electronegativity, among others (see supplementary information of Reference 5 for full list). The number of valence electrons is also used, as well as elemental HOMO/LUMO energies calculated from neutral atoms with DFT at the LDA level. The final feature is the bandgap center position calculated using the geometric mean of electronegativities as demonstrated by Nethercot.[6] All of these features are generated using the Matminer package.[7]

7.2.2 Model tuning

Gradient boosting regression (GBR) is used to build the model and two strategies are considered:

(i) A one-step process using the materials in the CMR dataset as a training set to train a model that predicts GLLB-sc bandgaps from compositions.

(ii) A two step process, firstly using the Materials Project (MP) database to train a model

that predicts PBE bandgaps from compositions, then the PBE bandgaps are correlated to GLLB-sc bandgaps.

If there is a strong correlation between PBE and GLLB-sc bandgaps, the overall accuracy for approach (ii) could be higher than for approach (i), as the MP database constitutes a larger dataset, which could improve performance of the ML model. Firstly, two training/testing scenarios are compared for approach (i):

(1) Only the *oxides* in the dataset are used to train the model, then *oxides* bandgaps are predicted during 10-fold CV.

(2) *All* the materials in the dataset are used to train the model, then *oxide* bandgaps are predicted during 10-fold CV.

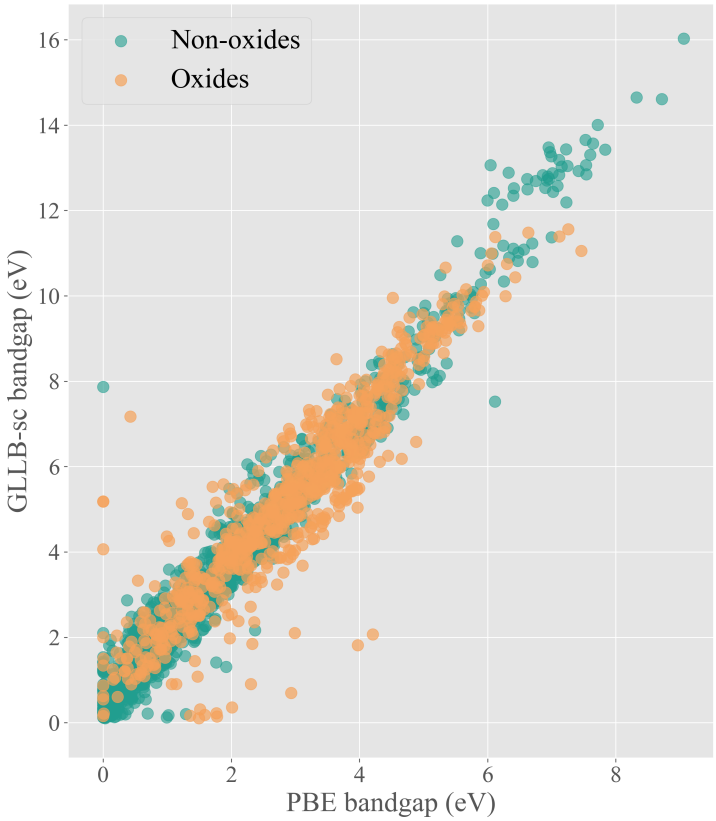

Figure 7.2: Correlation between bandgap calculated with the PBE and the GLLB-sc functionals for materials in the CMR dataset.

In each case, the root mean squared error (RMSE) of the predictions during 10-fold CV

is used as the loss function to quantify performance and hyperparameters are left at their default values as per the `scikit-learn` package.

For scenario (1), the RMSE is 1.18 eV, and for scenario (2) the RMSE is 1.20 eV, indicating that there is no improvement when including non-oxides in the training set. Figure 7.2 shows the correlation between PBE and GLLB-sc bandgaps for the CMR dataset. The GLLB-sc bandgaps are those in the CMR dataset, while the PBE calculations had been carried out on the same compounds as part of the Materials Project workflow.[8] The expected linear trend is observed, with a characteristic underestimation of the bandgap by the PBE functional.

There is a significant spread around the linear relationship and this spread is larger in general for oxides. The standard deviation is 0.85 eV, meaning that for approach (ii) to be advantageous over approach (i) the RMSE of the GBR model trained on the MP dataset would have to be unreasonably low (< 0.35 eV). It is therefore practical to instead opt for approach (i) and tune the hyperparameters of the learner to improve performance as much as possible.

Optimal hyperparameter values for this GBR model were found using the procedure outlined in Chapter 3 and are listed in Table 7.1. Using these parameters, as well as removing oxide gases such as CO_2 and SO_2, complex anions containing uncommon oxidation states such as phosphites and perphosphates, yields a final model with a RMSE of 0.95 eV.

Table 7.1: Hyperparameter values used in final GBR model.

Parameter	Optimal value
minimum compounds to split a node	65
maximum depth of tree	20
minimum compounds at a leaf	1
max features considered	86
fraction of compounds to fit each tree	0.9
learning rate	0.0145
number of boosting stages	8000

7.2.3 Model performance

The accuracy of any ML model that predicts bandgaps from composition alone is limited due to the influence of crystal structure and this is especially true of oxides, as the structural diversity of oxides results in a wide variety of local bonding arrangements. This phenomenon has been quantified by Walsh and Butler, who have demonstrated that for oxygen the Madelung site potential – a quantity that reflects the electrostatic potential of

an ion in a crystal by approximating ions as point charges – varies across all binary metal oxides with a striking range of 16 V.[9]

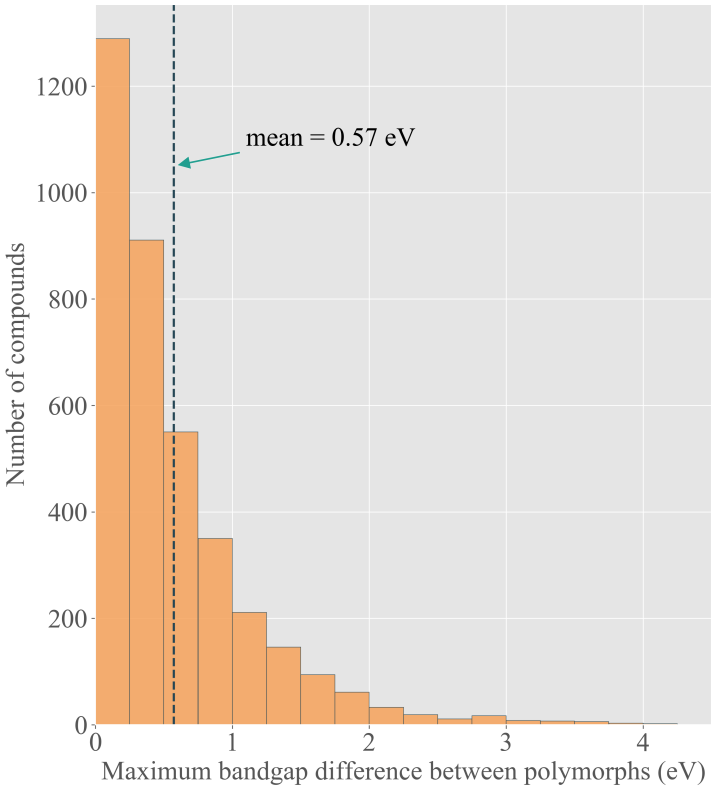

Figure 7.3: Distribution of maximum bandgap difference between polymorphs for oxides in the Materials Project (MP) database that exhibit polymorphism. Only compounds with an energy above the convex hull of < 0.1 eV and a maximum bandgap difference of > 0.05 eV are included. Bandgaps are calculated in the MP using the PBE XC functional.

Figure 7.3 shows the distribution of the maximum PBE bandgap difference between polymorphs for all oxide compositions in the MP database that exhibit polymorphism. While for a large number of oxides, polymorphism results in a bandgap difference of < 0.5 eV, the difference can be as large as 4.18 eV (e.g. $LiFePO_4$) and the mean difference is 0.57 eV. This highlights the extent to which crystal structure plays a role in determining bandgap, and that a model that considers chemical composition alone can only be used as a rough pre-screening filter. In this context, a model with a RMSE of 0.95 eV is suitable. Some examples of predicting bandgaps from composition have involved using larger datasets and lower RMSE values have been reported. For example, Zhuo *et al.*, have trained a support vector machine model on 2,458 compounds with experimental bandgaps and achieved a

RMSE of 0.45 eV.[10] However, given the effect of polymorphism on bandgap values, it is likely that such models would not perform so well outside of the training dataset.

It is also instructive to compare the performance of this model with the SSE approach used in Chapter 4. From Figure 7.4, it is clear that using SSE values to predict bandgaps of the oxides in the training set is not viable, as there is no correlation between the predicted bandgap and GLLB-sc calculated bandgap. Importantly, Figure 7.4a shows the GBR predicted bandgap for each oxide during cross-validation, i.e. when that compound was not used in the training of the model.

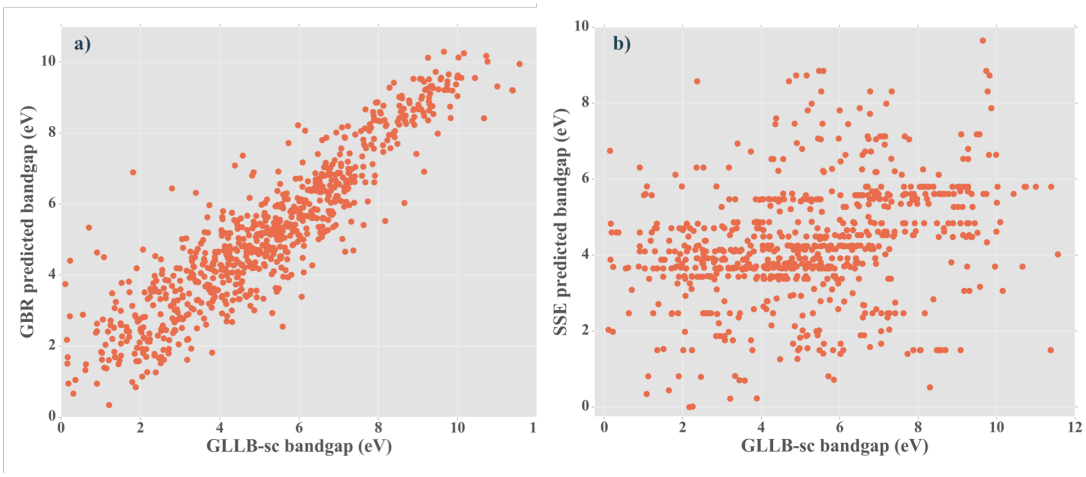

Figure 7.4: Ground truth (GLLB-sc) bandgaps vs a) bandgaps predicted using the gradient boosting regression (GBR) model and b) using the solid state energy (SSE) scale, for all oxides in the training data set.

Finally, we can inspect which features are most important in the final GBR model. Figure 7.5 shows that the mean melting temperature of the elements is the most important feature by a large margin. The range of values for volume per atom and mean absolute deviation of melting temperature are also relatively important. The extent to which this can be interpreted as meaningful depends on how highly correlated the features are. For example, we would expect covalent radius and volume per atom to be correlated to some degree, which makes it harder to decouple their contributions to the overall model. In general, a number of features contribute significantly to the final model and investigation into the effect of systematically removing correlated features and retraining the model would be an interesting avenue for further study.

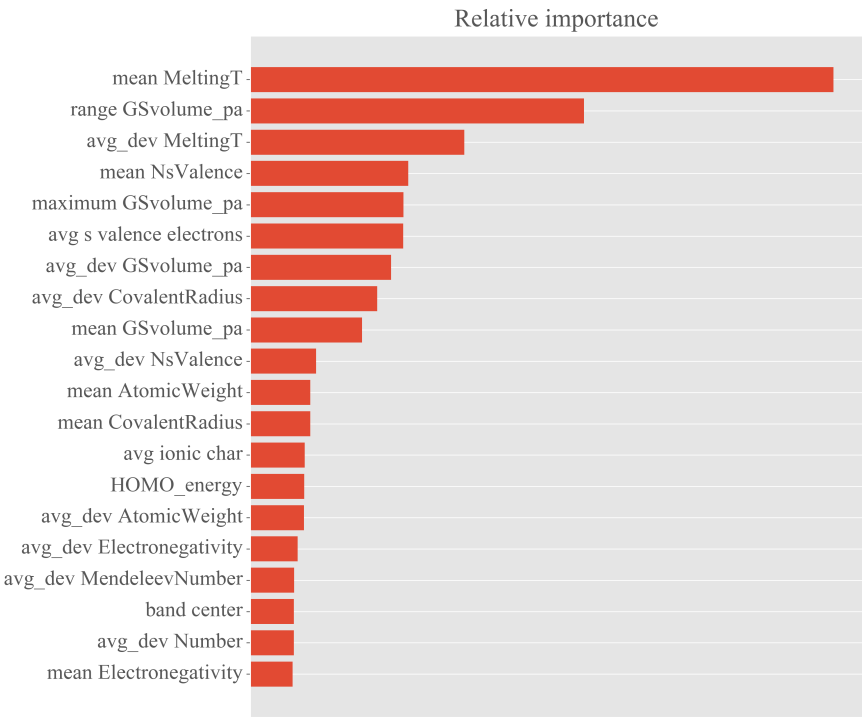

Figure 7.5: Relative importance of the 20 most important features in the final gradient boosting regression (GBR) model. *MeltingT* refers to melting temperature, *GSvolume_pa* refers to volume per atom from ground state $0K$ calculations, *NsValence* refers to number of valence s electrons, *ionic char* refers to Pauling's empirical ionic character between pairs of atoms calculated using electronegativities,[11] *band center* is the calculated using the approach by Nethercot,[6] and *Number* refers to atomic number.

7.3 Bandgap screening

We now use this model to search for promising candidates from a large search space (1.1 million) of hypothetical quaternary oxide compositions, that have been generated using the `SMACT` package, implementing the rules outlined in Chapter 4. The target bandgap window is 1.0 – 2.5 eV in order to capture potential photovoltaic materials as well as solar fuel materials, as the latter often require larger bandgaps in order to mitigate against the combination of loss mechanisms found in practical devices.[12,13]

Figure 7.6a shows the distribution of errors obtained using the GBR model. Materials predicted by the model to have a bandgap at the centre of the target window (1.75 eV) have a 60% probability of having a GLLB-sc bandgap within the window. This assumes the accuracy of the model is consistent across all bandgaps. By contrast, Figure 7.6b shows the distribution of bandgaps of all oxides in the CMR dataset and the probability of choosing one at random with a bandgap in the target window is just 8%.

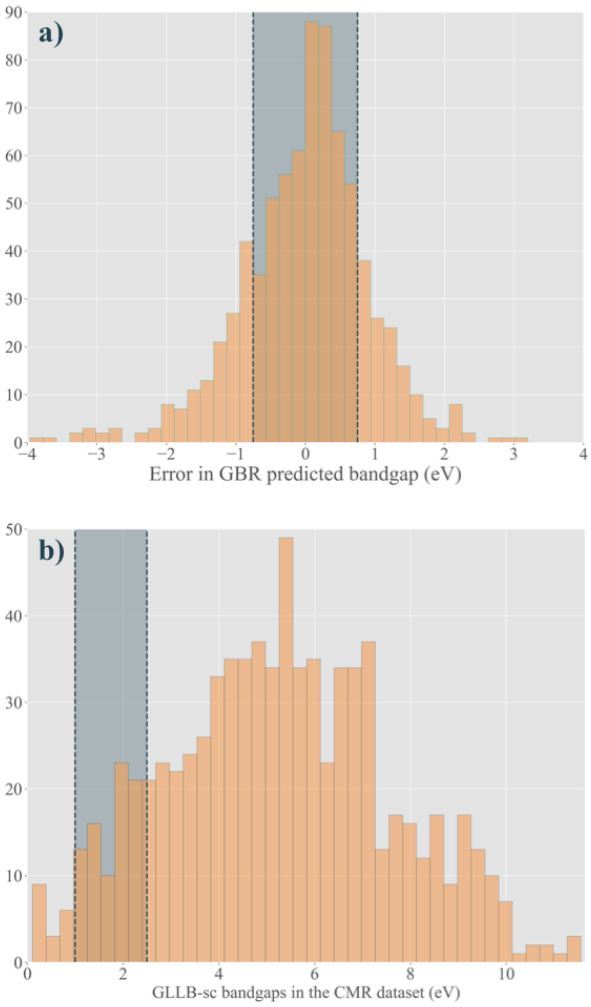

Figure 7.6: a) Distribution of error in predicted bandgap by the final GBR model. The shaded region corresponds to an error of ± 0.75 eV and encloses 60% of all predictions. b) Distribution of GLLB-sc bandgaps for oxides in the CMR training dataset. The shaded region corresponds to a band gap of 1.8 ± 0.75 eV, i.e. 1.0 – 2.5 eV.

Applying the GBR model to the search space of 1.1 million quaternary oxides, we filter out those which do not have a predicted bandgap of 1.75 ± 0.02 eV. This leaves 17,833 compositions. It should be emphasised that this approach does not aim to capture all the hypothetical compositions that fall between within the target bandgap window. Rather, those compositions that are most likely to have useful bandgaps according to the GBR model are targeted. This screening step corresponds to a greater than 60-fold reduction of the search space.

7.4 Crystal structure search

Compositions are ranked by sustainability using the HHI_R scale.[14] A database of all possible quaternary oxide crystal structures for the 1.1 million starting compositions is constructed using the structure substitution algorithm by Hautier *et al.*[15] and contains over 2 million compounds (QuOx DB in Figure 7.1). Starting with the most sustainable composition, a search is carried out on this database to find any matching compounds until **n** compositions have had at least one crystal structure assigned to them. **n** was initially set to 100, however this did not yield enough candidates later in the screening process, so was subsequently set to 135.

After checking that the oxidation states in the crystal structures that are identified for each composition are consistent with the oxidation states in the original composition generated by `SMACT`, the oxidation state probability model (Chapter 5) is applied to filter the compounds. A relaxed probability threshold of 0.05 is used so only very unlikely species combinations are eliminated. We also choose to eliminate Ti^{3+} compounds due to the d^1 electronic configuration being linked to fast electron-hole recombination. In addition, such compounds would not be amenable to a general high-throughput DFT workflow, due to the well-known challenges for electronic-structure modelling of highly correlated systems.[16] This results in 235 candidate structures, corresponding to 61 different compositions. We take these candidates forward to calculate their thermodynamic stability using DFT.

7.5 Thermodynamic stability

Competing phases are identified using the MP database and geometry optimisations are carried out on candidates and all competing phases using DFT at GGA level. This is done in high-throughput using the `Atomate`[17] and `Fireworks`[18] packages and calculation details are identical to those used to calculate thermodynamic stabilities in Chapter 5.

Of the 235 compounds, 27 are calculated to be within 100 meV/atom of the convex hull. Four of the 27 compounds were found to be structurally identical to one other compound in the set, leaving 23 unique compounds. The presence of identical structures can occur when different parent structures are found for a composition using the structure substitution algorithm which ultimately yield the same crystal structure after relaxation.

The relatively small proportion of stable and metastable compounds is unsurprising given the existence of a large number of stable binary and ternary oxides that act as competing phases. The energies above the convex hull for the candidate compounds are given in Table 7.2. Only one compound, Li_2MnSiO_5, has been previously reported in the MP database, but has not been synthesised experimentally to the author's knowledge. Shown in Figure 7.7a, the compound $ZrMnSi_2O_7$ is the only one predicted to be thermodynamically stable, while a second polymorph of $ZrMnSi_2O_7$ along with a Li_2TiMnO_4 structure are predicted to be < 10 meV/atom above the convex hull, as shown in Figure 7.7b and Figure 7.7c, respectively.

While three polymorphs of Li_2TiMnO_4 are in the MP database, including one that has been investigated as a possible active material for Li-ion battery applications,[19] none of the crystal structures adopted by the candidate compounds have previously been reported. The new phase of Li_2TiMnO_4 differs from the three previously reported polymorphs as the metals are in tetrahedral environments as opposed to octahedral. It also has a wide bandgap of 4.21 eV, as calculated using Hybrid DFT in the following section, whereas the previously reported compounds all have GGA bandgaps of less than 0.4 eV. No compounds have previously been reported for any of the other compositions listed in Table 7.2.

7.6 Bandgap calculations

The bandgaps of the 23 candidate compounds were calculated with hybrid DFT using the HSE06 functional.[20,21] The same calculation procedure was used that was employed to calculate bandgaps in Chapter 6, with Γ-centred homogeneous k-point meshes of density 64 Å^3 in the reciprocal lattice. The majority of compounds have a calculated bandgap of $>$ 4 eV, which is well outside the target bandgap window (Table 7.2). Four of the compounds are calculated to have bandgaps within the target window. The most thermodynamically stable compound with a useful bandgap is $MnAg(SeO_3)_2$ and is shown in Figure 7.8.

Encouragingly, the four compounds with useful bandgaps include three different compositions, while the compounds with too large a bandgap include five different compositions. Since the original GBR model is trained on composition alone, this preliminarily indicates that it is performing at a 37.5% success rate. For this set of compositions, the structure prediction algorithm was able to find more crystal structures for the compositions on which the GBR model performed poorly.

The success rate of 37.5% is not as high as the original 60% as indicated by the 10-fold CV

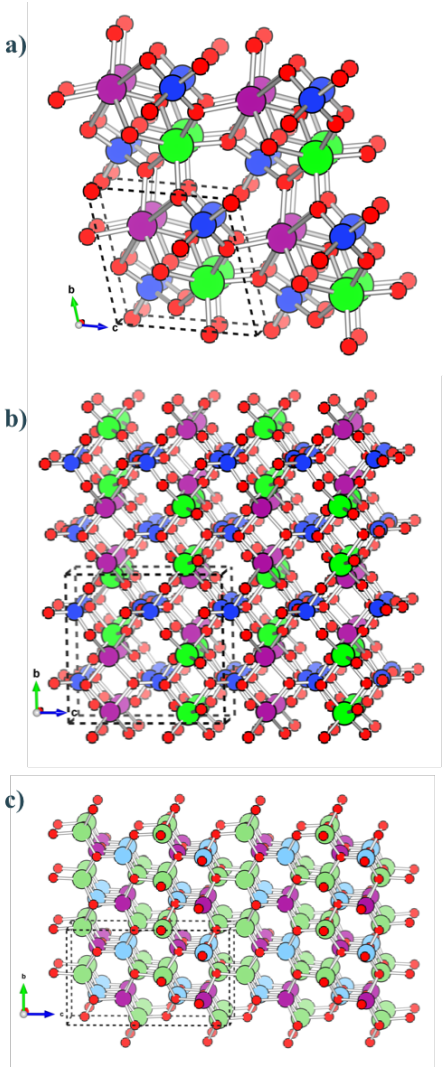

Figure 7.7: Three most stable compounds identified by the workflow. a) and b) are different polymorphs of $ZrMnSi_2O_7$ in which Si, Zr and Mn atoms are depicted as blue, green and purple circles, respectively. c) A Li_2TiMnO_4 structure in which Li, Ti and Mn atoms are depicted as green, blue and purple circles, respectively. O atoms are red circles in all three structures.

results. The latter should be considered a maximum achievable success rate when using this model predictively. This is indicative that the GBR model may exhibit high variance and is overfit to the training data to some extent. This is not something that can easily be mitigated against when training models on small datasets, as cross validation is the only option to guard against overfitting (there is not enough data to leave a third portion for a final test after hyperparameter tuning).

It is also important to note that the model was trained on bandgaps calculated using the GLLB-sc XC functional, while these bandgaps are calculated using the HSE06 XC functional as it is not possible to use the former functional in the present high-throughput DFT

Table 7.2: Summary of most stable compounds found after high-throughput DFT calculations. Bandgaps calculated with hybrid DFT that fall within the target window of 1.0 – 2.5 eV are shown in bold.

Number	Formula	space group symbol	E_{hull} (meV/atom)	Bandgap (eV)
1	$MgFe(SO_4)_2$	$P2_1/m$	99	4.07
2	$MgFe(SO_4)_2$	C2/m	11	4.15
3	Li_2MnSiO_5	P4/nmm	86	**2.24**
4	$MnCdGe_2O_6$	$P2_1/c$	99	**2.47**
5	$MnCdGe_2O_6$qua	C2/c	99	**1.76**
6	$ZrMnSi_2O_7$	C2	0	4.64
7	$ZrMnSi_2O_7$	P-1	40	4.32
8	$ZrMnSi_2O_7$	P-1	72	3.95
9	$ZrMnSi_2O_7$	$P2_1/m$	3	4.33
10	$ZrMnSi_2O_7$	$P2_1/c$	39	4.40
11	$ZrMnSi_2O_7$	$P2_1/c$	36	5.12
12	Na_2YFeO_4	Pc	79	4.27
13	Na_2YFeO_4	$Pmn2_1$	90	4.33
14	$MnAg(SeO_3)_2$	$Pna2_1$	36	**2.31**
15	Li_2TiMnO_4	$P2_1/c$	38	4.10
16	Li_2TiMnO_4	I-42m	96	4.05
17	Li_2TiMnO_4	$Pna2_1$	40	4.19
18	Li_2TiMnO_4	$Pmn2_1$	11	4.23
19	Li_2TiMnO_4	Pnma	4	4.21
20	Li_2TiMnO_4	$P2_1/c$	31	4.58
21	Li_2TiMnO_4	Pnma	60	4.05
22	$NaCaFeO_3$	$Pna2_1$	61	3.73
23	$NaCaFeO_3$	$P2_1/c$	60	2.87

approach. In the original work by Castelli *et al.* in which they create the dataset used for training here, they show that bandgaps calculated using HSE06 and GLLB-sc are generally in good agreement.[3] However, they also show that for lower bandgaps such as those considered here, the HSE06 functional has a tendency to overestimate as compared with the GLLB-sc functional. This could be another reason for getting a lower success rate and would also explain why no compounds had bandgaps calculated using HSE06 lower than the target window. Future work will include investigating to what extend the model can be improved by reducing variance, and to what extent the results would improve by using the GLLB-sc XC functional.

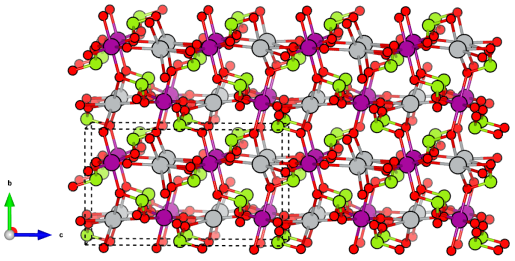

Figure 7.8: The most stable compound identified by the workflow with a bandgap within the target window,$MnAg(SeO_3)_2$. Mn, Ag, Se and O atoms are depicted as purple, silver, green and red circles, respectively.

7.7 Conclusion

In this chapter, a GBR model was built using a ML approach to predict bandgaps for quaternary oxide compositions. It was shown that the model performs reasonably well given the size of the training dataset and the extent to which it is possible to determine bandgap from composition alone, with an average RMSE from cross validation of 0.95 eV. The model is then used as part of a high-throughput screening workflow, in conjunction with the substitution algorithm to assign structures, and the oxidation state probability model to discard unlikely species combinations. A search is carried out on a space of 1.1 million quaternary oxide compositions generated using the `SMACT` package in order to identify new materials for solar applications. Using high-throughput DFT , 23 compounds are identified as falling within a stability window of < 100 meV/atom above the convex hull. Finally, four of these 23 compounds with three different compositions are calculated to have bandgaps that fall within a useful window of 1.0 - 2.5 eV. By using a combination of chemical heuristics and data-driven screening steps, the overall computational cost of the process is kept low. There is now ample scope for the remaining search space of oxide compounds predicted to have useful bandgaps to be further investigated using first-principles techniques.

Bibliography

[1] M. Kuisma, J. Ojanen, J. Enkovaara and T. T. Rantala, *Phys. Rev. B*, 2010, **82**, 115106.

[2] I. E. Castelli *et al.*, *Energy Environ. Sci.*, 2012, **5**, 5814–5819.

[3] I. E. Castelli *et al.*, *Adv. Energy Mater.*, 2015, **5**, 1400915.

[4] D. D. Landis *et al.*, *Comput. Sci. Eng.*, 2012, **14**, 51–57.

[5] L. Ward, A. Agrawal, A. Choudhary and C. Wolverton, *npj Comput. Mater.*, 2016, **2**, 16028.

[6] A. H. Nethercot, *Phys. Rev. Lett.*, 1974, **33**, 1088–1091.

[7] L. Ward *et al.*, *Comput. Mater. Sci.*, 2018, **152**, 60–69.

[8] *The Materials Project*, `https://materialsproject.org/-[Accessed: 01-01-2016]`.

[9] A. Walsh and K. T. Butler, *Acc. Chem. Res.*, 2014, **47**, 364–372.

[10] Y. Zhuo, A. Mansouri Tehrani and J. Brgoch, *J. Phys. Chem. Lett.*, 2018, **9**, 1668–1673.

[11] L. Pauling, *The Nature of the Chemical Bond*, Cornell University Press, Ithaca, 3rd edn., 1960.

[12] T. Bak, J. Nowotny, M. Rekas and C. Sorrell, *Int. J. Hydrogen Energy*, 2002, **27**, 991–1022.

[13] B. A. Pinaud *et al.*, *Energy Environ. Sci.*, 2013, **6**, 1983–2002.

[14] M. W. Gaultois *et al.*, *Chem. Mater.*, 2013, **25**, 2911–2920.

[15] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain and G. Ceder, *Inorg. Chem.*, 2011, **50**, 656–663.

[16] B. J. Morgan and P. A. Madden, *Phys. Rev. B*, 2012, **86**, 035147.

[17] K. Mathew *et al.*, *Comput. Mater. Sci.*, 2017, **139**, 140–152.

[18] A. Jain *et al.*, *Concurr. Comput.*, 2015, **27**, 5037–5059.

[19] M. Küzma *et al.*, *J. Power Sources*, 2009, **189**, 81–88.

[20] J. Heyd, G. Scuseria and M. Ernzerhof, *J. Chem. Phys.*, 2003, **118**, 8207–8215.

[21] A. V. Krukau, O. A. Vydrov, A. F. Izmaylov and G. E. Scuseria, *J. Chem. Phys.*, 2006, **125**, 224106.

Chapter 8

Summary

8.1 Key findings

Firstly, a method has been demonstrated for defining the composition space of 2-, 3- and 4-component stoichiometric inorganic materials using the heuristic tools of electronegativity and oxidation state. These tools are used to ensure that the compositions make some "chemical sense" and reduce the space of exhaustive numerical element combinations by roughly two orders of magnitude. We have seen that when the stoichiometry is limited to an integer that reflects a natural limit for most known structure types (i.e. 8), the number of existing compounds is but a tiny fraction of the search space for 3- and 4-component materials, which respectively exceed 10^7 and 10^{10} compositions. The fact that we are able to enumerate so many hypothetical compositions quickly on a desktop computer using Python is both promising and useful from a materials discovery viewpoint.

We have also seen in Chapter 5 that the extent to which oxidation states are considered to be accessible for different elements is open to some interpretation. By looking in more detail at existing oxidation states data, it is possible to build a screening tool that considers the likelihood of element combinations and further narrows down the search space, as demonstrated by an immediate 3-fold reduction of ternary metal halides, also in Chapter 5.

Secondly, screening for target properties based on compositional descriptors is shown to be a practical, low-cost, step to place at the beginning of a materials design workflow. Sustainability metrics based on element resources (such as the $\mathrm{HHI_R}$) do not require any structural information. We have seen that the SSE scale can be used to identify promising

candidates for photoelectrochemical water splitting and that ML techniques can be used to target compounds for solar energy applications by correlating compositional descriptors to bandgap.

These techniques assisted in the identification of $Cd_5(S_2Cl)_2$ for photoelectrochemical water splitting (Chapter 6) and $Mn_2Ag(SeO_3)_2$ for solar energy applications (Chapter 7), among other possibly metastable candidates. In both cases, only a small set of compounds (4 and 23, respectively) are carried all the way through to hybrid DFT calculations to determine accurate properties due to practical limitations. It is therefore difficult to draw any quantitative conclusions about the performance of each individual screening approach. Nevertheless, it is promising that potentially useful compounds are identified from unpremeditated regions of each search space at lower cost than a high-throughput DFT or experimental approach.

Finally, going from composition to crystal structure presents a significant challenge for the materials design approaches presented. The difficulty in predicting structure given knowledge of chemical composition is well documented and was described as a "scandal" by J. Maddox exactly 30 years ago.[1] In Chapter 6 we saw that it is possible to suggest crystal structures by using expensive evolutionary algorithms, driven by first-principles techniques, that have been developed in response to this challenge. We also so that more recently developed data-driven approaches can be used.

It was shown that the evolutionary algorithm consistently finds lower energy structures, sometimes by a significant margin such as 34 meV/atom for $Sn_5S_4Cl_2$, and sometimes only fractionally, e.g. 0.2 meV/atom for Cd_4SF_6. Again, the sample set is limited, and further work is needed to compare these two approaches across a range of chemistries. Importantly, only DFT total energy is used to assess stability in this work, as is often the case. As has been mentioned previously, thermodynamic stability does not necessarily guarantee that a material can be synthesised, nor that it will be dynamically stable if it can be synthesised.

8.2 Future work

Improvements to SMACT: The results in this thesis open up many interesting directions for future study: One opportunity is to further improve the `smact` library, adding features beyond those which build search spaces and screen based on heuristic rules. For example, the substitution model for predicting crystal structure is currently implemented

in the pymatgen code.[2] The advantage of this is that new structures can seamlessly be imported into other workflows that use some of the powerful features of this package. The disadvantage is that the structure substitution process itself is quite slow, taking between 3 and 4 minutes per composition to search a database of around 30,000 structures. While this is orders of magnitude faster than the evolutionary algorithm-based alternative, it could certainly be made faster by avoiding the heavy objects that are required elsewhere in pymatgen. If a similar algorithm were implemented in SMACT, the priority would be to create a simple database of existing structures, keyed by composition, that can be queried as quickly as possible. This speed up could allow for the generation of structures for many thousands of compositions on a desktop computer, without the need to turn to parallel computing to achieve the same goal.

There is also scope to improve upon the probabilistic oxidation states model that is built into SMACT. As discussed in the Remarks section of Chapter 5, more sophisticated models that factor in the presence of multiple anions and cations could be investigated. Additionally, this may be an appropriate problem for a supervised ML approach, where compositional descriptors are used to predict the oxidation states in a compound.

Further machine learning studies: The ML model used to predict bandgap from composition is another potential topic of investigation. There are many degrees of freedom involved at every stage of the model building process (Figure 3.1, Chapter 3) and this means that often quite arbitrary decisions are made without full investigation of the parameters at hand. For example, the choice of learner in this case was gradient boosting regression (GBR), in which decision trees are built sequentially. A similar approach is random forest (RF) in which trees are built in parallel and an average prediction from all trees is taken as a final prediction. RF models are formed of more fully grown trees so are less prone to bias (but more prone to overfitting) so it would be interesting to see how a RF model performed. The choice of features that represent compositions also introduces many more degrees of freedom and a systematic investigation into the optimal combination could give insights into which properties are most important. As a final example, the hyperparameter tuning process can also be automated using Bayesian optimisation and Gaussian processes or similar. In this kind of approach, the next set of hyperparameters to test is made by an acquisition function over a surrogate model which is much quicker to evaluate than testing the model itself. This is implemented in various packages including the skopt Python library.[3]

Given that a large number of hypothetical materials can be generated quickly using the structure substitution algorithm (a database of 2 million quaternary oxides was created

for the workflow in Chapter 7), a different approach could involve using these as input to a ML algorithm that uses structural features to predict properties. Representing inorganic crystalline solids to ML algorithms is a relatively new challenge and some methods are emerging.[4–6] However, most examples of learning properties from structure to date sidestep this issue by focusing on just one structure type at a time.[7,8] The application of new representation methods to predict real materials properties is likely to be an important area of development over the next few years.

Other application areas: Finally, we have so far only explored these new tools in the context of solar energy materials. This was done partly for convenience, as the bandgap constitutes a clear performance metric that can be directly computed from first principles. By adapting the workflows presented here, the composition space generated in `smact` could be searched for new materials for different application areas, from thermoelectrics, to ferroelectrics, to battery materials. For instance, there has been a recent rise in interest in anionic redox materials for battery cathodes. In these materials, the anion is also partially oxidised along with the cation, and this phenomenon has the potential to significantly increase battery capacity.[9] The tools presented here would be well suited for this application; ML techniques and chemical heuristics can link cheap descriptors to the required electronic structure features, `SMACT` and the structure substitution algorithm can provide a pool of new compounds to test, and high-throughput DFT can be used to validate and improve the ML model.

Bibliography

[1] J. Maddox, *Nature*, 1988, **335**, 201.

[2] S. P. Ong *et al.*, *Comput. Mater. Sci.*, 2013, **68**, 314–319.

[3] T. Head, *Scikit-optimize*, `https://scikit-optimize.github.io/-[Accessed: 05-09-18]`.

[4] K. T. Schütt *et al.*, *Phys. Rev. B*, 2014, **89**, 205118.

[5] L. Ward *et al.*, *Phys. Rev. B*, 2017, **96**, 024104.

[6] O. Isayev *et al.*, *Nat. Commun.*, 2017, **8**, 15679.

[7] F. A. Faber, A. Lindmaa, O. A. Von Lilienfeld and R. Armiento, *Phys. Rev. Lett.*, 2016, **117**, 135502.

[8] F. Legrain, J. Carrete, A. van Roekeghem, G. K. H. Madsen and N. Mingo, *J. Phys. Chem. B*, 2018, **122**, 625–632.

[9] G. Assat and J. M. Tarascon, *Nat. Energy*, 2018, **3**, 373–386.

Closing Remarks

As a community of chemists and materials scientists, we have barely scratched the surface of the inorganic composition space. Even for binary, ternary and quaternary stoichiometric systems, simply enumerating the number of possible compositions presents its own challenges and yields a search space that is intractable to high-throughput first-principles calculations, let alone experiment. Computational materials design is a rapidly advancing field that is rising to this challenge and there are many tools emerging from the fields of machine learning and big data that can be applied in a variety of ways to large search spaces. Meanwhile, chemical heuristics still have an important part to play, and their codification can provide intuitive links between descriptors and materials properties.

In this thesis, we have seen approaches to screening the inorganic composition space using a mixture of tools, leading to the successful prediction of feasible chalcohalide and oxide compounds with target bandgaps, as verified by DFT. One of the largest hurdles for composition-based screening is the ability to predict stable structures for those compositions. This is especially true when moving to higher-order compositions, such as quaternaries, as shown in Chapter 7 where 135 compositions led to just 23 compounds with DFT total energies placing them near the convex hull. Structure prediction is a challenge that has been well documented in solid state chemistry for a number of years and it would perhaps be beneficial for more studies to focus on predicting stable compounds, including experimental verification, rather than targeting specific properties at the outset.

Another limitation for materials design is the relatively small amount of high-quality data that is available, whether from experiment or calculation. Recent investments in open access databases have begun to lift this restriction, and both researchers and algorithms can now learn from more data than ever before. With such a rapidly changing definition of what can be calculated within practical time limits, the predictive power of computational techniques is growing. It is now hard to imagine a future where first-principles and data-driven methods do not become a critical aspect of the design of all new materials.

Appendix

1. Publication 1 Supplementary Information
2. Publication 2 Supplementary Information
3. Publication 3 Supplementary Information

Chem, Volume *1*

Supplemental Information

Computational Screening of All Stoichiometric Inorganic Materials

Daniel W. Davies, Keith T. Butler, Adam J. Jackson, Andrew Morris, Jarvist M. Frost, Jonathan M. Skelton, and Aron Walsh

I. SUPPLEMENTAL DATA ITEMS

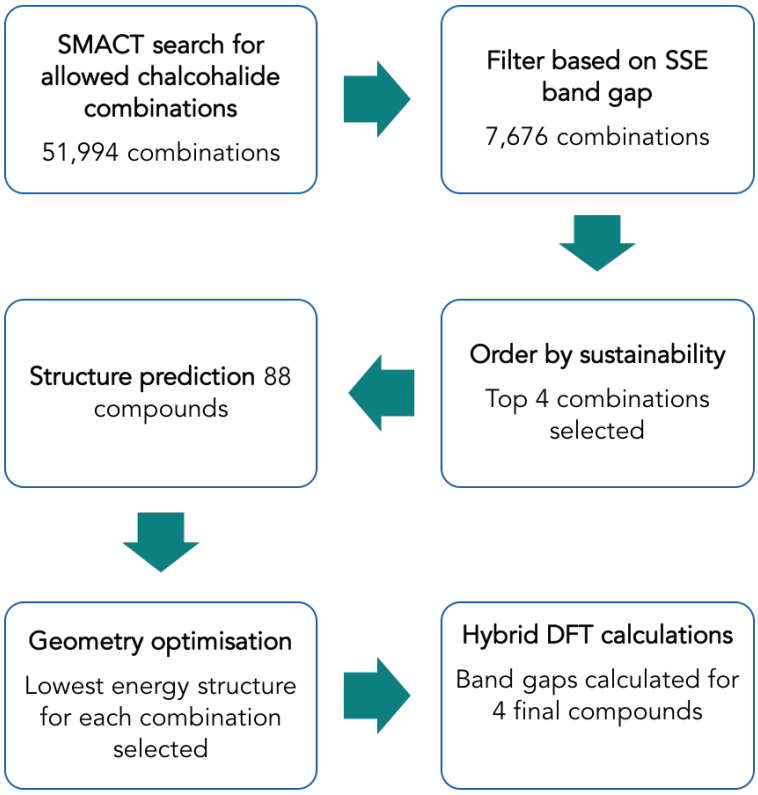

FIG. S1. Computational workflow: searching the combinatorial space for photoelectrochemical water splitting materials.

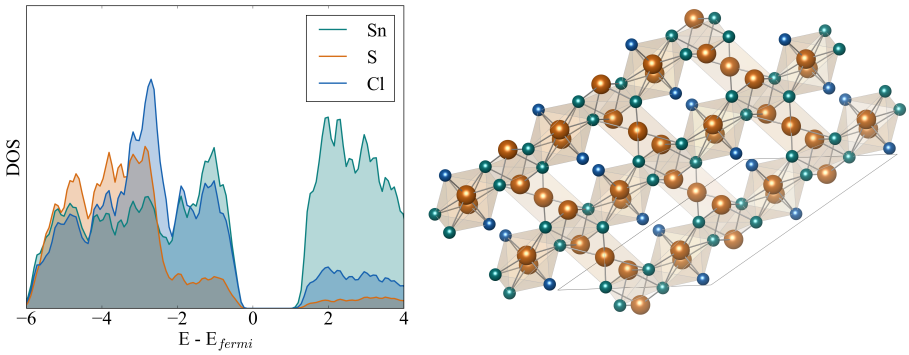

FIG. S2. (left) Electronic density of states and (right) predicted crystal structure of $Sn_5S_4Cl_2$.

TABLE S1: Comparison of measured bandgaps[1] (E_g^{exp}) against those predicted from the SSE model (E_g^{SSE}).

Material	E_g^{exp} (eV)	E_g^{SSE} (eV)
$MgSiP_2$	2.60	2.03
$ZnSiP_2$	1.70	2.00
$ZnSiAs_2$	1.00	1.93
$ZnGeN_2$	4.00	2.67
$ZnGeP_2$	1.70	2.14
$ZnGeAs_2$	1.00	1.15
$ZnSnP_2$	1.30	1.66
$ZnSnAs_2$	0.60	0.75
$ZnSnSb_2$	0.50	0.40
$CdSiP_2$	1.20	2.20
$CdSiAs_2$	0.50	1.55
$CdGeP_2$	1.20	1.73
$CdGeAs_2$	0.50	0.57
$CdSnP_2$	1.20	1.17
$CdSnAs_2$	0.50	0.26
$ZnGa_2S_4$	2.40	3.25
$ZnGa_2Se_4$	2.60	2.18
$ZnIn_2S_4$	1.80	2.87
$ZnIn_2Se_4$	2.00	1.68
$ZnIn_2Te_4$	1.40	1.35
$CdAl_2S_4$	1.90	3.40
$CdGa_2S_4$	1.90	3.16
$CdGa_2Se_4$	2.10	2.33
$CdGa_2Te_4$	1.50	1.50
$CdIn_2S_4$	1.80	2.21
$CdIn_2Se_4$	2.00	1.83
$CdIn_2Te_4$	1.40	1.15

$MgGa_2S_4$	2.50	3.40
$MgGa_2Se_4$	2.70	2.20
AsSBr	1.40	2.50
SbSI	1.50	1.88
SbSBr	1.50	2.26
SbSeBr	1.70	1.92
SbSeI	1.50	1.68
SbTeI	1.10	1.28

TABLE S2: Calculated bandgaps of top compounds identified by the screening procedure based upon density functional theory calculations (HSE06 functional) of the predicted crystal structures.

ABC combination	Formula	E_g^{calc} (eV)
CdSCl	$Cd_5S_4Cl_2$	2.96
CdSF	Cd_4SF_6	3.40
SnSCl	$Sn_5S_4Cl_2$	1.62
SnSF	Sn_4SF_6	3.00

II. SUPPLEMENTAL COMPUTATIONAL PROCEDURES

A. Validation of ternary bandgaps using the solid-state energy scale

The SSE dataset was initially built from binary compounds. In the original paper[2] the authors speculate about its applicability to ternary and higher order materials; however, we can find no reports of any such application. In order to assess whether the bandgap of a ternary material can be estimated from the difference between the highest anion and lowest cation SSE, we have tested this method against a set of well-characterised ternary semiconductor bandgaps.[1]. We compare to 35 materials, covering III-IV-V_2, II-III_2-VI_4 and V-VI-VII compounds, including metal halides, chalcogenides and pnictides. The agreement is reasonable, with a root-mean-squared deviation between of 0.66 eV. The data are presented in Table S1.

B. Workflow for selecting candidate photoelectrodes

The six step procedure that we adopt is shown schematically in Figure S1.

1. Allowed chalcohalide combinations

The constraints of charge neutrality and electronegativity are applied to all possible $A_xB_yC_z$ combinations with B = [O, S, Se, Te] and C = [F, Cl, Br, I]. Stoichiometry is restricted to $A_xB_yC_z$, where the integers $w, x, y, z \leq 8$. Additionally we limit the A cations to those with an SSE higher than the water reduction potential (approx. -4.5 V relative to the vacuum at pH = 0). This results in 51,994 combinations.

2. SSE bandgap filter

The elemental combinations with a bandgap outside the range of 1.5 – 2.5 eV according to the SSE scale are discarded. Since $\sim$ 2 eV would represent an ideal bandgap, the $\pm$0.5 eV range allows sufficient space to allow for the uncertainty in the predicted SSE values. This results in 7,676 allowed combinations.

3. Sustainability filter

The sustainability of the 7,676 $A_xB_yC_z$ combinations is assessed based on sum the HHI_R scores of the three elements. The 20 combinations with the smallest HHI_R scores are shown in Figure 2 and the four combinations with the smallest HHI_R scores are taken forward to the structure prediction stage.

4. Structure prediction

In order to ascribe three-dimensional structures to the four element combinations, we use the approach developed by Hautier *et al.*[3] based on structural analogy. It suggests probable structure types based on the likelihood of ionic substitutions in existing compounds with known crystal structures. This procedure enables a rapid screening step which returns possible compounds with an associated probability of each crystal structure being adopted. We use a probability threshold of 0.001 and the Materials Project as the database for existing compounds. This results in a total of 88 structures to be taken forward to the density functional theory (DFT) optimisation step.

5. Crystal structure optimisation

For the structural relaxations, we employ DFT with a projector-augmented plane wave basis[4] and the PBEsol exchange-correlation functional[5] as implemented in the Vienna Ab-initio Simulation Package (VASP)[6,7]. A Monkhorst-Pack k-point grid was generated for each calculation with k-point spacing of 0.242 $Å^{-1}$. The kinetic energy cutoff is set at 500 eV and the force on each atom is converged to within 0.01 $eVÅ^{-1}$. For each of the four element combinations, the lowest total energy structure of those for which a local minimum could be found was taken forward to the bandgap calculation step.

6. Electronic structure calculations

Semi-local exchange-correlation treatments such as the PBEsol functional provide an accurate description of crystal structures but tend to underestimate the electronic bandgaps of semiconductors. To overcome this issue, computations of bandgaps are performed by

using the hybrid non-local functional HSE06,[8] which includes 25% screened Hartree-Fock exact exchange. The calculated bandgaps of the four final materials are presented in Table S2.

REFERENCES

[1]O. Madelung, *Semiconductors: Data Handbook* (Springer-Verlag, Berlin, Heidelberg, 2004).

[2]B. D. Pelatt, R. Ravichandran, J. F. Wager, and D. A. Keszler, J. Am. Chem. Soc. **133**, 16852 (2011).

[3]G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder, Inorg. Chem. **50**, 656 (2011).

[4]G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).

[5]J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, Phys. Rev. Lett. **100**, 136406 (2008).

[6]G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996).

[7]G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).

[8]A. V. Krukau, O. A. Vydrov, A. F. Izmaylov, and G. E. Scuseria, J. Chem. Phys. **125**, 224106 (2006).

Materials Discovery by Chemical Analogy: Role of Oxidation States in Structure Prediction

Daniel W. Davies,[1] Keith T. Butler,[1] Olexandr Isayev,[2] and Aron Walsh[3,4,a)]

[1)] *Centre for Sustainable Chemical Technologies and Department of Chemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK*

[2)] *Laboratory of Molecular Modeling, Division of Chemical Biological and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599, USA.*

[3)] *Department of Materials Science and Engineering, Yonsei University, Seoul 03722, Korea*

[4)] *Department of Materials, Imperial College London, Exhibition Road, London SW7 2AZ, UK*

(Dated: 7 February 2018)

[a)] Electronic mail: a.walsh@imperial.ac.uk

I. SUPPLEMENTAL DATA ITEMS

A. All species fractions

Distribution of all metal species included in the dataset, normalised by the total number of compounds containing a given species. Anions on the x-axes are in order of decreasing electronegativity. Numbers to the left of each species on the y-axes show the raw number occurrences of each oxidation state for each metal. Continued on the next page.

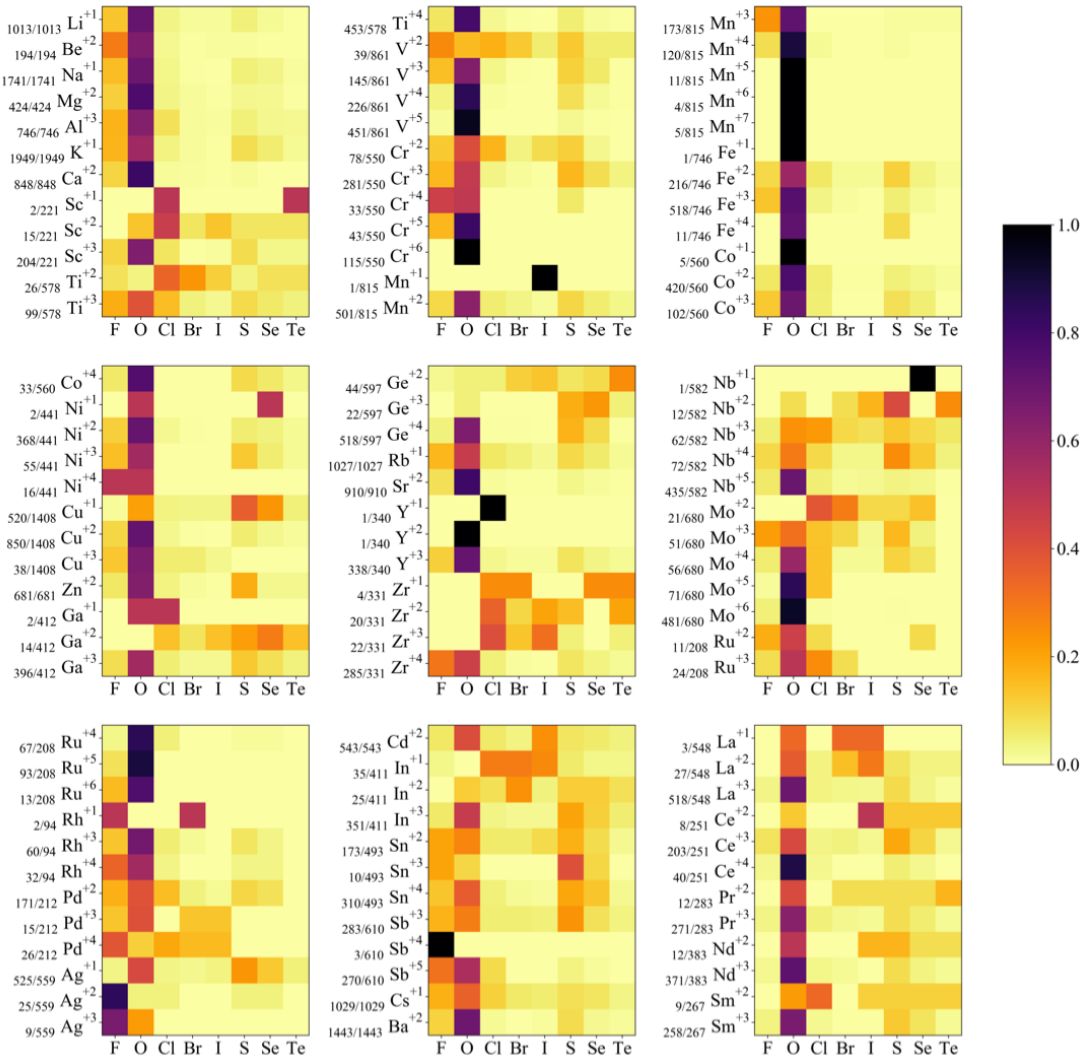

FIG. S1: Species Fractions part I.

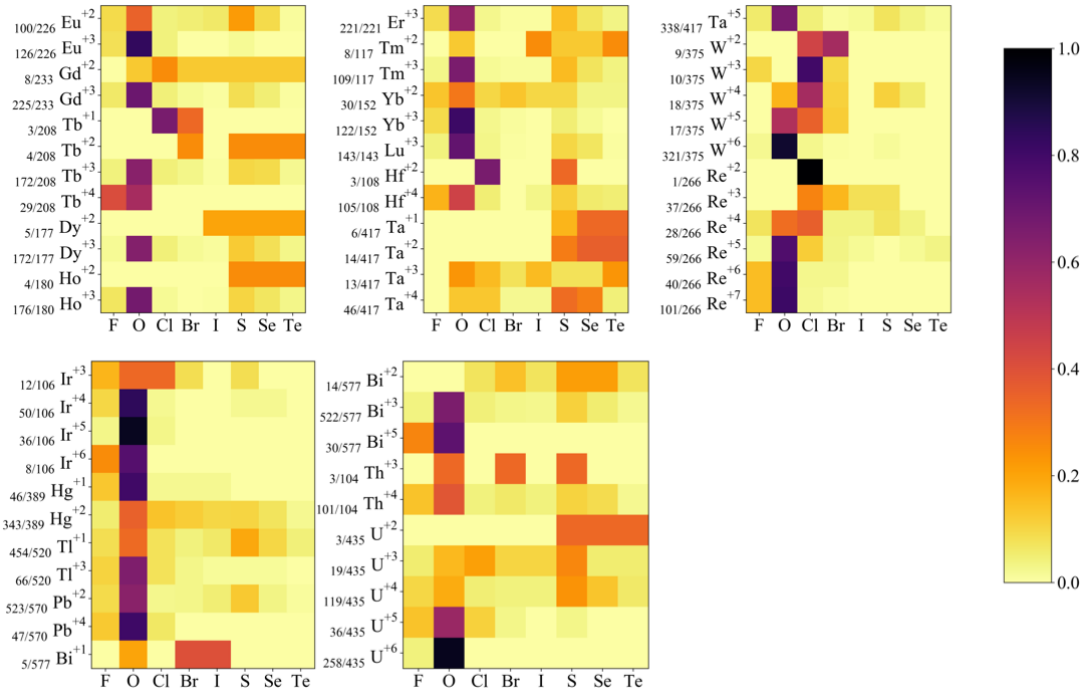

FIG. S2: Species Fractions part II.

B. Further species distributions

Additional plots of the distribution of some metal species in the dataset. The trends discussed in the main manuscript are also seen in the third row d-block metals (Figure S3) and the Lanthanides (Figure S4).

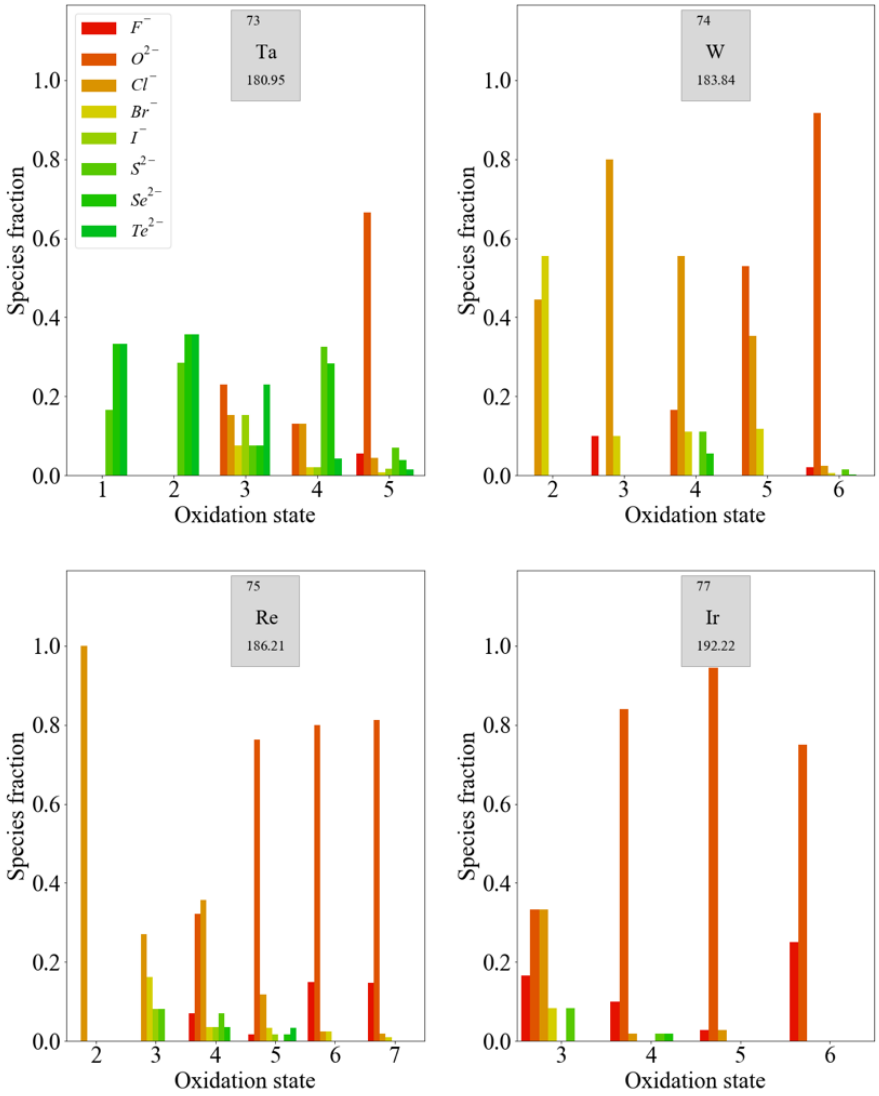

FIG. S3: Distribution of some third row transition metal species. The color scale represents the electronegativity of the most electronegative anion present in the compound from dark red (F, most electronegative) to dark green (Te, least electronegative).

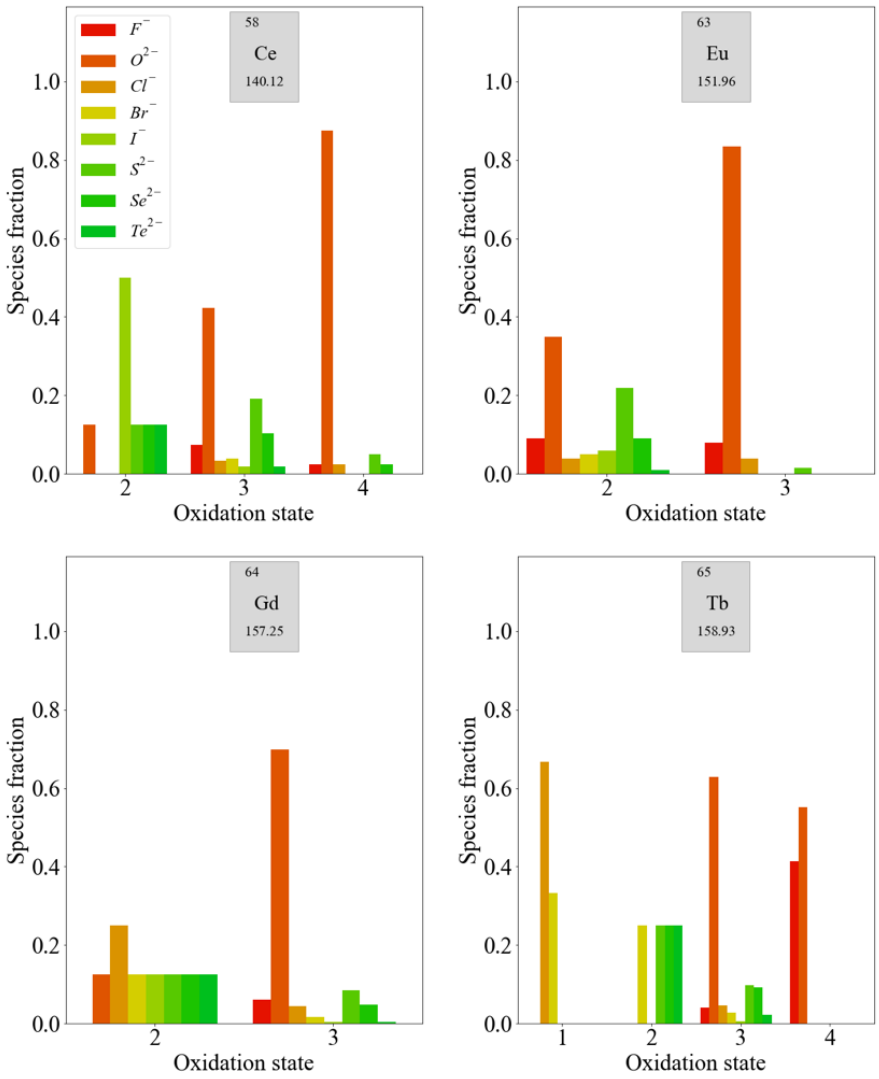

FIG. S4: Distribution of some Lanthanide metal species. The color scale represents the electronegativity of the most electronegative anion present in the compound from dark red (F, most electronegative) to dark green (Te, least electronegative).

C. All species–anion probabilities

Graphical representation of the lookup table used by the probabilistic model. The number of compounds containing a given species with the most electronegative anion is normalised by the total number of compounds containing the metal with the most electronegative anion. Anions on the x-axes are in order of decreasing electronegativity. Numbers to the left of each species on the y-axes show the raw number occurrences of each oxidation state for each metal. Continued on the next page.

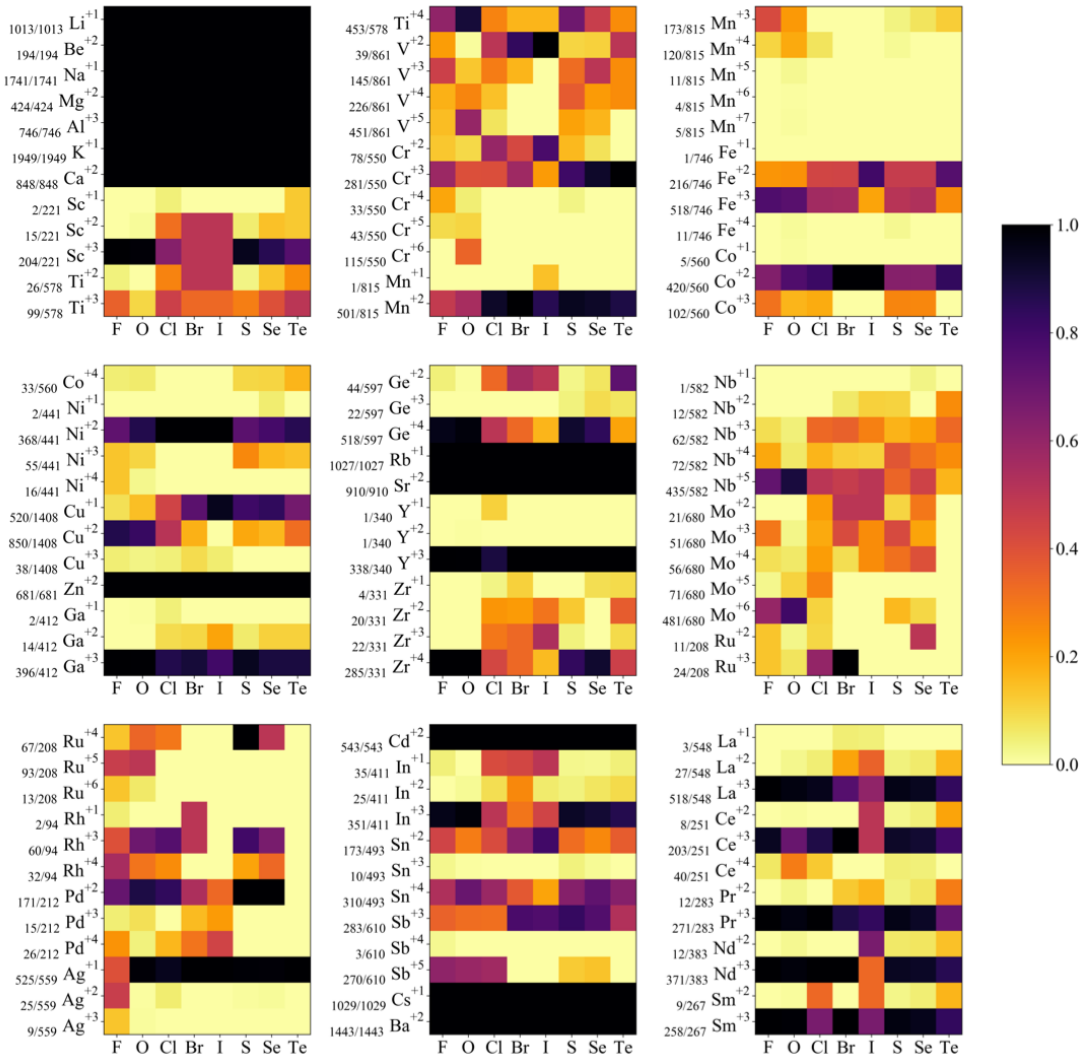

FIG. S5: Species–anion probabilities lookup table part I.

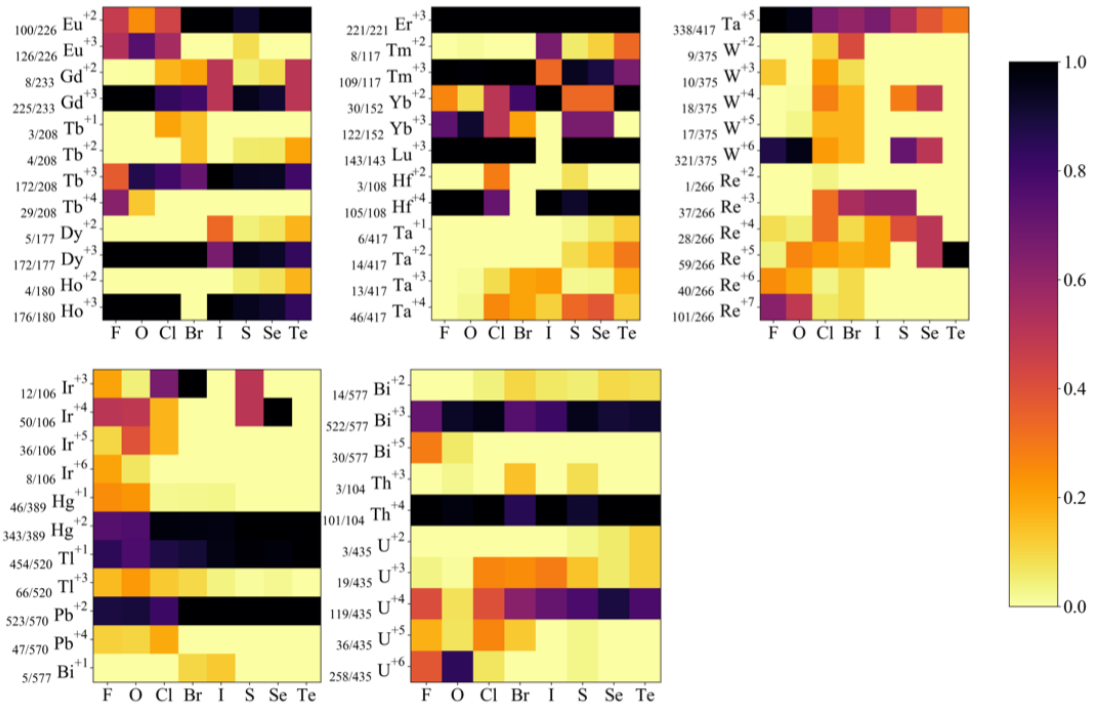

FIG. S6: Species–anion probabilities lookup table part II.

Design of Photoactive Metal Chalcohalide Semiconductors: From Composition to Structure with Data-mining and Global Optimisation

Daniel W. Davies,[1] Keith T. Butler[†],[1] Jonathan M. Skelton,[1] Congwei Xie,[2] Artem R. Oganov,[3,4,5] and Aron Walsh[6,7]

[1)] *Centre for Sustainable Chemical Technologies and Department of Chemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK*

[2)] *Science and Technology on Thermostructural Composite Materials Laboratory, International Center for Materials Discovery, School of Materials Science and Engineering, Northwestern Polytechnical University, Xian, Shaanxi 710072, Peoples Republic of China*

[3)] *International Center for Materials Discovery, School of Materials Science and Engineering, Northwestern Polytechnical University, Xian, Shaanxi 710072, Peoples Republic of China*

[4)] *Department of Geosciences, Center for Materials by Design, and Institute for Advanced Computational Science, State University of New York, Stony Brook, New York 11794-2100, USA*

[5)] *Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region 141700, Russia*

[6)] *Global E^3 Institute and Department of Materials Science and Engineering, Yonsei University, Seoul 120-749, Korea*

[7)] *Department of Materials, Imperial College London, Exhibition Road, London SW7 2AZ, UK*

(Dated: 15 November 2017)

I. SUPPLEMENTAL DATA ITEMS

A. Dynamic Stability

No negative frequency phonon modes were found at Γ for the structures predicted by global optimisation. The phonon densities of states are shown in Figure S1. The negative frequency phonon modes in the $Sn_5S_4Cl_2$ structure are at the Y, T and Z high symmetry points in the Brillouin zone. These imaginary modes would not be present if the structure was doubled along X (and Y as these are equivalent) and T were considered. Due to the practical limits of the size of the unit cell that can be considered for a global search, such a structure was not identified.

For the compounds predicted by analogy, the $Cd_5S_4Cl_2$ and Cd_4SF_6 structures both had modes with negative frequencies (imaginary modes) which indicate a lack of a restoring force when a ions are displaced along the collective mode coordinate. Although this can often indicate dynamical instability, mapping out of the modes in question can in each case provide a satisfactory explanation for their presence.

In the case of the $Cd_5S_4Cl_2$ structure, three imaginary modes were found. Mapping of the first reveals a double well potential energy surface (Figure S3a). The second and third reveal two extremely shallow degenerate double wells (Figure S3b). The structure did not relax into one of these wells during the DFT relaxation step due to limitations inherent to the numerical optimisers used for structure relaxation; if the structure is at a saddle point with some symmetry on the potential-energy surface, it is unlikely that the optimiser will break the symmetry to find a minimum, as the forces on the structure are balanced. Nudging the structure into the larger of the two wells in Figure S3a results in a slight reduction of total energy and elimination of all three imaginary phonon modes. Mapping of the one imaginary mode present for the Cd_4SF_6 structure reveals a wide, flat-bottomed potential (Figure S3c), which suggests the imaginary mode is due to anharmonicity rather than the system being a saddle point on the energy surface.

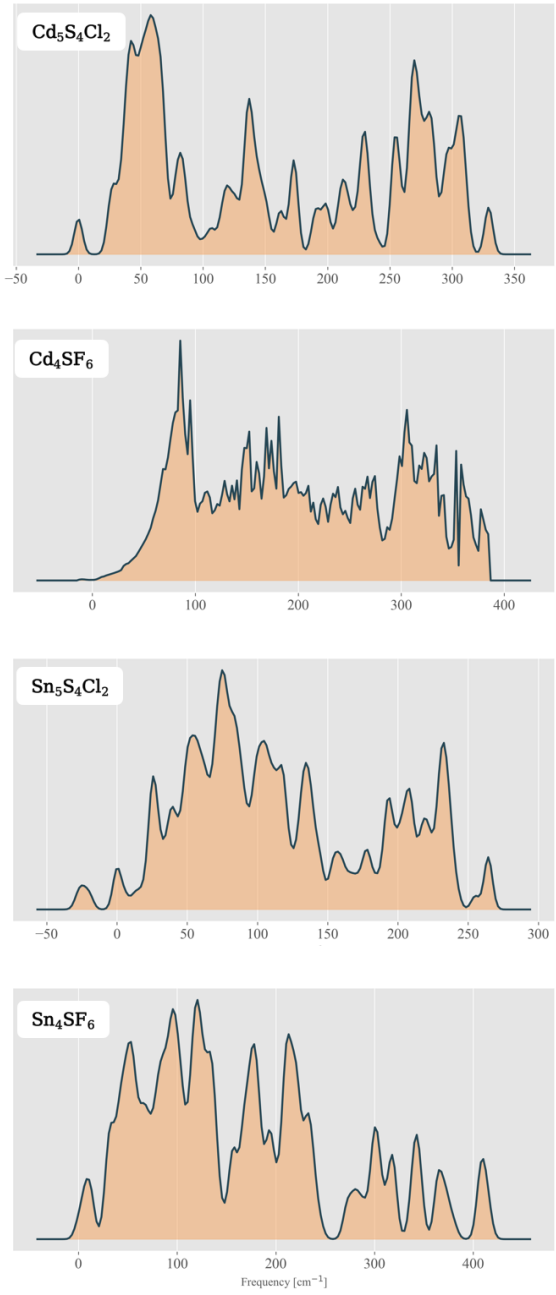

FIG. S1: Phonon densities of states for each of the structures found by global optimisation.

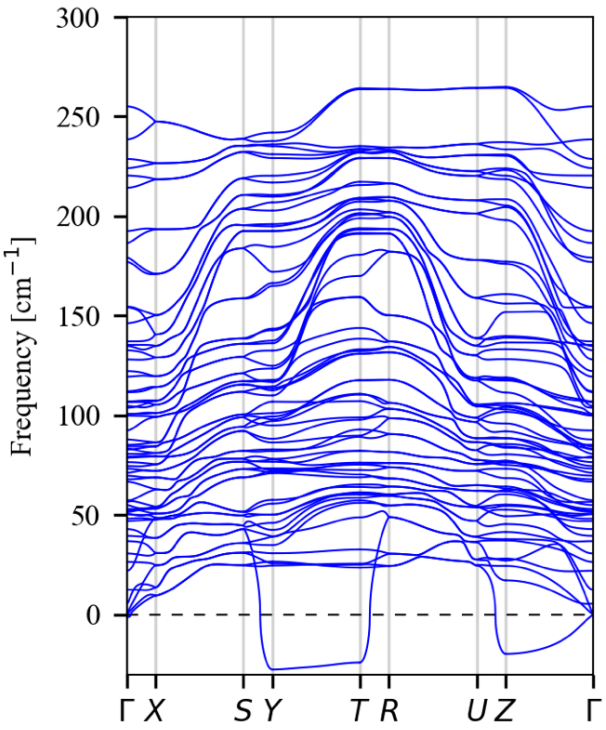

FIG. S2: Phonon band structure from a $1 \times 2 \times 2$ supercell of the $Sn_5S_4Cl_2$ crystal structure found by global searching.

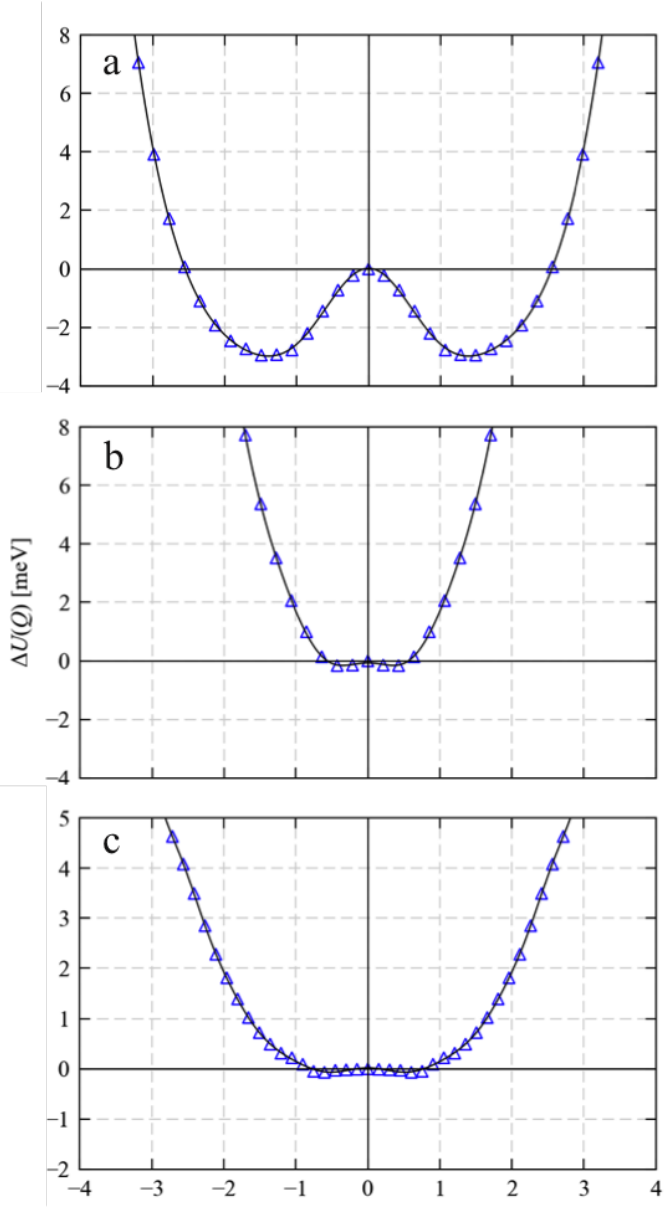

FIG. S3: Potential energy mapping of the imaginary phonon modes in the $Cd_5S_4Cl_2$ (a and b) and Cd_4SF_6 structures (c) found by analogy with known structure types.

B. Electronic Band Structures

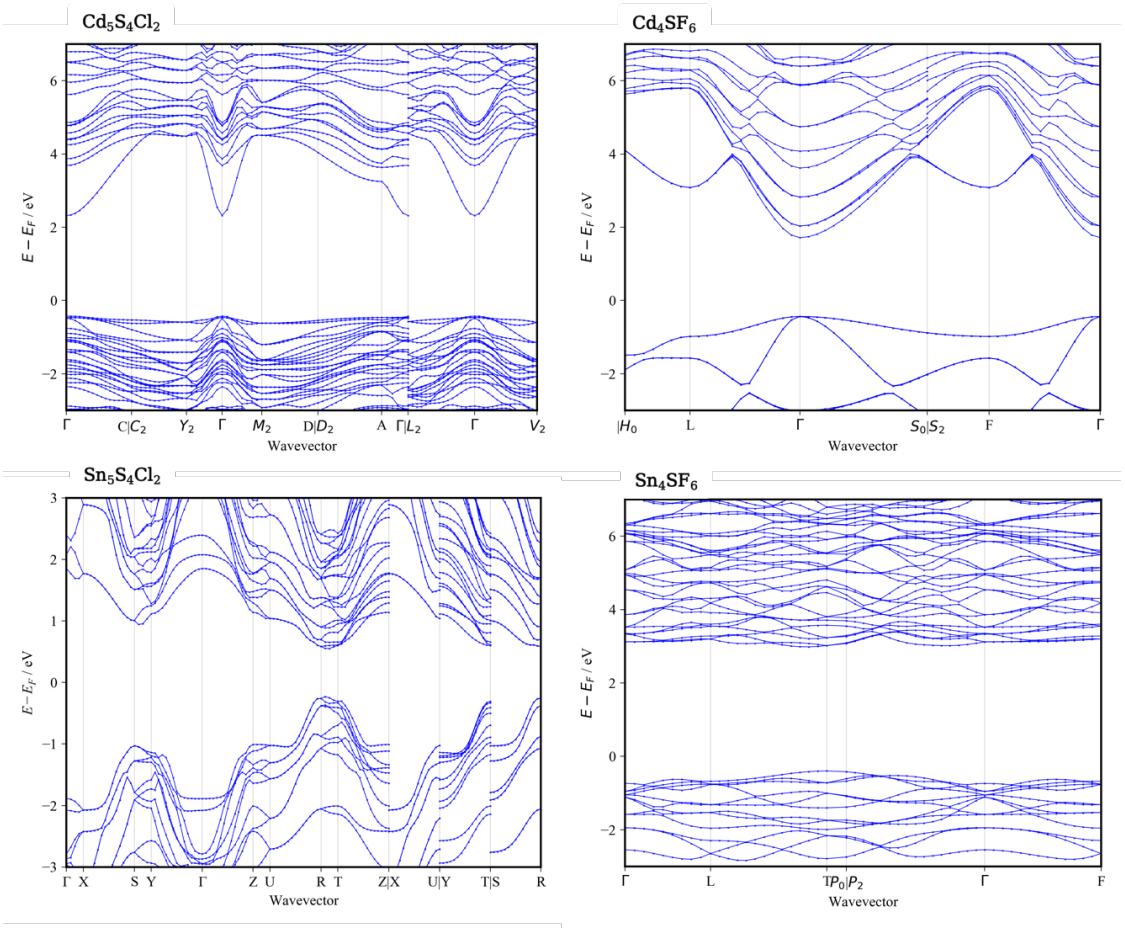

FIG. S4: Electronic band structures of the proposed chalcohalide compounds calculated using DFT and the HSE06 hybrid functional.