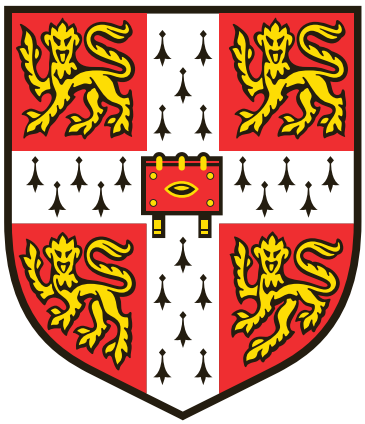
Accelerating Materials Discovery with Machine Learning



Rhys Edward Andrew Goodall

Department of Physics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Jesus College

November 2021

In memory of Sian Jones

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

The research described in this thesis was performed between October 2018 and November 2021, and was supervised by Dr Alpha A. Lee.

Rhys Edward Andrew Goodall
November 2021

Accelerating Materials Discovery with Machine Learning

by Rhys Edward Andrew Goodall

As we enter the data age, ever-increasing amounts of human knowledge are being recorded in machine-readable formats. This has opened up new opportunities to leverage data to accelerate scientific discovery. This thesis focuses on how we can use historical and computational data to aid the discovery and development of new materials.

We begin by looking at a traditional materials informatics task – elucidating the structure-function relationships of high-temperature cuprate superconductors. One of the most significant challenges for materials informatics is the limited availability of relevant data. We propose a simple calibration-based approach to estimate the apical and in-plane copper-oxygen distances from more readily available lattice parameter data to address this challenge for cuprate superconductors. Our investigation uncovers a large, unexplored region of materials space that may yield cuprates with higher critical temperatures. We propose two experimental avenues that may enable this region to be accessed.

Computational materials exploration is bottle-necked by our ability to provide input structures to feed our workflows. Whilst *ab-intio* structure identification is possible, it is computationally burdensome and we lack design rules for deciding where to target searches in high-throughput setups. To address this, there is a need to develop tools that suggest promising candidates, enabling automated deployment and increased efficiency. Machine learning models are well suited to this task, however, current approaches typically use hand-engineered inputs. This means that their performance is circumscribed by the intuitions reflected in the chosen inputs. We propose a novel way to formulate the machine learning task as a set regression problem over the elements in a material. We show that our approach leads to higher sample efficiency than other well-established composition-based approaches.

Having demonstrated the ability of machine learning to aid in the selection of promising compound compositions, we next explore how useful machine learning might be for identifying fabrication routes. Using a recently released data-mined data set of solid-state synthesis reactions, we design a two-stage model to predict the products of inorganic reactions. We critically explore the performance of this model, showing that

whilst the predictions fall short of the accuracy required to be chemically discriminative, the model provides valuable insights into understanding inorganic reactions. Through careful investigation of the model's failure modes, we explore the challenges that remain in the construction of forward inorganic reaction prediction models and suggest some pathways to tackle the identified issues.

One of the principal ways that material scientists understand and categorise materials is in terms of their symmetries. Crystal structure prototypes are assigned based on the presence of symmetrically equivalent sites known as Wyckoff positions. We show that a powerful coarse-grained representation of materials structures can be constructed from the Wyckoff positions by discarding information about their coordinates within crystal structures. One of the strengths of this representation is that it maintains the ability of structure-based methods to distinguish polymorphs whilst also allowing combinatorial enumeration akin to composition-based approaches. We construct an end-to-end differentiable model that takes our proposed Wyckoff representation as input. The performance of this approach is examined on a suite of materials discovery experiments showing that it leads to strong levels of enrichment in materials discovery tasks.

The research presented in this thesis highlights the promise of applying data-driven workflows and machine learning in materials discovery and development. This thesis concludes by speculating about promising research directions for applying machine learning within materials discovery.

Acknowledgements

I owe a large debt of gratitude to my supervisor, Dr Alpha A. Lee. A great irony in my life is that, in the original Welsh, my name means "enthusiasm". Fortunately, Alpha has an egregious oversupply of academic enthusiasm sufficient to counter even my own academic cynicism. Alpha has always placed enormous faith in me and helped me articulate my own research voice.

One of the most notable examples of this faith was in supporting my application to the Machine Learning for Physics and the Physics of Learning Long Program at the Institute for Pure & Applied Mathematics (IPAM) at the University of California, Los Angeles, when I was mere months into my PhD. I am enormously grateful to the Institute, the program organising committee, and other program members for making my time there so enjoyable. Above all, I would like to thank Kim, my office mate at IPAM, for his support, friendship, and ability to drive and Jonas for his advice and perspective on post PhD career choices.

Closer to home, I am grateful to my collaborators: Prof Judith L. MacManus-Driscoll, Dr Bonan Zhu, Dr Felix A. Faber, Dr Rickard Armiento, Shresth A. Malik, and Abhijith S. Parackal, for helping me formulate and translate my ideas into the research contained in this thesis. In the Lee Group, I would like to thank Janosh, Yunwei, Alex, Alwin, and "downstairsgang" for responding to my idiosyncrasies with humour and friendship. I have also enjoyed being a part of TCM, even if my sporadic attendance at social events came at the cost of being reminded of the vast gaps that exist in my grasp of theoretical physics.

Cambridge has been one of my homes for seven years, the first four years at Robinson College and the last three at Jesus College. I am grateful for all the friends I have made throughout my time here. In the last three years, I have enjoyed being part of the MCR community at Jesus – Joe, Adrian, Chris, Orla, Ahsan, Friederike, and Connor, amongst others. It's a shame that I ended up spending so much of my PhD away from Cambridge.

Waterpolo has also played a hugely important role in my time at Cambridge. Through the sport, I have met and trained alongside a diverse and engaging group of teammates. Whilst I may never have represented the pinnacle of sporting achievement, I am very proud to have been part of the Varsity first team on two occasions. I want to thank all my teammates over the years – Toberg, Santiago, Jo, Rob, Isaac, Ben, Patrick, Alex, Balint, etc. – for their memes and for forgiving my lack of commitment to exercise.

Lastly, I would like to thank my parents, Sian and Andy, for their love and continual support, the strength of which has become painfully apparent since my mother's passing.

Table of contents

Preface **1**

1 Introduction **3**

2 Methods in Computational Materials Discovery **11**
2.1 Machine Learning 11
2.2 Density Functional Theory 19

3 Structure-Function Relationships in Cuprate Superconductors **26**
3.1 Inferring Structural Parameters of Cuprates 30
3.2 Materials Informatics Reveals Unexplored Structure Space 36
3.3 Discussion 38

4 Representation Learning from Stoichiometry **40**
4.1 Representation Learning of Inorganic Materials 41
4.2 Data Sets 44
4.3 Evaluation of Sample Efficiency 45
4.4 Evaluation of Uncertainty Estimates 46
4.5 Transfer learning 50
4.6 Ablation Study 51
4.7 Discussion 55

5 Predicting the Products of Inorganic Reactions **58**
5.1 Solid-State Syntheses Data Set 60
5.2 Representation of Inorganic Reactions 61
5.3 Reaction Graph Model 63
5.4 Evaluation of Model Performance 66

5.5 Explaining Predictions via Reaction Similarity . . . 70
5.6 Identified Failure Modes . . . 73
5.7 Benchmarking and Ablation Study . . . 74
5.8 Discussion . . . 80

6 Wyckoff Representation Regression **83**
6.1 Wyckoff Representation Regression . . . 84
6.2 Data Sets . . . 88
6.3 Wyckoff Positions and Model Inputs . . . 89
6.4 Exploration of Structurally Diverse Chemical Systems . . . 93
6.5 Enrichment in Novel Chemical Systems . . . 101
6.6 Selecting Stable Materials from Diverse Chemical Spaces . . . 102
6.7 Computational Prospecting for Novel Stable Materials . . . 111
6.8 Discussion . . . 114

7 Outlook **117**
7.1 Directions for Future Research . . . 118

Appendix A Model Implementation Details **124**
A.1 Roost Architecture and Training . . . 124
A.2 Reaction Graph Architecture and Training . . . 125
A.3 Wren Architecture and Training . . . 127

Appendix B Density Functional Theory Settings **129**
B.1 Strain Dependence of Cuprates . . . 129
B.2 Stability of Novel Candidate Materials . . . 130

References **131**

Preface

Chapter 1 gives an introduction to bottlenecks that exist in the materials discovery pipeline and highlights how data-driven workflows can be used to address some of the challenges that exist.

In Chapter 2 we introduce several core concepts and methods in machine learning and density functional theory (DFT) needed to understand the following research.

Chapter 3 brings us to new research looking at high-temperature superconductivity in cuprate systems. Using historical data, we investigate structure-function relationships between the critical temperature and apical and in-plane Cu-O distances. This work was carried out with Dr Bonan Zhu, Prof Judith L. MacManus-Driscoll, and Dr Alpha A. Lee and resulted in the publication of the following article:

> Rhys EA Goodall, Bonan Zhu, Judith L MacManus-Driscoll, and Alpha A Lee. Materials informatics reveals unexplored structure space in cuprate superconductors. *Advanced Functional Materials*, page 2104696, 2021.

In this work, I carried out all of the data curation and analysis. Dr Bonan Zhu carried out the DFT calculations used to validate analysis assumptions. Prof Judith L. MacManus-Driscoll and Dr Alpha A. Lee supervised the work. I was the principal author of the article.

Chapter 4 introduces a new framework - representation learning from stoichiometry - for predicting the properties of inorganic materials when the crystal structure is unknown. This work was carried out with Dr Alpha A. Lee and resulted in the publication of the following article:

> Rhys E. A. Goodall and Alpha A. Lee. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nature Communications*, 11(1):1–9, 2020,

In this work, I developed and implemented the proposed framework, conducted the model characterisation, and was the article's primary author. Dr Alpha A. Lee supervised the work.

Following on from this work, Chapter 5 explores how similar techniques can be applied to inorganic reaction prediction. This work was conducted with Shresth A. Malik and Dr Alpha A. Lee and resulted in the publication of:

> Shreshth A Malik, Rhys E A Goodall, and Alpha A Lee. Predicting the outcomes of material syntheses with deep learning. *Chemistry of Materials*, 33(2):616–624, 2021,

In this work, I developed and implemented the core codebase. Shresth A. Malik and I devised the experiments and analysed the results. Shresth A. Malik wrote the scripts to carry out the experiments. Dr Alpha A. Lee and I co-supervised Shresth A. Malik's contributions on this work that he completed as part of his Part III masters project. Shresth A. Malik and I prepared the article's first draft and worked together on the revisions to the article.

In Chapter 6, we return to structure-agnostic property prediction and re-examine the problem drawing inspiration crystallography to propose a coarse-grained representation of crystal structures that can empower screening of novel materials space. This work was carried out in collaboration with Abhijith S. Parackal, Dr Felix A. Faber, Dr Rickard Armiento, and Dr Alpha A. Lee. In this work, I developed and implemented the proposed framework and conducted the main experiments. Abhijith S. Parackal carried out the prospective DFT calculations. Dr Felix A. Faber, Dr Rickard Armiento and Dr Alpha A. Lee supervised the work.

The final chapter provides a summary of the research presented and explores new directions for further research.

Chapter 1

Introduction

Over the last century, there has been steady progress in our exploration of materials space. These efforts have resulted in many breakthrough discoveries that have fundamentally changed humanity. Perhaps one of the most prominent modern examples of materials driven technological development are Li-ion batteries. Li-ion batteries hinge on the availability of high-voltage cathode and anode materials that enable the stable cycling of lithium ions. In 1991 $LiCoO_2$ became the first cathode material to be commercialised for Li-ion batteries alongside a graphite-based anode. This first cell chemistry has since been followed by a variety of other chemistries based on alternative cathode materials, for example, lithium manganese oxide - $LiMn_2O_4$, lithium iron phosphate - $LiFePO_4$, and lithium nickel oxide - $LiNiO_2$.

The ability to store power in a rechargeable, lightweight, dense and relatively low-risk medium has facilitated the development of mobile technologies whose impact touch all facets of society. Developing new chemistries and materials for Li-ion batteries remains an active area of research. For example, ongoing research into using cathodes based on cation-disordered rock-salt materials may offer up to a two-fold increase in the energy density of commercial lithium-ion batteries [4, 5].

The enormous success of Li-ion battery technologies, however, also highlights the folly of continuing with the prevailing "Edisonian"[1] approach to materials discovery. The structure of $LiCoO_2$ was first characterised in 1958 [6], 22 years before it was investigated for use as a cathode material for Li-ion batteries [7]. Similar stories crop up over and over again in material science, for example, correct reports of the structure of La_2CuO_4

[1]Thomas Edison reportedly carried out more than 6,000 experiments before discovering a viable filament material for his electric lightbulb.

came over a decade [8] before it was identified as a high-temperature superconductor [9]. The reality of the challenge is that the vastness of material space and the complexity of experimental characterisation renders materials discovery via exhaustive experimentation infeasible with current approaches – it is simply not possible to test every material for every feasible application.

Due to the constraint of finite resources, the traditional approach to materials discovery has been iterative, driven by explorations of groups of similar materials linked by heuristically defined design rules and human insights. However, as ever-increasing numbers of materials are investigated, the ability of an individual to be aware of, let alone formulate hypotheses connecting, all relevant known materials diminishes. In the fight against this burgeoning complexity, efforts to standardise and digitise as much of our understanding as possible are vital. One of the most prominent examples of such standardisation is the Crystal Information Framework (CIF) [10, 11]. The near-universal adoption of CIF files for reporting crystal structures has enabled the compilation of large databases for both organic [12] and inorganic materials [13], allowing researchers to readily reuse and build upon the work of others.

At the time of writing, the Inorganic Crystal Structure Database (ICSD) [13] contains CIF files for 200,809 experimental structures (although many of these are distinct realisations of the same or very similar materials[2]). Inorganic material science typically deals with materials containing four or fewer main elements. The building blocks for inorganic materials are usually the 98 naturally occurring elements. Combining these building blocks leads to 3,612,280 possible chemical systems [14]. Within each of these systems, stoichiometric and polymorphic freedoms cause the number of possible materials to balloon. It has been estimated that the total number of possible materials could be as large as $\mathcal{O}(10^{10})$ [15]. Consequently, whilst the ICSD offers extensive coverage of known materials, it covers a very limited region of the potential materials design space. This highlights a key challenge we explore in this thesis – how can we efficiently search materials space given the claim that the current "Edisonian" approach is not desirable?

[2]One of the great challenges of dealing with materials data is that it is highly heterogeneous, for example, investigating the meta-data in ICSD reveals that 84,901 structures are recorded as being determined from single-crystal samples and 104,834 from powder diffraction samples. This leaves 11,074 entries for which this meta-data is not available. Of these 7,794 do not have Digital Object Identifiers making it difficult to trace their sources. If these structures are characterised at different levels of accuracy, how reasonable is it to say that we have 200,809 examples? If we need to be more granular, how much more so?

For a search strategy to be efficient, it needs to find the "good" examples in a design space whilst carrying out as few expensive function evaluations (i.e. experiments) as possible – here a relevant function evaluation might be the synthesis and characterisation of a new material. Two main paradigms have emerged for tackling this problem. The first is to develop approaches that reduce the cost of experiments via combinatorial approaches [16–19], automation and synthesis planning. The second is to make use of surrogate models for high-throughput virtual screening. Virtual screening allows the number of experiments required to be significantly reduced by discarding candidates that are unlikely to exhibit desired properties before carrying out any real experiments. Current virtual screening efforts in inorganic material discovery are primarily based on *ab initio* calculations – most commonly density functional theory. This is a highly successful paradigm and has led to the discovery of several novel functional materials [20, 21]. Despite progress, the typical time frame from initial discovery through to commercialisation remains over a decade. To shrink this time frame, we need to focus on alleviating the bottlenecks that persist within current workflows.

Addressing Bottlenecks in Experiments

Whilst an in-depth review of high-throughput experiments is beyond the scope of this thesis, it is important to understand the birds-eye view and the challenges that exist. Doing so allows us to see how the interdisciplinary work presented in this thesis fits into the wider materials discovery pipeline.

The central idea behind combinatorial experiments is that if we can develop protocols that allow for many experiments to be carried out in parallel, then we can cover larger search spaces at once and identify promising materials more quickly [22, 23]. The most prevalent experimental platform for combinatorial synthesis of inorganic materials is the production of thin-film screening libraries. Experimental protocols such as sputtering-based approaches, pulsed laser deposition, metal-organic chemical vapour deposition and chemical solution deposition have all been used for this type of combinatorial synthesis. Whilst such thin-film synthesis methods enable the scaling of experiments to the high-throughput settings, they can result in poor quality (e.g. amorphous or high polycrystallinity) samples unless the appropriate method is chosen and the parameters are correctly tuned. Indeed, because of this, many early applications of combinatorial synthesis focused on simply identifying favourable synthesis parameters, for example,

conducting experiments under temperature gradients to quickly identify good growth conditions [24].

Whilst combinatorial synthesis is now a relatively mature idea, it has not been, by itself, the panacea some believed it could be. The practical restriction on its utility is that only certain types of materials can be made with amenable experimental protocols. Consequently, alternative and additional solutions are needed to reduce the cost of experiments. Amongst these, self-driving laboratories have been proposed to reduce the amount of human involvement and oversight required by automating the design of experiments (i.e. tuning of experimental parameters), bringing down costs and increasing efficiency. The key idea is that past experimental results can be used to train sequential learning models that select future experiments with the goal of maximising some objective. Algorithmic selection of parameters allows for the consideration of much higher dimensional design spaces, however, in workflows that involve human input, looking at high dimensional design spaces risks that the limited attention spans of human researchers may lead to searches being prematurely curtailed. This reality drives the need for end-to-end automation of the entire Design-Make-Test cycle for self-driving laboratories to be successful. Already promising demonstrations of the potential of automation have been realised, for example, Ref. [25] reports the deployment of a robot chemist to tune the formulation of an organic photocatalyst for hydrogen production, improving the hydrogen production rate 6-fold after eight days of searching, and Ref. [26] reports the development of equipment for high-throughput solid-state synthesis using dry powders which in principle could be used in conjunction with a robot chemist to enable fully automated solid-state synthesis.

Overall the outlook for combining combinatorial synthesis, self-driving laboratories, and automation is highly promising, however, these techniques are only applicable to combinations of material that can be produced with standard experimental protocols. For target materials that do not fit into this mould, i.e. novel material candidates, selection of experimental protocols for inorganic synthesis remains challenging. In general, the interplay between precursors, reaction-pathways and processing conditions is not well

understood[3]. This lack of understanding is primarily due to complexity but is also compounded by the fact that information about materials synthesis and preparation processes is often hard to access, index, and search. Unlike crystal structures, there is no widely accepted machine-readable format for reporting materials synthesis and preparation processes. Even if such a standard can be defined, there exist enormous challenges in converting historical knowledge into any such new format as the relevant data currently exists in a heterogeneous array of formats - primarily natural language text in proprietary laboratory notebooks at private companies or (potentially pay-walled) academic publications.

To begin to resolve these issues, efforts to produce data-mining software specifically designed for academic literature are being pursued [29, 30] – a task that is particularly challenging due to the lack of a consistent ontology for describing inorganic materials synthesis [31]. Looking at organic chemistry and cheminformatics, the availability of data-mined and freely available reaction data sets, most prominently the USPTO data set [32], and filtered derivatives thereof [33], have played a key role in driving innovation in reaction-prediction and retrosynthesis models for drug-like molecules. Making data available is key as it opens up research avenues to scientists without access to proprietary or commercial data sources. Recently an analogous data set of data-mined inorganic synthesis reactions has been released [34], whilst much smaller in scale than organic reaction data sets, the pure availability of data enables previously intractable research questions to be posed. In Chapter 5 we use this new data source to look at the challenges that exist in the construction of a forward reaction-prediction model for inorganic synthesis. Whilst data-mining technologies are incredibly valuable, information can be lost in translation. The availability of machine-readable data will hopefully spur the development of a common standard, akin to the CIF, for reporting materials synthesis and preparation processes, resulting in a positive feedback loop where that availability of data-mined data leads to direct reporting of standardised data in the future.

[3]Whilst *ab initio* calculations can be used to elucidate such mechanisms for individual systems [27], doing so requires significant computational resources and expertise. Recent work has explored how computational workflows can be made routine by directly building simulation tools that interface with electronic laboratory notebooks [28]. As this technology matures, it will hopefully become feasible to develop workflows that allow the calculation of properties relevant to the understanding of inorganic materials synthesis in an off-the-shelf manner.

Addressing Bottlenecks within Virtual Screening

A central challenge for virtual screening is how to build effective surrogate models for predicting materials' properties. For the most part, this has been a question of how do we build physical models, be they *ab initio*, semi-empirical or phenomenological, for the processes we are interested in. This choice of surrogate is typically viewed as a multi-objective optimisation problem where the cost and accuracy of different approaches are traded off against each other giving rise to a Pareto frontier of compelling methods. Whilst semi-empirical and phenomenological approaches were widespread in the past [35], the accuracy requirements of modern virtual screening programs and the maturation of density functional theory (DFT) have led to DFT being adopted as the workhorse of most virtual screening campaigns for inorganic materials.

The great filter on whether it might be possible to synthesise a material is its thermodynamic stability. As a back of the envelope calculation for how long it may take to screen the $\mathcal{O}(10^{10})$ materials design space for theoretically stable materials, we consider the Open Quantum Materials Database (OQMD) [36, 37]. In 2014 the database contained approximately $\sim$ 285,000 entries, by 2017 this was up to $\sim$ 470,000 entries, and at the time of writing the database contains $\sim$ 815,000 entries. This growth gives OQMD a doubling period of approximately three years. Assuming that Moore's law holds firm and this doubling period remains consistent, it would take approximately 30 years to grow from $\mathcal{O}(10^6)$ entries to $\mathcal{O}(10^{10})$, with the vast majority of this growth occurring within the last decade. This 30-year timeframe is highly optimistic as it assumes that the calculations do not increase in cost, i.e. that we use additional computational resources to run more calculations rather than moving to more costly exchange-correlation functionals or higher levels of theory for increased accuracy.

Exhaustive virtual screening workflows are best suited to small design spaces where it is reasonable to carry out all the required calculations. This is sometimes referred to as the "function follows form" paradigm, where first the entire space is mapped out before looking for regions with good properties. When we move to much larger design spaces, such as the $\mathcal{O}(10^{10})$ materials space considered above, the timescale for exhaustive screening becomes prohibitive. When dealing with such a large design space, a critical consideration is that whilst $\mathcal{O}(10^{10})$ materials might be reasonable, a substantially smaller subset are likely to be ordered, interesting and experimentally accessible[4]. Suppose it is possible to

[4]In OQMD only $\sim$ 55,000 out of 815,000 currently recorded materials sit on the convex hull.

develop tools that enable us to systematically sub-select more promising candidates from this $\mathcal{O}(10^{10})$ design space for *ab initio* study, then the number of calculations required to identify promising materials could be dramatically reduced. This alternative "form follows function" approach is more commonly known as inverse design.

A common misconception about inverse design is that it necessarily involves inverting forward calculations of properties, whilst this is sometimes the case, for example, in the inverse design of simple liquids [38, 39], in most instances inverse design is achieved indirectly by iteratively carrying out limited numbers of forward calculations under algorithmic guidance. For example, a pioneering demonstration of inverse design was in the exploration of the bandstructure-landscape of AlAs/GaAs and GaP/InP superlattices [40]. Here simulated annealing was used to select trial arrangements for which the electronic properties were then calculated using a semi-empirical model. The resulting workflow allowed arrangements with given electronic properties to be found using just $\mathcal{O}(10^4)$ evaluations in design spaces as large as $\mathcal{O}(10^{14})$.

More recently, there has been significant interest in the application of inverse design principles to organic molecules [41]. This field has progressed rapidly, benefiting from fertile cross-talk between the machine learning and cheminformatics communities. The key idea behind many of these models for inverse molecular design is to perform property optimisation within the latent space of a generative model. Generative models work by sampling embedding vectors from their latent spaces that can be passed through a decoder to return molecular structures. Separately a surrogate model for the functional property of interest is constructed in the latent space. This model is then used to suggest latent embeddings that it believes correspond to molecular structures with the desired functionality [42]. In principle, similar strategies should be applicable for inorganic materials. However, due to the additional complexity of more variable chemistries and the periodicity of bulk materials, this direction of research has not received as much attention [43].

The success of both generative and screening-based approaches for inverse design is contingent on the availability of accurate and cheap surrogate models to carry out the virtual screening. Machine learning algorithms are now becoming mature for many properties relevant virtual screening tasks [44], inducing a paradigm shift from physics-based models to data-driven models. However, for inorganic materials discovery, there remains a lack of accurate models for applications where crystal structures are *a priori*

unknown. This thesis attempts to tackle this bottleneck in two places. This problem is initially tackled in Chapter 4 where we construct machine learning models that operate using only composition-based information. These ideas are then extended in Chapter 6 to address challenges specific to high-throughout *ab initio* virtual screening using a model based on crystallographic Wyckoff positions.

Chapter 2

Methods in Computational Materials Discovery

2.1 Machine Learning

Machine learning (ML) refers to the construction of models and algorithms that perform tasks without being explicitly programmed. Instead, ML algorithms rely on optimisation techniques to infer what to return based on available data. In the following paragraphs, we provide a summary of the ML concepts and ideas needed to understand the following research. For brevity, we focus on supervised regression and neglect many important sub-fields. We recommend Refs. [45, 46] for readers interested in other topics, such as supervised classification, unsupervised learning, and generative models.

2.1.1 Shallow Learning

Shallow learning refers to machine learning models that take hand-crafted features as inputs. Features are constructed heuristically by experts based on domain knowledge of the target problem. Concretely, the aim of shallow learning is given features, x, to learn the parameters, $\hat{\theta}$, of a predicative model $\hat{y} = f(x, \hat{\theta})$ that minimise a given cost function, $\mathcal{L}(y, \hat{y})$, where y is the target variable and $\hat{y}$ is the predicted value, i.e. to find the solution

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \ \mathcal{L}(y, \hat{y}). \tag{2.1}$$

The most common loss function choice is the squared residuals,

$$\mathcal{L}(y, \hat{y}) = \sum_i (y_i - \hat{y}_i)^2. \tag{2.2}$$

The most ubiquitous example of shallow learning is linear regression where a model of the form $\hat{y}_i = \hat{W} x_i + \hat{b}$ is assumed. Large values of the feature weights $\hat{W}$ tend to be linked to overfitting, causing high generalisation errors. Therefore, it is common to add regularisation terms to the cost function that penalise either the L1 (LASSO) or L2 (Ridge/Tikhonov) norm of the weights to reduce the likelihood of overfitting. More advanced examples of shallow learning are ensemble-based models and kernel-based models described below.

Ensemble-based Models

The two most common types of ensemble models are Random Forests and Gradient-boosted models. Random Forests are a decision tree-based model that use an ensemble of multiple weak regressors to make predictions [47]. Each tree is constructed to find a series of decision boundaries that split the data to minimise the squared deviations between the samples and the sample mean in each branch or leaf of the tree. Predictions are made by averaging the outputs of the different trees when applied to new data. To overcome issues of over-fitting common to decision tree methods, Random Forests use bagging and random subspace projection to reduce the correlation between the trees, improving their generalisation performance.

Gradient-boosted models work by training a series of weak regressors, commonly decision trees, that attempt to correct the errors of the previous regressors. It is common to use a squared error loss function such that the task for the weak regressors is to fit the remaining residuals at each stage. This process is continued until the remaining residuals contain no signal. As with Random Forests, a variety of techniques, such as shrinkage and pruning [48], are used to improve performance.

Kernel-based Models

Linear machine learning algorithms can often be recast in terms of inner products between points in the feature space. The kernel trick involves replacing these inner products with

kernel functions that are easy to compute, for example, the squared exponential kernel,

$$K_{\mathrm{SE}}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{x}'||^2}{2\ell^2}\right), \tag{2.3}$$

where ℓ is the length scale of the kernel, that determines how quickly the function can change, and $\boldsymbol{x}$ and $\boldsymbol{x}'$ are the feature vectors of the points being considered in the original feature space. The squared exponential kernel corresponds to the inner product of a basis set of radial basis functions with infinitely many reference centers. Consequently, the kernel trick allows us to explore models that are, in theory, infinitely complicated with a finite amount of computation. Kernel ridge regression, support vector machines, and Gaussian processes are all examples of model classes that make use of the kernel trick [49]. Training kernel methods notoriously suffers from $\mathcal{O}(N^3)$ scaling, where N is the number of training examples. However, in practical applications – particularly within material science applications where kernel models are commonly used to approximate a potential energy surface for atomistic simulations – it is often reasonable to make sparse approximations that can reduce the scaling to $\mathcal{O}(NM^2)$, where M is the number of inducing points [50].

Descriptors for Material Science

The success of shallow machine learning depends heavily on the quality of input features. Consequently, as machine learning has attracted attention within computational material science, a zoo of structure-based descriptors have emerged for describing local environments in molecules and materials [51–57]. When constructing structure-based descriptors, incorporating symmetries present in physical systems is critical to improving data efficiency. In practice, this means ensuring descriptors are invariant, or in the case of tensorial properties equivariant, to *SO(3)* rotations and translations. It is also necessary to construct models that are permutationally invariant to the order in which sites are considered [52]. These descriptors have typically been combined with specific types of model, e.g. FCHL-18/19 with kernel ridge regression [56, 58] and SOAP with Gaussian process regression to give Gaussian Approximation Potentials [59, 53, 60].

For materials discovery, we often want to apply machine learning to material science problems where knowledge of the atomic coordinates is *a priori* unavailable. In such cases, it is common to resort to composition-based descriptors. Focusing on materials with a

small and fixed number of elements, pioneering works [61, 62] constructed descriptors by exhaustively searching through analytical expressions comprising combinations of atomic properties such as the ionic radius, electronegativity, and nuclear mass. This approach has been employed successfully for learning descriptors for the Gibbs energy [63], and the critical temperatures of conventional superconductors [64]. However, the computational complexity of this approach scales exponentially with the number of constituting elements and is not applicable to materials with different numbers of elements or dopants. To address this shortcoming, general-purpose material descriptors, hand-curated from the weighted statistics of chosen atomic properties across the elements in a material, have been proposed [65–67]. Of these, we highlight the *Magpie* feature set [65] which contains 145 features incorporating knowledge about the elements, stoichiometry, and electronic properties that we make use of in subsequent research. The *Magpie* feature set has been leveraged in accelerating the discovery of metallic glasses [68].

2.1.2 Deep Learning

In contrast to shallow learning, the deep learning revolution of the last decade is built around models that learn their representations from raw data inputs. The workhorse of deep learning is the neural network. At their heart, neural networks are compositions of feature maps that transform the raw input features, x, into a new set of features that are linearly related to their target, y. The prototypical example of a neural network is the multi-layer perceptron (MLP). A l-layer MLP approximates functions $f(x)$ using l successive non-linear feature maps constructed as compositions of affine transformations and non linearities, i.e.

$$f(x) \simeq \hat{W}_l \sigma(\hat{W}_{l-1} ... \sigma(\hat{W}_1 x + \hat{b}_1) ... + \hat{b}_{l-1}) + \hat{b}_l, \tag{2.4}$$

where $W_i \in \mathbb{R}^{M_i \times M_{i-1}}$ is the i^{th} weight matrix inferred from data, M_i is the number of units in layer i, and $\sigma(x)$ is a non-linear activation function – frequently a rectified linear unit (ReLU) activation of the form $\text{ReLU}(x) = \max(x, 0)$. Early neural networks typically made use of sigmoidal activation functions but these have by-and-large been replaced with ReLU-type activation functions, or smooth approximations thereof [69], due to issues with vanishing gradients [70].

The remarkable success of deep learning emerges from the fact that neural networks, such as MLPs, can be optimised effectively using first-order gradient-based approaches. Moreover, the necessary gradients can be efficiently calculated using the chain rule by back-propagation of the training loss. In practice, modern neural networks are implemented inside automatic differentiation frameworks that abstract away the technical burden of implementing back-propagation [71, 72]. In addition, these frameworks are designed to enable the necessary calculations to be carried out on hardware accelerators, such as graphical processing units (GPUs), that dramatically reduce the time for training.

The most simple gradient-based optimisation algorithm is gradient-descent. In gradient descent at each step the model's parameters are updated according to

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_t), \tag{2.5}$$

where the learning rate, η, is a hyper-parameter of the optimiser that determines the size of the parameter updates. In practice, the full-batch gradient of the loss, $\nabla_\theta \mathcal{L}(\theta_t)$ is replaced with a stochastic approximation of the gradient calculated on a mini-batch of data randomly sampled from all the available training data. Typically mini-batches are randomly drawn from the training data without replacement until all the training examples have been considered. Each complete cycle through the training set is referred to as an epoch. After each epoch, the training set is shuffled and the process is repeated until the loss has satisfactorily converged. Replacing full-batch gradient descent with mini-batch stochastic gradient descent significantly speeds up optimisation, reducing the amount of computation required to determine the gradient for each step and providing helpful regularisation effects that drive the optimisation towards flatter local basins of attraction in the loss landscape. Further improvements to model optimisation procedures can be achieved by incorporating additional terms such as momentum, learning rate schedules, or adaptive learning rates [73], and additional regularisation procedures such as weight decay, early-stopping, or dropout [74].

Inductive Biases and Differentiable Programming

In the above description of the MLP, the key requirement is that the model is end-to-end differentiable. This allows the parameters of the model to be optimised by the combination of back-propagation and gradient descent. Accordingly, provided we ensure that all operations in our models have defined derivatives, we can build up novel

neural networks architectures with specific inductive biases as compositions of custom differentiable building blocks – notable examples are recurrent neural networks (RNNs) that are designed to handle series data (e.g. Gated-Recurrent-Unit (GRU) [75] and Long-Short-Term-Memory (LSTM) networks [76]) and convolutional neural networks (CNNs) that build in translational invariance for computer vision applications (e.g. LeNet [77], AlexNet [78], and ResNets [79]).

Within materials science, this ability to compose differentiable building blocks into novel architectures has led to the development of a wide variety of message-passing neural networks that operate directly on the atomic coordinates of molecules and materials. Typically these models operate on "radius"-graphs of interconnected local environments determined using a cutoff radius – the resulting data structure closely mirrors that of Verlet lists used in atomistic simulations [80]. The nodes of the graphs encode atoms with edges encoding interactions or bonds. As with shallow descriptor-based models, it is important to encode the underlying symmetries of the problem into the network architecture. Earlier models ensured *SO(3)*-invariance by only including position information via the relative distances between connected sites [81–83]. More recently *SO(3)*-equivariant architectures have been proposed that include angular and higher-order information or relative displacement vectors between atoms to allow construction of message passing operations that maintain equivariance [84].

Whilst specific neural network architectures have been developed for structure-based material science applications, there has been minimal progress on the development of structure-agnostic or coordinate-free architectures. Chapters 4 and 6 focus on the development of such models. Here the principal inductive bias we need to incorporate into our architectures is permutational invariance. This leads to the development of model architectures that are closely related to Deep Sets [85] and Set Transformers [86].

Abstracting yet further, recent work has shown it is possible to implement entire physical simulators within automatic differentiation frameworks. This enables the approximations and fitted functions that are typically integral to the implementation of such simulators to be replaced with trainable neural networks. This paradigm of differentiable programming can circumvent the traditional limitations imposed by fixed approximations opening up new regions on the cost-accuracy Pareto frontier. Important examples within computational materials science include fully variational Hartree-Fock

calculations [87], exchange-correlation functionals within the Kohn-Sham equations [88], and highly expressive neural-ansatzes for variational quantum Monte-Carlo [89, 90].

Uncertainty Estimation and Deep Ensembles

When we apply neural networks in decision-making scenarios, e.g. of these candidates which are the most promising, having some notion of the risk or uncertainty associated with each choice facilitates more informed and potentially better decisions. Whilst some model classes naturally allow for uncertainty estimation, e.g. Gaussian Processes, the neural-network-based models that we turn to throughout this thesis do not provide uncertainty estimates out of the box.

In statistical modelling, there are two sources of uncertainty that are necessary to consider. Firstly, the aleatoric uncertainty, which is the variability due to the natural randomness of the process (i.e. the measurement noise). Secondly, the epistemic uncertainty, which is related to the variance between the predictions of plausible models that could explain the data. This uncertainty arises due to having an insufficient or sparse sampling of the underlying process such that many distinct but equivalently good models exist for explaining the available data. Epistemic uncertainty is often linked to the notion of distance from the data support - this is the basis of uncertainty estimation in Gaussian Processes.

Throughout this thesis, we make use of a *Deep Ensemble* approach [91] for uncertainty estimation. When dealing with neural networks, *Deep Ensembles* offer highly competitive uncertainty estimates without the complexities and additional hyper-parameters of other approximate inference approaches [92]. A *Deep Ensemble* is simply a collection of identical neural network architectures that are trained independently from distinct random initialisations. Within a *Deep Ensemble* individual models require a proper scoring rule [93] to be used as the training criterion.

To define a proper scoring rule for regression, we consider the aleatoric uncertainty as part of a heteroskedastic problem formulation where the measurement noise depends on the position in the input space. The model is made to predict two outputs corresponding to the predictive mean, $\hat{\mu}_\theta(x_i)$, and the aleatoric variance, $\hat{\sigma}_{a,\theta}(x_i)^2$ [94, 95]. By assuming a probability distribution for the measurement noise, we can obtain maximum likelihood estimates for the parameters of individual models by minimising a loss function proportional to the negative log-likelihood of the chosen distribution. For example choosing a

Laplace distribution gives rise to the loss function

$$\mathcal{L} = \sum_i \frac{\sqrt{2}}{\hat{\sigma}_{a,\theta}(x_i)} \ \|y_i - \hat{\mu}_\theta(x_i)\|_1 + \log\Big(\hat{\sigma}_{a,\theta}(x_i)\Big), \tag{2.6}$$

which is occasionally referred to as a robust L1 or robust MAE loss. Such loss functions are said to be robust as they allow the model to learn to attenuate the importance of potentially anomalous training points by assigning large aleatoric variance to those examples.

Due to the non-convex nature of the loss landscape, different initialisations typically end up in different local basins of attraction within the parameter space that have approximately equal losses [96]. An ensemble of W models therefore results in W sets of plausible model parameters, $\{\hat{\theta}_1, ..., \hat{\theta}_W\}$. In the *Deep Ensemble* these samples can be used to make estimates for the expectation of the model, $\hat{y}(x_i)$, and the epistemic contribution to its variance, $\hat{\sigma}_e^2(x_i)$,

$$\begin{aligned} \hat{y}(x_i) &= \int P(\hat{\theta}|x,y)\hat{\mu}_\theta(x_i)\ d\hat{\theta} \\ &\simeq \frac{1}{W}\sum_w^W \hat{\mu}_{\theta_w}(x_i), \end{aligned} \tag{2.7}$$

$$\begin{aligned} \hat{\sigma}_e^2(x_i) &= \int P(\hat{\theta}|x,y)\Big(\hat{y}(x_i) - \hat{\mu}_\theta(x_i)\Big)^2\ d\hat{\theta} \\ &\simeq \frac{1}{W}\sum_w^W \Big(\hat{y}(x_i) - \hat{\mu}_{\theta_w}(x_i)\Big)^2, \end{aligned} \tag{2.8}$$

where $P(\hat{\theta}|x,y)$ is the hypothetical distribution of models that could explain the data. The effective marginalisation of $P(\hat{\theta}|x,y)$ from using an ensemble of models not only provides a way to estimate the epistemic uncertainty but also invariably leads to lower average errors. The total uncertainty of the ensemble expectation is, in accordance with the law of total variance, simply the sum of the epistemic contribution and the average of the aleatoric contributions from each model in the ensemble,

$$\hat{\sigma}^2(x_i) = \hat{\sigma}_e^2(x_i) + \frac{1}{W}\sum_w^W \hat{\sigma}_{a,\theta_w}^2(x_i). \tag{2.9}$$

2.2 Density Functional Theory

Much of the data that underpins the research presented in this thesis is generated using Density Functional Theory (DFT). To apply data-driven techniques effectively, it is necessary to have a good physical understanding of the data. Consequently, the following sections outline the core concepts behind DFT and high-throughput computational material science databases.

2.2.1 Kohn-Sham equations

In general, our aim in *ab initio* computational material science is to solve the full time-independent many-body Schrödinger equation,

$$\hat{H}\Psi_n(\{R\},\{r\}) = E_n\Psi_n(\{R\},\{r\}). \tag{2.10}$$

This equation gives the energy, E_n, of the n^{th} state of a system of atomic nuclei at positions $\{R\}$ and electrons at positions $\{r\}$ that are described by a wavefunction $\Psi_n(\{R\},\{r\})$ under a Hamiltonian, $\hat{H}$. Sadly, directly solving this equation is computationally intractable for systems with more than a few particles. The first simplifying approximation required to efficiently solve the above equation is that, due to the disparity in masses, nuclei can be approximated as stationary on the time scales of electron motion. This approximation is known as the Born-Oppenheimer approximation and allows the Schrödinger equation to be solved solely in terms of electron wavefunctions,

$$\hat{H}_e(\{R\})\Psi_n(\{r\}) = E_n\Psi_n(\{r\}). \tag{2.11}$$

In the Born-Oppenheimer approximation the nuclei positions now enter the equation as parameters of the many-electron Hamiltonian, $\hat{H}_e$. To make further progress the many-electron Hamiltonian is typically broken apart into electron kinetic energy, $\hat{T}_e$, electron-electron interaction energy, $\hat{V}_{ee}$, and external energy (e.g. electron-nuclei interactions), $\hat{V}_{ext}$, operators,

$$\hat{H}_e = \hat{T}_e + \hat{V}_{ee} + \hat{V}_{ext}. \tag{2.12}$$

In contrast to other quantum mechanical theories, the central quantity of DFT is not the many-electron wavefunction but the electron density. To get from the many-electron Hamiltonian in terms of many-electron wavefunctions to DFT relies on the Hohenberg-

Kohn (HK) theorems. The first HK theorem states that the external potential is a unique functional of the electron density $n(r)$. This allows us to write the total energy as a functional,

$$E_{tot}[n(r)] = \int d^3r \; V_{ext}(r) \; n(r) + F_{HK}[n(r)] \tag{2.13}$$

where $F_{HK}[n(r)] = T_e[n(r)] + V_{ee}[n(r)]$ is the HK functional. Critically, the HK functional is a universal functional completely independent of the system being studied. The second HK theorem states that the ground-state density uniquely minimises the total energy functional.

Whilst the HK theorems tell us that a universal HK functional exists, the explicit form that it takes remains unknown. Determining approximations for the HK functional is the basis for much of the research into DFT. A significant step forward was presented in the formulation of the Kohm-Sham (KS) equations. Kohm and Sham rewrote the HK functional in terms of the differences to a reference system without electron-electron interactions, such that

$$F_{HK}[n(r)] = T_s[n(r)] + E_H[n(r)] + E_{xc}[n(r)], \tag{2.14}$$

where $T_s[n(r)]$ is the kinetic energy functional for a non-interacting set of electrons with the density $n(r)$, $E_H[n]$ gives the energy arising from the Hartree potential (the classical Coulomb interaction between two distributions of charges with density $n(r)$), and $E_{xc}[n(r)]$ is the exchange correlation functional capturing the differences in the kinetic and electron-electron potential energies between interacting and non-interacting systems. Recasting the equation in this way allows the total energy to be written as

$$E_{tot}[n(r)] = T_s[n(r)] + \int d^3r \; V_{ext}(r) \; n(r) + E_H[n(r)] + E_{xc}[n(r)]. \tag{2.15}$$

The first three terms in this equation are known, leaving the exchange-correlation functional as the only potential source of approximation in an otherwise exact framework. Given an approximation for the exchange-correlation functional, Kohn and Sham proposed that Equation 2.15 could be solved by the reintroduction of wavefunction-like single-electron KS orbitals, $\psi_i(r)$. These KS orbitals are constrained to be orthonormal and satisfy $n(r) = \sum_i |\psi_i(r)|^2$. Applying the variational method to minimise the total energy and using Lagrange multipliers to ensure the constraints are satisfied leads to the KS

equations

$$\hat{H}_{KS}[n(r)]\psi_i(r) = \left(-\frac{\nabla^2}{2} + \hat{V}_{KS} \right)\psi_i(r) = \epsilon_i \psi_i(r), \tag{2.16}$$

where $\hat{V}_{KS} = \hat{V}_{ext} + \hat{V}_H + \hat{V}_{xc}$, the Hartree potential, V_H, is the functional derivative of $E_H[n(r)]$ with respect to n(r), and the exchange-correlation potential, $\hat{V}_{xc}$, is the functional derivative of $E_{xc}[n(r)]$ with respect to $n(r)$. Since $\hat{V}_{KS}$ depends on the KS orbitals the KS equations need to be solved in a self-consistent manner until converged to an acceptable tolerance. Once satisfactorily converged the total energy can then be calculated as

$$E_{tot} = \sum_i \epsilon_i - E_H[n(r)] - \int d^3r \; V_{xc}(r) \; n(r) + E_{xc}[n(r)]. \tag{2.17}$$

2.2.2 Exchange Correlation Functionals - Jacob's Ladder

The formulation of DFT and the KS equations would be exact if the universal exchange-correlation functional were known. The great success of DFT in practice is that many compelling approximations for the exchange-correlation functional have been developed. These functionals exist in a hierarchy of computational cost and accuracy commonly referred to as Jacob's Ladder.

The bottom rung of the ladder is the Local Density Approximation (LDA). The LDA approximates the exchange-correlation energy at a given point using the exchange-correlation energy of a uniform electron gas of the same charge density. In the LDA, the exchange contribution is determined exactly, and the correlation contribution is obtained from a parameterisation of Monte-Carlo reference calculations.

The second rung of the ladder consists of Generalised Gradient Approximations (GGA). GGA functionals are parameterised both in terms of the charge density and its gradient. Of attention for this thesis is the Perdew–Burke-Ernzerhof (PBE) functional [97] as it is the functional of choice for the high-throughput data sources we make use of [36, 37, 98]. PBE is a so-called "physicists' functional" as its parameterisation is constructed to satisfy a set of theoretical limits. This can be contrasted against "chemists' functionals" which are typically parameterised to more accurately reproduce experimental observations.

Beyond GGAs exist meta-GGAs and hybrid-GGAs. Meta-GGAs seek to further improve accuracy by including more gradient information. For numerical stability, meta-

GGAs often use the kinetic energy density instead of the Laplacian charge density. Amongst meta-GGAs, the strongly constrained and appropriately normed (SCAN) functional [99], and regularised versions thereof [100, 101], have recently attracted significant interest for high-throughput applications due to their ability to achieve corrected accuracies (see below) as low as ~ 30 meV per atom [102]. The computational cost of using the SCAN functional in a high-throughput setting has been reported to approximately five times that of PBE [103]. Hybrid-GGAs seek to improve accuracy by including a fraction of the exact Hartree-Fock exchange term. An important example of a hybrid functional is the Heyd-Scuseria-Ernzerhof (HSE) functional [104] which includes only the short-range component of the Hartree-Fock exchange term. HSE is of interest as it gives better bandgaps predictions than either LDA or PBE, which suffer from systematic underestimation issues. This improved accuracy comes at a computational cost as the two-electron integrals in the Hartree-Fock exchange term canonically scale as $\mathcal{O}(N^4)$, where N is a measure of the system size, as opposed to the $\mathcal{O}(N^3)$ scaling of lower rungs on Jacob's ladder.

2.2.3 Periodicity, Plane-waves and Pseudopotentials

In this thesis, we are primarily interested in inorganic crystalline materials. To improve the efficiency of calculations, crystals are typically broken down into minimal repeating units called unit cells. Periodic boundary conditions are then applied to allow the calculation of bulk properties. In this periodic setup the KS orbitals are naturally represented using a plane-wave basis set that satisfies Bloch's theorem,

$$\psi_k(r) = \sum_G c_{k+G} e^{i(k+G)\cdot r}. \tag{2.18}$$

where G is a reciprocal lattice vector and k is a vector in reciprocal space within the brillouin zone. The size of the plane-wave basis set is selected using an energy cutoff for the kinetic energy of the plane waves,

$$\frac{\hbar^2}{2m_e}|k+G|^2 < E_{cut}. \tag{2.19}$$

When carrying out DFT calculations, to obtain reliable results, we need to increase the energy cutoff, E_{cut}, to the point where the total energy has converged as a function of

the cutoff. When treating all the electrons independently, the valence KS orbitals display oscillatory behaviour in the core region due to the requirement that they are orthogonal to core KS orbitals. Consequently, a large energy cutoff is required to resolve these oscillations, increasing the computational cost. To reduce these costs, whilst maintaining accuracy, pseudopotentials are typically used. Pseudopotentials are constructed by freezing the core electrons and combining the nuclei and core-electron potentials into a single smooth potential. Pseudopotential use reduces the cost of solving the KS equations in two distinct ways: firstly, reducing the number of KS orbitals required to describe the system, and secondly, allowing much lower energy cutoffs to be used by reducing the oscillatory behaviour of valence KS orbitals.

2.2.4 High-Throughput Databases

As DFT has matured as an approach for the *ab initio* calculation of the energies and properties of materials, best practise frameworks have been developed to enable high throughput calculations for large numbers of materials. Of these, the most notable efforts are the *Materials Project* (MP) [98] - which focuses on calculating the properties of known materials with workflows for the calculation of a variety of properties, including elastic, dielectric, and piezoelectric tensors, the *Open Quantum Materials Database* (OQMD) [36, 37] - which targets discovery of novel materials and has undertaken extensive screening of common structure prototypes such as Heuslers and perovskites, and the *Automatic-FLOW for Materials Discovery* (AFLOW) [105] - which contains an extensive exploration of binary alloys as well as Cobalt, Nickel and Iron-based ternary superalloy systems.

As ever-increasing amounts of data are generated, the vast majority will fall outside of the large high-throughput efforts listed above. Accordingly, significant complementary efforts to facilitate the standardisation and dissemination of computational materials science data and meta-data are being pursued, notable examples of such work are: the *Materials Cloud* [106], the *Novel Materials Discovery Repository and Archive* (NOMAD) [107], and the *Open Databases Integration for Materials Design* (OPTIMADE) API [108].

2.2.5 Formation Energies, Correction Schemes and Stability

A useful property that can be calculated using DFT is the formation energy. The formation energy of a material is defined as the energy required to form a compound

from its constituent elements,

$$\Delta E_f = E_{tot} - \sum_i \mu_i x_i \tag{2.20}$$

where μ_i and x_i are the chemical potential and abundances of element i in the material. The chemical potentials are typically determined from the DFT total energy of the elemental ground state. The accuracy of GGA DFT in predicting the formation energy across diverse chemistries is believed to be $\sim$ 100 meV per atom [37]. To improve the accuracy of calculations with respect to experiments, high-throughput workflows often include element-specific settings or post-hoc correction schemes to account for the shortcomings of DFT. For example, DFT struggles to reproduce highly localised orbitals expected for d and f-valence electrons in some transition metal, lanthanide, and actinide chemistries. A relatively inexpensive approach to tackling this issue is the DFT+U method that adds a Hubbard-like term to the Hamiltonian [109]. Both MP and OQMD use the simplified (rotationally invariant) DFT+U scheme introduced by Dudarev et al. [110] as implemented in the *Vienna Ab initio Simulation Package* (`VASP`)[1] [113]. Considering post-hoc corrections, the most important effect that correction schemes address is that the experimental reference phases for some systems are in a different state to the DFT reference phases obtained at the static lattice level. Notably, several elements are gaseous at STP, e.g. Oxygen, Nitrogen, Chlorine, Fluorine. Using DFT+U, where appropriate, in conjunction with post-hoc correction schemes based on either composition [114–116] or local environments [117, 102] can reduce the error in the formation energy with respect to experiment to $\sim$ 50 meV per atom when using a GGA functional. This is comparable with the uncertainty of experimental measurements of the formation energy, reported to be around $\sim$ 80 meV per atom [37].

One of the major aims of computational materials discovery is identifying structures that are likely to be experimentally accessible. Whilst a negative formation energy is a prerequisite for thermodynamic stability, it is not sufficient. To assess whether a material

[1] Due to the choice of projector functions the choice of U parameters in DFT+U schemes are not comparable between different DFT implementations. This is one of the factors that prevents the deterministic calculation of material properties using nominally equivalent settings within different simulation packages. Whilst the differences are normally small, this can be a significant consideration for the work conducted in this thesis because databases we leverage, like OQMD and MP, are built on top of `VASP`. Consequently, if we wish to replicate their results in a machine learning pipeline without a covariate shift, we cannot use other simulation packages such as `CASTEP` [111] or `Quantum ESPRESSO` [112] that have free academic licenses.

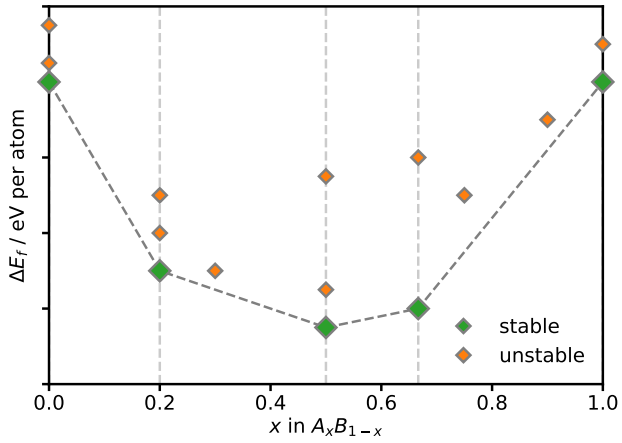


Figure 2.1: **Schematic Representation of a Convex Hull in the A-B Chemical system.** The green points represent the lowest energy structures of stable compositions. The orange points are either higher energy structures of these stable compositions or correspond to structures of compositions which are not stable to decomposition.

is stable against decomposition we need to compare it against other materials within the same chemical system. The most convenient approach to do this is via the construction of the convex hull of materials within the given chemical system, Figure 2.1 shows a toy 2D example. Materials lying on or close to the convex hull are the most likely to be experimentally accessible.

One of the potential pitfalls of using high-throughput databases is that their exploration of materials space is sparse. As a consequence, many under-explored patches exist. Convex hulls calculated for these under-explored regions can be unphysical due to missing stable materials. A useful construction to identify whether this might be a problem is the phase separation energy or decomposition energy [118, 119]. The phase separation energy is the energy to the convex hull calculated when materials with the same composition as the target material are excluded from the construction of the convex hull. If stable materials have very large phase separation energies it is a strong indicator that the chemical system in question is under-explored.

It is important to note that even in chemical systems that have been extensively sampled, the hulls constructed using the DFT total energy are likely to differ from the actual hulls for many reasons. The choice of exchange-correlation functional and often neglected effects of finite temperatures and zero-point fluctuations can cause changes in the relative stability of materials.

Chapter 3

Structure-Function Relationships in Cuprate Superconductors

This chapter is based on Rhys EA Goodall, Bonan Zhu, Judith L MacManus-Driscoll, and Alpha A Lee. Materials informatics reveals unexplored structure space in cuprate superconductors. *Advanced Functional Materials*, page 2104696, 2021. Reproduced with permission from Wiley-VCH GmbH.

Superconductivity is an uncommon physical phenomenon whereby on cooling a material below a critical transition temperature (T_c), the electrical resistance of the material vanishes. Harnessing this effect has enabled the development of various superconducting technologies such as superconducting-electromagnets and superconducting quantum interference devices (SQUIDs). The primary industrial use of superconducting-electromagnets is in MRI scanners for healthcare applications. However, these devices also play a central role in modern physics, facilitating large scale experiments such as the Large Hadron Collider, where powerful magnetic fields are used to steer the particle collider's beam, and the International Thermonuclear Experimental Reactor, where the magnetic field will be used to contain the high-temperature plasma and hopefully enable a sustained nuclear fusion reaction resulting in the net release of energy.

The most widely used superconductors in industrial applications are Nb-Ti alloys and Nb_3Sn with maximum T_c's of ~ 9 K and ~ 18 K respectively. Accordingly, liquid helium, with a boiling point of 4.2 K, is required to cool these materials below their critical temperatures. Nb-Ti alloys and Nb_3Sn are examples of conventional superconductors – those that can be well explained by the BCS theory of superconductivity. Whilst

conventional superconductors with higher T_c's exist at ambient pressure, e.g. MgB_2 has $T_c \sim 39$ K, other materials requirements currently limit the construction of practical devices. Furthermore, such devices would still require liquid helium to be used as the coolant. Recently there have been reports of conventional superconductivity in some hydride materials at giga-pascal pressures [120, 121], whilst these materials phases are unlikely to be stabilised at ambient pressures, they highlight how when we understand the underlying physics, we can identify materials with higher T_c [122, 123].

In 1986, whilst investigating the electronic properties of oxide ceramics, J. Georg Bednorz and K. Alex Mueller discovered superconductivity in the La-Ba-Cu-O system at ~ 30 K [9], work for which they were later awarded the Nobel prize in 1987. This discovery was quickly followed by the discovery of higher T_c's in related cuprate materials, notably Y-Ba-Cu-O based cuprates which were the first materials reported with critical temperatures above 77 K. This is a significant breakpoint as it is the boiling temperature of liquid nitrogen, meaning that Y-Ba-Cu-O based cuprates no longer require a liquid helium coolant.

Over the decades since these breakthroughs, significant amounts of research has been devoted to tuning, doping, and growing new cuprates to understand their physics and optimise their properties – both their T_c and other important properties such as the coherence length of the superconducting charge carriers which controls the ability of super-currents to exist in polycrystalline materials. This work has uncovered a variety of scaling laws [124–126] and structure-function relationships [127–129] that provide insight into the origin of superconductivity in cuprates and other superconducting systems. However, the many different variables that influence T_c cannot all be optimised at the same time. Furthermore, some are interdependent, making it difficult to establish causation versus correlation for these trends. In particular, bond distances cannot be independently tuned in the three orthogonal directions.

The most common way to tune the structure of cuprates is via thin-film epitaxy, where the choice of substrate is used to tune the in-plane lattice parameter. However, the out-of-plane lattice parameter is not fixed in such a setup and it responds elastically to the in-plane strain. Recently, experimental techniques based on nano-engineering [130, 131] have been reported that enable independent tuning of lattice distances. These techniques open up new experimental pathways to increase T_c through the exploitation of structure-function relationships.

Several structure-function relationships are now well established in the literature for high-temperature superconductivity (HTS) in cuprates. The most important factors believed to be relevant for the T_c are:

i) The concentration of charge carriers in the conduction planes,

ii) The nature of bonding in the charge-reservoir layers,

iii) The in-plane Cu-O distance, and

iv) The apical Cu-O distance (here we do not consider electron-doped superconducting cuprates without apical oxygens, i.e. those adopting a T' structure).

This work focuses on the structure-dependent relationships between T_c and the apical and in-plane Cu-O distances. The influence of the apical Cu-O distance on T_c is attributed to its effect on the localisation of charge carriers in the CuO_2 planes [132]. The in-plane Cu-O distance is believed to be important due to its impact on Cu-O-Cu super-exchange in the CuO_2 planes. Such attempts to couple experimentally observed trends to physically relevant mechanisms provide qualitative understanding and insight into the nature of superconductivity in these systems. Unfortunately, whilst progress has been made on theories describing the percolative nature through which superconductivity emerges in cuprates [133–135] and the importance of inhomogeneity in such processes, the consolidation of satisfactory, quantitative theories for HTS in cuprates that reflect known structure-function relationships and are capable of making predictions about T_c, remains a challenge [136, 137].

In the absence of an accepted mechanistic theory, the potential availability of large amounts of historical data and high impact applications has led some researchers towards data-driven phenomenological approaches, i.e. machine learning. Thus far, most work in this area has focused on building models for predicting T_c given a set of easy to evaluate descriptors that represent the materials in question [138–140]. The hope is that such models may enable the discovery of new families of high-temperature superconductors by first detecting abstract empirical patterns in featurisations of materials currently known to display superconductivity, and then screening new materials based on similarity to these identified patterns. However, questions exist about whether such approaches will be fruitful when tested experimentally as the evaluation metrics used in proof-of-concept workflows are often not reflective of real materials discovery workflows [141].

Two fundamental failings are shared by many of these machine learning approaches. Firstly, data sets of superconducting materials are highly correlated – researchers will often conduct series experiments where doping is varied slightly – when this is coupled to studies that report cross validation results under random splitting, it is trivial for models to achieve high accuracies as they are asked to make predictions about materials that are incrementally different from those in the training set. Secondly, when these models are used to screen materials libraries to identify novel superconducting candidates, they often neglect to carry out uncertainty estimation. As these data sets are biased towards materials with high T_c and rarely contain negative examples, the mean T_c of the training data is high (typically in excess of 30K). For most machine learning models, we expect that the model should predict the mean with high uncertainty far away from the data support. Consequently, only predictions with low uncertainty and T_c higher than the mean are worth investigating. If we apply a simple uncertainty estimation approach based on the impurity and diversity of leaves [142] to the Random Forest model introduced in Ref. [138], we see that out of the 35 candidate materials highlighted only 6 materials, $KBa_6Zn_4Ga_7O_{21}$, $RbSeO_2F$, $CsSeO_2F$, $KTeO_2F$, $CsBeAsO_4$, and $K_{0.8}Li_{0.2}Sn_{0.76}O_2$, are inconsistent with the training set mean at 1-σ uncertainty. We note that none of these 6 are predicted to have T_c greater than the mean.

In contrast to materials screening applications involving machine learning models, the application of materials informatics approaches, in conjunction with careful physical insights, has shown promise for probing our understanding of systems already known to display superconductivity [143]. In this chapter, we adopt a materials informatics approach to investigate the importance of the apical and in-plane Cu-O distances for the T_c of HTS cuprates. By combining high-resolution data structural data where the T_c is unknown with coarse-resolution data where T_c is known, we show the existence of unexplored regions of the materials space defined by the apical and in-plane Cu-O distances that are ripe for further experimental investigation. Our results highlight how materials informatics can play an important role in helping to guide experimental efforts in material science.

3.1 Inferring Structural Parameters of Cuprates

The principal data source for the work in this chapter is the *SuperCon* database compiled and distributed by the Japanese National Institute for Material Science. Whilst *SuperCon* records the critical temperatures of an extensive range of superconducting materials, the information available for each composition is minimal – structural information is only available for a small proportion of the entries, and it is limited to the lattice parameters when available. Unfortunately, it is the common structure shared between different cuprates, characterised by the apical and in-plane distances of the superconducting CuO_2 planes, that is most interesting for examining trends.

Whilst the lattice parameters can be determined relatively easily via x-ray diffraction, directly measuring the atomic positions, needed to determine the apical and in-plane distances, typically involves much more specialised x-ray diffraction apparatus or neutron diffraction experiments. Fortunately, many cuprate structures are already recorded in the ICSD [13]. However, the critical temperatures of these materials are not recorded alongside their structures, therefore, limiting the utility of the ICSD as a data source for studying T_c-structure trends.

Consequentially, whilst large amounts of data are available on cuprates, the structure of that data is incomplete in terms of the information required to look at structure-function relationships. As a result, trends are often examined between relatively small numbers of hand-selected data points for which the information necessary for the analysis of a proposed structure-function hypothesis is available. Selection of data in this way has the potential to lead to bias and overconfidence in T_c-structure trends observed. In this chapter, we attempt to overcome this issue of incomplete data by obtaining estimates for the apical and in-plane distances from the knowledge of the more readily available lattice parameters. Whilst the accuracy of such an approach is diminished and prevents truly quantitative analysis, using estimates allows for a far greater number of examples to be considered, therefore ensuring the robustness of any trends that remain.

For cuprates, the a-lattice parameter is closely related to the in-plane distance. Assuming that the CuO_2 planes are approximately square planar this entails that the in-plane distance can be estimated as $a/2$ for tetragonal phases and $a/2\sqrt{2}$ for octahedral phases. This assumption breaks down for cuprates under pressure where the CuO_2 planes generally tend to buckle to relieve pressure on the structure.

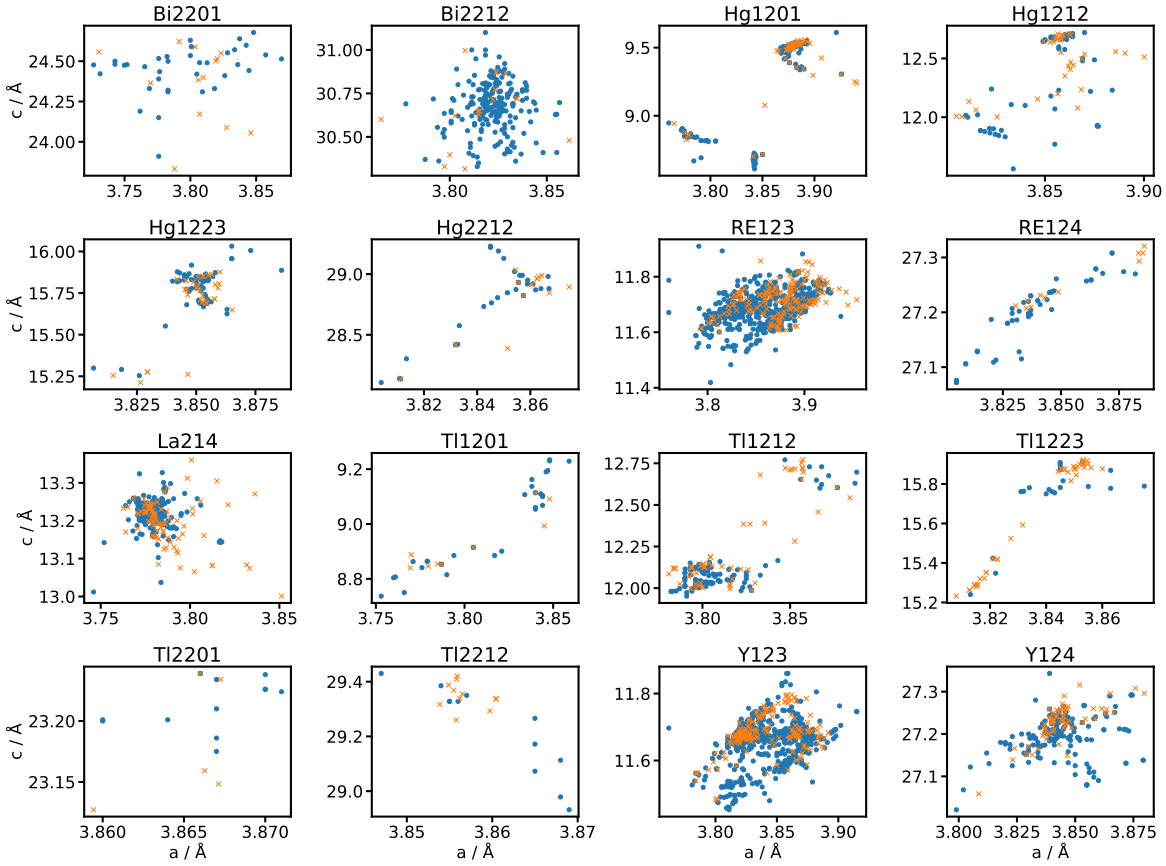


Figure 3.1: **SuperCon and ICSD show good density overlap for cuprate lattice parameters.** The figure shows the overlap in the lattice parameters between the SuperCon data set (blue dots) and the ICSD reference data (orange crosses). See Table 3.1 for a list of abbreviations.

Whilst the c-lattice parameter is often considered as a proxy for the apical distance, there is little correlation between the two – the c-lattice parameter also depends on the thickness of the charge-reservoir layers, which can vary significantly between different families of cuprates. Consequentially, to estimate the apical distances, each family has to be treated independently. The approach adopted here is to use linear calibration models constructed on reference data that relate the a and c-lattice parameters to the apical distance. Figure 3.1 shows a scatter plot of the structure space defined by the a and c-lattice parameters for both the source data (*SuperCon*) and the reference data (ICSD). The 16 cuprate families investigated here were selected due to having greater than 5 examples in both ICSD and *SuperCon*. We see that both data sets assign density to similar regions of the structure space for these families.

Table 3.1: **Abbreviations for cuprate families.** We make use of the A-jk(n-1)n four-digit notation for cuprates of the form $A_jB_kS_{n-1}Cu_nO_{j+k+2n}$ described in Ref. [144] for Tl, Hg and Bi based cuprates. In the RE123 and RE124 abbreviations RE denotes a range of rare earth metals i.e. RE={Nd, Gd, Pr, etc.}. The representative formulas given are not exclusive and the materials contained in both SuperCon and ICSD contain a variety of dopants, e.g. Y for Ca in Bi2212 to reduce cation disorder [145], and elemental substitutions, e.g. Sr for Ba in Hg1201 to produce Ba-free cuprates [146].

Family Label	Representative Formula
La214 (T)	$La_2CuO_4/La_{2-x}Sr_xCuO_4$
Y123	$YBa_2Cu_3O_7$
RE123	$YBa_2Cu_3O_7$
Y124	$YBa_2Cu_4O_8$
RE124	$YBa_2Cu_4O_8$,
Tl1201	$TlBa_2CuO_5$
Tl1212	$TlBa_2CaCu_2O_7$
Tl1223	$TlBa_2Ca_2Cu_3O_9$
Tl2201	$Tl_2Ba_2CuO_6$
Tl2212	$Tl_2Ba_2CaCu_2O_8$
Hg1201	$HgBa_2CuO_5$
Hg1212	$HgBa_2CaCu_2O_6$
Hg1223	$HgBa_2Ca_2Cu_3O_9$
Hg2212	$Hg_2Ba_2YCu_2O_8$
Bi2201	$Bi_2Sr_2CuO_6$
Bi2212	$Bi_2Sr_2CaCu_2O_8$

The variation in the apical distance within families is due to the chemical pressure that arises from doping or atomic substitutions. The simplest model is that the pressure imparts a uniform stress along the c axis that strains the material. If the total strains are in the Hookean regime, the strain along the c-axis should then be directly proportional to the strain in the apical distance - the implication being that we can approximate the materials' layered structure with hypothetical slabs of constant Young's modulus. The strains along a and c will also be related by the Poisson effect. Therefore, the minimal linear calibration model for the apical distances, $\hat{d}_{\text{apical}}$, that also includes this effect is:

$$\hat{d}_{\text{apical}} = \alpha c + \beta a^* + \gamma, \tag{3.1}$$

where α, β and γ are the parameters of the calibration model that need to be fitted for each family. In each case, a robust linear model based on the Huber penalty [147] was used to reduce the effect of outliers when fitting the calibration models of the form (3.1).

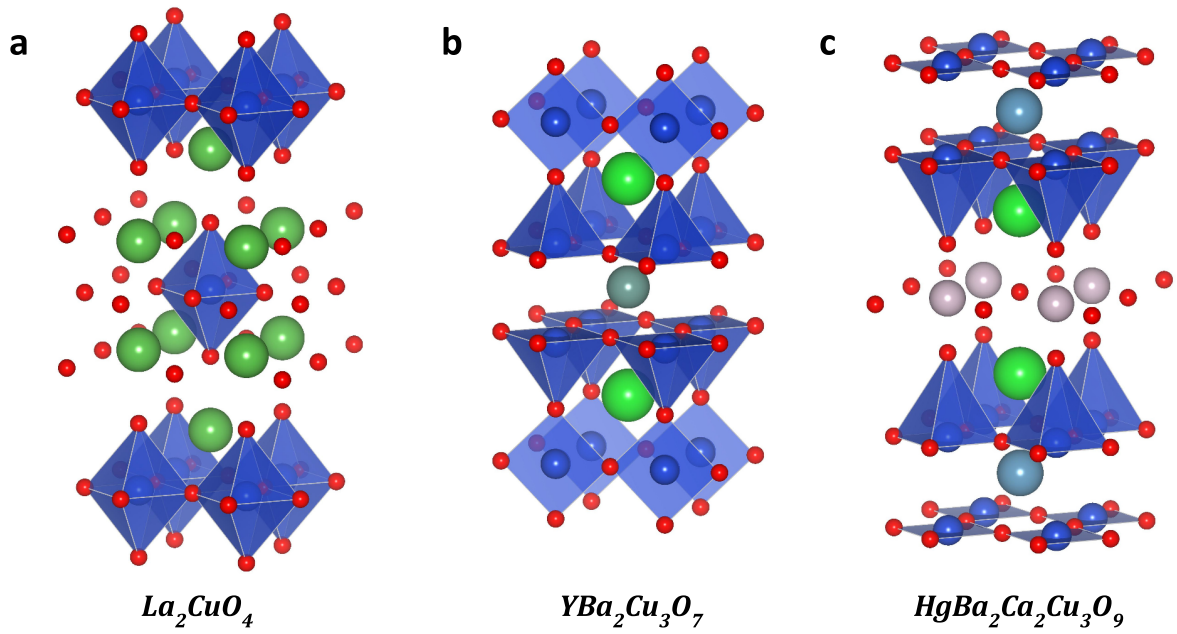


Figure 3.2: **One, two, and three layer cuprates share common structural motifs.** Prototypical crystal structures for: **a** - La214 (La_2CuO_4), **b** - Y123 ($YBa_2Cu_3O_7$), and **c** - Hg1223 ($HgBa_2Ca_2Cu_3O_9$). The blue polyhedra show the coordination of copper atoms (blue) in the structures by nearby oxygen atoms (red).

Models were fitted for the following families: La214 (T), Y123, RE123, Y124, RE124, Tl1201, Tl1212, Tl1223, Tl2201, Tl2212, Hg1201, Hg1212, Hg1223, Hg2212, Bi2201, and Bi2212 (see Table 3.1 for a list of abbreviations). To increase the amount of available reference data, both neutron and x-ray scattering structures recorded in ICSD were used. Ideally, only structures derived from neutron scattering experiments would be used as the oxygen positions for structures derived from x-ray experiments can be affected by systematic errors. However, here we believe the increased abundance and diversity of reference data outweigh this potential loss of accuracy.

To check the validity of this simple approach, we employ density functional theory (DFT) to investigate how the apical Cu-O distance changes as the lattice is strained. We only use DFT as a proxy to look for qualitative trends that are likely to be mirrored in real systems. This is due to well-documented discrepancies in the lattice parameters between the structures of cuprates as reconstructed from neutron and x-ray scattering experiments and the relaxed structures returned from DFT calculations. Details of the DFT settings used are given in Appendix B.1.

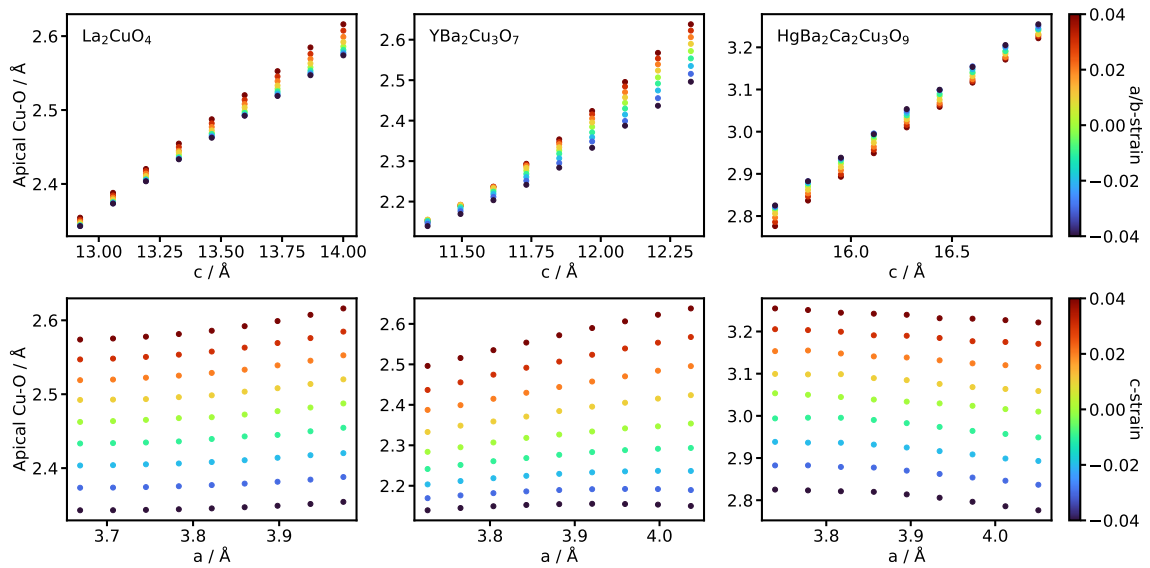


Figure 3.3: **Linear relationship exists between apical distance and lattice parameters for DFT structures.** Variations in the apical distance, as calculated via density functional theory, for three prototypical cuprate systems that have strained along the c axis and in the a/b plane. The plots show broadly linear trends for the behaviour of the apical distance as strains are applied. These results validate the use of linear models to estimate the apical distance from experimental lattice measurements.

We take La_2CuO_4, $YBa_2Cu_3O_7$, $HgBa_2Ca_2Cu_3O_9$ as prototypical test cases being illustrative of one, two and three-layer cuprates respectively (crystal structures for these materials are shown in Figure 3.2). Figure 3.3 shows that for tensile and compressive strains of up to 4%, monotonic and broadly linear trends exist between the a and c-lattice parameters and the apical distance for the prototype systems explored. The dependence on the c-lattice parameter is stronger than on the a-lattice parameter, as expected. As the strains become larger, some degree of non-linearity does appear. However, the maximal strains examined here are significantly larger than the spread in the experimental data. Whilst a more complicated model could be used to fit this non-linearity in the clean data obtained via DFT, for the experimental data other factors such as systematic variations between different experimental setups cannot be accounted for. Therefore, adding additional terms to the calibration models, without strong physically-motivated priors for their inclusion, is undesirable.

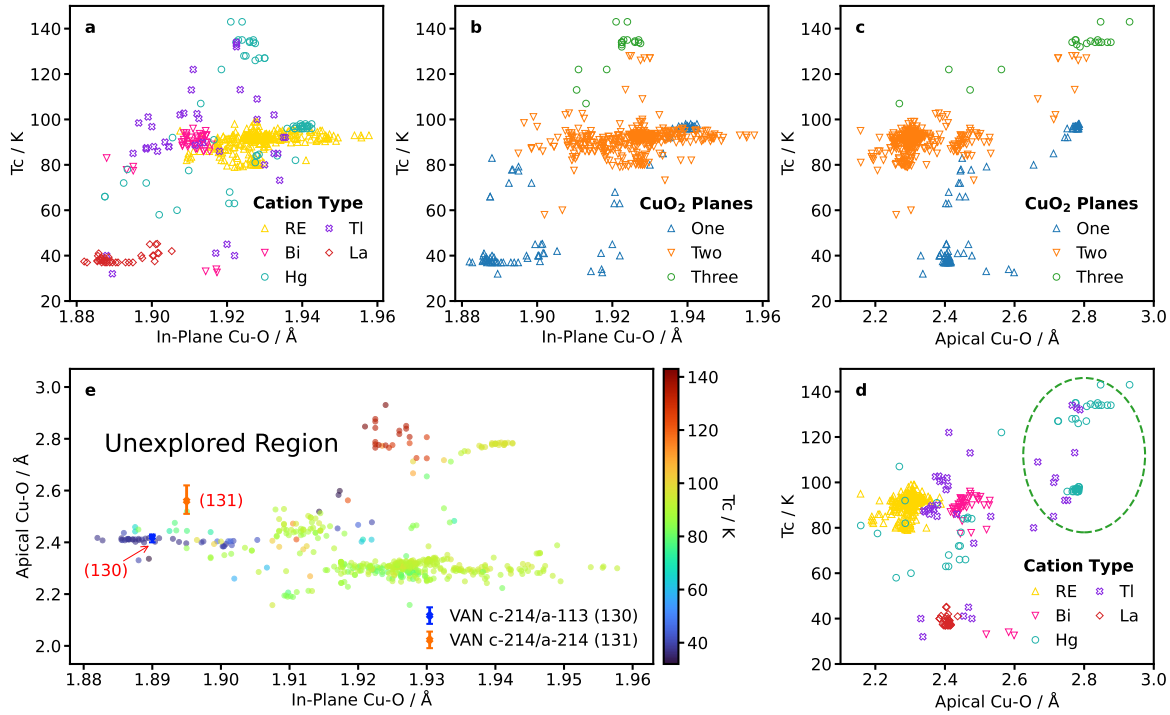


Figure 3.4: **Structure-function maps highlight the existence of an unexplored region of materials space.** Panels **a-d** show the trends between the critical temperature and the apical and in-plane distances stratified by the number of CuO_2 planes and cation type. In panels **a** and **b** we see the apparent optimum in the in-plane distance of 1.92 Å for Hg and Tl-based materials. Panels **b** and **c** clearly highlight the trend that T_c increases with the number of CuO_2 planes. The green circled region in **d** shows that the trend of T_c increasing with apical distance is only apparent due to Hg and Tl-based materials with high apical distances. **e** shows the variation of T_c with the apical and in-plane distances. A large region, labelled "Unexplored Region", of high apical Cu-O distance and low in-plane Cu-O distance is apparent. Experiments that probe this region are likely to provide useful insight into the nature of T_c-structure trends in cuprates. We highlight the vertically aligned nanocomposite (VAN) samples from Refs. [130, 131] as examples of nano-engineering approaches to investigate the region. In the key VAN c-214/a-113 refers to the $La_2CuO_{4-\delta}/LaCuO_3$ interface from Ref. [130] and VAN c-214/a-214 refers to the c-aligned $La_2CuO_{4-\delta}$/a-aligned $La_2CuO_{4-\delta}$ interface from Ref. [131].

3.2 Materials Informatics Reveals Unexplored Structure Space

Beyond the apical and in-plane distances, other important factors known to influence T_c need to be considered. Unfortunately, many of these are typically harder to quantify, for example, the nature of bonding in the charge-reservoir layers. Perhaps the most important of these factors is that achieving optimal oxygen-doping is necessary to maximise T_c. Given the aim of maximising T_c, we are generally only interested in trends between materials characterised in optimal states. Unfortunately, materials in *SuperCon* are commonly reported with unknown oxygen concentrations. This is problematic because achieving optimal oxygen doping for a given composition within a family depends on which or whether other dopants are present. As a result, naively selecting the materials with the highest T_c's would end up discarding much of the diversity in the data set, making it difficult to establish trends. To ensure that we maintain as much structural diversity as possible, we first perform k-means clustering [148] on the data in the a and c-lattice parameter space and then take the top 20% of each cluster by T_c. This selection strategy ensures that the data points considered in the subsequent analysis are diverse in terms of their apical and in-plane Cu-O distances but are also likely to be close to optimal in terms of the other relevant factors for optimising T_c. Stratifying the remaining data set into different sub-groups, we observe the following trends:

1. If cuprate materials are differentiated according to the number of CuO_2 planes in the unit cell there is a clear separation between the different groups (Figures 3.4**b** and 3.4**c**). The observed increases in T_c with the number of planes are well understood considering intralayer interactions between CuO_2 planes [149]. A similar separation is also clearly visible in the Uemura relation [124], where the saturation and suppression of T_c occurs at different relaxation rates depending on the number of CuO_2 planes.
2. Whilst grouping by the number of CuO_2 planes emphasises a strong positive correlation observed between T_c and the apical distance (Figure 3.4**c**), grouping materials via the main cation (Figure 3.4**d**) shows that all the highest T_c's come from Hg and Tl-based materials with high apical distances (Circled in green in Figure 3.4**d**). We note that higher critical temperatures have been achieved in both two and three-layer Hg-based cuprates via the application of hydro-static pressure

[150, 151] which is known to decrease the apical distance [152] with minimal impact on the in-plane distance. However, these increases in T_c have been attributed to the effect of pressure on the position of Ba atoms in the buffer layer [153, 154].

3. There is an apparent optimum in-plane distance for Hg and Tl-based cuprates around 1.92Å (Figure 3.4**a**). In contrast, Rare-Earth-based (RE) cuprates show minimal variation in T_c as the in-plane distance changes. Looking at the highest T_c Bi-based materials, there is a slight increase in T_c as the in-plane distance approaches 1.92Å. However, as there are no Bi-based materials reported with in-plane distances above 1.92Å in the data sets examined, it is not apparent whether a drop off in T_c, as is the case for Hg and Tl-based materials, would be observed. The slight increase in T_c with the in-plane distance for these Bi-based materials could perhaps be attributed to changes in the multi-layer structure between the high T_c Bi2201 type materials ($Bi_{2+x}Sr_{2-x-y}Ca_yCuO_{6+\delta}$) [155] and the Bi2212 ($Bi_2Sr_2CaCu_2O_{8+\delta}$) family. This suggests that for constant apical distance there may be no strong dependence on the in-plane distance for the Bi-based materials.

Figure 3.4**e** shows the apical distance versus in-plane distance. The data points are coloured according to their T_c, with red colours indicating higher T_c and blue colours indicating lower T_c. It is clear that large areas of the apical versus in-plane materials space, potentially yielding higher T_c materials, remain unexplored (This region is labelled "Unexplored Region" in Figure 3.4**e** and it occurs for high apical distance, i.e. above 2.5Å).

Past efforts have only been able to sample in small regions around known systems due to the limitations of perturbing systems with mechanical and chemical pressures. A key limitation is that due to Poisson effects, such methods influence both the a and c-lattice parameters, preventing the exploration of trends in a one-factor-at-a-time manner. Recently, new experiments have shown that it is possible to tune the a and c-lattice parameters independently, allowing for unexplored regions of the materials space to be investigated [130, 131]. This vertically aligned nanocomposite (VAN) approach has led to enhanced T_c's of 50 K [130] and up to ∼120 K from magnetic measurements [131] in nano-engineered $La_2CuO_{4-\delta}$ films relative to 40 K in the bulk (These points are highlighted in Figure 3.4**e**). In Ref. [130] a T_c of 50K and a and c-lattice parameters of 3.79-3.76 Å and 13.20-13.28 Å are reported for the $La_2CuO_{4-\delta}/LaCuO_3$ (c-214/a-113) interfacial region. The presence of domain matching in the structure suggests uniform

stress along the c-axis, therefore, the apical distance can be estimated using the linear calibration model for the La214 (T) family. This gives an estimate of 2.40-2.43 Å for the apical distance. In Ref. [131] a weak magnetic signature for superconductivity at 120K is reported for a c-aligned $La_2CuO_{4-\delta}$/a-aligned $La_2CuO_{4-\delta}$ (c-214/a-214) interface. Here there is La-block matching at the interface, rather than domain matching, suggesting a non-uniform stress. The La-block is believed to be 4.00 ± 0.01Å at the interface - much larger than the average of 3.67 Å for the ICSD La214 (T) reference data. As this estimate requires a large degree of extrapolation, we cannot justify a linear model. Instead, we derive an estimate for the apical distance from considering the offset from the top of the La-block to the apical oxygen. This offset is strongly peaked around 0.56 Å giving an estimate of 2.56 Å with a 90% confidence interval of 2.51-2.62 Å. These examples support the hypothesis that the unexplored region may yield higher T_c systems.

3.3 Discussion

Having established the existence of a largely unexplored region of the structure space of cuprate superconductors, we believe that novel experimental approaches that allow for new regions of the apical versus in-plane materials space to be probed would be fruitful to further understand structure-T_c trends in cuprates and potentially increase T_c's.

From the results presented here, we believe that the 3D strain engineering of Bi2201 or Bi2212 systems are of particular interest because they lie at the base of the unexplored region with >2.5Å apical distances (Figure 3.4**e**). 3D strain engineering using VAN is suitable for Bi2201 and Bi2212 because they can be made in-situ via epitaxial growth methods [156, 157]. Such experiments would allow greater insight into whether the high T_c's of Hg and Tl based cuprates are due to structural effects from the large apical distance or intrinsic electronic effects of the Hg and Tl cations. When attempting to optimise T_c, it should be noted that both Bi2201 and Bi2212 are known to benefit from substitutional doping [145]. This potential need for substitutional doping is important to consider when selecting suitable materials for the substrate and matrix within the VAN setup. Bi2212 is also known to naturally exhibit crystal "super-modulation" which manifests as large variations in the Cu-O apical distance at the unit cell level [158].

Of interest also is whether it would be possible to increase T_c's in Hg1212 or Hg1223 thin films by preparing samples in a manner that reduces the in-plane Cu-O distance

whilst constraining the apical Cu-O distance to remain high. This would allow us to move into the "Unexplored Region" of Figure 3.4**e** from its right-hand edge (from the cluster of red points at the highest apical distance values), rather than moving up from its bottom edge as proposed for Bi-based systems. This approach is likely to be more challenging than the strain engineering of Bi2201 and Bi2212 due to issues that arise when growing Hg-Ba-Ca-Cu-O thin films [159].

Finally, we note that more systematic integration of existing data sources, deposition of new data, and novel data mining efforts [29] are desirable to improve superconductivity databases. Consolidating the vast amount of information available in the literature into a comprehensive source containing critical temperatures alongside atomic structures would facilitate the improved application of materials informatics approaches. Critically, such resources would enable direct consideration of how distributions of bond distances and variations in bond angles, e.g. buckling of the CuO_2 planes, affect T_c.

Chapter 4

Representation Learning from Stoichiometry

This chapter is based on Rhys E. A. Goodall and Alpha A. Lee. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nature Communications*, 11(1):1–9, 2020. Reproduced with permission from Springer Nature.

The discovery of new materials is vital to making technologies cheaper, more functional, and more sustainable. In our efforts to accelerate materials discovery, virtual screening has shown great potential by allowing promising candidates to be prioritised for time-consuming lab-based experiments based on the predictions of *ab initio* calculations. However, a critical challenge exists in that high-throughput virtual screening is limited to materials whose structures we can provide to our *ab initio* methods. Whilst *ab initio* crystal structure prediction is well-established [160–162] it remains computationally costly, presenting a significant challenge for high-throughput workflows seeking to explore general chemistries. Whilst alternative strategies such as prototyping from known crystal structures [37, 163, 164] can be employed to manoeuvre around this bottleneck, identifying new stable compounds in a timely manner remains an important goal for computational material science.

A recent trend of research for circumventing the structure bottleneck has been the development of machine-learning models that only depend on the composition of materials [2, 65, 165, 166]. Composition-based machine learning models have been fruitful when tackling various problems in material science such as the identification of bulk

metallic glasses [68], Li-ion conductors [167], and superhard materials [168]. However, these pioneering works all rely on hand-curated descriptors whose predictive power is circumscribed by the intuitions behind their construction.

In this chapter, we develop a novel composition-based machine learning framework that learns the stoichiometry-to-descriptor map directly from data. Our key insight is to reformulate the stoichiometric formula of a material as a dense weighted graph between its elements. A message-passing neural network is then used to directly learn material descriptors. The advantage of this approach is that the descriptor is systematically improvable as more data becomes available. Our approach is inspired by breakthrough methods in chemistry that directly take a molecular graph as input and learn the optimal molecule-to-descriptor map from data [169, 170].

We show that our model achieves lower errors and higher sample efficiency than commonly used models. Moreover, its learnt descriptors are transferable, allowing us to use data-abundant tasks to extract descriptors that can be used in data-poor tasks. We highlight the important role of uncertainty estimation for applications in material science and show that the construction of a *Deep Ensemble* [91] produces useful uncertainty estimates.

4.1 Representation Learning of Inorganic Materials

To eschew the hand-engineering required by current structure-agnostic descriptor generation techniques, we represent each material's composition as a dense weighted stoichiometry graph. The nodes in this graph represent the different elements present in the composition, with each node being weighted by the fractional abundance of the corresponding element. This novel representation for the stoichiometries of inorganic materials allows us to leverage neural message passing [170]. The message passing operations are used to update the representations of each of the element nodes such that they are contextually aware of the types and quantities of other elements present in the material. This process allows the model to learn material-specific representations for each of the constituent elements in a material and pick up on physically relevant effects such as co-doping [171] that would otherwise be obscured within the construction of hand-engineered materials descriptors. We refer to this approach as *Roost* (**R**epresentati**o**n

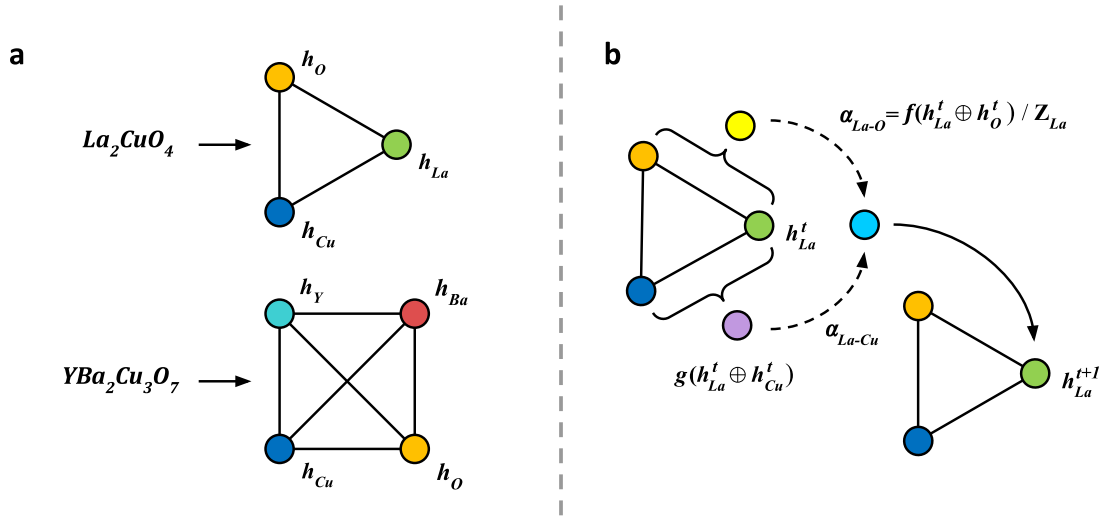


Figure 4.1: **Schematic representation of dense weighted stoichiometry graphs and graph update rule.** Panel **a** shows example stoichiometry graphs for La_2CuO_4 and $YBa_2Cu_3O_7$. Panel **b** shows a graphical representation of the update function for the La representation in La_2CuO_4. The pair dependent perturbations, shown as the cyan and purple nodes, are weighted according to their attention coefficients, α_{ij}, before being used to update the La representation.

Learning from **St**oichiometry). In the following paragraphs, we introduce a specific model based on this idea.

To begin, each element in the model's input domain is represented by a vector. Whilst the only requirement is that each element has a unique vector, it can improve performance, particularly when training data is scarce, to embed elements into a vector space that captures some prior knowledge about correlations between elements [172, 173]. These initial representations are then multiplied by a n by $d-1$ learnable weight matrix where n is the size of the initial vector, and d is the size of the internal representations of elements used in the model. The final entry in the initial internal representation is the fractional weight of the element. A message-passing operation is then used to update these internal representations by propagating contextual information about the different elements present in the material between the nodes in the graph. Figure 4.1 shows a schematic representation of this process. The mathematical form of the update process is

$$\boldsymbol{h}_i^{t+1} = U_t^{(h)}(\boldsymbol{h}_i^t, \boldsymbol{\nu}_i^t), \tag{4.1}$$

where $\boldsymbol{h}_i^t$ is the feature vector for the i^{th} element after t updates, $\boldsymbol{\nu}_i^t = \{\boldsymbol{h}_\alpha^t, \boldsymbol{h}_\beta^t, \boldsymbol{h}_\gamma^t, ...\}$ is the set of other elements in the material's composition, and $U_t^{(h)}$ is the element update function for the $t+1^{th}$ update. For this work, we use a weighted soft-attention mechanism for our element update functions. In general, attention mechanisms tell models how important different features are for their given tasks. Soft-attention builds upon this concept by allowing the function that produces the attention coefficients to be learnt directly from the data. The soft-attention mechanism is the crux behind many state-of-the-art sequence-to-sequence models used in machine translation and natural language processing [174]. Recently, attention mechanisms have also shown good results on graphs [175] and in some material science applications [176, 166]. In this domain, the attention mechanism allows us to capture important materials concepts beyond the expressive power of older approaches, e.g. that the properties and thus the representation of metallic atoms in a metal oxide should depend much more on the fact that oxygen is present than other metallic dopants being present.

The first stage of the attention mechanism is to compute unnormalised scalar coefficients, e_{ij}, across pairs of elements in the material,

$$e_{ij}^t = f^{t,m}(\boldsymbol{h}_i^t \oplus \boldsymbol{h}_j^t), \tag{4.2}$$

where $f^{t,m}(...)$ is a single-hidden-layer neural network for m^{th} head of the $t+1^{th}$ update, the j index runs over all the elements in $\boldsymbol{\nu}_i^t$, and $\oplus$ is the concatenation operation. The coefficients e_{ij} are directional depending on the concatenation order of $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$. These coefficients are then normalised using a weighted softmax function,

$$a_{ij}^{t,m} = \frac{w_j \exp\left(e_{ij}^{t,m}\right)}{\sum_k w_k \exp\left(e_{ik}^{t,m}\right)}, \tag{4.3}$$

where j indexes a given element from $\boldsymbol{\nu}_i^t$, the k index runs over all the elements in $\boldsymbol{\nu}_i^t$, and w_j are the fractional weights of the elements in the composition. The elemental representations are then updated in a residual manner [79] with learnt pair-dependent perturbations weighted by these soft-attention coefficients,

$$\boldsymbol{h}_i^{t+1} = \boldsymbol{h}_i^t + \sum_{m,j} a_{ij}^{t,m} g^{t,m}(\boldsymbol{h}_i^t \oplus \boldsymbol{h}_j^t), \tag{4.4}$$

where $g^{t,m}(...)$ is a single-hidden-layer neural network for m^{th} head of the $t+1^{th}$ update and the j index again runs over all the elements in $\boldsymbol{\nu}_i^t$. Throughout the update process we make use of multiple attention heads, indexed m, to stabilise the training and improve performance. The number of times the element update operation is repeated, T, as well as the number of attention heads per update, M, are hyper-parameters of the model that must be set before training.

A fixed-length representation for each material is determined via another weighted soft-attention-based pooling operation that considers each element in the material in turn and decides, given its learnt representation, how much attention to pay to its presence when constructing the material's overall representation. Finally, these material representations are taken as the input to a feed-forward output neural network that makes target property predictions. Using neural networks for all the building blocks of the model ensures the whole model is end-to-end differentiable. This allows for its parameters to be trained via stochastic gradient-based optimisation methods. Whilst the rest of this chapter focuses on regression tasks, the model can be used for both regression and classification tasks by adapting the loss function and the architecture of the final output network as required. Details of the *Roost* architecture and the hyper-parameters used are given in Appendix A.1.

4.2 Data Sets

For this work, we consider a selection of experimental and *ab initio* data sets. The Open Quantum Materials Database (OQMD) data set contains the formation enthalpy per atom calculated via density functional theory [36]. For comparison purposes, we take the subset of 256,620 materials from Ref. [165]. This subset contains only the lowest energy polymorph for each composition. The Materials Project (MP) data set we look at contains the band gaps for 43,921 non-metals present in the Materials Project catalogue [98, 37]. As before, we take only the lowest energy polymorph for each composition to ensure that the composition-to-property map is well defined. Finally, we consider a much smaller experimental data set consisting of 3,895 non-metals for which the band gap has been measured experimentally (EX) as used in Ref. [66].

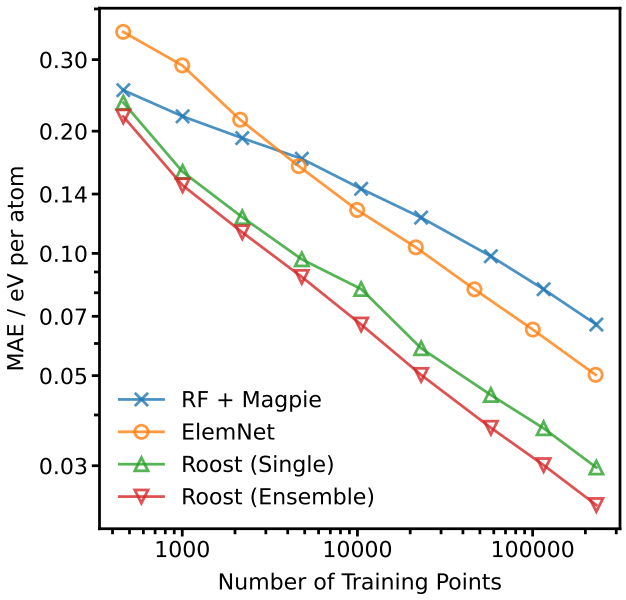


Figure 4.2: **Sample efficiency learning curve on OQMD.** The figure shows learning curves for the OQMD data set as the amount of training data is varied for a fixed test set. Plotted on log-log scales the trends follow inverse power law as expected from statistical learning theory. Results for *ElemNet* are taken from Ref. [165].

4.3 Evaluation of Sample Efficiency

Materials discovery workflows are often data limited. As a result, the sample efficiency of models is of critical importance. The sample efficiency can be investigated by looking at how the model's performance on a fixed test set changes as the model is exposed to more training data. From statistical learning theory, one can show that, in the large data limit, the average error for a model approximately follows an inverse power law relationship with the amount of training data [177, 56]. As such the gradient and intercept on a log-log plot of the training set size against the model error indicate the sample efficiency of the model.

Figure 4.2 shows such learning curves for the OQMD data set, and Table 4.1 records the benchmark results for when all the training data is used. In this case, 10% of the available data was held back from the training process as the test set. We compare our approach against a baseline model consisting of a Random Forest applied to *Magpie* embeddings [65] and *ElemNet* [165], an alternative neural network-based model that also takes the atomic fractions of each element as input. The comparison shows that the inductive biases in the representation learning approach lead to a much higher sample

Table 4.1: **Performance benchmarks on OQMD.** The table shows the mean absolute error (MAE), and root mean squared error (RMSE) for the baseline and proposed models on 10% of the data that was randomly sampled and withheld as a test set. The bracketed numbers show the standard deviation in the last significant figure.

	MAE / eV per atom	RMSE / eV per atom
RF + *Magpie*	0.067	0.121
ElemNet [165]	0.055	
Roost (Single)	0.0297(7)	0.0995(16)
Roost (Ensemble)	0.0241	0.0871

efficiency. Indeed the crossover where *Roost* begins to outperform the *Magpie* plus Random Forest baseline occurs for $O(10^2)$ data points – a size typical of experimental databases collated for novel material classes [178, 179] – as opposed to $O(10^3)$ for *ElemNet*.

4.4 Evaluation of Uncertainty Estimates

A major strength of structure-agnostic models is that they can be used to screen large data sets of combinatorially generated candidates, amortising the cost of more time-consuming and expensive experiments or calculations that may otherwise have been carried out. However, most machine learning models are designed for interpolation tasks, thus predictions for materials that are out of the training distribution are often unreliable. During a combinatorial screening of novel compositions, we cannot assume that the distribution of new materials matches that of our training data. In such applications, well-behaved uncertainty estimates can allow for models to be used with greater confidence. Consequently, beyond simply building more sample-efficient models (e.g. by designing improved architectures or leveraging techniques such as transfer learning), we are interested in whether we can construct models that know when they do not know. Figure 4.3 highlights this idea on the OQMD data set. The plot shows how the test set error varies as a function of the confidence percentile. The error for a confidence percentile of X is determined by re-calculating the average error of the model after removing the $X\%$ of the test set assigned the highest uncertainty by the model. Additional illustrative curves are included to show what would happen if the data was restricted in a random order or if the data was restricted according to the size of the model's error.

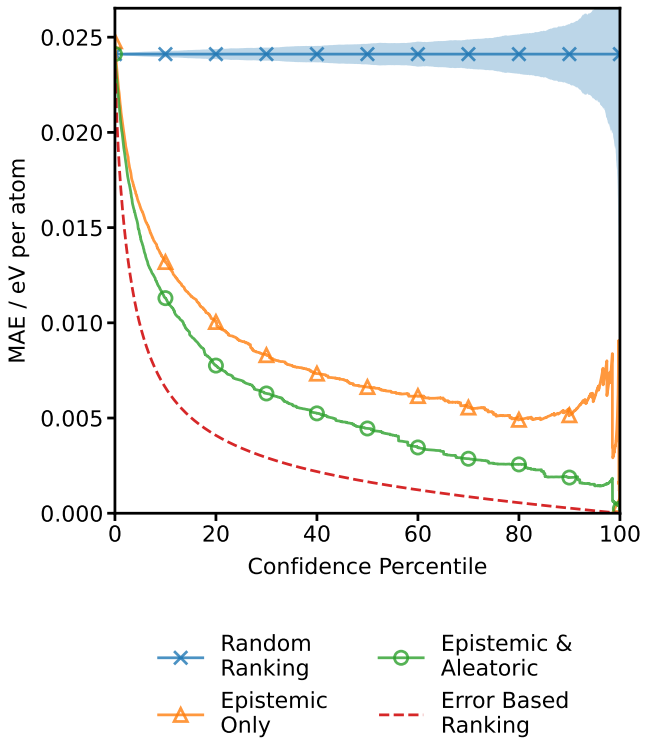


Figure 4.3: **Confidence-error curves on OQMD.** The figure shows confidence-error curves on the OQMD test set. The curves show how the average model error changes as the data points that the model is most uncertain about are removed sequentially. The random ranking-based curve in blue serves as a reference showing the result if all points are treated as having equal confidence. The blue shaded area highlights the curve's standard deviation computed over 500 random trials. The error-based ranking curve shown in red gives a lower bound on the uncertainty-based curves.

The added value of any form of uncertainty estimation is evident in large differences between the random ranking and the uncertainty-based curves – points with large uncertainties do, on average, have larger errors. On the other side, the error-based ranking curve provides a useful lower bound for comparison about how good those uncertainty estimates are. It should be noted, however, that optimal uncertainties would not result in exact coincidence with this error-based ranking curve. This is due to instances where the model might make accurate predictions despite those predictions not being well supported by the training data, in which case the model should have high uncertainty despite its low error. These points would be removed early in any uncertainty-based curve but late in the error-based ranking curve resulting in the uncertainty-based curve being higher than the error-based ranking curve.

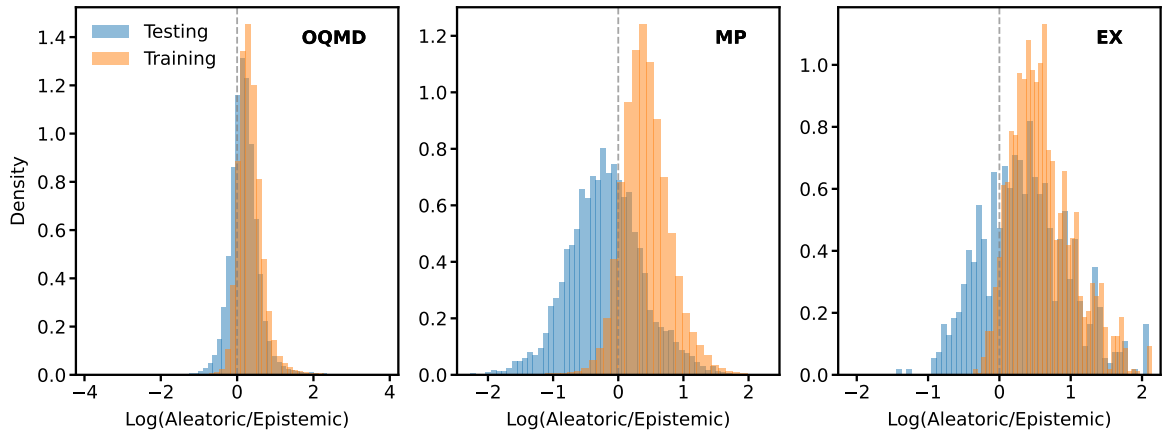


Figure 4.4: **Histograms of the log ratio of the aleatoric and epistemic contributions to the uncertainty for the OQMD, MP and EX data sets.** The figure shows histograms of the relative magnitudes of the aleatoric and epistemic contributions to the uncertainty. As expected, the epistemic uncertainty on the training set is generally smaller, resulting in the distributions being shifted towards the right. Looking at the testing sets for OQMD and EX, the aleatoric uncertainty tends to be larger, whilst for MP, the epistemic uncertainty tends to be larger. This result may suggest that the MP test set requires more extrapolation, whilst the OQMD and EX test sets are more interpolative.

To highlight the benefit of using the full framework for estimating the uncertainty, one that considers both aleatoric and epistemic uncertainties, we compare our *Deep Ensemble* against a purely epistemic alternative based on an ensemble of similar models that only estimate a predictive mean and are trained using an L1 loss function. We see that whilst the two ensembles have comparable errors over the whole data set, the full framework gives more reliable uncertainty estimates shown by the curve for the full framework (green curve with circular markers) decreasing more steeply than the curve for the epistemic-only alternative. Within the full framework, the relative magnitudes for the epistemic and aleatoric components vary depending on the data set being investigated and the extent to which the model is being tested in an interpolative regime (see Figure 4.4). This implies that the different forms of uncertainty do in fact capture different effects in the data and further supports using a full framework in cases where the overall error metrics across the entire data set are similar.

Uncertainty estimates produced using the *Deep Ensemble* approach [91] for regression tasks are typically not calibrated out-of-the-box (see Figure 4.5), depending on the application post-hoc calibration approaches can be applied to attempt to correct for this

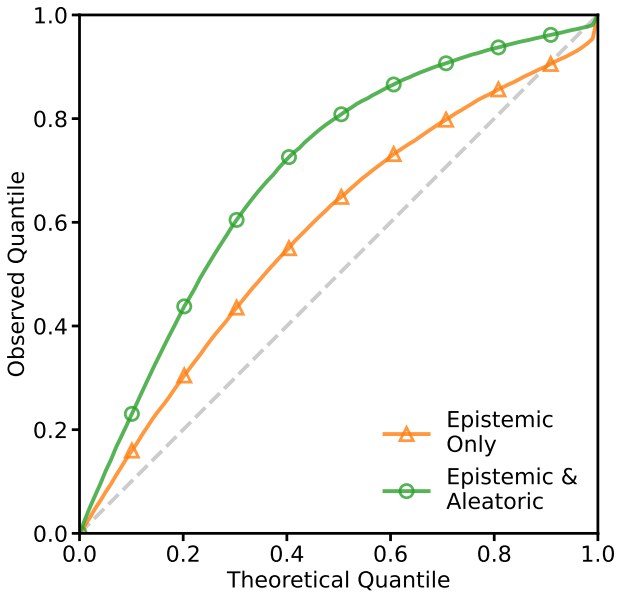


Figure 4.5: **Gaussian quantile-quantile plot on OQMD.** The figure shows how the quantiles of the distribution of uncertainty normalised residuals differ from the quantiles of a Gaussian distribution for the epistemic only ensemble and the *Deep Ensemble*, where both epistemic and aleatoric effects are accounted for. The epistemic only model has a distribution of normalised residuals closer to that of a Gaussian, but as shown in Figure 4.3 better calibration alone does not imply more useful uncertainty estimates.

[180–182]. Fortunately, within materials discovery, uncertainty estimates are typically made use of via active learning workflows [141, 183–185] that are often robust against miscalibration of this sort. In active learning workflows, an acquisition function is used to select candidates or batches of candidates to test. Most common acquisition functions [186] contain hyper-parameters that allow for the exploration-exploitation trade-off of the search process to be tuned. Often selecting such hyper-parameters is akin to adjusting the temperature of the uncertainty distribution, i.e. equivalent to a post-hoc calibration of the uncertainty. As selecting these hyper-parameters is non-trivial, even for perfectly calibrated uncertainties, the miscalibration of the model uncertainties is not prohibitive for materials discovery workflows provided the uncertainty estimates are well-behaved. Indeed, due to the difficulty of selecting good hyper-parameters, search strategies that acquire batches of points by selecting multiple candidates under a range of hyper-parameters have been proposed for effectively searching chemical spaces [187].

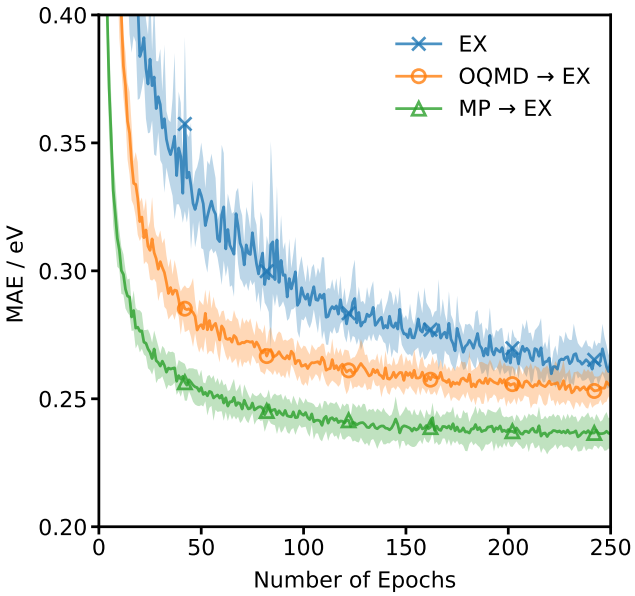


Figure 4.6: **Test set error during training on EX.** The figure shows how the MAE on the test set changes throughout the training of the model for different transfer learning scenarios. The curves show the average MAE over 10 independent randomly-initialised models with the shaded area corresponding to the standard deviation of the models at each point.

4.5 Transfer learning

For experimental data sets with smaller numbers of data points, shallow machine learning methods based on ensemble or kernel models have historically tended to perform comparably to, if not better than, deep neural network-based models. However, a strength of neural network-based models over such methods is that they are much more amenable to transfer learning [188]. Transfer learning focuses on using knowledge gained from one problem to achieve faster optimisation and lower error on another problem.

As a result of substantial efforts, data-sets derived via high-throughput *ab initio* workflows can be many times larger than their experimental cousins, making them ripe for transfer learning [189]. To investigate the extent to which transfer learning helps our model, we train three sets of models on the EX data set. The first set is directly trained on EX, the second is first trained on OQMD then fine-tuned on EX (OQMD → EX), and the third is trained on MP before fine-tuning on EX (MP → EX). Due to the similarity of the MP and EX tasks, to ensure any changes in performance observed are not artefacts of the experimental design, we remove all materials from the MP data set

Table 4.2: **Transfer learning benchmarking on EX.** The table shows the ensemble mean absolute error (MAE), and root mean squared error (RMSE) for the three transfer learning scenarios and baselines on 20% of the data that was randomly sampled and withheld as a test set.

	MAE / eV	RMSE / eV
RF + *Magpie* EX	0.277	0.460
SVM EX [66]		0.45
Roost EX	0.243	0.422
Roost OQMD → EX	0.240	0.404
Roost MP → EX	0.219	0.364

that are also found in the EX data set such that the two are independent. For all these experiments, the same 20% of EX was withheld as an independent test set.

A benefit of learning material descriptors is that similarity between the descriptors of different materials learnt for a given task should be relevant for other tasks, allowing non-cognate transfer learning. We see this in Figure 4.6, where transfer learning from OQMD leads to faster convergence and slightly lower errors on the EX data set than direct training despite the mismatch between the tasks. If the tasks are cognate, as is the case between MP and EX, the benefits of transfer learning are even more pronounced. Here, in addition to the benefits of having pre-trained the message passing sections of the model, the pre-trained weights of the output network give a strong inductive bias for fitting the materials descriptor-to-property mapping resulting in notably lower predictive errors (Table 4.2).

4.6 Ablation Study

The proposed reference model incorporates many different ideas to build upon previous work in the materials informatics and machine learning communities. Therefore, we have conducted an ablation study to show which parts of the model are most important for its enhanced performance. We examined the following design choices:

1. The use of an element embedding that captures correlations between elements versus a OneHot embedding of elements,

2. The use of a robust loss function based on the negative log-likelihood of a Laplace distribution (Equation 2.6) against the use of a standard L1 loss function,

3. Whether it is best to include the fractional element weights as node-level features, within the pooling operation, or in both places,

4. The use of our weighted soft-attention-based pooling throughout the architecture versus an alternative mean-based pooling mechanism,

5. The use of residual architectures for both the message passing and output neural networks, and

6. The impact on model performance from only using the message passing section of the model without an output network.

The combinations of design choices examined are shown in Table 4.3. We train 10 randomly-initialised models for each design choice. We look at both the statistics across these single models as well as their ensembled performance to allow for the significance of different choices to be understood. We repeat the ablation study for both the EX and OQMD data sets to understand how different design choices trade-off in the small and large data limits. The results are shown in Table 4.4.

The primary conclusion from the ablation study is that whilst the design choices made in the reference architecture described do lead to slight improvements in performance, all models from the ablation study, with the exception of Model 3, which does not include any information about the element weights, still significantly out-perform alternative models such as *ElemNet* or the Random Forest plus *Magpie* baseline on the OQMD data set. As such, it is apparent that it is the *Roost* framework's approach of reformulating the problem as a set regression task, and not specific architectural details, that is responsible for the observed improvements. This observation has been confirmed by the more recently introduced *CrabNet* [166] and *AtomSets* [190] architectures. *CrabNet* leverages a QKV-attention mechanism [174], whilst *AtomSets* use a set2set mechanism [191], within otherwise equivalent set regression frameworks, with both achieving similar performance to *Roost*.

Comparing the reference model and Model 1, we see that the choice of an element embedding that captures chemical correlation leads to improved model performance on the smaller EX data set but does not result in significant differences for the larger OQMD data set. This suggests that the models can learn to compensate for the lack of domain knowledge if sufficiently large amounts of data are available [192]. This result supports

Table 4.3: **Model design choices for ablation study.** The table shows the different model architectures based on the Roost framework studied in the ablation study.

	Reference	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Matscholar Element Embedding [173]	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
OneHot Element Embedding		✓									
Robust Loss Function	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Weights on Nodes	✓	✓	✓			✓	✓	✓	✓	✓	✓
Weights in Pooling	✓	✓	✓		✓			✓	✓	✓	✓
Soft Attention Pooling	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
Mean Pooling							✓				
Residuals when Message Passing	✓	✓	✓	✓	✓	✓	✓		✓		✓
Residuals in Output Network	✓	✓	✓	✓	✓	✓	✓	✓			
No Output Network											✓

Table 4.4: **Ablation study model performances.** The table shows the how the performance varies for different model architectures based on the Roost framework. Numbers in parentheses are used to show the standard error in the last significant figure. The lowest values in each row are highlighted in bold.

		Reference	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
EX	MAE	**0.264(3)**	0.279(2)	0.269(2)	0.329(2)	0.274(3)	**0.269(3)**	0.269(2)	0.284(3)	0.271(2)	0.292(3)	0.273(2)
	RMSE	**0.448(4)**	0.476(4)	**0.447(4)**	0.529(4)	0.476(6)	0.454(4)	0.459(5)	0.477(7)	0.465(5)	0.490(4)	0.470(5)
OQMD	MAE	**0.0297(2)**	**0.0295(1)**	0.0310(2)	0.1673(2)	0.0317(2)	0.0300(2)	0.0313(3)	0.0320(3)	0.0306(4)	0.0330(3)	0.0299(1)
	RMSE	0.0995(5)	0.0992(3)	0.0979(4)	0.2710(3)	0.1022(6)	0.0999(5)	**0.0959(6)**	0.1025(6)	0.1017(9)	0.1040(6)	0.0981(4)
EX ens	MAE	**0.243**	0.260	0.249	0.312	0.251	0.251	0.250	0.259	0.251	0.265	0.255
	RMSE	**0.422**	0.445	0.423	0.501	0.450	0.428	0.435	0.442	0.435	0.451	0.444
OQMD ens	MAE	**0.0241**	**0.0241**	0.0248	0.1644	0.0256	0.0243	0.0248	0.0253	0.0247	0.0259	0.0249
	RMSE	**0.0871**	0.0878	0.0882	0.2682	0.0898	0.0874	0.0875	0.0880	0.0877	0.0885	0.0911

our claim that end-to-end featurisation continuously improves as the model is exposed to more data.

The robust loss function (Equation 2.6) performs comparably on the EX data set to a more conventional L1 loss function (Model 2). Given that they offer similar average errors, the use of a robust loss function is highly compelling, even for single models, as it also provides an estimate of the aleatoric uncertainty with minimal computational overhead. Looking at the OQMD data set, the distinction between the two different loss functions is more apparent. The attenuation effect of the robust loss – that it can suppress the need to fit outliers – is observed in how the reference model achieves a lower MAE but a higher RMSE than Model 2. When proceeding to ensemble the single models, the validity of such a mechanism becomes apparent as both the MAE and RMSE are lower for the reference model in the ensembled case. This can be attributed to the cancellation of errors amongst predictions on the outlying (high squared-error) data points when ensembling.

Models 3, 4 and 5 from the ablation study look at how including the fractional element weights in different ways influences the model performance. As expected, we see that omitting the element weights entirely in Model 3 leads to an order of magnitude decrease in performance on the OQMD data set. However, whilst there is still a significant decrease in performance for the EX data set the error is still relatively comparable to that achieved by the reference model. This is due to a lack of diversity within different chemical sub-spaces in the EX data set. As a consequence, the EX data set is perhaps a less discriminative benchmark than OQMD despite the challenges associated with data scarcity. Including the weights on both the nodes and via the pooling operation gave the best results, marginally better than solely including the element weights on the nodes. Only including the weights via the pooling operation gave slightly worse results. This can be explained from the relative lack of information as the weighted soft-attention-based pooling (Equation 4.3) only includes the weights of the second element in the pairing as opposed to both elements if the weights are included as node features.

Whilst we primarily make use of a soft-attention-based pooling mechanism, alternative pooling mechanisms are feasible. In Model 6 we replace the pooling operations with a mean-pooling mechanism of the form

$$\boldsymbol{h}_i^{t+1} = \boldsymbol{h}_i^t + \frac{1}{J}\sum_{j=1}^{J} g^t(\boldsymbol{h}_i^t \oplus \boldsymbol{h}_j^t), \tag{4.5}$$

where $\boldsymbol{h}_i^t$ is the internal representation of the i^{th} element after t updates, $g^t(...)$ is a single-hidden-layer neural network for the $t+1^{th}$ update, $\oplus$ is the concatenation operation and the j index runs from 1 to J over the set $\boldsymbol{\nu}_i^t$ which contains the other elements in the material's composition. This model achieves a lower RMSE but has a higher MAE when considering individual models. However, when the models are ensembled the soft-attention-based pooling mechanism achieves both lower MAE and RMSE. This suggests that there is scope to tailor the reference model presented here for different applications by conducting neural architecture searches. However, this is an extremely computationally expensive process beyond the scope of this work.

Comparing Models 7, 8, and 9, we see that using residual architectures in both the message passing stages and the output network led to improved performance. Interestingly we see that replacing the output network with a single linear transformation (Model 10) does not significantly impact the performance of single models on the OQMD data set but does result in worse performance from the ensembled models. A potential explanation for this comes from considering the effective prior of the model without an output network. The addition of the output network changes the form of the prior hypothesis space of the model and as a result the distribution of distinct local basins of attraction [193]. The reduced benefits of model averaging within the ensemble for models without output networks could potentially be due to changes in the loss landscape meaning that such models are more likely to end up in correlated basins of attraction.

4.7 Discussion

In this chapter, we have proposed a novel and physically motivated machine learning framework for predicting material properties using just the material's compositions as input. Our key methodological insight is to represent the compositions of materials as dense weighted graphs. We show that this formulation significantly improves the sample efficiency of the model compared to other composition-based approaches.

Through modelling both the uncertainty in the physical process and our modelling processes, the model produces useful estimates of its uncertainty. We demonstrate this by showing that as we restrict, according to our uncertainty estimates, the confidence percentile under consideration, we observe steady decreases in the average error on the

test set. Such behaviour is important if we wish to use our model to drive an activate learning cycle.

We show that the representations learnt by the model are transferable, allowing us to leverage data-abundant databases, such as those obtained by high-throughput *ab initio* workflows, to improve model performance when investigating smaller experimental data sets. The ability of the model to transfer its learnt descriptors suggests that self-supervised learning may be a viable avenue to bolster model performance [194, 195].

We have conducted an extensive ablation study to examine the model. We show that it is the reformulation of the problem as a set regression task, such that both the descriptor and the fit are learnt simultaneously, that results in the improved performance, not the specific details of the message passing architecture used. We believe that recasting more problems in material science into this language of set regression, using the same message passing framework as our *Roost* approach or other frameworks [85, 86], provides an exciting new area for the development of novel machine learning methods. In Chapter 5 we look at how we can apply a set regression architecture, based in part on the *Roost* framework, to the task of predicting the major products of inorganic synthesis reactions.

Whilst *Roost* achieves highly compelling accuracies compared to other composition-based approaches, several questions remain about whether this improved performance is sufficient for materials discovery [119]. As highlighted in Ref. [119] accurate predictions of the formation energy, i.e. models with errors to DFT on par with the magnitude of the errors reported from DFT to experiment, are not as useful as DFT for identifying stable materials because DFT's errors are often systematic, whilst machine learning models have random error distributions. When constructing convex hulls, the systematic errors of DFT often cancel, resulting in predicted hulls that closely match reality, whereas the random errors of composition-based machine learning models tend to compound, adversely impacting the quality of stability predictions. *CGCNN* [82], a structure-based model, is shown not to suffer from this issue to the same extent as *Roost*. In part, this is because structure-based models are more accurate for predicting formation energies, leaving less scope for random errors to affect stability results, but also there appears to be some degree of beneficial error cancellation not seen for composition-based models. This observation informs the work in Chapter 6 where we look at problem formulations and models that maintain the critical requirement of having computably enumerable inputs, such that we can conduct combinatorial screening campaigns, whilst also seeking

to include some degree of structural information, which we believe may lead to increased accuracy and facilitate the beneficial cancellation of errors when calculating stabilities.

Chapter 5

Predicting the Products of Inorganic Reactions

This chapter is based on Shreshth A Malik, Rhys E A Goodall, and Alpha A Lee. Predicting the outcomes of material syntheses with deep learning. *Chemistry of Materials*, 33(2):616–624, 2021. Reproduced with permission from the American Chemical Society.

In the previous chapter, we introduced a new framework for materials property prediction that only requires material compositions as input. Using such a model, it is possible to generate large sets of promising materials candidates worthy of experimental investigation. In recent years many virtual screening efforts have done precisely this. However, due to the challenges associated with material synthesis, only a tiny proportion of the candidates suggested by virtual screening campaigns have been produced in the laboratory and experimentally validated [196, 168].

The most common experimental route for producing inorganic materials is solid-state synthesis [197], this typically involves calcining a mixture of solid reactants. Despite the conceptual simplicity of producing materials by heating and cooling the correct elements in the proper ratio, the reality of inorganic reactions is far more complicated. Reactions occur through solid-state diffusion of ions, thus reactants are often milled and mixed to improve the reaction kinetics. The interplay between thermodynamically driven energy minimisation and kinetic factors results in products that are often difficult to predict – the field lacks the well-understood reaction mechanisms in organic chemistry. Particular procedures or reactants result in the formation of crystals with unique morphologies and

compositions as they form through metastable states. Thus researchers frequently rely on a trial-and-error process of synthesis and evaluation, this means that the development of viable, reliable synthesis pathways can take months, if not years. Consequently, materials synthesis is a principal bottleneck in the discovery pipeline. The development of tools to propose and evaluate the most promising synthetic pathways is, therefore, a significant challenge facing materials discovery [198–201].

A recently released inorganic synthesis data set has now enabled the possibility of using a data-driven approach to predict the outcomes of inorganic synthesis [34]. To our knowledge, there exists only one study which explores this opportunity [202], which addresses the inverse problem of retrosynthesis – predicting precursors that can react to yield a target product. However, save for large scale experimental validation, the accuracy of retrosynthesis models cannot be quantitatively benchmarked because there are almost infinitely many ways to synthesise a material, and reactions reported in the literature are not necessarily the best synthetic routes. Take, for example, the simple anonymous composition ABC, plausible synthesis routes could be from pure elemental species, i.e. A + B + C $\rightarrow$ ABC, binary species, i.e. A_2C + B_2C $\rightarrow$ 2 ABC, a mixture of binary compounds and elemental species, i.e. AB + C $\rightarrow$ ABC, or even via reactions containing foreign elements not present in the target product i.e. AB_2D_2 + AC_2 $\rightarrow$ 2 ABC + D_2. There are a host of reasons why different pathways may not be viable, for example, when attempting a solid-state synthesis, the thermal activation barrier for a particular set of precursors may require temperatures above the thermal decomposition temperature of the product ABC. Alternatively, in synthesis procedures based on solution processing, the solubility of the precursors in the solvent will control the viability of different pathways.

In this chapter, we explore forward reaction prediction models for inorganic synthesis. This work is a building block towards an inorganic retrosynthesis planner that can interface directly with computational tools to help close the materials Design-Make-Test cycle. As there are multiple pathways through which a material can be synthesised, and the pathways reported in the literature are not necessarily optimal, the construction of reliable retrosynthesis models is dependent on the ability to accurately predict forward reactions such that we can check for self-consistency.

5.1 Solid-State Syntheses Data Set

We can define a generalised solid-state synthesis procedure for a target material as a sequence of processes (*actions*) performed on a set of starting materials (*precursors*). A recent study has extracted detailed information for over 19,000 such synthesis procedures from academic literature [34]. Each synthesis procedure has four relevant fields for this problem,

1. **Target:** The stoichiometric formula of the target material for the synthesis.
2. **Precursors:** Defined as starting materials that share at least one element with the target material, excluding "abundant" compounds, i.e. those found in the air.
3. **Processing Actions:** The sequence of synthesis actions performed on the precursors, including process conditions where available.
4. **Balanced Chemical Equation:** Balanced chemical equations for the formation of the target compound indicate the relative amounts of precursors in the reaction mixture.

To our knowledge, this data set is the most extensive and highest-fidelity available for inorganic syntheses. The chemical information (precursors and targets) has been extracted with high accuracy (93%) [34]. To obtain balanced equations, the original authors assume a set of "abundant" compounds that can be released or absorbed during solid-state synthesis (O_2, CO_2, N_2, etc.) could be present in all the reactions. These "abundant" compounds are not counted as precursors.

However, some notable limitations must be considered when utilising this data set for modelling. Firstly, some procedures have actions and conditions with missing or incorrect data. Over 10% of reactions have no or only one recorded action. This is primarily because standard procedures are often referenced in the source literature, e.g. "The samples were prepared following John Smith, et al. (2017)", consequently the text does not contain explicitly stated synthesis details, and therefore the named entity recognition model used for data mining has no actions to classify. Secondly, only 46% of actions have associated conditions, i.e. solvents, timings, etc. The accuracy of the complete procedures (where all actions, conditions, and chemical information are correctly extracted) is 51% [34]. As a result, due to difficulties in representing action data of mixed fidelity and

completeness, we only make use of coarse-grained high-level action descriptions in this work. These are reported with an F1 score of 0.90 [34]. Further, there is no structural information for the products, and so a purely composition-based approach must be used. This lack of structural information about the crystal structures and micro-structures of the precursors and products limits the physical analysis of the synthesis reactions, for example, the role of nucleation in product selection. Finally, data set balance and bias towards popular products and syntheses techniques is also a potential issue to be noted.

Nevertheless, this data set provides a wealth of structured information on successful solid-state syntheses, and machine learning models offer a unique way to assimilate empirical patterns from data. To apply machine learning in this instance, we consider that the target material is the output of an abstract but learnable function of the precursors and processing procedure, and then attempt to fit a model for that function.

5.2 Representation of Inorganic Reactions

Machine learning models generally operate on fixed-length inputs. However, our understanding of inorganic reactions is typically expressed using variable-sized sets of precursors and sequences of processing actions. Bridging this discrepancy is the central challenge that needs to be addressed when constructing machine learning algorithms for inorganic reaction prediction. Here we consider how to obtain fixed-length representations for the precursors and the processing sequence.

5.2.1 Representation of Precursors

A naive approach would consider representing materials as sparse vectors with components proportional to the fractional amounts of its constituent elements (its stoichiometry). However, such a representation fails to capture critical correlations between different elements. Here we once again make use of the general-purpose *Magpie* embeddings [65] to represent the precursors. In principle, we could have made use of the *Roost* framework introduced in Chapter 4 to learn task-specific embeddings, however, doing so would have dramatically increased the complexity of the model architecture presented. Consequently, it was decided to first investigate the use of hand-curated descriptors before investigating whether using *Roost* to learn the composition-to-descriptor map may improve performance in future work.

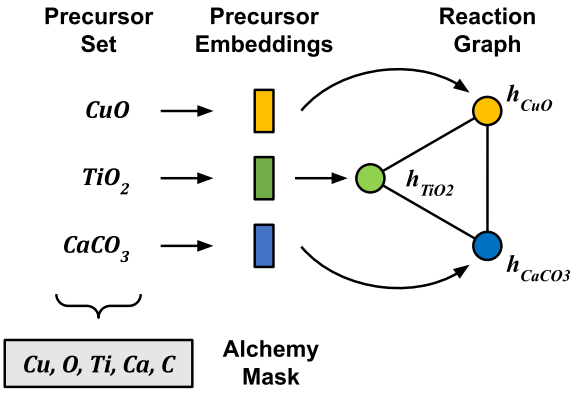


Figure 5.1: **Representation of precursors and building the Reaction Graph.** *Magpie* embeddings [65] are used to represent the precursors as fixed-length vectors. These are used as features for nodes on a dense graph. An *alchemy mask*, used to constrain the predictions of the model, is constructed from the set of elements in the precursors.

5.2.2 Representation of Processing Actions

To allow reactions with different numbers of processing actions to be combined into mini-batches for training, we need to obtain a fixed-length representation for the action sequences. Here we utilise a single-layer Long Short-Term Memory unit (LSTM) [76] trained as part of an autoencoder architecture to encode the variable-length action sequences.

The autoencoder architecture is composed of an encoder-decoder pair trained to compress and then reconstruct their inputs from a learnt low dimensional representation. Once trained, the encoder produces low-dimensional fixed-length representations of the action sequences that are used in other parts of our model.

Currently, we only consider the seven different high-level classes of action presented in the data set. We use a one-hot encoding to represent this vocabulary of actions for our LSTM model. Previous works have made use of vocabularies with more action types than the seven considered here. In Ref. [202] a more extensive vocabulary of 50 actions was curated by clustering the word embeddings of synthesis actions based on their cosine similarity and then assigning each action a one-hot encoding based on the nearest cluster. We explored this idea but saw that it did not lead to a significant change in performance.

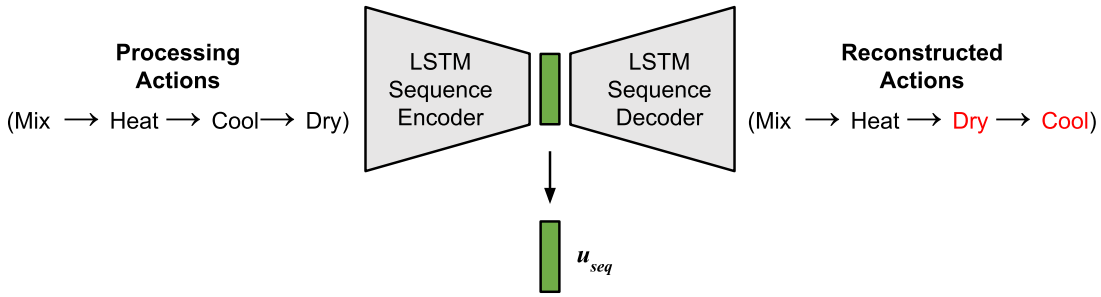


Figure 5.2: **Encoding variable-length action sequences.** The processing action sequences are encoded into a fixed-length representation using an LSTM encoder. The encoder is trained using an autoencoder architecture where the variable-length sequence is first encoded to a fixed-length representation before a second LSTM model is used as a decoder to reconstruct an action sequence. The encoder and decoder are trained simultaneously to minimise a reconstruction loss based on the cross-entropy between the action labels. This loss is minimised when both the input action sequence and the reconstructed action sequence are the same. In the schematic, we show a case where the autoencoder fails to correctly reconstruct the sequence with the final two actions being swapped.

5.3 Reaction Graph Model

In solid-state syntheses, precursors are inherently "mixed" together to form a product. Given the precursor embeddings, $\boldsymbol{h}$, and processing sequence embedding, $\boldsymbol{u}_{seq}$, we make use of an architecture for set regression based on the *Roost* model [2] that mimics this behaviour. The model treats this process as a series of message-passing stages in which precursors appear as nodes on a dense weighted graph. The principal difference between this model and the *Roost* architecture is that we also concatenate an action embedding into each of the update steps (see Figure 5.3**.a**). Concretely, the equivalent equations to Equation 4.2 and Equation 4.4 for the reaction graph update process are

$$e_{ij}^{t,m} = f^{t,m}(\boldsymbol{h}_i^t \oplus \boldsymbol{h}_j^t \oplus \boldsymbol{u_{seq}}), \tag{5.1}$$

where $f^{t,m}(...)$ is a single-hidden-layer neural network for m^{th} head of the $t+1^{th}$ update, the j index runs over all the precursors in the reaction excluding the i^{th} precursors, and $\oplus$ is the concatenation operation, and

$$\boldsymbol{h}_i^{t+1} = \boldsymbol{h}_i^t + \sum_{m,j} a_{ij}^{t,m} g^{t,m}(\boldsymbol{h}_i^t \oplus \boldsymbol{h}_j^t \oplus \boldsymbol{u}_{seq}), \tag{5.2}$$

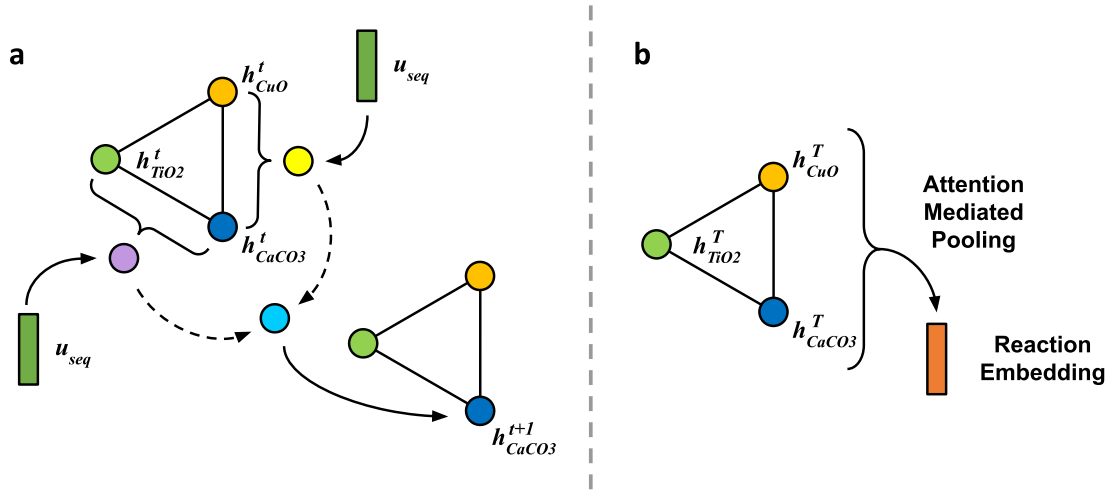


Figure 5.3: **Update and pooling operations for the Reaction Graph model. a** - The reaction graph goes through a series of message-passing operations between nodes. The action embedding is used as a global state in all the message-passing steps. The panel shows a schematic illustration of the t^{th} update for $CaCO_3$ in the reaction graph. **b** - After T update iterations the final node representations are then pooled, weighted by attention coefficients, to obtain a fixed-length representation of the reaction that we refer to as a reaction embedding.

where $g^{t,m}(...)$ is a single-hidden-layer neural network for m^{th} head of the $t+1^{th}$ update and $a_{ij}^{t,m}$ are the normalised attention coefficients obtained from the softmax of $e_{ij}^{t,m}$ over the j index. The indices m and t run to M and T and index the number of heads and number of update steps respectively. M and T are hyper-parameters of the model that must be selected before training.

A fixed-length representation for the reaction is derived from this graph using an attention-mediated graph-pooling operation (see Figure 5.3**.b**). Pooling to a fixed-length representation allows for generalisation to arbitrary numbers of precursors. We refer to this fixed-length representation as the reaction embedding.

The model is broken apart into a two-stage architecture. Firstly the reaction embedding is passed through a feed-forward neural network that is trained to perform multi-label element classification for elements in the target product of the reaction (Figure 5.4**.a**). This first stage of the model is trained in an end-to-end manner with the message-passing stages. In materials synthesis, the products cannot contain elements that are not present in the precursors. We add this inductive bias in the form of a binary *alchemy mask* on the output of the network. Oxygen is included in all masks irrespective of its presence in the precursors as it is assumed to be abundant. For inference, a threshold of 0.5 is

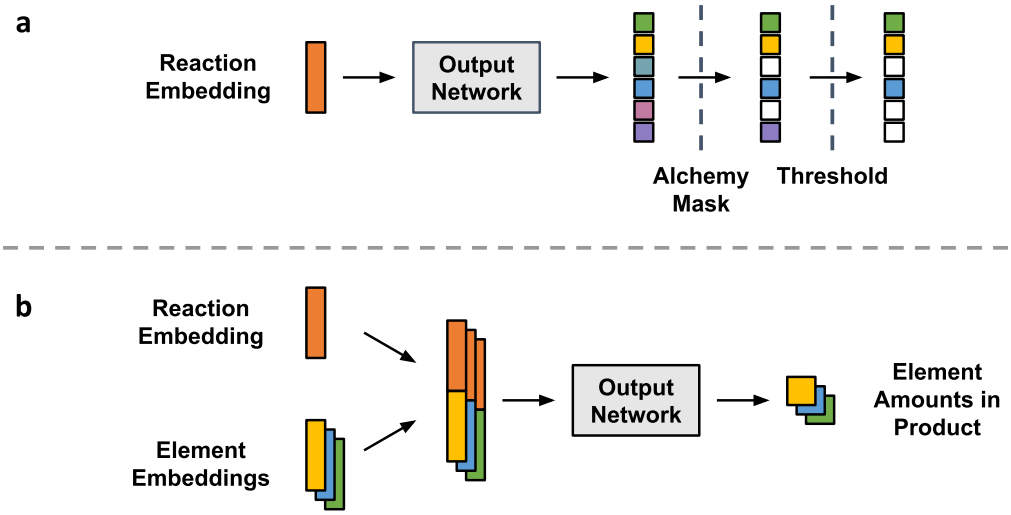


Figure 5.4: **Element Prediction and Regression Schematics.** **a** - The learned reaction embedding is used to predict which elements are present in the major product. The *alchemy mask* ensures that only elements present in the precursors can be predicted as present in the product. **b** - The amount of each predicted element is regressed with its fractional stoichiometry from the reaction embedding. *Matscholar* element embeddings [173] are concatenated with the reaction embedding to query the element in question.

applied after a sigmoidal non-linearity on the output to classify whether a given element is present in the product. If the stoichiometry were directly predicted, an arbitrary threshold stoichiometry would have to be used to select which precursor elements are actually present in the product. The two-stage approach used here is therefore necessary to differentiate between products with and without trace elements.

Once the elements in the product have been predicted, the second stage of the model is to predict the relative amounts of each element in the product. Here, we use *Matscholar* element embeddings, $\boldsymbol{\epsilon}_i$, to represent and differentiate between constituent elements [173]. These are concatenated with the reaction embedding, $\boldsymbol{r}$, to provide a contextualised query and fed through another feed-forward neural network, h, to predict the relative amount of element i,

$$s_i = h\left(\boldsymbol{\epsilon}_i \oplus \boldsymbol{r}\right). \tag{5.3}$$

These are then normalised using a softmax function to arrive at a fractional stoichiometry

$$w_i = \frac{\exp(s_i)}{\sum_{j=1} \exp(s_j)}, \tag{5.4}$$

where w_i is the normalised stoichiometry of element i in the product and the index j runs over all the elements in the input domain. A schematic for this stage is shown in Figure 5.4**.b**.

We again look at ensembles of 5 models to increase accuracy and allow estimation of the epistemic uncertainty. For the first part of the model, comprising the message passing stages and the element classifier, we allow each ensemble member to produce its own reaction embedding. In the second part of the model, the element regressor, we take the reaction embeddings from the first ensemble and average them. This averaged embedding is used as the input to all the members of the second ensemble. This design choice was made to simplify the nearest neighbour searching by having a single embedding for each reaction instead of 5 different embeddings. Averaging in this way could potentially degrade performance as different regions, and directions in the learnt embedding spaces of the models in the ensemble will encode different concepts. Consequently, averaging could therefore lead to destructive interference between models. In this work, preliminary results did not show any significant differences in performance, therefore, we opted to use the averaged reaction embedding for the second half of the overall model. We refer to these ensembled models as the reaction graph model.

5.4 Evaluation of Model Performance

The reaction graph model was tested on solid-state reactions with up to ten precursors. The model predicted the elements in the product with a subset accuracy of 0.940 – the subset accuracy indicates the percentage of samples that have all their labels classified correctly in a multi-label classification. This result can be compared to a Null baseline accuracy of 0.338 which would be achieved if all elements in the precursors were assumed to be present in the product. There is evidently a bias towards positive element labelling, 83.9% of elements in the precursors are also present in the product. Another metric of interest is the F1 score. The average F1 score achieved was 0.991, where the score for each element is weighted by the number of true occurrences to account for the elemental imbalance.

Results for stoichiometry prediction can be seen in Figure 5.5. To access the quality of our stoichiometry predictions, we examine the mean L1, L2, and element-movers (EleMD) distances between the predicted and ground truth fractional stoichiometry vectors across

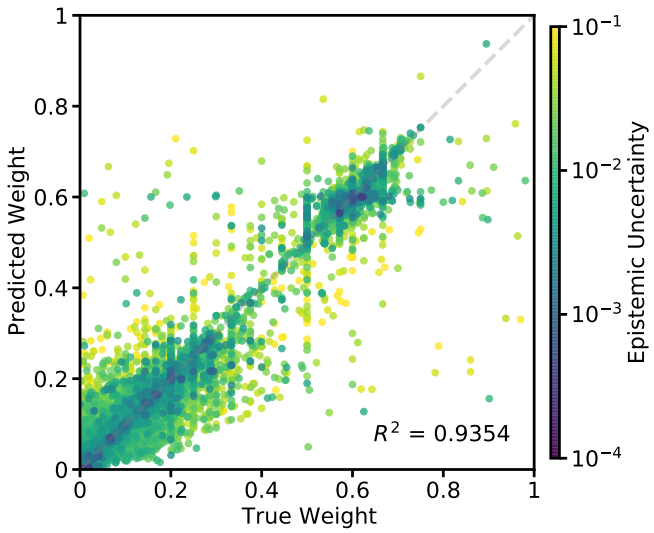


Figure 5.5: **Parity plot for product stoichiometry prediction.** The fractional amount of each element in the target product is plotted against its predicted value. The data points are coloured according to their epistemic uncertainty, estimated from the variation in predictions from the ensemble. We observe that there is a clear banding structure in the true weights that arises out of the fact that material science is often clustered around integer stoichiometries.

the held-out test set. The results for the proposed model (with and without the action sequences) and a Null baseline are summarised in Table 5.1.

The L1 distance is the sum of the absolute deviations between the predicted fractions of each element and the true fractions, and the L2 distance is the square root of the sum of the squared deviations. The EleMD is a recently developed metric for assessing compositional similarity through calculating the minimal amount of work taken to transform one distribution of elements to another along the modified Pettifor scale

Table 5.1: **Performance of Reaction Graph Model.** Results for the predictions of the proposed reaction graph model (including and excluding action embeddings) on reactions with up to ten precursors.

	Subset Accuracy	F1 Score	Mean L1	Mean L2	Mean EleMD	R^2
Null Baseline	0.338	0.986	0.262	0.148	4.98	0.840
Reaction Graph (excl. Actions)	0.933	0.990	0.136	0.079	2.44	0.930
Reaction Graph (inc. Actions)	0.940	0.991	0.126	0.073	2.29	0.935

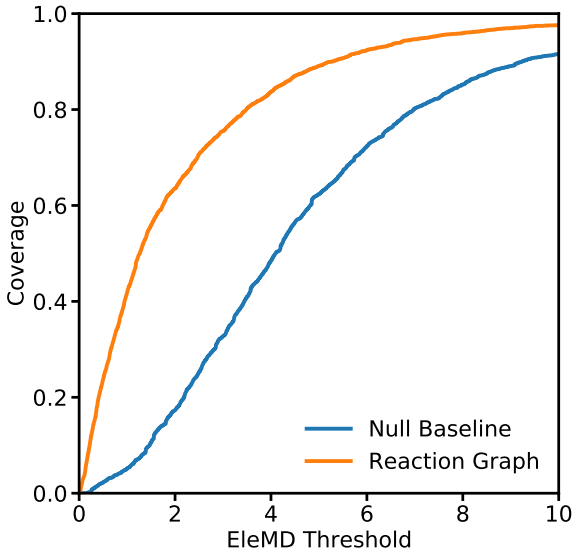


Figure 5.6: **Coverage of Reaction Graph predictions.** The cumulative proportion of stoichiometries correctly predicted to within a given EleMD threshold is plotted against the threshold. A Null baseline result is shown for comparison.

[203, 204]. The Null baseline assumes all elements in the precursors are present in the product and are present in their average relative amounts derived from the training data.

To rationalise the results in Table 5.1 we need to understand what is an acceptable level of error for the model to be practically useful. The central question is whether the model achieves a level of accuracy that enables us to distinguish between similar materials. To fix ideas, we consider iron oxides as a toy system: The L1 and L2 distances between Fe_2O_3 and FeO are 0.2 and 0.14, between Fe_2O_3 and Fe_3O_4 are 0.06 and 0.04, and between FeO and Fe_3O_4 are 0.14 and 0.10 respectively. Accordingly, we suggest a rule of thumb of $O(10^{-1})$ as being the point below which these metrics might start to become "chemically discriminative". For the EleMD, it is more challenging to motivate a threshold based on toy examples as the magnitude of the difference in the EleMD depends on the elements being considered, not simply the relative weightings. The discretisation of the modified Pettifor scale [203] further complicates this as elements differing by the same amount under the scale can have very different levels of chemical dissimilarity. Despite these limitations, the EleMD remains a valuable metric because it captures some notion of chemical similarity and avoids the double-counting we see in the L1 and L2 distances. For this chapter, we will consider an EleMD of below 2 as

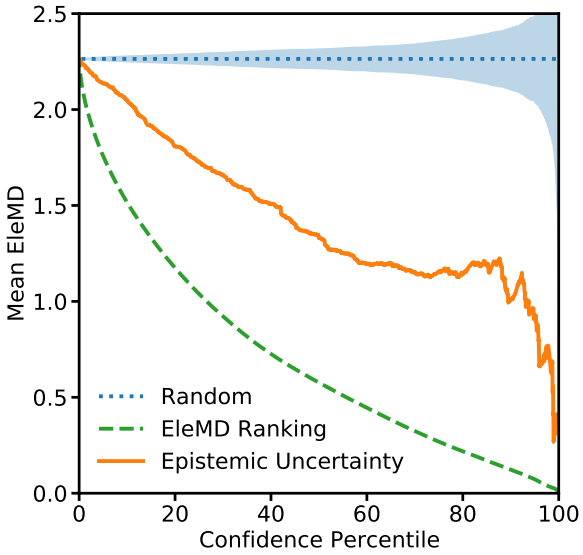


Figure 5.7: **Epistemic uncertainty helps identify unreliable model predictions.** Confidence-error plot quantifying the decrease in Mean EleMD between the predicted stoichiometry and the product stoichiometry as the least confident predictions are removed. Additional reference lines are shown for ranking and removal based on the size of the EleMD. The blue shaded region gives the standard deviation of the blue curve computed over 500 random orderings. The monotonic decrease in the mean EleMD shows that the uncertainty estimates produced are well-behaved, with points with larger errors on average having larger uncertainties.

being "chemically discriminative". Figure 5.6 highlights that approximately 60% of the reactions are predicted within this threshold EleMD of 2 compared to less than 20% for the Null baseline.

The results in Table 5.1 show that our model achieves accuracy around these levels but not low enough to claim that we are "chemically discriminative". The estimation of uncertainty allows our model to be used more confidently than these average metrics might suggest, as the unreliable predictions that inflate the average metric can be flagged up by looking at the uncertainty of the model (Figure 5.7 illustrates this idea). Here we use the square root of the total variance, defined as the trace of the covariance matrix, for the uncertainty of the model. The variance in the composition itself is always equal to 0 due to the constraint that the sum of the fractional element weights is equal to 1 causing the off-diagonal terms of the covariance matrix to exactly cancel the total variance.

5.5 Explaining Predictions via Reaction Similarity

The predictions of the model can be explained by considering the reaction embeddings produced by the model – similar feature vectors will result in similar predictions. Therefore, we can understand why specific predictions are made by assessing the cosine similarity between the reaction embedding of a given set of precursors and those of reactions in the training data. As chemists using the model are likely only to be interested in a single or few similar reactions for a given project, being able to infer which reactions in the data set are relied upon by the model to make a prediction is critical to its utility. Attribution enables users to assess the fidelity of the source literature from which the data was drawn and therefore make more informed judgements on the model's predictions. We illustrate this by looking at reactions that are not in the training set and showing that we can identify similar reactions in the training set to explain the model's predictions.

Firstly we investigate the formation of $NaNiO_2$, often used as a cathode material, from a peroxide, [205]

$$0.5\ Na_2O_2 + NiO \rightarrow\ NaNiO_2. \tag{5.5}$$

The product composition was predicted accurately within an L1 error of 0.015 and L2 error of 0.0096. To analyse why the prediction was accurate, we assess the similarity of the learned reaction embedding derived from the precursors and processing action sequence to other reactions in the training set:

Nearest Neighbors to (5.5)	**Similarity**
$0.5\ Na_2O_2 + 0.333\ Fe_3O_4 \rightarrow NaFeO_2 + 0.167\ O_2$	99.0%
$0.5\ Li_2O_2 + 0.333\ Co_3O_4 \rightarrow LiCoO_2 + 0.167\ O_2$	98.9%

Here we see that the embedding is very similar to other AMO_2 syntheses from peroxides where A is an alkali metal and M is a transition metal. Despite differing cations appearing in the precursors, the model has learnt the general form of such reactions and the expected ratios of elements in the product. The model triangulates a prediction from similar observed reactions and the elemental context.

The role of air can be important in solid-state reactions as it can result in changes in the oxidation state of elements. In the synthesis of a columbite, studied for its optical, magnetic and dielectric properties [206],

$$Nb_2O_3 + MnO_2 + 0.5\ O_2 \rightarrow MnNb_2O_6 \tag{5.6}$$

the precursors are oxidised by oxygen in the atmosphere. This product was accurately predicted to an L1 error of 0.036 and L2 error of 0.022. The reactions in the data set are automatically balanced for common compounds released and absorbed (e.g. oxygen, carbon dioxide, water) given only the precursors and product [34]. Similarly, our model is only given information about the precursors and not the amounts of oxygen or other common compounds involved in the reaction. Nevertheless, the model accurately infers the redox chemistry that results in the absorption of oxygen and the change in the oxidation state of the elements in the product.

We can again understand why this prediction was made by analysing the reaction embedding. Looking at the most similar reactions in the training set, we find additional syntheses that involve redox chemistry:

Nearest Neighbors to (5.6)	**Similarity**
$0.393\ Y_2O_3 + 0.108\ Ta_2O_3 + 0.108\ O_2 \rightarrow Y_{0.785}Ta_{0.215}O_{1.715}$	98.6%
$0.35\ Y_2O_3 + 0.15\ Ta_2O_3 + 0.15\ O_2 \rightarrow Y_{0.7}Ta_{0.3}O_{1.8}$	98.6%

The reaction embeddings also serve as a means to cross-validate the data set and analyse why certain predictions are incorrect through comparison with similar reactions in the training set. As many of these failure modes can be identified via estimation of the epistemic uncertainty, which can indicate when there is a lack of support from the data set, we are primarily interested in explaining the worst predictions – those which have large error but low uncertainty. Reaction (5.7) is an example of one such case.

$$1.5\ Na_2CO_3 + 0.333\ Co_3O_4 \rightarrow Na_3CoO_2 + 1.5\ CO_2 + 0.417\ O_2 \tag{5.7}$$

Target	Na: 0.5000	Co: 0.1667	O: 0.3333
Prediction	Na: 0.1985	Co: 0.2692	O: 0.5323

In the original paper [207], Na_3CoO_2 is the nominal product but the paper actually reports the formation of different phases of Na_xCoO_2 in their experiment. As such, the model prediction is not unreasonable. The model reached that prediction by inferring from similar reactions observed in the training data (but not in Zhou et al. [207]):

Nearest Neighbors to (5.7)	**Similarity**
0.375 Na_2CO_3 + 0.333 Co_3O_4 + 0.146 O_2 → $Na_{0.75}CoO_2$ + 0.375 CO_2	99.9%
0.5 Na_2CO_3 + 0.333 Co_3O_4 + 0.083 O_2 → $NaCoO_2$ + 0.5 CO_2	99.9%

As a second example, we explore a different subspace of materials. Perovskite manganites with the general formula $RE_{1-x}A_xMnO_3$ (where RE is a trivalent rare earth and A is a divalent alkaline earth) have attracted interest since the discovery of large negative magneto-resistance in such structures [208]. When faced with the *supposed* reaction,

$$0.05\ MnO_2 + 0.15\ CaCO_3 + 0.3\ La_2O_3 + 0.01\ BaCO_3 \rightarrow \\ La_{0.6}Ca_{0.15} + 0.05\ Ba_{0.2}MnO_3 + 0.16\ CO_2 + 0.505\ O_2, \quad (5.8)$$

the model performs badly. For this reaction the reaction embedding was observed to be most similar to manganite syntheses in the training set:

Nearest Neighbors to (5.8)	**Similarity**
0.083 Pr_6O_{11} + 0.47 $CaCO_3$ + 0.03 $BaCO_3$ + MnO_2 → $Pr_{0.5}Ca_{0.47}Ba_{0.03}MnO_3$ + 0.5 CO_2 + 0.208 O_2	97.6%
0.05 $SrCO_3$ + 0.15 $CaCO_3$ + 0.5 Mn_2O_3 + 0.05 $BaCO_3$ + 0.375 La_2O_3 + 0.063 O_2 → $La_{0.75}Ca_{0.15}Sr_{0.05}Ba_{0.05}MnO_3$ + 0.25 CO_2	97.5%

Accordingly, the model was interpolating this and predicting a manganite for the reaction resulting in a large error given the nominal products were $La_{0.6}Ca_{0.15}$ and $Ba_{0.2}MnO_3$ in the ratio 1:0.05. Flagging this discrepancy between the high level of similarity and the high error, we examined the source for the reaction and established that actual reaction reported in the original manuscript did in fact result in a manganite [209],

$$MnO_2 + 0.2\ CaCO_3 + 0.3\ La_2O_3 + 0.2\ BaCO_3 \rightarrow \\ La_{0.6}Ca_{0.15\cdot\ 0.05}Ba_{0.2}MnO_3 + 0.4\ CO_2 + 0.55\ O_2. \quad (5.9)$$

Given the correct reaction products, the model's actual L1 and L2 errors for this example were 0.017 and 0.0082 respectively. The product composition was incorrectly parsed and

extracted as a mixture of $La_{0.6}Ca_{0.15}$ and $Ba_{0.2}MnO_3$ due to the abnormal dot notation in the calcium stoichiometry that was missed by the correction rules in the original data extraction workflow that are used to fix most common errors. This example shows how the idea of a reaction embedding can be helpful as a way to cross-validate reactions extracted from literature, thus improving data set quality with minimal labour.

5.6 Identified Failure Modes

5.6.1 Issues in Element Prediction

In general, through the inclusion of the *alchemy mask*, which prevents chemical inconsistencies, the prediction of which elements are present in the product is very accurate. The small number of erroneous predictions made by the model can typically be attributed to two areas:

	Carbon False Negatives
Test Set Reaction [210]	0.5 Cs_2CO_3 + 0.5 Li_2CO_3 → $LiCsCO_3$
Target Elements	Li, Cs, O, C
Prediction	Li, Cs, O
Analysis	This synthesis involves the formation of a mixed alkali carbonate. Products containing carbon are rare in the data set as most reactions with carbonates involve decomposition reactions. The model has learnt this bias, predicting that carbon is not present in the product.

	Data Processing of Dopants
Test Set Reaction [211]	2 $SrCO_3$ + SiO_2 → Sr_2SiO_4 + 2 CO_2; with additives Eu via Eu_2O_3
Target Elements	Sr, Si, O
Prediction	Sr, Si, O, Eu
Analysis	Some products in the data set have additives and dopants which do not appear explicitly in the target stoichiometry. However, the relevant precursors are still present in the data (Eu_2O_3 in this case). In these cases, the model accurately predicts the presence of these elements in the product, but the additive is not accounted for in the product stoichiometry data. This is relatively common for materials with an undefined amount of doping.

5.6.2 Issues in Stoichiometry Prediction

Analysing erroneous stoichiometry predictions reveals that the failure cases with the largest errors occur when the precursors include elemental materials, for example:

	Pure Precursors
Test Set Reaction [212]	8 Ni + 86 Al + 6 La $\rightarrow$ $Al_{86}Ni_8La_6$
Target	Ni: 0.0800 La: 0.0600 Al: 0.8600
Prediction	Ni: 0.6859 La: 0.1497 Al: 0.1642
Analysis	This reaction is for the formation of an amorphous aluminium alloy anode through arc melting. *Magpie* features are not designed to work with pure elemental materials. This prevents the model from being able to learn reasonable mappings involving pure precursors and thus leads to more erroneous predictions.

5.6.3 Issues from Lack of Similar Examples

The lack of similar reactions is observed to lead to higher uncertainty and more erroneous results. An example of this is the formation of the ceramic phase Ti_2AlC in a matrix composite from powdered Ti, Al and stearic acid ($C_{18}H_{36}O_2$) [213]. As discussed in previous sections, the model cannot handle pure metals nor carbon products as effectively. Further, the model has not had exposure to similar reactions involving weak acids in the training data. A similarity search returns reactions for the formation of metal oxides. These factors compound and result in a highly erroneous prediction.

5.7 Benchmarking and Ablation Study

To place the accuracy of our model into context, we benchmark our reaction graph model against a baseline approach and conduct an ablation study on its key features.

The baseline model we use takes the *Magpie* precursor embeddings and concatenates them to be used as the input to a neural network with the same architecture as the output network of the reaction graph model. As with the reaction graph model, we use an ensemble of 5 individual models for this baseline.

While the reaction graph model can handle variable numbers of precursors, this *Magpie* baseline approach is inherently restricted to a particular number of precursors.

Table 5.2: **Ablation performance for reactions with up to three precursors.** The Null baseline result assumes all elements in the precursors are present in the product and assumes that elements are present in their average amounts from the training data. Providing inductive bias to the model in the form of the alchemy mask improves accuracies across the board. Removing the action context does not have a notable effect on model accuracy for reactions with up to three precursors.

Reaction Representation	Alchemy Mask	Subset Accuracy	Weighted F1 Score	Mean L1	Mean L2	Mean EleMD	R^2
Magpie Baseline		0.580	0.882				
Reaction Graph (excl. Actions)		0.773	0.950				
Reaction Graph (inc. Actions)		0.770	0.948				
Null Baseline	✓	0.416	0.988	0.223	0.133	4.45	0.841
Magpie Baseline	✓	0.944	0.992	0.160	0.097	2.94	0.898
Reaction Graph (excl. Actions)	✓	0.960	0.994	0.130	0.079	2.38	0.915
Reaction Graph (inc. Actions)	✓	0.960	0.994	0.133	0.081	2.40	0.913

Thus we need to build separate models to handle different numbers of precursors. For the ablation study, we consider reactions with 2 or 3 precursors where data is most abundant, making up 28% and 42% of the up to ten precursor data set respectively.

The ablation study compares reaction graph models with and without the processing actions and the alchemy mask to discern which parts of the model design are significant. The results from the study are summarised in Table 5.2. A Null baseline result is given for comparison. The Null baseline assumes all elements in the precursors are present in the product and that they are present in their average relative amounts derived from the training data.

5.7.1 Reaction Representation Learning

For product element prediction, the reaction graph models are observed to perform significantly better than using *Magpie* features alone. For stoichiometry prediction, Figure 5.8 shows the proportion of predictions that are within given EleMD tolerances. A clear improvement is seen through incorporating the graph representation learning framework.

Figure 5.9 shows plots of actual against predicted fractional stoichiometry. This figure shows how predictions are distributed about the equivalence line and their relative confidence as measured through ensemble variance. Similar to the Null baseline, the *Magpie* baseline model predictions show a lack of parity about the equivalence line. In

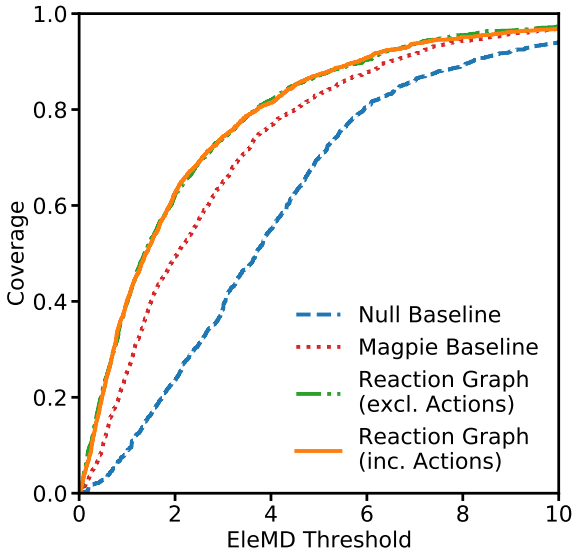


Figure 5.8: **Coverage of Predictions for Ablation Study Models.** The cumulative proportion of stoichiometries correctly predicted to within a given EleMD threshold is plotted against the threshold for the Null baseline, *Magpie*, and Reaction Graph models considered in the ablation study. Both Reaction Graph models, including and excluding processing actions, outperform the *Magpie* baseline. The inclusion of the processing actions appears to make minimal difference to the accuracy in this case.

contrast, the learned reaction graph embeddings show a significant improvement, both in absolute terms and in better prediction of the model uncertainty. In general, there is more parity, and the most certain predictions are closest to the equivalence line, indicating that the model has assimilated some understanding of where it lacks chemical knowledge.

5.7.2 Alchemy Mask

The results in the lower half of Table 5.2 are achieved through the inclusion of the alchemy mask. This addition provides a notable increase in accuracy across the board for element prediction. The role of the alchemy mask is to prevent the model from splitting probability mass between elements that might have similar representations in materials space but are not present in the precursors. For example, without the mask, the model may perform alchemy and predict that a reaction with only sodium-based materials in its precursors might have both sodium and potassium in the product due to the alkali metals having similar chemistry and therefore similar contributions to the reaction embeddings.

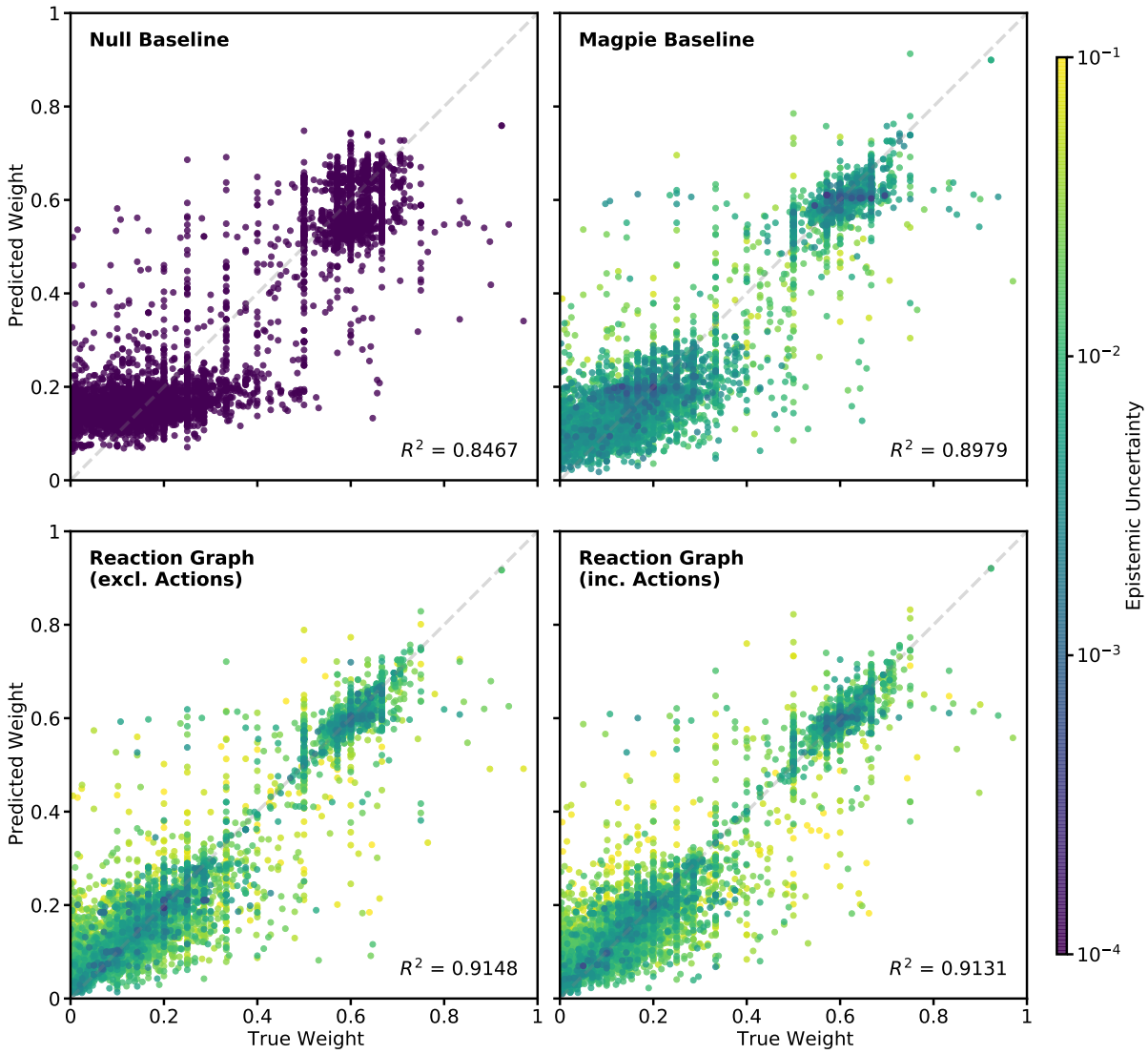


Figure 5.9: **Parity plots comparing Reaction Graph models with the *Magpie* and Null baselines.** Ablation study parity plots for product stoichiometry prediction for reactions with up to three precursors. An improvement in both correlation and uncertainty estimation is seen when comparing the Reaction Graph models against the *Magpie* and Null baselines.

Another detail to consider is that we assumed oxygen abundance across all of the data set when constructing the alchemy masks. However, 0.6% of the up to ten precursor data set (91 examples) consists of reactions in which oxygen is not present in either the precursors or the reaction atmosphere. Importantly, even though oxygen is included in the alchemy mask for these reactions, we observe that the model correctly predicts that oxygen is not present in the product for the examples in the test set (19 examples). This suggests that the model is not biased to always predict the presence of oxygen even though the oxygen is present in 95% of the products in the data set.

5.7.3 Processing Action Sequences

The synthesis procedure is crucial in determining the major product of an inorganic reaction. Different procedures on the same precursors can lead to differing products due to the interplay between thermodynamics and kinetics.

When we consider the complete data set, including reactions with up to ten precursors, there was an observable improvement in accuracy through incorporating the processing actions sequences. A mean L1 error of 0.126 was achieved with the actions compared to a mean L1 error of 0.136 without the actions (see Table 5.1). However, Table 5.2 shows that the addition of the action sequences in the reaction embedding has a detrimental effect – if any – on model performance when considering up to three precursors.

The lack of improvement shown on the restricted data set highlights the limitations of using coarse-grained concepts to describe the processing sequences. Different product compositions and morphologies can be formed from slight changes in temperature or pressure. Consequently, our inability to capture such subtleties in our input may inhibit the model from extracting meaningful information from the processing sequences in the restricted data set.

The improvement seen with the full data set may be explained by considering the greater complexity and diversity in the synthesis procedures recorded. In this expanded setting, even the coarse-grained processing sequences can still capture relevant contextual information leading to improved performance. These observations suggest that incorporating the processing sequence in some manner is important but that using coarse-grained representations limits the utility of this addition. Improved representations of processing actions should therefore be investigated further for their potential to improve model performance.

The principal issue with using coarse-grained action representations is that they do not capture important information about the conditions and timings of processing actions. We believe that finding a way to include such information could improve model performance. For instance, we would hypothesise that performance on metastable structures is likely to be worse than for stable structures (assuming stable structures are more reported more frequently) as the formation of metastable structures is often induced using specific processing actions with particular conditions, e.g. quenching. Therefore, it seems plausible that the lack of this information might be causing the model to interpolate between reactions that produce metastable products and those that produce stable products, reducing the performance for both. It is currently not possible to check this hypothesis as we do not have a systematic way to identify the reactions that result in metastable products in the data set.

We explored approaches to include more of the information from the data set but were unable, as of yet, to find a workable solution. As conditions are not available in all cases, the critical challenge is how to differentiate coarse and detailed descriptions (e.g. "heated in furnace until melt formed" versus "sintered at 1300K for 5 and a half hours") but to do so without exposing the model to issues associated with incomplete and mixed fidelity data. The most promising idea explored was based on named entity recognition and word-embeddings efforts for materials science that have recently been presented in the literature [30, 173, 214]. The envisaged solution would be to identify vector directions for each condition (including directions for continuous conditions), e.g. a vector direction for an hour, a vector direction for 1 K, a vector for using an inert atmosphere when heating, etc. The cited works establish that operations such as addition and subtraction are meaningful in the vector spaces spanned by the word-embeddings. Therefore finding such directions would enable this additional contextual information to be included by simply summing the embeddings for the action with those of its conditions. The strength of this approach would be that it would still provide a meaningful representation of actions in the absence of conditions. Sadly, we could not identify a prior work that had incorporated conditions into these continuous vector embeddings for actions. Nonetheless, we believe this is a promising avenue to investigate in future work both to improve the accuracy of this approach but also in general for the digitisation of synthesis procedures.

5.8 Discussion

In this chapter, a purely data-driven approach to inorganic reaction prediction has been investigated. A physically-motivated model was developed to predict the major products of solid-state reactions from precursors and synthesis procedures. The key feature of this method is the generation of a learned reaction embedding which effectively encodes the reaction context for product prediction. This framework is shown to approach a level of accuracy believed to be chemically discriminative and produce uncertainty estimates that indicate when the model can be used with confidence. An ablation study was conducted to quantify the model's salient features, showing that this approach predicts products more accurately than a competitive baseline method.

The learned reaction representations have been shown to provide a degree of explainability to novel predictions as similarities with known reactions in the literature can be analysed based on the cosine similarity of the representations. Whilst we do not reach the degree of accuracy required for the construction of a retrosynthesis planner for inorganic materials synthesis, we demonstrate how the model can be used to identify errors in the source database via cross validation. As such databases grow, tools that can improve data quality with minimal human labour are desirable.

There remain several directions to improve the presented model that could be tackled as possible next steps,

1. We would like to train all parts of the model, both the element classifier and stoichiometry regressor, in an end-to-end manner. At present, only the element classifier is trained in an end-to-end fashion with the message passing stages. Conceivably this limits the model performance. This might be possible via architectural changes or via alternating between the regression and classification tasks during training.

2. The present work uses makes use of high-level concepts when encoding processing actions. Ideally, we would be able to encode both timings and conditions. Addressing this is particularly challenging as the problem is multi-fidelity spanning from highly discrete processing conditions, i.e. "sinter at 700 K for 7 hours", to continuous or adaptive processing profiles. The most promising solutions to this challenge exist in improved word embeddings for material science [173].

3. Replacing the *Magpie* embeddings with a learnt embedding by constructing a hierarchical model that uses *Roost* to featurise the precursors may be beneficial. At present, this option is unlikely to yield any meaningful improvement as the available data set is only moderate in size. However, as the data set sizes increase over time, this option will become more attractive.

4. Currently, the element regressor treats elements individually. An alternative would be to group all the elements together and then use a distribution based loss. The Jensen–Shannon distance, Hellinger distance and the Sinkhorn approximation [215] to the EleMD are examples of alternative losses that could be worthy of exploration.

Whilst we believe that these pathways will enable us to improve the performance of the current approach, the key question to consider is whether these improvements are likely to be sufficient to achieve the level of performance required to be useful for our desired goal of developing a viable retrosynthesis model for inorganic synthesis?

In the previous chapter, we saw how identifying the correct representation facilitated the development of an improved machine learning model – by representing the compositions of materials as dense weighted graphs, we could then apply message-passing algorithms that were able to learn composition-to-property mappings more efficiently. For inorganic synthesis prediction, it remains unclear how best to represent products and precursors. Whilst we demonstrate that the reaction graph approach outperforms other regression baselines, the formulation of our predictions as a regression task is not necessarily the best design choice. For instance, material science has a bias towards parent materials with integer stoichiometries, leading to a quasi-discrete distribution of fractional weights in materials compositions. Alternative approaches such as auto-regressive text-based generative models [202] are more likely to replicate this quasi-discrete bias. However, text-based approaches do not reflect the fact that materials compositions are permutationally invariant, i.e. that $YBa_2Cu_3O_7$ is the same as $Ba_2Cu_3O_7Y$. More exploratory work is needed to investigate the strengths and weaknesses of our approach compared to different predictive paradigms.

If we limit ourselves to integer stoichiometries, we can also formulate the task of retrosynthesis as a database-centric search problem instead of a predictive problem. This paradigm is much closer to how retrosynthesis models for organic chemistry work, identifying synthesis routes from databases of known precursors that can be reacted together to produce the desired target molecules. For inorganic materials, a key strength

of database-centric approaches is that they allow us to leverage *ab initio* thermochemistry data to aid the identification of promising reaction pathways. The two most developed approaches involve approximating terms from classical nucleation theory to predict pathways that might favourably nucleate the desired product [200] and applying shortest-path algorithms to chemical reaction networks [199]. However, such database-centric approaches have two main limitations. Firstly, many functional materials are doped to induce functionality and are therefore not amenable to database-centric methods that are restricted to integer stoichiometries. In such instances, database-centric planning may still be helpful for identifying novels pathways for synthesising parent structures, potentially facilitating the identification of new synthesis routes for the desired doped materials. Secondly, database-centric approaches are inherently limited in regions of chemical space that have not been adequately explored. There is likely to be substantial overlap between such regions and the regions where we have the least insight regarding how to synthesise new materials. Consequently, expansion of high-throughput *ab initio* databases helps both in the identification of promising targets and potentially in the selection of viable synthesis routes by filling in underexplored patches of materials space. In the next chapter, we explore how we can construct accurate machine learning models specifically tailored to the demands of expanding high-throughput *ab initio* databases in an efficient manner.

Chapter 6

Wyckoff Representation Regression

Finding a needle in a haystack is often used as an analogy for materials discovery. Only a small proportion of viable material compositions (believed to be of the order 10^{10} [15]) will have thermodynamically stable polymorphs that are experimentally accessible. Most approaches to tackle this challenge focus on predictive models for materials properties – metaphorical sieves that filter out the hay. This chapter investigates an alternative approach: Can we significantly cut down materials space by changing how we represent materials – making most of the hay disappear?

Our approach is motivated by a concept ubiquitous in science: coarse-graining. Taking molecular chemistry, for example, chemists typically build intuitions about chemical properties using molecular graphs. Molecular graphs are a coarse-grained representation of molecules, with each graph corresponding to a unique ensemble of atomic coordinates. Searching in the enumerable space of molecular graphs, as opposed to the innumerable space of possible atomic coordinates, has enabled the development of powerful computational tools [169, 216] as well as efforts that exhaustively enumerate chemical space [217, 218].

In materials science, however, an analogous coarse-grained representation of crystal structures is missing. Thus, we are left confronting the innumerable search space problem. Composition-based approaches can somewhat overcome this challenge [219, 65, 2, 166] but do so at the cost of discarding all information about the crystal structures of the materials being considered. As such, either extensive computational crystal structure searching [160–162] or lab-based experiments are required to validate predictions.

One avenue to manoeuvre around this challenge has been to explore restricted classes of structure prototypes using novel descriptors, e.g. Perovskites [220–222], quaternary

Heuslers [223], or Elpasolites [224]. Specifying the prototype avoids the need for crystal structure searching, empowering more extensive screening campaigns as the computational cost of validation is significantly reduced.

In this chapter, we introduce an approach that generalises these prototype-restricted models by considering Wyckoff representations – coordinate-free sets of symmetry-related positions in a crystal. This framework allows us to develop accurate machine learning models for materials discovery tasks where the relaxed crystal structure is *a priori* unknown. We first illustrate our approach by considering the problem of exploring Ti-Zn-N, Zr-Zn-N and Hf-Zr-N ternary systems, showing that our model finds low energy structures in the phase diagram with significantly lower computational effort. We then test the ability of our model to identify novel stable materials across a diverse range of chemistries, showing that it has a precision ~ 3 times higher than the state-of-the-art. Finally, we develop a materials exploration pipeline that, starting from an initial nucleus of known materials, screens nearby materials space and allows the efficient discovery of new stable materials. We identify 1,558 hitherto unknown materials that are below the known convex hull of previously calculated materials from just 5,675 attempted *ab initio* calculations.

6.1 Wyckoff Representation Regression

Building an accurate machine learning model hinges on identifying model inputs that are sufficiently informative to allow the target variable to be predicted. However, for a machine learning model to be useful in practice, these inputs need to be significantly cheaper to obtain than the cost of labelling data. In the context of materials discovery, previous works have shown that virtual screening workflows based on Kohn-Sham density functional theory (DFT) can be used to identify novel functional materials [20, 21]. Separately it has been shown that accurate machine learning models can be built for the formation energies of inorganic crystals calculated via DFT using the DFT-relaxed crystal structure as the model input [82, 81, 83, 225]. Variations on these models are often used to construct potential energy surface models for molecular dynamics applications. These models typically require $\mathcal{O}(10^3)$ samples per system in order to build accurate models. The high upfront cost of collecting this training data is then offset by the fact that inference using these models is significantly cheaper than the DFT calculations

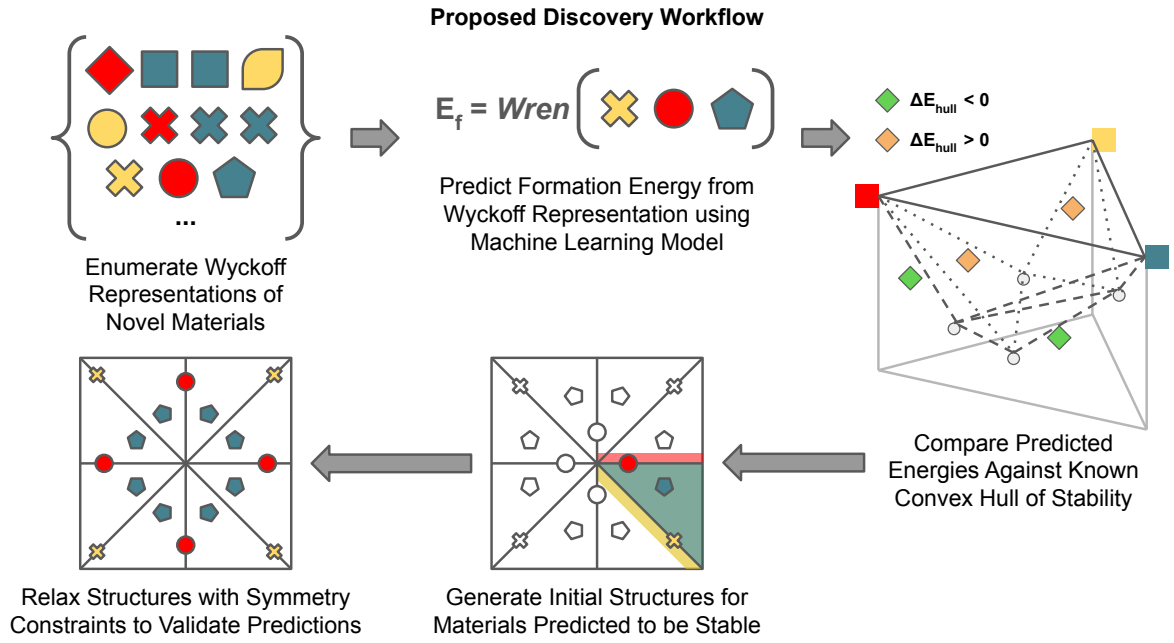


Figure 6.1: **Coarse-graining materials space using Wyckoff representations enables efficient data-driven materials discovery.** A machine learning powered materials discovery workflow that takes advantage of the benefits of the proposed Wyckoff representation. The workflow uses a machine learning model to predict formation energies for candidate materials from an enumerated library of Wyckoff representations (shapes are used to denote different Wyckoff positions and colours to denote different element types). These predicted formation energies are then compared against the known convex hull of stability. Structures satisfying the required symmetries are then generated and relaxed for materials predicted to be stable. The calculated energies of the relaxed structures can then be compared against the known convex hull to confirm whether the candidate is stable.

they approximate, enabling the exploration materials properties in larger systems and over longer timescales than previously possible [226–228]. Sadly the application of such structure-based models to materials discovery applications is circular because arriving at a DFT-relaxed structure necessitates calculating the energy using DFT multiple times.

To construct a model that is useful for materials discovery, the cost of DFT needs to be entirely avoided when generating model inputs. Looking towards molecular systems, this has been achieved via generating low energy conformers using a lower level of theory, e.g. empirical [229, 230] or semi-empirical methods [231]. These conformers can then be treated using a variety of methods such as Δ-learning [232, 233] or RMSD-weighted Gaussian process regression [60] to learn the energies for the conformers after relaxation at

a higher level of theory, e.g. semi-empirical → DFT, DFT → Coupled Cluster with Singles and Doubles (and connected Triples) (CCSD(T)). To our knowledge, however, empirical or semi-empirical methods for obtaining approximately low energy inorganic crystal structures have not been suitably parameterised for the general inorganic chemistries considered in materials discovery applications. We also note that even if such methods were available, generating the approximately low energy crystal structures for large-scale screening campaigns could still require significant computation in contrast to the enumerable inputs we describe below.

Several groups have therefore proposed to use composition-based inputs [2, 65, 166, 219], which avoid the upfront need for structure identification. However, the composition is not expressive enough to differentiate polymorphs. This is a significant shortcoming as different polymorphs can have radically different properties, most famously the example of diamond and graphite. In this work, we propose model inputs that can distinguish polymorphs whilst also avoiding the cost of DFT. Such models can be used to triage which DFT calculations are carried out in materials discovery workflows, allowing for more efficient use of computational resources.

In crystallography, one way to completely specify the crystal structure of a material is via a combination of:

1. The spacegroup of the structure,
2. The dimensions of its unit cell, and
3. A set of Wyckoff positions with the elements that sit on them.

The Wyckoff positions describe sites that map onto equivalent sites under the symmetry transformations of the given spacegroup [234]. As a consequence, a single Wyckoff position can encode the positions of multiple atoms. To construct model inputs from sets of Wyckoff positions, we discard the information about the exact positions and lattice parameters. In the resulting coordinate-free representation, that we refer to as the Wyckoff representation, each Wyckoff position is labelled by a Wyckoff letter and the element at that position. Consequently, as the Wyckoff representation is discrete, it is possible to computationally enumerate Wyckoff representations that represent candidate materials for use in screening campaigns.

The procedure of obtaining the Wyckoff representation from a crystal structure can be viewed as a coarse-graining process that takes us from an unsymmetrised initial parameter

space of size $4N + 6$, through the symmetrised Wyckoff position space of maximum size $5M + 6$, to the much smaller coordinate-free space of Wyckoff representations with size $2M$, where N is the number of sites in the unit cell and the corresponding number of Wyckoff positions M satisfies $M \leq N$. The back mapping from the coarse-grained Wyckoff representation to the full structure can often be satisfactorily obtained via a single symmetry constrained DFT-relaxation of a prototype structure (see Figure 6.1).

To use the Wyckoff representation as the input for a machine learning model, we formulate the task of property prediction as a multi-set regression problem. A message-passing neural network architecture based on the *Roost* architecture [2] introduced in Chapter 4 is used to do this – whereas the *Roost* model performs materials property prediction via set regression on the weighted set of elements in a material's composition we now consider the multi-set of coordinate-free Wyckoff positions in the material's Wyckoff representation.

The principal idea behind the model architecture is to embed the coordinate-free Wyckoff positions of a given material into a vector space. The representations in this embedded space are then updated via message passing operations that consider all directed pairwise combinations of members in the multi-set. The messages propagate contextual information between Wyckoff positions leading to the emergence of material-specific representations. These message passing stages are repeated multiple times before a permutation invariant pooling operation is applied to the multi-set to get a fixed-length representation. As the labelling of Wyckoff positions includes several choices of setting (see Table 6.4), we carry out on-the-fly augmentation of equivalent Wyckoff representations. We average the fixed-length representations for these equivalent inputs to ensure invariance to this choice. These averaged fixed-length representations are then fed into a feed-forward output neural network that returns the models predictions.

This chapter focuses primarily on models that predict the formation energy of inorganic crystalline materials with a view towards applications in materials discovery. However, the proposed framework and inputs are applicable to any material property. We call the proposed model *Wren* (**W**yckoff **Re**presentation regressio**N**). Throughout this work, we train *Deep Ensembles* consisting of 10 *Wren* models starting from different random initialisations [91], allowing us to estimate the model's uncertainty as well as providing better point estimates. Details of the *Wren* architecture and the hyper-parameters used are given in Appendix A.3 .

6.2 Data Sets

TAATA Data Set

The first data set we consider is the TAATA data set [235] which consists of 3 highly sampled *ab initio* phase diagrams for the Hf-Zn-N, Ti-Zn-N and Zr-Zn-N ternary systems. The ternary systems studied in the TAATA data set were investigated for their potential in piezoelectric devices and energy harvesting applications. The TAATA data set contains a diverse range of stable and unstable structures for each composition.

After applying a canonicalisation and cleaning treatment (see below), we are left with 3,128 entries over 520 unique compositions in the Ti-Zn-N phase diagram, 2,722 entries over 449 compositions in the Zr-Zn-N phase diagram, and 3,406 entries over 595 compositions in the Hf-Zn-N phase diagram.

To obtain training and test sets for the TAATA data set, we perform a group split based on composition prototypes (e.g. {Ti, Zr, Hf}Zn_2N, {Ti, Zr, Hf}ZnN_4 etc.) and then in each instance randomly assign two of the chemical systems (if present in the data) to the training set and one to the test set.

Materials Project Compatible Data Sets

The second data source we consider is the Materials Project (MP) database [98] which is a highly curated database of high-throughput DFT calculations. At the time of access, the Materials Project database contained approximately 140k crystal structures. We apply the same canonicalisation and cleaning procedure to this MP data set applied to the TAATA data set. This process leaves a final MP data set containing approximately 105k distinct materials.

In addition to the official Materials Project database, we also consider an additional source of Materials Project compatible data from Ref. [164]. In this work, the authors calculated the energies and properties of a large number of crystal structures that were generated through the substitution of chemically similar elements into known crystal structures [203]. We refer to this data set as the WBM data set. After de-duplication and cleaning, the WBM data set contains approximately 220k materials.

6.3 Wyckoff Positions and Model Inputs

6.3.1 Evaluation of Wyckoff Positions

We use `spglib` [236] to determine the spacegroup and Wyckoff positions for the structures in the data sets. We set the tolerance thresholds as 0.1 Å for positions and 5° for angles (Note, these are the exact tolerances used by the Materials Project to calculate the spacegroup). In real materials, we often observe some degree of off-site relaxation away from high-symmetry sites. Depending on the level of anisotropy and the symmetry finder's tolerance threshold, this might result in materials being classed as $P1$ – having no discernible symmetry. As lower symmetry Wyckoff representations encode less information about the structure, the symmetry finder's tolerance is seemingly an important hyper-parameter to bear in mind. However, preliminary investigations showed that varying the tolerance threshold between typical values of 0.01 Å and 0.1 Å did not significantly impact model accuracy on the TAATA data set. We note that as an alternative to manual selection of tolerance hyper-parameters, symmetry finders with adaptive tolerances, such as `aflow-sym` [237], could be used for the identification of the spacegroup and Wyckoff positions. However, given that we did not observe any appreciable improvement in accuracy using `aflow-sym` when testing it on the TAATA data set, and adaptive schemes are typically associated with greater computational cost, `spglib` was picked over other symmetry finders due to its speed.

6.3.2 Canonicalisation and Cleaning

All the data used to train models in this chapter went through a canonicalisation and cleaning process. Tables 6.1, 6.2 and 6.3 show how much data is discarded at each stage.

The check for valid Wyckoff representations looks at whether the multiplicities of the Wyckoff positions returned correctly sum up to give the material composition. Invalid Wyckoff representations are believed to arise due to overlapping Wyckoff positions within the tolerance of the symmetry finder used. As the number of invalid Wyckoff representations is small compared to the data sets' overall size, invalid structures were simply discarded.

We use the AFLOW prototype encyclopedia [238, 239] scheme to canonicalise our model inputs and then discard higher energy structures that have equivalent Wyckoff representations to other structures in the data set. The majority of the structures

Table 6.1: Table showing the impact of cleaning criteria on the TAATA data set.

Filter	Number
Full Data Set	12,815
Valid Wyckoff Representation from `spglib`	12,801
Lowest Energy Canonical Representations	9,775
Less than 16 Wyckoff Positions	9,552
Less than 64 sites in Crystal Structure	9,256
Volume per site less than 500 Å^3	9,256

Table 6.2: Table showing the impact of cleaning criteria on the MP data set.

Filter	Number
Full Data Set	139,367
Valid Wyckoff Representation from `spglib`	138,927
Lowest Energy Canonical Representations	128,904
Less than 16 Wyckoff Positions	108,378
Less than 64 sites in Crystal Structure	104,791
Volume per site less than 500 Å^3	104,610

Table 6.3: Table showing the impact of cleaning criteria on the WBM data set.

Filter	Number
Full Data Set	257,489
Valid Wyckoff Representation from `spglib`	250,302
Lowest Energy Canonical Representations	250,280
Less than 16 Wyckoff Positions	250,239
Less than 64 sites in Crystal Structure	250,227
Volume per site less than 500 Å^3	250,226
After Removal of Duplicates	220,751

removed by canonicalisation are triclinic as the lack of symmetries in triclinic systems results in many distinct structures mapping to the same Wyckoff representation.

When we take the union of the MP and WBM data sets, we remove duplicates from the WBM data set that have since been included in the MP data set. For duplicated elemental structures, we kept the structures recorded in the MP data set to ensure that our endpoints for calculating formation energies were consistent. During de-duplication, 2,446 additional materials were removed from the MP data set, leaving 322,915 materials in the union of the WBM and MP data sets.

6.3.3 Wyckoff Position Embedding

The coordinate-free Wyckoff positions in the Wyckoff representation need to be represented with vectors before they can be fed into the proposed model. These embedding vectors have two parts, the elemental embedding and the Wyckoff embedding (see Figure 6.2**a**). The elemental information is encoded using the *Matscholar* embedding from Ref. [173]. The Wyckoff embedding we use is made up of 3 parts,

1. A one-hot encoding of the crystal system (of which there are 6),
2. A one-hot encoding of the Bravais lattice centring (of which there are 5), and
3. An embedding constructed from the sum of multi-hot encodings of the equivalent sites within a given Wyckoff position.

To construct the multi-hot encodings, we first collate all the sites within all the allowed Wyckoff positions as recorded on the Bilbao crystallographic server [240]. In total, across the 230 crystallographic spacegroups in three-dimensional space, there are 1731 different Wyckoff positions. Each site can be broken apart into its offset and algebraic terms, e.g.

$$(-x + y + 1/4, y, z + 3/4) = (1/4, 0, 3/4) + (-x + y, y, z). \tag{6.1}$$

From this, we construct separate one-hot encodings for the unique algebraic and unique offset positions. We end up with 185 unique algebraic positions and 248 unique offset positions. A Wyckoff position is then represented by a sum of the embeddings of all the allowed sites. The resulting embedding has a dimensionality of 444, with the 1731 Wyckoff positions being encoded into 1038 unique embeddings. This embedding is designed to try and expose as many correlations as possible that might exist between different Wyckoff positions.

As an illustrative example the embeddings for the "f" Wyckoff position of $Fm\bar{3}$ (No. 202), $F432$ (No. 209), and $Fm\bar{3}m$ (No. 225) are all the same. This arises as they are all face-centred cubic lattices that describe the positions of 32 atoms within the unit cell at ((0, 0, 0), (0, 1/2, 1/2), (1/2, 0, 1/2), (1/2, 1/2, 0)) $\oplus$ ((x, x, x), (-x, -x, x), (-x, x, -x), (x, -x, -x), (-x, -x, -x), (x, x, -x), (x, -x, x), (-x, x, x)), where x is a free-coordinate of the Wyckoff positions. The embedding vector has 4's in the positions corresponding to the (x, x, x), (-x, -x, x), (-x, x, -x), (x, -x, -x), (-x, -x, -x), (x, x, -x), (x, -x, x), (-x, x, x) algebraic terms and 8's in the positions corresponding to the (0, 0, 0), (0, 1/2, 1/2), (1/2,

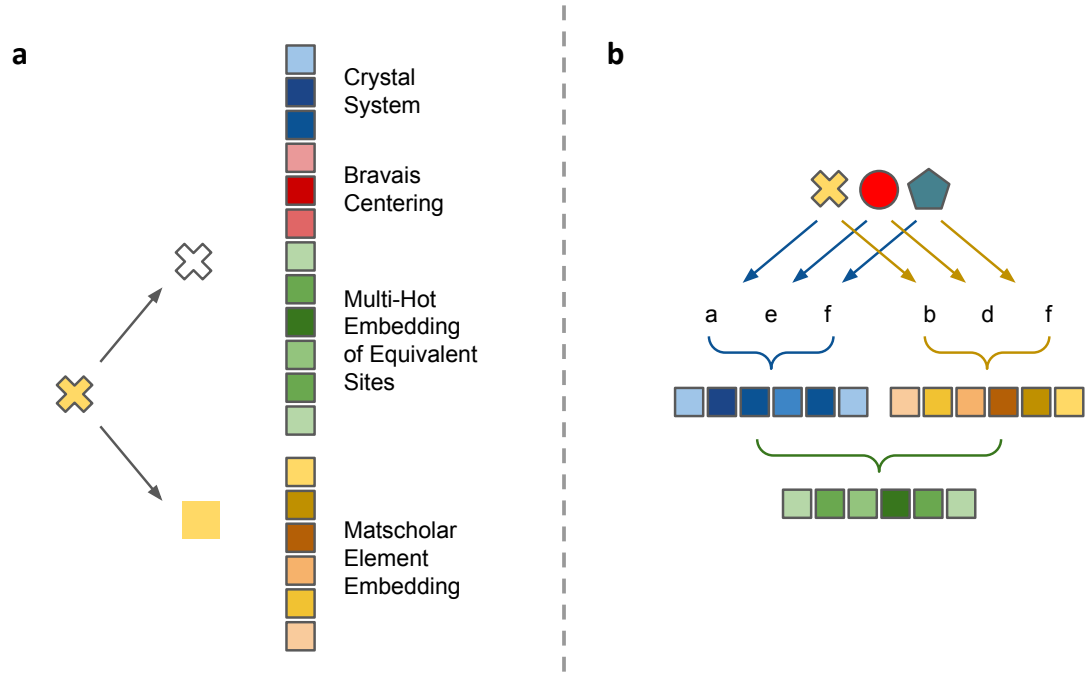


Figure 6.2: **Schematic of Wyckoff position embeddings and on-the-fly augmentation of equivalent Wyckoff representations. a** - The Wyckoff position embeddings are made up of two parts. First, the Wyckoff proportion of the embedding comprises three subsections encoding the crystal system, Bravais centring and equivalent sites in the Wyckoff positions. Second, the elemental embedding for which we take the *Matscholar* embedding from Ref. [173]. **b** - The labelling of Wyckoff positions includes a choice of setting. In order to ensure that our model is invariant to these choices, we perform on-the-fly augmentation of all equivalent Wyckoff representations and then average the augmented embeddings before they are fed into the output network.

0, 1/2), (1/2, 1/2, 0) offset terms. In principle further engineering of this embedding could be carried out to encode more prior knowledge. However, for the sizes of data set considered in this work the benefit of doing so is likely to be marginal.

6.3.4 Invariance to Equivalent Wyckoff Representations

The categorisation of Wyckoff positions depends on a choice of origin [241]. As such, there is not a unique mapping between the crystal structure and the Wyckoff representation (see Table 6.4). To ensure the model is invariant to the choice of origin, we perform on-the-fly augmentation of Wyckoff positions with respect to this choice of origin (see Figure 6.2**b** and Table 6.5). The augmented representations are averaged at the end of the message passing stage to provide a single representation of equivalent Wyckoff

Table 6.4: Wyckoff Sets of Space Group $R\bar{3}$ (No. 148).

Wyckoff Letter	Multiplicity	Site Symmetry	Rep.	Equivalent WP
f	18	1	(x, y, z)	f
e	9	-1	(1/2 , 0, 0)	de
d	9	-1	(1/2 , 0, 1/2)	de
c	6	3.	(0, 0, z)	c
b	3	-3.	(0, 0, 1/2)	ab
a	3	-3.	(0, 0, 0)	ab

Table 6.5: Transformation of the Wyckoff Positions of $R\bar{3}$ (No. 148) under the coset representatives of its affine normaliser.

Coset Representative	Geometrical Interpretation	Transformed WP
x, y, z	1	a b c d e f
x, y, z+1/2	t (0, 0, 1/2)	b a c e d f
-y, -x, z	m x, -x, z	a b c d e f
-y, -x, z+1/2	c x, -x, z	b a c e d f

representations to the output network. By pooling at this point, we ensure that the model is invariant to the choice of origin and that its training is not biased towards materials for which many equivalent Wyckoff representations exist.

6.4 Exploration of Structurally Diverse Chemical Systems

6.4.1 Differentiation of Polymorphs

We start by examining the performance of *Wren* on the TAATA data set. To put the *Wren* model into context we compare against alternative composition-based and structure-based models. For the composition-based alternative we use the *Roost* model [2] (see Chapter 4), which has been shown in a recent benchmark [119] to outperform other composition-based machine learning models for formation energy prediction. For the structure-based alternative we use *CGCNN* [82], a message-passing neural network [170] that operates on a graph representation of local environments within a crystal. We train *CGCNN* using the relaxed training set structures and examine the test set performance on both the relaxed structures and the original pre-relaxation structures from the TAATA workflow.

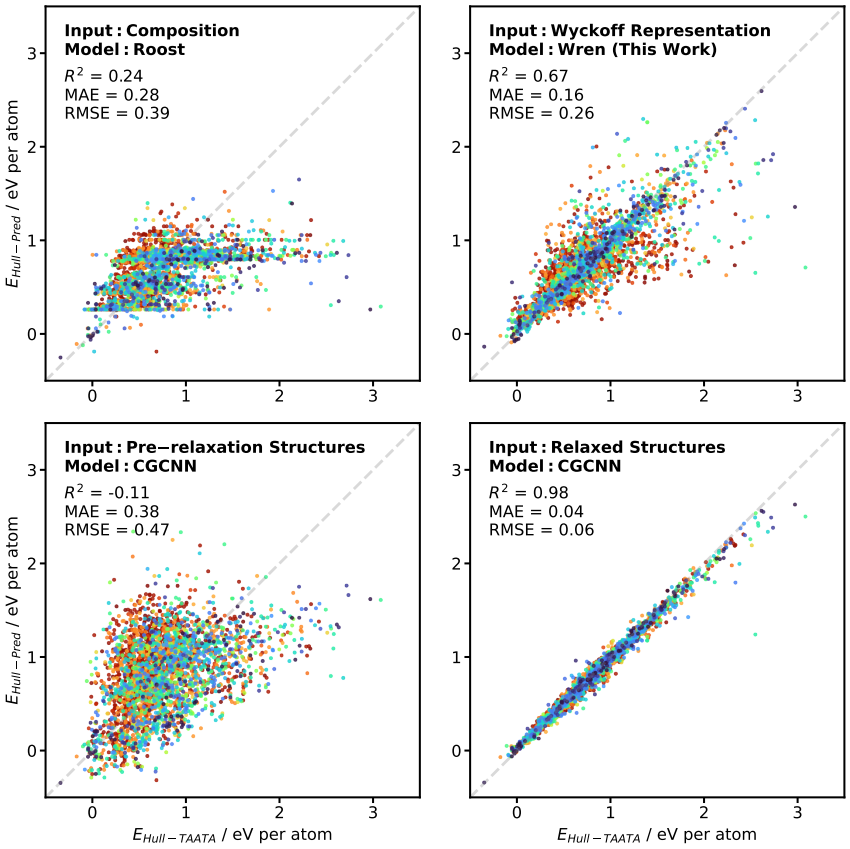


Figure 6.3: **Using Wyckoff representations enables coordinate-free models that distinguish polymorphs.** Parity plots of the energy from the convex hull of the TAATA training set for materials in the TAATA test set against the model's predicted equivalent for *Roost*, *Wren* and *CGCNN* models. The coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE) on the test set are given for each model. Points are coloured according to their spacegroup using the international numbering scheme as a scale. Red colours are used for lower spacegroup numbers and blue colours for higher spacegroup numbers. Looking at the parity plot for the Wren model, we see that the materials that the model performs worst on generally have lower spacegroup numbers. Intuitively this makes sense as the Wyckoff representation encodes much less information about the final structure in low symmetry structures. Whilst *CGCNN* is highly predictive when asked to predict the relaxed energies given the relaxed structures, using the energy predicted on pre-relaxation structures as a surrogate for the relaxed energy is highly inaccurate.

The architectural similarity between *Roost* and *Wren* allows us to attribute any significant differences in model performance to the change in inputs. Figure 6.3 presents parity plots for materials in the TAATA test set, showing the predicted energies above the convex hull of the training set against the energy calculated using the DFT energy of the relaxed structure. Despite being coordinate-free, *Wren* can distinguish polymorphs. This is in stark contrast to *Roost* where a clear banding is observed since multiple polymorphs exist for many of the compositions studied. This pathological behaviour is why previous composition-based models have only been tested on nominal ground-state structures [2, 119, 165]. Considering pairwise comparisons of structures with the same composition, the *Wren* model correctly ranks the energy of the polymorphs 79% of the time compared to just 57% of the time for *CGCNN* on the pre-relaxation structures. *Wren*'s performance is significantly higher than the 50% expectation under random ranking.

6.4.2 Comparisons against using Structure-based Models for Materials Discovery

When trained and tested on relaxed structures, *CGCNN* exhibits a very high coefficient of determination, $R^2 = 0.98$. However, due to the inability to generate relaxed structures without incurring the costs of DFT, we cannot use relaxed structures as the inputs to structure-based models in materials discovery applications. The most straightforward baseline strategy is to substitute the relaxed structures for pre-relaxation structures [225]. When the same model is deployed on the pre-relaxation structures, we observe a negative R^2 score of -0.12 between the actual energy above the hull and that calculated using the energy predicted for the pre-relaxation structure. This lack of correlation can be attributed to two related effects. Firstly, in prototype-based substitution workflows, the distributions of bond distances seen before relaxation can be highly unphysical. This arises as the original species often have very different atomic radii from the substituted species. Consequently, structure-based models are asked to make predictions about chemical environments that are structurally distinct to those seen in the training data, resulting in low accuracy. Secondly, even if the model accurately predicts the energy of the pre-relaxation structure, this energy can differ significantly from that of the relaxed structure. Consequently, there is a substantial task discrepancy between how the model is trained and how it is deployed.

CGCNN encodes local environments based on distance information, therefore, descriptor sets that also encode local environments based on 2-body (and 3-body) distances [53, 56] are likely to suffer similar pathologies. A conceptually distinct set of descriptors are Voronoi crystal structure attributes [242]. These features are constructed using the areas of the Voronoi facets around a given site in such a way that they are invariant to changes in volume. This is a compelling property as one of the principal changes that occurs when relaxing a prototype structure after chemical substitution is the change in the unit cell volume. Another related approach is the symmetry-labelled Voronoi graph convolutional neural network of [243] which also discards explicit distance information. Here, instead of labelling edges in the graph according to the distances between sites, the edges are labelled by the approximate 2D symmetry of the Voronoi facets shared between sites.

To explore the first hypothesis about the poor performance of *CGCNN* on pre-relaxation structures being due to the differences in the distributions of bond distances seen before and after relaxation, we constructed an additional structure-based baseline model consisting of a Random Forest that takes the Voronoi crystal structure attributes [242] and *Magpie* composition features [65] as inputs as proposed in Ref. [242]. We will refer to this as the *Voronoi* model. To explore the second hypothesis about the poor performance being due to the shift between the testing and training tasks, we consider an alternative problem formulation where all the structures in any given basin of attraction map to the energy of the basin minima. We call the resulting surface the discrete potential energy surface.

Whilst there is no physical interpretation for this discrete potential energy surface setup, it more closely matches how we want to use the model in practice – If we relax this candidate structure, will the resulting structure be stable? A key consideration is that, in practice, we do not have access to all the pre-relaxation structures that map to a given local minimum. Nonetheless, we can make a first approximation to this setup by training models that map both the relaxed and pre-relaxation TAATA structures to the relaxed structure's energy. We will refer to such models as *D*-variant models. We do this for both *CGCNN* and the *Voronoi* models and refer to these models as *CGCNN-D* and *Voronoi-D* respectively.

In Figure 6.4 we see that, when trained on relaxed structures and tested on pre-relaxation structures, the *Voronoi* model performs more poorly than *CGCNN* on the

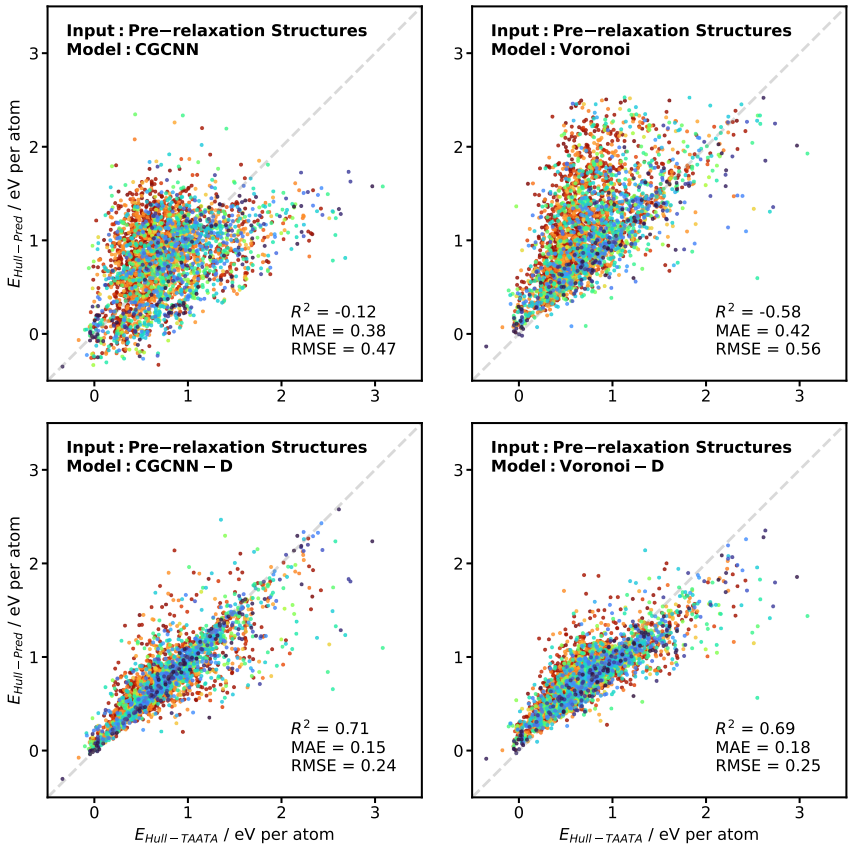


Figure 6.4: **Structure-based models make poor predictions of local energy minima on pre-relaxation structures.** Parity plots of the energy from the convex hull of the TAATA training set for materials in the TAATA test set against the model's predicted equivalent for *Voronoi*, *CGCNN*, *Voronoi-D* and *CGCNN-D* models. The coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE) on the test set are given for each model. Points are coloured according to their spacegroup using the international numbering scheme as a scale. Red colours are used for lower spacegroup numbers and blue colours for higher spacegroup numbers. The plots on the top row show that using models trained on relaxed structures to estimate the stability of relaxed structures from their pre-relaxation structures performs badly. The second row of plots shows a significant improvement in performance if we train models that map both the pre-relaxation and relaxed training set structures to the relaxed energies, with both the *Voronoi-D* and *CGCNN-D* models offering performance comparable to *Wren*.

equivalent task. The *Voronoi* model systematically over predicts the energy, resulting in a mean error of 0.35 eV per atom compared to the mean error of 0.04 eV per atom seen for the *CGCNN* model on pre-relaxation structures. This systematic overprediction could potentially be desirable as if we have a model that captured the true potential energy surface, the predicted energies on the pre-relaxation structures should be systematically higher than the actual energies of the relaxed structures. Sadly, the initial energies for the pre-relaxed TAATA structures are not readily available, therefore, it is not possible to test whether the systematic over predictions correspond to physical differences in the potential energy surface.

The *Voronoi-D* and *CGCNN-D* models have MAEs of 0.18 and 0.15 eV per atom and RMSEs of 0.25 and 0.24 eV per atom predicting on pre-relaxation structures. These results are comparable to the MAE of 0.16 eV per atom and RMSE of 0.26 eV per atom seen for *Wren*. That the *D*-variant models return results closer to *Wren* is unsurprising as the reformulated task closely matches the task of fitting the Wyckoff-representation-to-relaxed-energy mapping. However, despite empirical performance on par with *Wren*, the use of *D*-variant models is somewhat unsatisfactory as the choice of pre-relaxation structure is under-determined – many sensible pre-relaxation structures will exist in any given local basin of attraction. Consequently, the misfit between the pre-relaxation and relaxed structures in both the training and test sets is inconsistent. We will return to the discussion of *D*-variant models later in the chapter.

As an alternative to the *D*-variant models and discrete potential energy surface paradigm described above, another approach for improving the performance of structure-based models for stability prediction is to construct machine learning workflows that quasi-relax the pre-relaxation structures according to surrogate energy models. By quasi-relaxing structures, we hope to reduce the task discrepancy between the training and deployment scenarios, resulting in smaller errors when using structure-based models to make stability predictions without access to relaxed structures. Such models could also compliment *D*-variant approaches by producing more consistent pre-relaxation structures resulting in a better defined problem akin to Δ-learning [232].

The *BOWSR* algorithm [244] is a recently introduced quasi-relaxation method that uses Bayesian optimisation to adjust the free parameters of symmetrised pre-relaxation structures, i.e. they attempt to use a quasi-relaxation approach to optimise the free parameters that are coarse-grained in the Wyckoff representation. They highlight that

using Bayesian optimisation for the quasi-relaxation is superior to gradient-based approaches such as L-BFGS. This most likely arises out of the fact that the original authors make use of a *MEGNet* model [83] trained only on the energies of relaxed structures in the Materials Project as their surrogate model. By training exclusively on relaxed structures, there is no signal to the model that the examples it has seen sit at the bottoms of their local basins of attraction. Consequently, the force estimates of such a model are unsupported and unreliable. The obvious way to remedy this shortcoming would be to provide additional training data within each local basin of attraction, for example, by training on the relaxation trajectory. Directly training on the forces, in addition to the energies, should also improve model performance in this regard [245, 246]. However, as high-throughput relaxation workflows often involve multiple relaxation runs with different settings, for example, the OQMD workflow [36, 37] carries out both coarse and fine relaxation runs with system-dependent energy cut-offs before a final relaxation with a consistent energy cut-off, the structural diversity of suitable relaxation trajectories – those carried out with the same settings – is likely to be limited.

Whilst machine-learning-based quasi-relaxation approaches are certainly an exciting area for future investigation, we have not benchmarked against them here as in the quasi-relaxation paradigm screening a material requires several minutes of computation as opposed to the millisecond inference time of single-pass machine learning approaches such as the *Wren* or the *CGCNN-D* model.

6.4.3 Symmetry Changes During Symmetry Constrained Relaxations

When carrying out a symmetry constrained relaxation we can see the merger of distinct Wyckoff positions into higher symmetry Wyckoff positions. In such cases, the Wyckoff representation changes between the pre-relaxation and relaxed structures (see Figure 6.5). To ensure that we have a well-defined map, we train *Wren* models to learn the target properties as a function of the relaxed Wyckoff representation. However, we use the pre-relaxation Wyckoff representation for testing purposes as this matches how we intend to use the model in practice.

For the TAATA test set, the Wyckoff representation of the material changes during relaxation for 271 out of 3,087 systems. A change in Wyckoff representation was also observed in 21,688 out of 220,751 materials in the WBM data set, indicating that we

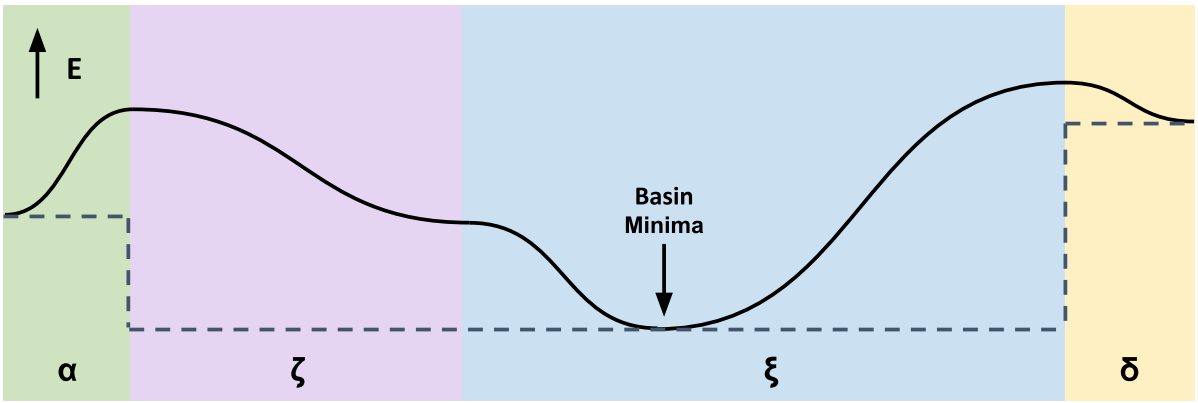


Figure 6.5: **Increasing symmetry can cause changes in Wyckoff representations during structure relaxation.** A toy potential energy surface is shown in black for a salient reaction coordinate along the x-axis. The coloured blocks labelled α, ζ, ξ, and δ denote different phases with distinct Wyckoff representations. The grey dashed line gives the discrete potential energy surface mapping each input within a basin of attraction to the energy of the basin minima. In this example, a structure initialised in the ζ-phase will relax into the ξ-phase by increasing its symmetry – such changes were seen to occur in $\sim 10\%$ of cases in both the TAATA and WBM data sets.

might expect such symmetry changes to occur in $\sim 10\%$ of candidates produced by prototype-based screening workflows. For the MP data set, where a significant proportion of initial structures were sourced from the ICSD, only 4,232 out of 104,610 materials in the data set exhibit changes in Wyckoff representation on relaxation.

If we consider the 271 materials in the TAATA test set where the Wyckoff representation changed during relaxation, using the pre-relaxation Wyckoff representations, as opposed to the relaxed Wyckoff representations, results in an increase in the MAE from 0.19 eV per atom to 0.21 eV per atom for these materials. The MAE of the whole test set changes increases to 0.156 eV per atom from 0.155 eV per atom. An important factor to consider when weighing the importance of this behaviour is that a sufficiently extensive enumeration of the input domain of Wyckoff representations would contain both the pre-relaxation and post-relaxation Wyckoff representations. We note that the model correctly ranks these closely related prototypes in only 130 out of 271 cases, implying that it is no better than random for discriminating between these edge cases.

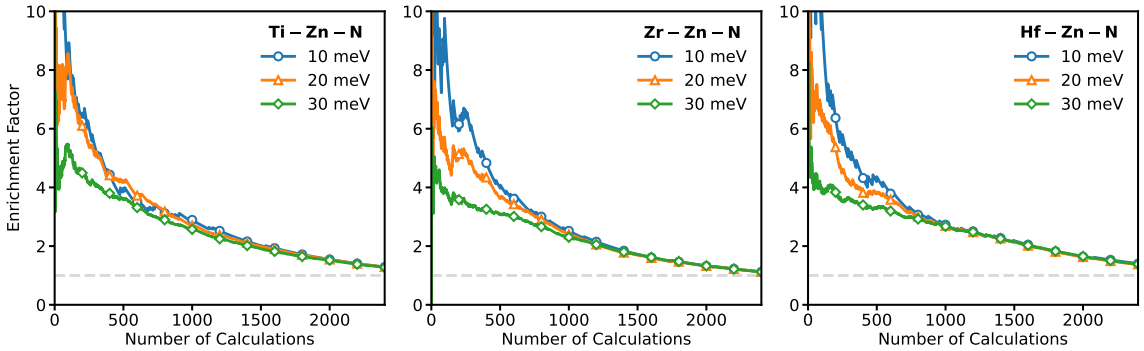


Figure 6.6: ***Wren* accelerates the discovery of low energy structures in unseen chemical systems.** The figures show how the enrichment factor varies as we use *Wren* to direct exploration of the Ti-Zn-N, Zr-Zn-N and Hf-Zn-N chemical systems. The enrichment factor is the ratio of candidates found satisfying a given triage criterion to the number we would expect to find via a random search. The enrichment factor is plotted for candidates within 10, 20 and 30 meV per atom from the convex hull of the full explored system. A light-grey guideline is included to show the performance expected from a random model – an enrichment factor of 1. The plots demonstrate that using *Wren* leads to a significant degree of early enrichment of low energy structures.

6.5 Enrichment in Novel Chemical Systems

From an applications perspective, researchers are often interested in exploring a single or small number of chemical systems that have not previously been studied [247, 235, 248]. Typical crystal structure prediction approaches for computationally mapping out the convex hull of novel chemical systems [161, 162] are highly expensive, often requiring thousands of structures to be relaxed. Figure 6.6 shows how *Wren* can accelerate phase diagram exploration, recovering low energy structures in unseen chemical systems with far fewer computations than an exhaustive search over prototypes.

To demonstrate this ability, we trained *Wren* on the MP data set but excluded all the tertiary compounds from the Ti-Zn-N, Zr-Zn-N and Hf-Zn-N chemical systems. This model was then used to predict the energies of tertiary compounds in the TAATA data set. Convex hulls for the different phase diagrams were constructed with these predicted energies, and compounds were ranked based on their energy to these predicted convex hulls. To assess how effective our model is at accelerating materials discovery, we look at the enrichment factor as a function of the number of calculations. The enrichment factor describes the ratio between the number of candidates found satisfying a target criterion when using a virtual screening strategy and the number of positive candidates

that hypothetically would have been found if the candidates were screened randomly. Enrichment factors are frequently reported for virtual screening campaigns in drug discovery applications [249, 250]. Figure 6.6 shows up to 10-fold early enrichment when carrying out a limited number of calculations. Considering materials within 20 meV per atom of the convex hull as our target criterion, we have an enrichment factor of 5.4 in the Ti-Zn-N chemical system, 5.1 for the Zr-Zn-N chemical system, and 4.5 for the Hf-Zn-N chemical system after 250 calculations when using the *Wren* model. The data used to train *Wren* in this instance contains 4,477 Ti-based materials, 2,680 Zr-based materials but only 1,749 Hf-based materials. These abundances may partially explain why the model offers the best performance on the Ti-Zn-N chemical system and slightly worse performance on the Zr-Zn-N and Hf-Zn-N chemical systems.

6.6 Selecting Stable Materials from Diverse Chemical Spaces

To accelerate the screening of materials space for novel stable materials, a model must reduce the expected number of calculations needed to find a candidate below the known convex hull. Here we consider the convex hull of the MP data set before cleaning. Using the MP and WBM data sets we can interrogate our model's ability to screen for novel stable materials in diverse chemical spaces. To do this, we make predictions for the formation energies of the materials contained in the WBM data set using a *Wren* model trained on the MP data set. We then assess how well the *Wren* model selects potentially stable materials from the WBM data set by calculating the predicted energy above the convex hull and comparing it to the equivalent energy calculated using the DFT formation energy. In order to better replicate a prospective materials discovery workflow, we use the pre-relaxation Wyckoff representations to screen materials in the WBM data set (see earlier discussion about symmetry changes during relaxation in Section 6.4.3).

For this cut-off based classification task the relevant metrics are,

- **Precision** – the proportion of the predictions of potentially stable materials that are correct,

$$PPV = \frac{TP}{TP + FP}, \tag{6.2}$$

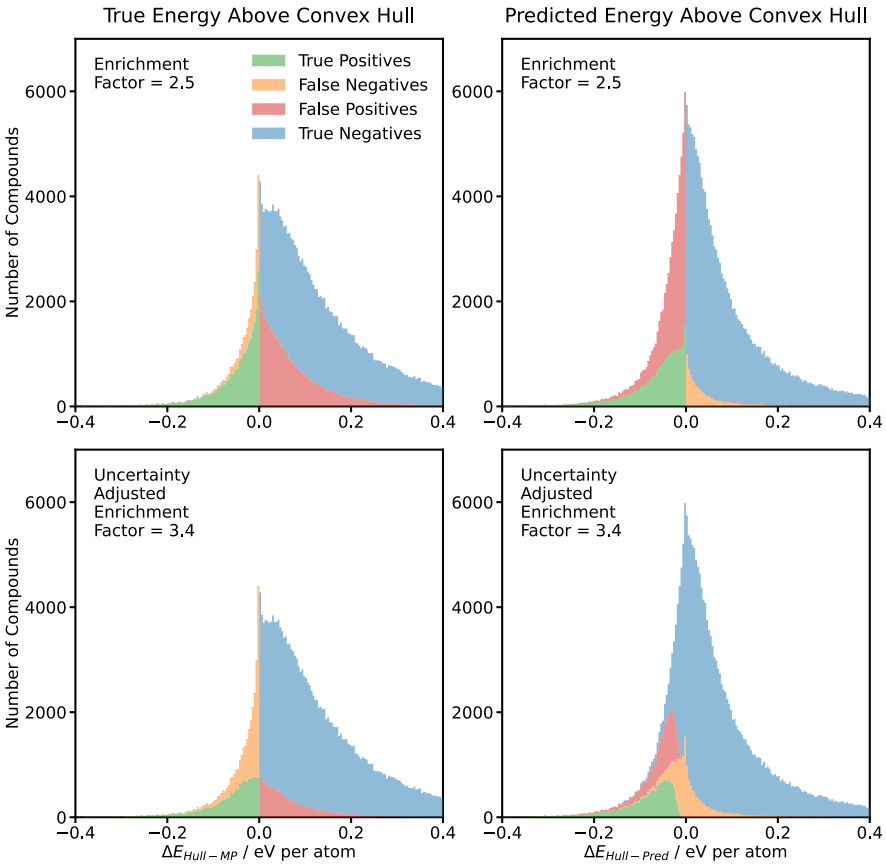


Figure 6.7: ***Wren* recalls the majority of stable structures when screening diverse chemical spaces.** Histograms of the energy to the convex hull for materials in the WBM data set. The histogram is broken down into true positives, false negatives, false positives, and true negatives based on whether the *Wren* model predicts candidates to be below the known convex hull. The top row shows the results when using the criterion that the predicted energy above the hull is less than zero. Using this criterion *Wren* exhibits a high recall, with the majority of materials below the convex hull being correctly identified by the model. The bottom row shows the results when using the uncertainty adjusted criteria. The uncertainty adjusted criteria results in greater enrichment but lower recall, meaning we miss more materials below the known hull. The left-hand column shows the histograms for the true energies above the convex hull, whilst the right-hand column shows the predicted energies above the convex hull. Comparing the two columns, it is apparent that the model routinely underpredicts the energy of unstable materials, a result that we attribute to a bias in the MP data set towards stable materials.

- **Recall** – the proportion of the materials actually below the known convex hull that are found,

$$TPR = \frac{TP}{TP + FN}, \tag{6.3}$$

- **Prevalence** – the proportion of materials below the known convex hull across the entire candidate pool,

$$\mathcal{P} = \frac{TP + FN}{TP + FP + FN + TN}, \tag{6.4}$$

where TP, FP, FN, and TN are the numbers of true positives, false positives, false negatives, and true negatives respectively. In this setup, the ratio of the precision and the prevalence gives the enrichment factor.

Figure 6.7 shows stacked histograms of the true and predicted energies to the convex hull of the full MP data set for the materials from the WBM data set. The histogram sections are coloured according to whether the model correctly predicts a candidate to be below the known convex hull given its Wyckoff representation. The precision using the *Wren* model to triage calculations is 38.7%. Consequently, given that the prevalence of theoretically stable materials in the WBM data set is 15.7%, using *Wren* leads to an enrichment factor of 2.5. We also observe a high recall of 75.3%, meaning that *Wren* misses relatively few potentially stable materials. As enrichment here is computed with respect to the active search strategy of Ref. [164] this translates into a significant improvement in efficiency over random or exhaustive search strategies as our improvements compound multiplicatively with theirs. Consequently, triaging screening workflows based on *Wren* should enable more materials below the known hull to be identified with limited computational resources.

The screening performance of the model can be tuned by adjusting our triage criteria. For example, an alternative triage criterion would be to select candidates for which $\Delta\hat{E}_{\text{Hull-Pred}} + \hat{\sigma} < 0$, where $\Delta\hat{E}_{\text{Hull-Pred}}$ is the predicted distance of a candidate material from the known convex hull and $\hat{\sigma}$ is the predictive uncertainty of the model. This uncertainty adjusted criterion encourages the model to suggest candidates it is more certain about, leading to an increased precision of 53.1%. The enrichment factor for the uncertainty adjusted criterion is 3.4. However, as we are screening a fixed candidate pool, the recall falls to 48.5%. Consequently, the choice of triage criteria should depend on the aims of a given workflow – the expected opportunity cost of false negatives versus false

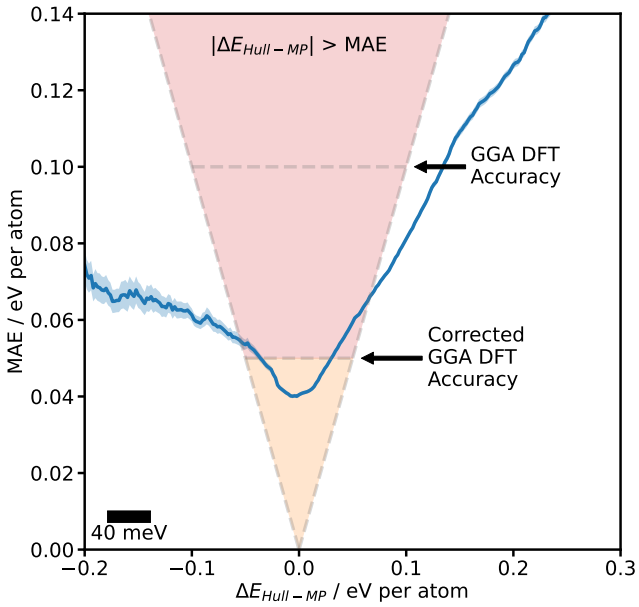


Figure 6.8: ***Wren*'s average error is below DFT error in the region around the stability threshold.** Rolling mean absolute error on the WBM data set as the energy to the convex hull is varied for *Wren* on the pre-relaxation Wyckoff representations. A scale bar is shown for the windowing period of 40 meV per atom used when calculating the rolling average. The standard error in the mean is shaded around each curve. The highlighted V-shaped region shows the area in which the average absolute error is greater than the energy to the known convex hull – this is the region where the model is most at risk of misclassifying structures. In the majority of this region *Wren*'s accuracy is well below the 100 meV per atom threshold considered to be the accuracy of GGA DFT across diverse chemistries [37] and comparable to the $\sim$ 50 meV per atom threshold characteristic of fitted correction schemes [114, 102, 116].

positives, the availability of experimental or computational resources, and how easy it is to expand the candidate pool.

6.6.1 Model Performance as a Function of Stability

The strong performance of *Wren* can be understood by looking at how the mean absolute error varies as a function of the distance from the known convex hull. Figure 6.8 shows that near to the stability threshold, $\Delta\hat{E}_{\text{Hull-MP}} = 0$, *Wren* makes highly accurate predictions of the formation energy. Larger errors are seen for materials far above and far below the hull. However, in these regions, the average error is less than the energy to the convex hull, indicating that the model's classifications are still reliable. We believe

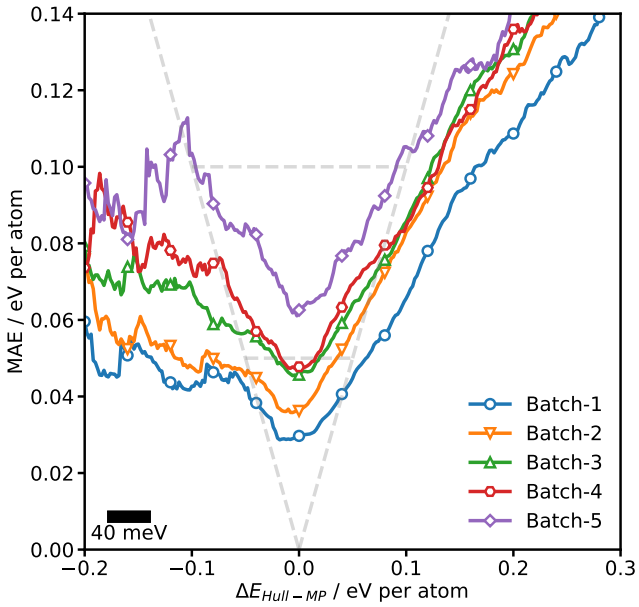


Figure 6.9: **Accuracy of *Wren* falls when making larger extrapolations in chemical space.** Rolling mean absolute error of *Wren* on the batches of the WBM data set as the energy to the convex hull is varied. We take the pre-relaxation WBM Wyckoff representation as input. A windowing period of 40 meV per atom is used when calculating the rolling average. The curves show that later batches, believed to be more likely to be chemically dissimilar to the training data, incur higher average errors.

the existence of this effect for materials far below the hull is partially responsible for the reported improvements over random search of using structure-based models to screen pre-relaxation structures of previous works [225]. The large errors far above the hull are due to the routine underestimation of the formation energy of unstable structures. This underestimation is a manifestation of a bias in the MP data set towards structures with low formation energies. The bias arises from the fact that large numbers of the initial structures in the MP data set are sourced from the ICSD [13]. This result highlights the importance of negative examples for building generally applicable machine learning models [251–253].

The WBM data set was generated using an iterative workflow where successful candidates from the first batch were included when generating candidates for the second batch and so forth. Consequently, candidates considered in later batches are likely to be less similar to materials contained in the MP data set. We can make use of this stratification to probe how *Wren*'s performance changes as it is asked to make larger and

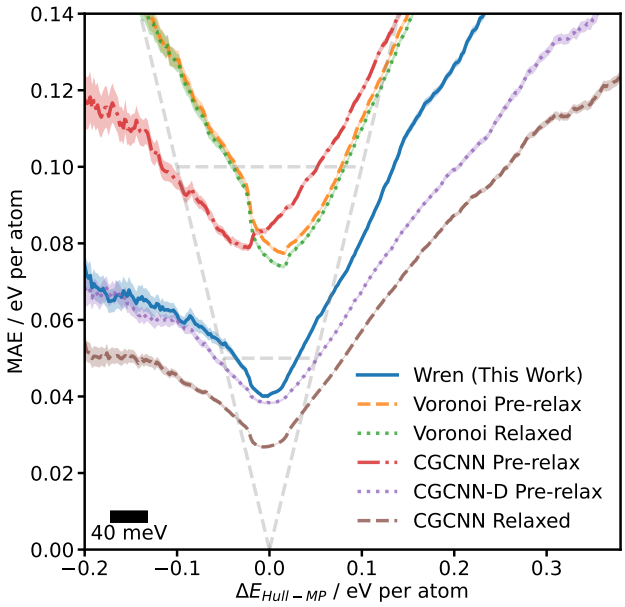


Figure 6.10: **Comparison of the rolling MAE of different models on the WBM data set.** Rolling mean absolute error of the *Wren*, *Voronoi*, *CGCNN* and *CGCNN-D* models on the WBM data set as the energy to the convex hull is varied. A scale bar is shown for the windowing period of 40 meV per atom used when calculating the rolling average. Grey guidelines highlight 100 meV per atom, 50 meV per atom and the MAE $= |\Delta E_{\text{Hull-MP}}|$. As expected, *Wren* is less accurate than *CGCNN* on relaxed structures. The *Wren* model is more accurate than both the *Voronoi* model and *CGCNN* but is slightly outperformed by *CGCNN-D* on pre-relaxation structures.

larger extrapolations by looking at how the mean absolute error changes as a function of the distance from the convex hull for the different batches. In Figure 6.9 we see that, whilst the overall shape of the curves remains the same as the V-shape seen for the full data set, there is an offset between the batches with later batches incurring higher errors on average.

6.6.2 Comparisons Against Structure-Based Models

To contextualise its performance, we compare *Wren* against several structure-based baseline models. First, we look at the *Voronoi* model. The *Voronoi* model is nominally robust against changes that occur during crystal structure relaxation as the Voronoi crystal structure attributes [242] are invariant to changes in volume. This robustness of the *Voronoi* model is confirmed in the observation that the relaxed and pre-relaxation

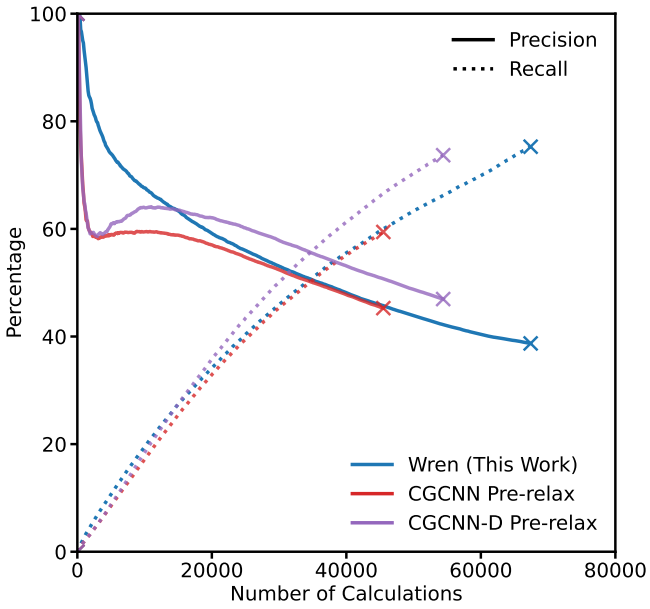


Figure 6.11: ***Wren* displays higher early enrichment than both *CGCNN* and *CGCNN-D*.** Precision and recall on the WBM data set for *Wren*, *CGCNN*, and *CGCNN-D* as the number of calculations increased. Crosses are used to mark the termination points where the models no longer believe there are anymore materials in the WBM data set below the MP convex hull. Initially both *CGCNN* and *CGCNN-D* have much lower precision than *Wren*. The performance of *CGCNN* appears to be upper-bounded by that of the *Wren* model. Following the initial sharp fall, there is a subsequent rise in the precision of the *CGCNN-D* model as the number of calculations carried out increases. After $\sim$ 15,000 calculations there is a crossover point after which *CGCNN-D* gives higher precision than *Wren*.

structures make similar predictions on the WBM data set after having been trained on the MP data set (see Figure 6.10). A mean absolute deviation of 0.019 eV per atom between the pre-relaxation and relaxed structure predictions is seen for the *Voronoi* model, which is relatively small in comparison to the MAEs of 0.144 and 0.141 eV per atom on the pre-relaxation and relaxed structure versions of the WBM data set. Despite being structure-based and the aforementioned robustness, the *Voronoi* model results in a much larger error than *Wren* even when screening the relaxed structures. This suggests that the Wyckoff representation is sufficiently informative to capture significant amounts of information about the relaxed crystal structure. As *Wren* outperforms the *Voronoi* model on relaxed structure inputs we do not benchmark against the *Voronoi-D* model on this task.

Next, we look at the performance of the *CGCNN* and *CGCNN-D* models. As expected and in keeping with the results of previous work [82, 83, 225], we see that using *CGCNN* to predict the stability of candidate materials based on their relaxed structures is highly accurate, yielding an MAE of 0.06 eV per atom. However, as previously noted, in prospective workflows we do not have access to relaxed crystal structures. Therefore, we are much more interested in the performance of models on pre-relaxation structures. When we compare the stability predictions obtained when using *CGCNN* to estimate the stability of candidate materials using the energies predicted for their pre-relaxation structures, we see that *CGCNN* is again outperformed by *Wren*. Figure 6.11 shows that this difference in accuracy manifests in much lower early precision and lower final recall. Turning to the *CGCNN-D* model, we see that the early precision is again much lower than for *Wren*. However, the precision of *CGCNN-D* recovers as more calculations are undertaken. After around $\sim$ 15,000 calculations *CGCNN-D* gives higher precision than *Wren*. Both *Wren* and *CGCNN-D* result in overall recalls of $\sim$ 75%.

In Figure 6.12 we break down the performance of *CGCNN-D* according to the WBM batches. Comparing against Figure 6.9 shows that for the early batches, batches 1 and 2, *Wren* is more accurate than *CGCNN-D* but as we move to the later batches *CGCNN-D* offers better generalisation performance. Further work is needed to investigate whether iteratively retraining the models on each batch would lead to *Wren* outperforming *CGCNN-D* overall after all batches have been considered. Iterative retraining has previously been shown to offer improved performance relative to one-shot approaches, such as the approach adopted here, when exploring binary systems [254]. Batches 1 and 2 contain $\sim$ 55k and $\sim$ 45k crystal structures respectively, corresponding to $\sim$ 2 million core-hours assuming 20 core-hours per relaxation. Training a *Wren* ensemble takes $\sim$ 80 GPU-hours and is embarrassingly parallelisable as the models in the ensemble are trained independently. These timings suggest that iteratively retraining the *Wren* model between batches would not significantly affect the total level of computation if such a strategy proved worthwhile.

Given the compelling performance of the *CGCNN-D* model and the simplistic way in which the *D*-variant models were constructed, the lack of existing literature exploring models trained to predict relaxed structure energies of inorganic materials from their pre-relaxation structures is surprising. As well as the earlier point about the concept of a pre-relaxation structure being under-determined, potentially the lack of research in

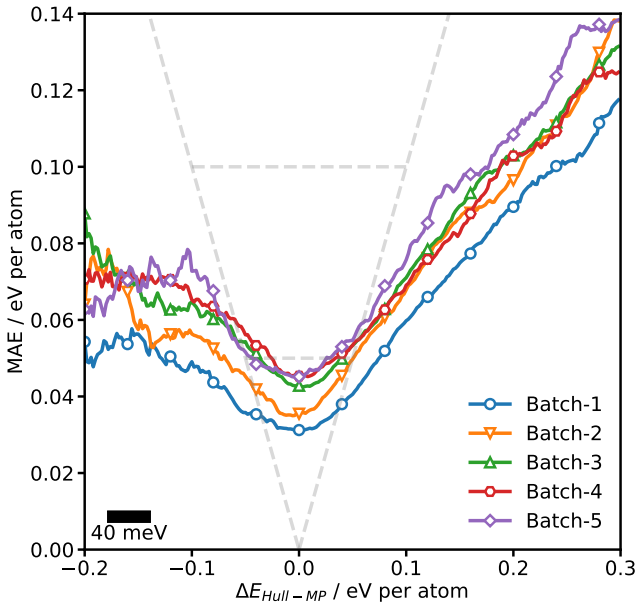


Figure 6.12: ***CGCNN-D* displays better generalisation performance than *Wren* over WBM batches.** Rolling mean absolute error of *CGCNN-D* on the batches of the WBM data set as the energy to the convex hull is varied. We take the pre-relaxation structure as input. A windowing period of 40 meV per atom is used when calculating the rolling average. As with the *Wren* model, the curves show that later batches, believed to be more likely to be chemically dissimilar to the training data, incur higher average errors. However, *CGCNN-D* displays a smaller deterioration in performance than *Wren* when asked to extrapolate.

this direction arises from the fact that, whilst sharing relaxed structures is increasingly commonplace, it is uncommon for researchers to share pre-relaxation structures or relaxation trajectories for *ab initio* crystal structure optimisations. In particular, the APIs of most high-throughput *ab initio* databases are not configured to allow easy access to this data. For example, to conduct the experiments in this chapter, it was necessary to individually query each pre-relaxation structure record from the Materials Project database. Processing all these queries took over 12 hours of wall-time, approximately 30 times the wall-time required to get hold of the equivalent relaxed structures despite the amount of data transferred being equivalent. Given the practical limitations on the availability of pre-relaxation training data, it is worth noting that the *Wren* approach only requires access to the relaxed structures. Assuming that data infrastructure can be changed to more readily accommodate sharing of pre-relaxation structures and relaxation trajectory data, further work should explore the limits of the *D*-variant models and the

discrete potential energy surface paradigm by looking at whether providing additional pre-relaxation structures from the relaxation trajectory improves model performance.

6.7 Computational Prospecting for Novel Stable Materials

Having established the promise of Wyckoff representation regression in predicting the stability of unseen materials, we deployed *Wren* on the prospective challenge of discovering new theoretically stable materials. For this stage, we trained *Wren* on the union of the MP and WBM data sets. This combined data set contains approximately 323k materials after canonicalisation and cleaning. We randomly sampled 5% of the data set to use as a test set and trained on the remaining 95%. The resulting model has a mean absolute error of 31 meV per atom on this test set, which is below the commonly quoted chemical accuracy level of 1 kcal per mol ($\sim$ 43 meV per atom) [44]. The model's accuracy as a function of training set size is shown in Figure 6.13, revealing a power-law relationship. Reassuringly, the model does not appear to saturate in performance, suggesting that the representation is rich enough and further increases in model performance can be unlocked given more data [177, 54, 56].

Whilst the coarse-grained space of Wyckoff representations is computably enumerable and far smaller than the infinite space of atomic coordinates, attempting materials discovery by exhaustively screening all possible Wyckoff representations is computationally inefficient as the prevalence of stable materials remains vanishingly low even in the coarse-grained space. To construct a design space with a higher expected prevalence [223], we draw inspiration from previous work [163, 164] and generate candidates for screening by making elemental substitutions in crystal structures that are near to the known convex hull.

To obtain our substitution probabilities, we extracted 39,164 ordered structures from the ICSD [13] and binned them according to their Wyckoff representations. Within each prototype, all pairs of structures are compared and we count which element substitutions (including self substitutions) would be needed to change one structure into the other [203]. We only consider substitutions where all Wyckoff positions sharing one element-type are changed simultaneously and not per position substitutions. Once normalised, the rows of the count matrix can be interpreted as substitution probabilities for each element.

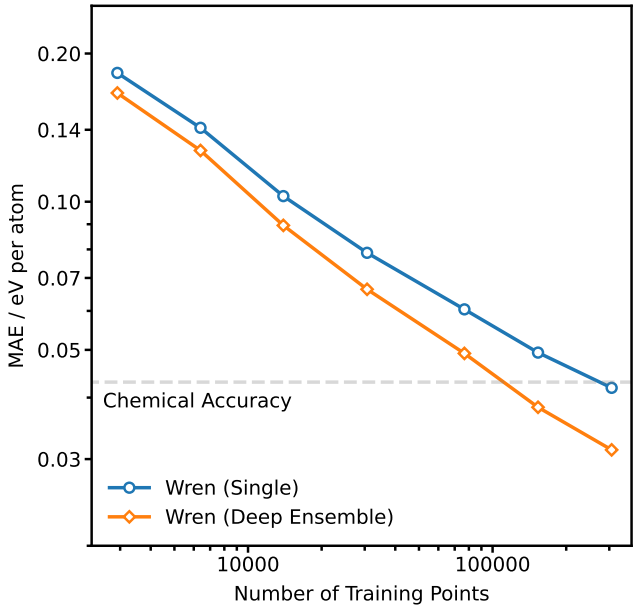


Figure 6.13: **Learning curve for the *Wren* model with increasing training data.** Learning curves for both single and *Deep Ensembles* of *Wren* models. The curves are produced by plotting the error on a fixed test set (here 5% of the union of the MP and WBM data sets) as the amount of data used to train the model is increased. A grey guideline indicates the level of chemical accuracy. The learning curve shows a power-law relationship between the amount of training data and the MAE of the trained model.

Using these data-mined probabilities, we generated a screening library of materials by substituting different elements into structures taken from the MP data set. We only consider initial structures from the MP data set with energies above the convex hull less than 100 meV per atom. This choice of this threshold means that we should be including most metastable structures within the MP data set. We consider ten different substitutions for each initial structure. Candidates that have the same composition as materials already present in the union of the MP and WBM data sets are removed from the library. Lanthanide and actinide-based materials and materials containing Noble elements were also excluded. This workflow produced a screening library of approximately 415k candidates.

Despite constraining our screening set to be close to known materials, it is likely that we are still asking the model to make predictions in areas of materials space where it lacks support from the training data. As shown on the WBM data set, uncertainty estimation allows us to reduce the risk in our materials screening process by factoring

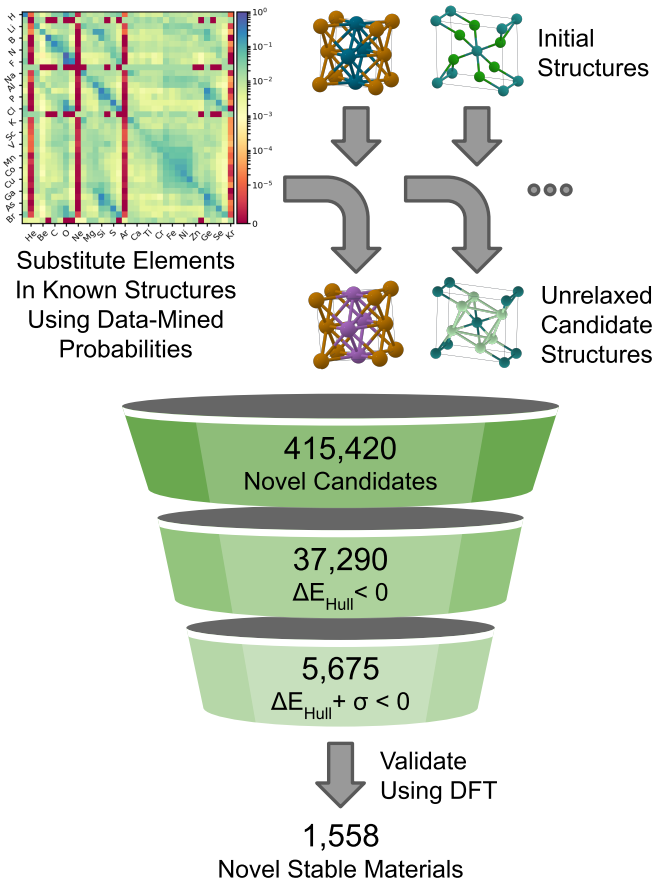


Figure 6.14: ***Wren* enables automated computational prospecting of new stable materials.** Data-mined substitution probabilities are used to generate novel candidates for screening. A heat-map of the data-mined log substitution probabilities for the first 36 main group elements is shown in the top left. The matrix captures known chemical trends e.g. that halogens can often be substituted for each other in crystal structures. Using the Wyckoff representation regression allows far more unrelaxed candidates to be considered than possible in conventional DFT-led high-throughput workflows. The funnel diagram shows the number of unrelaxed candidates that pass the different stability criteria. In total 4,721 out of 5,675 validation calculations completed. Of these 1,558 were below the MP convex hull, giving a precision of 33% amongst the completed calculations.

our model's uncertainty into our triage criterion. For simplicity we use the same simple uncertainty adjusted criterion considered previously, $\Delta\hat{E}_{\text{Hull-Pred}} + \hat{\sigma} < 0$. In total 5,675 candidates satisfied this screening criterion (see Figure 6.14).

We ran all selected candidates through a high-throughput DFT workflow which resulted in 4,721 completed calculations. Details of the DFT settings used are given in Appendix B.2. Of these, 1,558 were confirmed to be below the convex hull of the MP data set and 1370 below the convex hull of the union of the MP and WBM data sets. The precision for the completed calculations being below the convex hull of the MP data set was 33% and 29% for them being below the convex hull of the union of the MP and WBM data sets. These high precisions confirm the workflow's ability to accelerate materials discovery.

6.8 Discussion

Efforts to expedite the search for a wide variety of industrially desirable materials, e.g. thermoelectrics [255], piezoelectrics [235], fast-ion conductors [256], high voltage multivalent cathode materials [257], and caloric materials [258], are crucial for the transition to a low-carbon economy. DFT offers the tools we need to carry out this search *in silico* but does not scale readily to the enormous design spaces that exist in material science. The great filter on the viability of materials is their thermodynamic stability. Accordingly, being able to focus our calculation efforts on materials that are likely to be stable (or low enough in energy to be meta-stable) will facilitate more rapid discovery. In this chapter, we explored how we can use the concept of using coarse-graining to frame ideas about machine learning models for virtual screening. Drawing insight from organic chemistry and crystallography, we developed the framework of Wyckoff representation regression, *Wren*, and applied it to predict the formation energy of inorganic materials. *Wren* collapses the infinite search space of atomic coordinates into a combinatorially enumerable search space, enabling efficient data-driven exploration of materials space. On a set of challenging tasks curated from the literature, we find that our approach can accurately map the phase diagrams of unseen chemical systems with up to 10-fold early enrichment and offers an additional $\sim$ 3-fold improvement in precision when combined with materials screening strategies that use the chemical similarity of substituted elements to identify novel theoretically stable materials.

A common criticism of machine learning work is that it is often retrospective in nature – "amongst this data set of calculations we can rediscover these materials that have previously been discovered more efficiently". Whilst such work is valuable for advancing the state of the field, prospective discovery should be the principal goal of applying machine learning to material science. To this end, we developed a materials prospecting pipeline using *Wren*. As a proof-of-concept, we generated an extensive library of novel materials candidates by making chemical substitutions into materials recorded in the Materials Project. *Wren* was used to screen this library and candidates satisfying an uncertainty adjusted screening criterion were then relaxed using DFT. Using this workflow, 1,558 new materials below the known convex hull of the Materials Project were identified from just 5,675 attempted calculations. These results demonstrate that leveraging Wyckoff representation regression allows for more efficient and extensive expansion of computational material science databases.

Throughout this chapter, we explored how *Wren* compares against structure-based baselines. The central takeaway is that using structure-based models trained on equivalent data to screen pre-relaxation structures is less accurate than *Wren*. Whilst exploring the connection between *Wren*'s predictions and the true potential energy surface, we framed the idea of hypothetical discrete potential energy surface models that would map each structure within a local basin of attraction to the energy of its minima. We approximate this setup by training so-called *D*-variant models that map both the pre-relaxation and relaxed structures for materials in our training sets to the energies of the relaxed structure. Combining this simple idea with the *CGCNN* architecture [82] yields a model, referred to as *CGCNN-D*, that is comparable to *Wren* in our experiments on the WBM and TAATA data sets. Notably, *Wren* offers better early performance, but *CGCNN-D* gave better overall performance when the entirety of the WBM data set was considered.

Thus far, we have only investigated *D*-variant models using the initial structure and the final structure from the relaxation trajectory. If the *D*-variant framework is well constructed, the model performance should improve if provided with additional pre-relaxation structures during training. The most accessible source of additional relaxation structures would be to take intermediate structures from the relaxation trajectory. However, these structures are unlikely to capture the diversity of structures in the local basin of attraction. Collecting multiple relaxation trajectories for each structure would improve diversity. However, further work would be needed to assess whether the

performance improvements from generating additional relaxation trajectories outweigh the benefit of using that same computational budget to screen new materials directly.

The training setup for the *D*-variant models is closely linked to the idea of data-augmentation – processing labelled data to generate additional inputs that are assigned the same training label. When looking at image data in computer vision applications, it is common to use data-augmentation pipelines that crop, rotate, colour distort, or even warp training images to improve model generalisation performance. Most data augmentation strategies involve additional processing that has a negligible cost compared to labelling additional data. For crystal structures shaking atom positions or stretching lattices could be employed as strategies to induce diversity and potentially robustness at negligible cost. However, unlike with images, where a cropped image of a dog is still an image of a dog, the risk with crystal structures is that such actions will cause the structures to jump into different basins within the global potential energy surface, potentially harming model performance. This risk could be alleviated by weighing augmented examples according to some notion of distance between the pre-relaxation and relaxed structures [60]. Potential sources of distance metrics between crystal structures include AFLOW-XtalFinder [259], CMPZ [260] or COMPSTRU [261]. Despite the open questions that remain, *D*-variant models appear to be highly promising and deserving of future exploration to ascertain how their strong performance arises and how they might be improved.

Chapter 7

Outlook

Throughout this thesis, we have explored how data-driven approaches, in particular those based on machine learning, can be leveraged in materials discovery workflows. Although we have explored a broad range of topics and applications, the underlying question is always the same – how can we come up with data structures, models and strategies that allow us to handle the complexities of materials data?

In Chapter 3 this question materialised as how to bridge the gap between the available data on the lattice parameters of cuprates and the desired data about the apical and in-plane Cu-O distances needed to examine structure-function relationships. A simple calibration approach was used to estimate the apical and in-plane distances and explore how they are connected to T_c. The existence of an unexplored region of materials space was established that, if experimentally accessible, may yield higher T_c cuprate materials.

Chapter 4 grappled with how to represent materials compositions in a manner that enabled us to leverage improvements in modern neural network architectures. The proposed solution was to represent material compositions as dense weighted graphs. This choice of data structure allowed the use of a message-passing neural network architecture. The resulting *Roost* model was shown to demonstrate improved sample efficiency compared to models that represent materials compositions using fixed-length vectors.

Whilst machine learning has facilitated rapid improvement in retrosynthetic planning for small molecules and fine chemicals [262], there has been minimal progress on predictive synthesis for inorganic materials. Chapter 5 explored whether a pure machine learning solution was sufficient to predict the products of inorganic materials synthesis reactions from their precursors and processing sequences. While superior to the baseline models

considered, the model developed here fell short of the accuracy required to be chemically discriminative for forward reaction prediction. Nonetheless, the model was shown to have learnt meaningful embeddings of chemical reactions that are useful in their own right. In particular, these embeddings can be used to flag erroneous database entries, facilitating improved and automated data cleaning and curation.

An important stepping stone to the realisation of inverse design are models capable of reliably generating stable atomic structures. In Chapter 6 the idea of Wyckoff representation regression is explored, resulting in the *Wren* model. The Wyckoff representation is a coordinate-free coarse-grained representation of a material that builds upon crystallographic concepts. Unlike the infinite space of atomic coordinates, the Wyckoff representation is discrete and computably enumerable. The *Wren* model is shown to accelerate materials discovery in both retrospective and prospective experiments, predicting the formation energies of inorganic materials with an MAE as low as 31 meV per atom without access to the relaxed structure.

7.1 Directions for Future Research

The work presented in this thesis is broad in its scope. Inevitably there are many questions and lines of inquiry exposed but left unanswered within this thesis. Some of the most prominent directions that future work should seek to address are outlined below.

Knowledge Graphs for Materials Synthesis

Knowledge graphs are data structures that use an ontology to organise and integrate data from multiple sources [263]. There have already been several efforts that attest to the potential of employing knowledge graphs to rationalise and extract new insights from materials data [264, 265, 198, 199]. One interesting line of work has been to recast the idea of the convex hull in a graph framework where edges represent tie-lines between stable phases [265]. Ideas such as the nobility of materials can then be extracted in a purely data-driven manner by analysing the structure of the resulting phase stability network. In Ref. [198] the evolution dynamics of the phase stability network are analysed. They show that an accurate model for predicting whether hypothetical structures are likely to be synthesiseable at a given point in time can be constructed given the local topological features of the network.

If we adopt a top-down view, the only reasonable data structure to represent a large corpus of information about materials synthesis is a knowledge graph. However, as of yet there have been no publicly reported or accessible realisations of a knowledge graph for materials synthesis. Most synthesis procedures involve the mixing of reactants, therefore the framework needs to handle directed hyperedges that connect multiple precursors to multiple products. Building on the forward reaction models explored in Chapter 5, a key challenge will be developing a hyperedge schema that can handle variable length processing sequences. The top level of the hyperedge schema should correspond to the type of synthesis, i.e. solid-state, sol-gel, or hydrothermal synthesis [266]. Below this, the synthesis procedure should be recorded as a directed acyclic chain of operations carried out by experimentalists [31]. Detailed attributes describing the operation conditions and equipment would then be associated with the backbone operations in a hierarchical manner.

The largest outstanding issue to address in the construction of a knowledge graph is how to connect different synthesis graphs into a cohesive knowledge graph. The natural connection points are nodes that represent materials, but this only applies in the case of materials with integer stoichiometries. One strategy to deal with non-stoichiometric materials would be to introduce an additional node type to represent stoichiometric parent structures. Non-stoichiometric products could then be connected to these parent structures in order to facilitate data mining. The inclusion of such nodes for stoichiometric parent structures provides a natural place to incorporate *ab initio* data into the proposed knowledge graph. Given that synthesis is controlled by the interplay of thermodynamics and kinematics, the availability of *ab initio* thermodynamics data should allow more accurate reasoning models to be constructed from the graph.

We are still in the rudimentary stages of constructing an ontology suitable for describing the complexities of materials synthesis. Nonetheless, future work should attempt to realise a preliminary knowledge graph for materials synthesis using the data set produced in Ref. [34] in conjunction with *ab initio* data from the Materials Project [98]. This graph would allow for questions about synthesis to be formulated as link prediction tasks, e.g. given a set of precursors and products that can be joined by a hydrothermal synthesis hyperedge, can they also be joined by a solid-state synthesis hyperedge?

Prediction of Novel Theoretically Stable Materials

In Chapter 6 we began to explore *D*-variant models – structure-based models trained to map both the pre-relaxation structure and relaxed structure to the energy of the relaxed structure. Ongoing work should explore this paradigm in greater depth, in particular whether meaningful connections can be made between the discrete potential energy surface and disconnectivity graphs that are often used to visualise energy landscapes [267, 268].

One of the key conceptual strengths of the *Wren* approach is that a direct mapping exists between the target property and a single input. This is not the case with the *D*-variant approach as multiple pre-relaxation structures exist. Consequently, an interesting direction to explore is how well *CGCNN* can predict the Wyckoff representations of both relaxed and pre-relaxation structures. If the model can solve this highly multi-class classification problem, it would suggest that the model could learn the equivalent mapping to *Wren*, albeit indirectly. The Wyckoff representation is non-local, whereas CGCNN-type models leverage only local information. We saw that initially *Wren* outperformed *CGCNN-D* when screening the WBM data set, both in terms of accuracy on earlier batches and initial precision when screening the entire WBM data set. This is a clear demonstration of the fact that the models are learning different signals in the data. Consequently, employing a selection by committee strategy that uses both *Wren* and *CGCNN-D* in parallel may lead to additional increases in precision.

Progress in machine learning research is often driven by benchmarking. The availability of challenging and informative benchmark tasks can help spur advances in the field. For molecular systems, the QM9 [232] and MD17 [269] data sets are the go-to benchmark tasks. Unfortunately, there is no such standardisation for inorganic systems. High-throughput databases tend to be dynamic, with additional data being added over time. Therefore to benchmark methods in a consistent manner, authors must re-implement and produce results for competing methods, opening the door to unconscious biases that can suppress the performance of competing methods, e.g. less thorough hyper-parameter searching. A notable attempt at standardisation is the *MatBench* suite [270]. *MatBench* looks at accuracy across 13 regression and classification tasks spanning a range of data set sizes. Despite its strengths as a general-purpose benchmark suite, *MatBench* fails to directly address the intricacies of accelerated materials discovery, where the key metric is not only accuracy but the precision and enrichment when selecting new candidates –

note how in Chapter 6 *CGCNN* gave similar enrichment to *Wren* despite *Wren* being noticeably more accurate. New benchmark tasks that focus on enrichment, potentially building on the WBM [164] and TAATA [235] experiments presented in this thesis, are needed to focus efforts towards developing models that directly address the challenges of materials discovery.

Curation efforts for new benchmark tasks should seek to provide complete relaxation trajectory data. These trajectories are generated as a by-product of obtaining related structures but are informative in their own right. Accessing this data is not a trivial task. At present, none of the major high-throughput *ab initio* databases allow programmatic access to this data. In some instances, this is because there may be potential copyright issues with sharing it en masse, e.g. in cases where ICSD structures have been taken as the initial inputs. In cases where the relaxation trajectories are composed of different runs conducted with different settings, the potential of stitching together the different runs based on energy differences between ionic steps should be explored. The stitching process would introduce systematic errors into the energy labels of structures in earlier runs. However, it is plausible that the increase in accuracy due to the additional diversity and abundance of training data should compensate for this imperfection, making this a worthwhile direction of research.

Discovering Pareto Optimal Functional Materials

Moving beyond identifying novel stable materials, the ultimate goal of materials discovery is to predict and then experimentally realise novel functional materials. Many of the most useful classes of functional materials are characterised by trade-offs between competing physical effects, for example, good thermoelectric materials require high electrical conductivity to co-exist with low thermal conductivity, and good dielectric materials require highly polarisable materials that are also strong insulators. Whilst several works have explored optimisation of a single functional property, for example, hardness [168, 244], multi-objective optimisation of competing properties is much more challenging. Attempting to solve challenging problems exposes the principal bottlenecks in current workflows and allows us to focus our efforts on how to resolve them. To this end, we have begun exploratory work investigating the use of *Wren* to search for high-performance dielectric materials.

Early results indicate that the major bottleneck on our ability to identify new high-performance dielectric materials is our ability to reliably predict material bandgaps. The key pathology is the prediction of false insulators – metals predicted to have large bandgaps. As the amount of training data is increased, the average error of bandgap predictions fall, but false metal and false insulator misclassifications remain as critical failure modes. Consequently, alternative approaches for estimating the bandgap should be explored, for example, predicting the electronic density of states and then estimating the bandgap from the predicted spectra. More generally, effective architectures for predicting the spectral properties of materials would be valuable across a wide range of material science applications [271].

Afterword

The work contained in this thesis required $\sim$ 5,000 GPU hours and $\sim$ 100,000 CPU hours of computational resources. The majority of the GPU compute was carried out on Nvidia P100 cards with a TDP of 250W provided by the Cambridge Service for Data Driven Discovery (CSD3). The majority of the CPU resources were provided by the Swedish National Infrastructure for Computing (SNIC) Tetralith cluster, which uses 16 core Intel Xeon Gold 6130 processors with a TDP of 125W. In total, we estimate that a maximum of 2 MWh of electrical work was used to enable this research. This figure is an upper bound as it assumes maximum power draw at all times. In the United Kingdom, the carbon cost of electricity is $\sim$ 0.25 kg of CO_2 per KWh. Therefore, our electricity usage corresponds to a carbon cost of approximately 0.5 tonnes. The other significant carbon cost of this research was attending the Machine Learning for Physics and Physics of Learning Long Program at the Institute for Pure & Applied Mathematics (IPAM) at the University of California, Los Angeles. The carbon cost of a return trip to the west coast of America from the United Kingdom is approximately 3.5 tonnes at the time of writing. In total, we conclude that our research has led to the emission of approximately 4 tonnes of CO_2 – about 6 months of emissions for the average person in the United Kingdom at the time of writing.

Appendix A

Model Implementation Details

A.1 Roost Architecture and Training

In Chapter 4, we adopted the same architecture and hyper-parameters for all the problems investigated. These choices were made based on heuristic ideas from other graph convolution-based architectures.

We use the *Matscholar* embedding from [173] for which $n = 200$. We chose an internal representation size of $d = 64$ based on the *CGCNN* model [82].

We opted to use 3 message passing layers based on the default configuration of the *MEGNet* model [83]. For the neural networks inside our weighted soft-attention-based pooling function, we drew inspiration from the *GAT* architecture presented in Ref. [175] and used single-hidden-layer neural networks with 256 hidden units and LeakyReLU activation functions for $f^{t,m}(...)$ and $g^{t,m}(...)$. For the reference model, we used 3 attention heads in each of message passing layers.

The output network used for the reference model is a deep neural network with 5 hidden layers and ReLU activation functions. The number of hidden units in each layer is 1024, 512, 256, 126, and 64 respectively. Skip connections were added to the output network to help tackle the vanishing gradient problem [79].

The sizes of various networks were selected to ensure that our model was appropriately over-parameterised for the OQMD data set. Our reference model has 2.4 million parameters – approximately 10x the size of the OQMD training set used.

We configure the model to predict both a mean and aleatoric contribution to the variance such that we can use a robust L1 loss [94]. For numerical reasons the model is

made to predict $\log\left(\hat{\sigma}_a(x_i)\right)$ which is then exponentiated to get $\hat{\sigma}_a(x_i)$ when estimating the aleatoric uncertainty. We use ensembles of $W = 10$ to estimate the epistemic contribution within the *Deep Ensemble* framework.

All transfer learning experiments were conducted using warm-restarts with all of the model parameters being re-optimised given the new data. This was observed to give better performance than freezing the message passing layers and only re-optimising the weights of the output neural network.

The mean-based pooling function in the ablation study used single-hidden-layer neural networks with 256 hidden units and LeakyReLU activation functions for $g^t(...)$.

All the neural network-based models examined in both the main results and the ablation study were trained using the Adam optimiser and fixed learning rate of 3×10^{-4}. A mini-batch size of 128 and weight decay parameter of 10^{-6} were used for all the experiments. The models were trained for 250 epochs.

A.1.1 Baseline Models

For our baseline models we use the Random Forest implementation from `scikit-learn` [272] and use `Matminer` [273] to generate the *Magpie* features. The max features and number of estimators for the Random Forest were set to 0.25 and 200 respectively.

A.1.2 Data Availability

The OQMD data set used in Chapter 4 was collated from the openly available Open Quantum Materials Database at http://oqmd.org [36, 37]. We use the subset of OQMD studied in Ref. [165].

The MP data set used in Chapter 4 was collated using the Materials API [274] from the openly available Materials Project database at https://materialsproject.org [98].

The EX data set used in Chapter 4 is available alongside Ref. [172].

A.2 Reaction Graph Architecture and Training

Reactions with non-stoichiometric or organic precursors and targets, and reactions with pure products or only one precursor were removed from the dataset. This reduced the dataset size to 16,231 reactions with up to ten precursors, and 11,083 reactions for the

ablation study with up to three precursors. This was randomly split into training and test sets with an 80:20 ratio.

Processing action sequences ranged from 0 to 16 steps in length and were represented by 7 unique action types: {Dry Mixing, Mixing in solution, Quenching, Shaping, Heating, Drying, Liquid Grinding}.

Precursor stoichiometries were extracted along with their molar ratios from the balanced chemical equation for each reaction. Where precursors or targets were themselves a mixture of different materials/phases, their stoichiometries were added, weighted by their amounts.

Targets were extracted as 81-dimensional stoichiometric vectors. The set of elements present in the precursors was used to construct the alchemy mask.

The model was implemented in `PyTorch` [72] and `Matminer` [273] was used to generate the *Magpie* features for the precursors. The element movers distance (EleMD) [204] was calculated using the `pot` package.

An embedding dimension of 8 and a hidden state dimension of 32 were used for the processing actions autoencoder. This led to a sequence reconstruction accuracy of $\sim 87\%$. Reconstruction accuracies of over 95% were achieved using a hidden dimension of 64 but these longer embeddings were not used as they did not noticeably improve product prediction accuracy but did significantly increase the computational cost of training.

The learnable precursor embedding size was set to 128, and 5 message-passing layers were used, each with 3 attention heads. The hidden layer dimension was chosen to be 256 for both f and g. The output network of the product element prediction model had hidden layer dimensions of 256, 512, 512 and 256 respectively with ReLU activation. The threshold output probability for elemental presence after applying a sigmoidal non-linearity on the output was 0.5. The stoichiometry prediction model had hidden layer dimensions of 256, 256, 256, 256, 128, 128 and 64 respectively with ReLU activation and skip connections between layers.

Learning rates were chosen using a heuristic search where we choose a rate one magnitude lower than that which causes the training loss to diverge. A mini-batch size of 256 and the *Adam* optimiser [73] were used to train the reaction graph models.

The ensembles used to estimate the epistemic uncertainty in both the *Magpie* baseline and reaction graph models were constructed using 5 individual models trained on the same data with different initialisation.

A.2.1 Data Availability

The exact data used in Chapter 5 can be found at https://doi.org/10.6084/m9.figshare.9722159.v3. The original authors release more recent versions of the synthesis data set at https://github.com/CederGroupHub/text-mined-synthesis_public.

A.3 Wren Architecture and Training

The bulk of the *Wren* architecture directly mimics that of *Roost* [2]. The principle difference between the two architectures comes in that the nodes on the dense graph now represent different Wyckoff positions rather than just the different elemental species. The elemental information is encoded using the 'matscholar' embedding from [173] which has a dimensionality of $d_{el} = 200$. The remainder of the node embedding comprises the Wyckoff embedding plus the fractional multiplicity of that Wyckoff position within the unit cell. The combined dimensionality of the Wyckoff proportion of the embedding is $d_{wyk} = 445$.

To reduce the total dimension of the node embeddings, we project both the elemental and Wyckoff embeddings into lower-dimensional spaces using learnt affine transformations. The low dimension embeddings are then concatenated to give the node embeddings. In this work we chose $d^*_{el} = 32$ and $d^*_{wyk} = 32$ giving a total dimensionality of $d = 64$ for the node embeddings.

We use 3 message passing layers, each with a single attention head. We chose single-hidden-layer neural networks with 256 hidden units and LeakyReLU activation functions for both parts of the soft-attention mechanism. The output network consists of a feed-forward neural network with skip connections and ReLU activation functions. The output network used has 4 hidden layers containing 256, 256, 128, and 64 hidden units respectively.

Throughout this work, we train Deep Ensembles of 10 models starting from different random initialisations for each data setup and architecture considered.

All the models examined in this work were trained using the *AdamW* optimiser [275] with a fixed learning rate of 3×10^{-4}. A mini-batch size of 128 and a weight decay parameter of 10^{-6} were used for all the experiments. The models were trained for 400 epochs.

A.3.1 Baseline Models

The *Roost* baseline used the architectural hyper-parameters detailed in Appendix A.1. The *CGCNN* and *CGCNN-D* baseline models used the default architectural hyper-parameters suggested in Ref. [82]. All baseline models based on neural networks underwent identical training procedures to that of the *Wren* model.

The *Voronoi* and *Voronoi-D* baseline models made use of `Matminer` [273] to evaluate the Voronoi crystal structure attribute and *Magpie* composition features. The `scikit-learn` [272] Random Forest [47] implementation was used to train the model. The number of estimators was set to 150 and default values were used for the other hyper-parameters following the example in `Matminer`.

A.3.2 Data Availability

The relaxed structures for the TAATA data set, along with the ICSD references for the source structures, are available from the original authors upon request [235].

The MP data set used for Chapter 6 was queried from https://materialsproject.org [98] using the Materials API [274]. The data was queried from Supplemental Database Release V2021.03.22.

The relaxed structures for WBM data set used for Chapter 6 are available from https://archive.materialscloud.org/record/2021.68. The pre-relaxation structures were obtained from the original authors on request [164].

The ICSD [13] was accessed under license from FIZ Karlsruhe—Leibniz Institut for Information Infrastructure at https://icsd.products.fiz-karlsruhe.de.

Appendix B

Density Functional Theory Settings

B.1 Strain Dependence of Cuprates

In Chapter 3 Kohn-Sham DFT with the plane wave pseudopotential code `CASTEP` [111] was used to model how La_2CuO_4, $YBa_2Cu_3O_7$, and $HgBa_2Ca_2Cu_3O_8$ cuprates might behave under strain. Monkhorst-Pack grids were used for sampling the reciprocal space with k-point spacing less than $2\pi \times 0.05$ Å^{-1}. On-the-fly generated core-corrected ultrasoft pseudopotentials [276] from the `CASTEP C18` library were used with the PBE exchange-correlation functional [97] . A plane wave cut off energy of 700 eV was used. The equilibrium cell volumes of the structures were optimised with residual stresses less than 0.05 GPa. Once the equilibrium cell volumes were obtained, following optimisations were performed with fixed-cell sizes corresponding to strains in the c and a-b directions ranging from -4% to +4%. The ionic positions were relaxed until the maximum force was less than 0.01 eV Å^{-1} in all calculations. The `AiiDA` framework was used to manage and automate the calculations [277].

The calculations presented in this chapter were conducted by Dr. Bonan Zhu using resources provided by the Cambridge Service for Data-Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service, provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council.

B.2 Stability of Novel Candidate Materials

The validation of predictions in our materials prospecting pipeline presented in Chapter 6 was carried out using Kohn-Sham DFT with the plane wave pseudopotential code `VASP` [278, 279]. Projector augmented wave type pseudopotentials [280, 281] were used with the PBE generalised gradient approximation exchange-correlation functional [97]. All calculations were done using a 520 eV plane-wave energy cutoff. The pseudopotentials and Hubbard-U values were selected to ensure compatibility with data contained in the Materials Project. The `MaterialsProjectCompatibility` scheme as implemented in `pymatgen` [282] was applied to allow the mixing of GGA and GGA+U calculations [115]. We used the High-Throughput Toolkit (`httk`) introduced in Ref. [283] to manage the calculations.

The prospective calculations presented were conducted by Abhijith S. Packaral using computational resources provided by the Swedish National Infrastructure for Computing (SNIC) at NSC partially funded by the Swedish Research Council through grant agreement No. 2018-05973.

References

[1] Rhys EA Goodall, Bonan Zhu, Judith L MacManus-Driscoll, and Alpha A Lee. Materials informatics reveals unexplored structure space in cuprate superconductors. *Advanced Functional Materials*, page 2104696, 2021.

[2] Rhys E. A. Goodall and Alpha A. Lee. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nature Communications*, 11(1):1–9, 2020.

[3] Shreshth A Malik, Rhys E A Goodall, and Alpha A Lee. Predicting the outcomes of material syntheses with deep learning. *Chemistry of Materials*, 33(2):616–624, 2021.

[4] Huiwen Ji, Alexander Urban, Daniil A Kitchaev, Deok-Hwang Kwon, Nongnuch Artrith, Colin Ophus, Wenxuan Huang, Zijian Cai, Tan Shi, Jae Chul Kim, et al. Hidden structural and chemical order controls lithium transport in cation-disordered oxides for rechargeable batteries. *Nature Communications*, 10(1):1–9, 2019.

[5] Dongchang Chen, Juhyeon Ahn, and Guoying Chen. An overview of cation-disordered lithium-excess rocksalt cathodes. *ACS Energy Letters*, 6(4):1358–1376, 2021.

[6] WD Johnston, RR Heikes, and D Sestrich. The preparation, crystallography, and magnetic properties of the $Li_xCo_{1-x}O$ system. *Journal of Physics and Chemistry of Solids*, 7(1):1–13, 1958.

[7] K Mizushima, PC Jones, PJ Wiseman, and John B Goodenough. Li_xCoO_2 (0<x<-1): A new cathode material for batteries of high energy density. *Materials Research Bulletin*, 15(6):783–789, 1980.

[8] JM Longo and PM Raccah. The structure of La_2CuO_4 and $LaSrVO_4$. *Journal of Solid State Chemistry*, 6(4):526–531, 1973.

[9] J George Bednorz and K Alex Müller. Possible high T_c superconductivity in the Ba-La-Cu-O system. *Zeitschrift für Physik B Condensed Matter*, 64(2):189–193, 1986.

[10] Sydney R Hall, Frank H Allen, and I David Brown. The crystallographic information file (CIF): A new standard archive file for crystallography. *Acta Crystallographica Section A: Foundations of Crystallography*, 47(6):655–685, 1991.

[11] I. D. Brown and B. McMahon. CIF: The computer language of crystallography. *Acta Crystallographica Section B: Structural Science*, 58(3):317–324, June 2002.

[12] Colin R. Groom, Ian J. Bruno, Matthew P. Lightfoot, and Suzanna C. Ward. The Cambridge Structural Database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 72(2):171–179, April 2016.

[13] Mariette Hellenbrandt. The Inorganic Crystal Structure Database (ICSD) – present and future. *Crystallography Reviews*, 10(1):17–22, January 2004.

[14] Aron Walsh. The quest for new functionality. *Nature Chemistry*, 7(4):274–275, April 2015.

[15] Daniel W Davies, Keith T Butler, Adam J Jackson, Andrew Morris, Jarvist M Frost, Jonathan M Skelton, and Aron Walsh. Computational screening of all stoichiometric inorganic materials. *Chem*, 1(4):617–627, 2016.

[16] Gabriel Briceno, Hauyee Chang, Xiaodong Sun, Peter G Schultz, and X-D Xiang. A class of cobalt oxide magnetoresistance materials discovered with combinatorial synthesis. *Science*, 270(5234):273–275, 1995.

[17] Shiyan Ding, John M Gregoire, Joost J Vlassak, and Jan Schroers. Solidification of Au-Cu-Si alloys investigated by a combinatorial approach. *Journal of Applied Physics*, 111(11):114901, 2012.

[18] Martin L Green, CL Choi, JR Hattrick-Simpers, AM Joshi, I Takeuchi, SC Barron, E Campo, T Chiang, S Empedocles, JM Gregoire, et al. Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Applied Physics Reviews*, 4(1):011105, 2017.

[19] Alfred Ludwig. Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods. *npj Computational Materials*, 5(1):70, 2019.

[20] Gerbrand Ceder, Y-M Chiang, DR Sadoway, MK Aydinol, Y-I Jang, and Biying Huang. Identification of cathode materials for lithium batteries guided by first-principles calculations. *Nature*, 392(6677):694–696, 1998.

[21] Anubhav Jain, Yongwoo Shin, and Kristin A Persson. Computational predictions of energy materials using density functional theory. *Nature Reviews Materials*, 1(1):1–13, 2016.

[22] X-D Xiang, Xiaodong Sun, Gabriel Briceno, Yulin Lou, Kai-An Wang, Hauyee Chang, William G Wallace-Freedman, Sung-Wei Chen, and Peter G Schultz. A combinatorial approach to materials discovery. *Science*, 268(5218):1738–1740, 1995.

[23] Earl Danielson, Josh H Golden, Eric W McFarland, Casper M Reaves, W Henry Weinberg, and Xin Di Wu. A combinatorial approach to the discovery and optimization of luminescent materials. *Nature*, 389(6654):944–948, 1997.

[24] Hideki Minami, Kenji Itaka, Parhat Ahmet, Daisuke Komiyama, Toyohiro Chikyow, Mikk Lippmaa, and Hideomi Koinuma. Rapid synthesis and scanning probe analysis of $Ba_xSr_{1-x}TiO_3$ composition spread films on a temperature gradient Si (100) substrate. *Japanese Journal of Applied Physics*, 41(2a):L149, 2002.

[25] Benjamin Burger, Phillip M Maffettone, Vladimir V Gusev, Catherine M Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M Alston, Buyi Li, Rob Clowes, et al. A mobile robotic chemist. *Nature*, 583(7815):237–241, 2020.

[26] Shuang Shuang, Honghua Li, Gang He, Yong Li, Jiangtao Li, and Xiangmin Meng. High-throughput automatic batching equipment for solid state ceramic powders. *Review of Scientific Instruments*, 90(8):083904, 2019.

[27] Akira Miura, Christopher J Bartel, Yosuke Goto, Yoshikazu Mizuguchi, Chikako Moriyoshi, Yoshihiro Kuroiwa, Yongming Wang, Toshie Yaguchi, Manabu Shirai, Masanori Nagao, et al. Observing and modeling the sequential pairwise reactions that drive solid-state ceramic synthesis. *Advanced Materials*, page 2100312, 2021.

[28] Kevin Maik Jablonka, Michaël Zasso, Luc Patiny, Nicola Marzari, Giovanni Pizzi, Berend Smit, and Aliaksandr V Yakutovich. Connecting lab experiments with computer experiments: Making "routine" simulations routine. *ChemRxiv preprint chemrxiv:2021-h3381-v2*, 2021.

[29] Matthew C Swain and Jacqueline M Cole. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904, 2016.

[30] Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of Chemical Information and Modeling*, 59(9):3692–3702, 2019.

[31] Edward Kim, Kevin Huang, Olga Kononova, Gerbrand Ceder, and Elsa Olivetti. Distilling a materials synthesis ontology. *Matter*, 1(1):8–12, 2019.

[32] Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, University of Cambridge, 2012.

[33] Wengong Jin, Connor W. Coley, Regina Barzilay, and Tommi S. Jaakkola. Predicting organic reaction outcomes with Weisfeiler-Lehman network. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2607–2616, 2017.

[34] Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data*, 6(1):203, 2019.

[35] AR Miedema, R Boom, and FR De Boer. On the heat of formation of solid alloys. *Journal of the Less Common Metals*, 41(2):283–298, 1975.

[36] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher M Wolverton. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *JOM*, 65(11):1501–1509, November 2013.

[37] Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1(1):1–15, 2015.

[38] R. B. Jadrich, J. A. Bollinger, B. A. Lindquist, and T. M. Truskett. Equilibrium cluster fluids: pair interactions via inverse design. *Soft Matter*, 11:9342–9354, 2015.

[39] Rhys EA Goodall and Alpha A Lee. Data-driven approximations to the bridge function yield improved closures for the Ornstein-Zernike equation. *Soft Matter*, 17(21):5393–5400, 2021.

[40] Alberto Franceschetti and Alex Zunger. The inverse band-structure problem of finding an atomic configuration with given electronic properties. *Nature*, 402(6757):60–63, 1999.

[41] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[42] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

[43] Juhwan Noh, Jaehoon Kim, Helge S Stein, Benjamin Sanchez-Lengeling, John M Gregoire, Alan Aspuru-Guzik, and Yousung Jung. Inverse design of solid-state materials via a continuous representation. *Matter*, 1(5):1370–1384, 2019.

[44] Felix A. Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S. Schoenholz, George E. Dahl, Oriol Vinyals, Steven Kearnes, Patrick F. Riley, and O. Anatole von Lilienfeld. Prediction errors of molecular machine learning models lower than hybrid DFT error. *Journal of Chemical Theory and Computation*, 13(11):5255–5264, November 2017.

[45] Christopher M. Bishop. *Pattern recognition and machine learning, 5th Edition*. Information science and statistics. Springer, 2007.

[46] Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning*. MIT Press, 2016.

[47] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[48] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 785–794. ACM, 2016.

[49] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.

[50] Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

[51] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14), April 2007.

[52] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108(5):058301, January 2012.

[53] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

[54] Felix Faber, Alexander Lindmaa, O. Anatole von Lilienfeld, and Rickard Armiento. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry*, 115(16):1094–1101, August 2015.

[55] Haoyan Huo and Matthias Rupp. Unified representation of molecules and crystals for machine learning. *arXiv preprint arXiv:1704.06439*, 2017.

[56] Felix A Faber, Anders S Christensen, Bing Huang, and O Anatole von Lilienfeld. Alchemical and structural distribution based representation for universal quantum machine learning. *The Journal of Chemical Physics*, 148(24):241717, 2018.

[57] Michael J. Willatt, Félix Musil, and Michele Ceriotti. Atom-density representations for machine learning. *The Journal of Chemical Physics*, 150(15):154110, April 2019.

[58] Anders S Christensen, Lars A Bratholm, Felix A Faber, and O Anatole von Lilienfeld. FCHL revisited: Faster and more accurate quantum machine learning. *The Journal of Chemical Physics*, 152(4):044107, 2020.

[59] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian Approximation Potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 104(13), April 2010.

[60] Albert P Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R Kermode, Gábor Csányi, and Michele Ceriotti. Machine learning unifies the modeling of materials and molecules. *Science Advances*, 3(12):e1701816, 2017.

[61] Luca M. Ghiringhelli, Jan Vybiral, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler. Big data of materials science: Critical role of the descriptor. *Physical Review Letters*, 114(10):105503, March 2015.

[62] Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, and Luca M. Ghiringhelli. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials*, 2(8):083802, August 2018.

[63] Christopher J Bartel, Christopher Sutton, Bryan R Goldsmith, Runhai Ouyang, Charles B Musgrave, Luca M Ghiringhelli, and Matthias Scheffler. New tolerance factor to predict the stability of perovskite oxides and halides. *Science Advances*, 5(2):eaav0693, 2019.

[64] Stephan R. Xie, Gregory R. Stewart, James J. Hamlin, Peter J. Hirschfeld, and Richard G. Hennig. Functional form of the superconducting critical temperature from machine learning. *Physical Review B*, 100(17):174513, 2019.

[65] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):16028, August 2016.

[66] Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the band gaps of inorganic solids by machine learning. *The Journal of Physical Chemistry Letters*, 9(7):1668–1673, April 2018.

[67] Ekin D Cubuk, Austin D Sendek, and Evan J Reed. Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data. *The Journal of Chemical Physics*, 150(21):214701, 2019.

[68] Fang Ren, Logan Ward, Travis Williams, Kevin J Laws, Christopher Wolverton, Jason Hattrick-Simpers, and Apurva Mehta. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Science Advances*, 4(4):eaaq1566, 2018.

[69] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[70] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[71] Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18:153:1–153:43, 2017.

[72] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan

Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019.

[73] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[74] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[75] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL, 2014.

[76] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997.

[77] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551, 1989.

[78] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012.

[79] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[80] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*. Elsevier, 2001.

[81] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. SchNet: A deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

[82] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14):145301, April 2018.

[83] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, May 2019.

[84] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[85] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep Sets. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3391–3401, 2017.

[86] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3744–3753. Pmlr, Pmlr, 2019.

[87] Teresa Tamayo-Mendoza, Christoph Kreisbeck, Roland Lindh, and Alán Aspuru-Guzik. Automatic differentiation in quantum chemistry with applications to fully variational Hartree–Fock. *ACS Central Science*, 4(5):559–566, 2018.

[88] Li Li, Stephan Hoyer, Ryan Pederson, Ruoxi Sun, Ekin D Cubuk, Patrick Riley, Kieron Burke, et al. Kohn-Sham equations as regularizer: Building prior knowledge into machine-learned physics. *Physical Review Letters*, 126(3):036401, 2021.

[89] Jan Hermann, Zeno Schätzle, and Frank Noé. Deep-neural-network solution of the electronic Schrödinger equation. *Nature Chemistry*, 12(10):891–897, 2020.

[90] David Pfau, James S. Spencer, Alexander G. D. G. Matthews, and W. M. C. Foulkes. Ab initio solution of the many-electron Schrödinger equation with deep neural networks. *Physical Review Research*, 2:033429, September 2020.

[91] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[92] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, June 2016. Pmlr.

[93] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[94] David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pages 55–60. Ieee, 1994.

[95] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5574–5584, 2017.

[96] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[97] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865, 1996.

[98] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, July 2013.

[99] Jianwei Sun, Adrienn Ruzsinszky, and John P Perdew. Strongly constrained and appropriately normed semilocal density functional. *Physical Review Letters*, 115(3):036402, 2015.

[100] Albert P Bartók and Jonathan R Yates. Regularized SCAN functional. *The Journal of Chemical Physics*, 150(16):161101, 2019.

[101] James W Furness, Aaron D Kaplan, Jinliang Ning, John P Perdew, and Jianwei Sun. Accurate and numerically efficient r2SCAN meta-generalized gradient approximation. *The Journal of Physical Chemistry Letters*, 11(19):8208–8215, 2020.

[102] Rico Friedrich, Demet Usanmaz, Corey Oses, Andrew Supka, Marco Fornari, Marco Buongiorno Nardelli, Cormac Toher, and Stefano Curtarolo. Coordination corrected ab initio formation enthalpies. *npj Computational Materials*, 5(1):1–12, 2019.

[103] Eric B Isaacs and Chris Wolverton. Performance of the strongly constrained and appropriately normed density functional for solid-state materials. *Physical Review Materials*, 2(6):063801, 2018.

[104] Jochen Heyd, Gustavo E Scuseria, and Matthias Ernzerhof. Hybrid functionals based on a screened Coulomb potential. *The Journal of Chemical Physics*, 118(18):8207–8215, 2003.

[105] Stefano Curtarolo, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kesong Yang, Richard H. Taylor, Lance J. Nelson, Gus L. W. Hart, Stefano Sanvito, Marco Buongiorno-Nardelli, Natalio Mingo, and Ohad Levy. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science*, 58:227–235, June 2012.

[106] Leopold Talirz, Snehal Kumbhar, Elsa Passaro, Aliaksandr V Yakutovich, Valeria Granata, Fernando Gargiulo, Marco Borelli, Martin Uhrin, Sebastiaan P Huber, Spyros Zoupanos, et al. Materials Cloud, a platform for open computational science. *Scientific Data*, 7(1):1–12, 2020.

[107] Luca M. Ghiringhelli, Christian Carbogno, Sergey Levchenko, Fawzi Mohamed, Georg Huhs, Martin Lüders, Micael Oliveira, and Matthias Scheffler. Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats. *npj Computational Materials*, 3(1):46, November 2017.

[108] Casper W. Andersen, Rickard Armiento, Evgeny Blokhin, Gareth J. Conduit, Shyam Dwaraknath, Matthew L. Evans, et al. OPTIMADE, an API for exchanging materials data. *Scientific Data*, 8(1):217, Aug 2021.

[109] Vladimir I Anisimov, Jan Zaanen, and Ole K Andersen. Band theory and Mott insulators: Hubbard U instead of Stoner I. *Physical Review B*, 44(3):943, 1991.

[110] SL Dudarev, GA Botton, SY Savrasov, CJ Humphreys, and AP Sutton. Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA+U study. *Physical Review B*, 57(3):1505, 1998.

[111] Stewart J. Clark, Matthew D. Segall, Chris J. Pickard, Phil J. Hasnip, Matt I. J. Probert, Keith Refson, and Mike C. Payne. First principles methods using CASTEP. *Zeitschrift für Kristallographie*, 220(5-6-2005):567–570, 2005.

[112] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, et al. QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed matter*, 21(39):395502, 2009.

[113] A Rohrbach, J Hafner, and G Kresse. Electronic correlation effects in transition-metal sulfides. *Journal of Physics: Condensed Matter*, 15(6):979, 2003.

[114] Vladan Stevanović, Stephan Lany, Xiuwen Zhang, and Alex Zunger. Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies. *Physical Review B*, 85(11):115104, 2012.

[115] Anubhav Jain, Geoffroy Hautier, Shyue Ping Ong, Charles J Moore, Christopher C Fischer, Kristin A Persson, and Gerbrand Ceder. Formation enthalpies by mixing GGA and GGA+U calculations. *Physical Review B*, 84(4):045115, 2011.

[116] Amanda Wang, Ryan Kingsbury, Matthew McDermott, Matthew Horton, Anubhav Jain, Shyue Ping Ong, Shyam Dwaraknath, and Kristin A. Persson. A framework for quantifying uncertainty in DFT energy corrections. *Scientific Reports*, 11(1):15496, Jul 2021.

[117] Muratahan Aykol and C Wolverton. Local environment dependent GGA+U method for accurate thermochemistry of transition metal compounds. *Physical Review B*, 90(11):115105, 2014.

[118] Christopher J Bartel, Alan W Weimer, Stephan Lany, Charles B Musgrave, and Aaron M Holder. The role of decomposition reactions in assessing first-principles predictions of solid stability. *npj Computational Materials*, 5(1):1–9, 2019.

[119] Christopher J Bartel, Amalie Trewartha, Qi Wang, Alexander Dunn, Anubhav Jain, and Gerbrand Ceder. A critical examination of compound stability predictions from machine-learned formation energies. *npj Computational Materials*, 6(1):1–11, 2020.

[120] AP Drozdov, MI Eremets, IA Troyan, Vadim Ksenofontov, and Sergii I Shylin. Conventional superconductivity at 203 K at high pressures in the sulfur hydride system. *Nature*, 525(7567):73–76, 2015.

[121] AP Drozdov, PP Kong, VS Minkov, SP Besedin, MA Kuzovnikov, S Mozaffari, L Balicas, FF Balakirev, DE Graf, VB Prakapenka, et al. Superconductivity at 250 K in lanthanum hydride under high pressures. *Nature*, 569(7757):528–531, 2019.

[122] Neil W Ashcroft. Metallic hydrogen: A high-temperature superconductor? *Physical Review Letters*, 21(26):1748, 1968.

[123] NW Ashcroft. Hydrogen dominant metallic alloys: high temperature superconductors? *Physical Review Letters*, 92(18):187002, 2004.

[124] YJ Uemura, GM Luke, BJ Sternlieb, JH Brewer, JF Carolan, WN Hardy, R Kadono, JR Kempton, RF Kiefl, SR Kreitzman, et al. Universal correlations between T_c and n_s/m^* (carrier density over effective mass) in high-T_c cuprate superconductors. *Physical Review Letters*, 62(19):2317, 1989.

[125] Satoshi Nakamura, Toru Moriya, and Kazuo Ueda. Spin fluctuation-induced superconductivity in two and three dimensional systems. *Journal of the Physical Society of Japan*, 65(12):4026–4033, 1996.

[126] CC Homes, SV Dordevic, M Strongin, DA Bonn, Ruixing Liang, WN Hardy, Seiki Komiya, Yoichi Ando, G Yu, N Kaneko, et al. A universal scaling relation in high-temperature superconductors. *Nature*, 430(6999):539–541, 2004.

[127] Y Ohta, T Tohyama, and S Maekawa. Apex oxygen and critical temperature in copper oxide superconductors: Universal correlation with the stability of local singlets. *Physical Review B*, 43(4):2968, 1991.

[128] C. N. R. Rao and A. K. Ganguli. Structure–property relationship in superconducting cuprates. *Chemical Society Reviews*, 24(1):1–7, January 1995.

[129] Judith L. MacManus-Driscoll and Stuart C. Wimbush. Future directions for cuprate conductors. *IEEE Transactions on Applied Superconductivity*, 21(3):2495–2500, June 2011.

[130] Eun-Mi Choi, Angelo Di Bernardo, Bonan Zhu, Ping Lu, Hen Alpern, Kelvin H. L. Zhang, Tamar Shapira, John Feighan, Xing Sun, Jason Robinson, Yossi Paltiel, Oded Millo, Haiyan Wang, Quanxi Jia, and Judith L. MacManus-Driscoll. 3D strain-induced superconductivity in $La_2CuO_{4+\delta}$ using a simple vertically aligned nanocomposite approach. *Science Advances*, 5(4):eaav5532, April 2019.

[131] Eun-Mi Choi, Bonan Zhu, Ping Lu, John Feighan, Xing Sun, Haiyan Wang, and Judith L. MacManus-Driscoll. Magnetic signatures of 120 K superconductivity at interfaces in $La_2CuO_{4+\delta}$. *Nanoscale*, 12:3157–3165, 2020.

[132] E Pavarini, I Dasgupta, T Saha-Dasgupta, O Jepsen, and OK Andersen. Band-structure trend in hole-doped cuprates and correlation with T_c max. *Physical Review Letters*, 87(4):047003, 2001.

[133] Petar Popčević, Damjan Pelc, Yang Tang, Kristijan Velebit, Zachary Anderson, Vikram Nagarajan, Guichuan Yu, Miroslav Požek, Neven Barišić, and Martin Greven. Percolative nature of the direct-current paraconductivity in cuprate superconductors. *npj Quantum Materials*, 3(1):1–6, 2018.

[134] Damjan Pelc, Marija Vučković, Mihael S Grbić, Miroslav Požek, Guichuan Yu, Takao Sasagawa, Martin Greven, and Neven Barišić. Emergence of superconductivity in the cuprates via a universal percolation process. *Nature Communications*, 9(1):1–10, 2018.

[135] D Pelc, Z Anderson, B Yu, C Leighton, and M Greven. Universal superconducting precursor in three classes of unconventional superconductors. *Nature Communications*, 10(1):1–6, 2019.

[136] Nikolay Plakida. *High-Temperature Cuprate Superconductors: Experiment, Theory, and Applications*, volume 166. Springer Science & Business Media, 2010.

[137] Eduardo Fradkin, Steven A Kivelson, and John M Tranquada. Colloquium: Theory of intertwined orders in high temperature superconductors. *Reviews of Modern Physics*, 87(2):457, 2015.

[138] Valentin Stanev, Corey Oses, A. Gilad Kusne, Efrain Rodriguez, Johnpierre Paglione, Stefano Curtarolo, and Ichiro Takeuchi. Machine learning modeling of superconducting critical temperature. *npj Computational Materials*, 4(1):29, December 2018.

[139] Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, November 2018.

[140] Zhong-Li Liu, Peng Kang, Yu Zhu, Lei Liu, and Hong Guo. Material informatics for layered high-T_c superconductors. *APL Materials*, 8(6):061104, 2020.

[141] Bryce Meredig, Erin Antono, Carena Church, Maxwell Hutchinson, Julia Ling, Sean Paradiso, Ben Blaiszik, Ian Foster, Brenna Gibbons, Jason Hattrick-Simpers, Apurva Mehta, and Logan Ward. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Molecular Systems Design & Engineering*, 3(5):819–825, 2018.

[142] Frank Hutter, Lin Xu, Holger H Hoos, and Kevin Leyton-Brown. Algorithm runtime prediction: Methods & evaluation. *Artificial Intelligence*, 206:79–111, 2014.

[143] Sooran Kim, Xi Chen, William Fitzhugh, and Xin Li. Apical charge flux-modulated in-plane transport properties of cuprate superconductors. *Physical Review Letters*, 121(15):157001, October 2018.

[144] Charles P. Poole, Ruslan Prozorov, Horacio A. Farach, and Richard J. Creswick. *Superconductivity*. Elsevier, 2014.

[145] H. Eisaki, N. Kaneko, D. L. Feng, A. Damascelli, P. K. Mang, K. M. Shen, Z.-X. Shen, and M. Greven. Effect of chemical inhomogeneity in bismuth-based copper oxide superconductors. *Physical Review B*, 69(6):064512, February 2004.

[146] Neil C Hyatt, MO Jones, I Gameson, M Slaski, GB Peacock, JA Hriljac, and Peter P Edwards. Synthesis and structure of $Hg_{1-x}Cr_xSr_2CuO_{4+\delta}$ mercurocuprates. *Journal of Superconductivity*, 11(1):141–142, 1998.

[147] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in Statistics*, pages 492–518. Springer, 1992.

[148] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–136, 1982.

[149] Daniele Di Castro and Giuseppe Balestrino. Superconductivity in interacting interfaces of cuprate-based heterostructures. *Superconductor Science and Technology*, 31(7):073001, 2018.

[150] L Gao, YY Xue, F Chen, Q Xiong, RL Meng, D Ramirez, CW Chu, JH Eggert, and HK Mao. Superconductivity up to 164 K in $HgBa_2Ca_{m-1}Cu_mO_{2m+2+\delta}$ (m = 1, 2, and 3) under quasihydrostatic pressures. *Physical Review B*, 50(6):4260, 1994.

[151] Ayako Yamamoto, Nao Takeshita, Chieko Terakura, and Yoshinori Tokura. High pressure effects revisited for the cuprate superconductor family with highest critical temperature. *Nature Communications*, 6(1):1–7, December 2015.

[152] WIF Armstrong, AR anFd David, I Gameson, PP Edwards, JJ Capponi, P Bordet, and M Marezio. Crystal structure of $HgBa_2Ca_2Cu_3O_{8+\delta}$ at high pressure (to 8.5 GPa) determined by powder neutron diffraction. *Physical Review B*, 52(21):15551, 1995.

[153] Rafael Gatt, JS Olsen, Leif Gerward, I Bryntse, A Kareiva, Itai Panas, and Lars-Gunnar Johansson. Pressure effect in the Hg-based superconductors: A structural study. *Physical Review B*, 57(21):13922, 1998.

[154] LM Volkova, SA Polishchuk, SA Magarill, and AN Sobolev. Correlation between T_c and the structural parameters of the cation sublattice in the perovskite layer of $HgBa_2CuO_{4+\delta}$ and $Tl_2Ba_2CuO_{4+\delta}$. *Inorganic Materials*, 36(9):919–928, 2000.

[155] Ryozo Yoshizaki, Tetsuya Nakajima, and Masayoshi Tange. Ca substitution effect for Sr upon superconductivity of $Bi_{2.1}Ca_ySr_{1.9-y}CuO_{6+\delta}$. *Japanese Journal of Applied Physics*, 46(3l):L167, 2007.

[156] S Zhu, DH Lowndes, BC Chakoumakos, JD Budai, DK Christen, X-Y Zheng, E Jones, and B Warmack. In situ growth of epitaxial $Bi_2Sr_2CaCu_2O_{8-x}$ and $Bi_2Sr_2CuO_{6-x}$ films by pulsed laser ablation. *Applied Physics Letters*, 63(3):409–411, 1993.

[157] Onur Nane, Bekir Özçelik, and Doğan Abukay. The effects of the post-annealing time on the growth mechanism of $Bi_2Sr_2Ca_1Cu_2O_{8+\delta}$ thin films produced on MgO (100) single crystal substrates by pulsed laser deposition (PLD). *Ceramics International*, 42(5):5778–5784, 2016.

[158] JA Slezak, Jinho Lee, M Wang, K McElroy, K Fujita, BM Andersen, PJ Hirschfeld, H Eisaki, S Uchida, and JC Davis. Imaging the impact on cuprate superconductivity of varying the interatomic distances within individual crystal unit cells. *Proceedings of the National Academy of Sciences*, 105(9):3203–3208, 2008.

[159] D De Barros, L Ortega, Ch Peroz, F Weiss, and P Odier. Effect of substrate's nature on the growth of HgBCCO thin films. *Physica C: Superconductivity*, 440(1-2):45–51, 2006.

[160] Artem R Oganov, Andriy O Lyakhov, and Mario Valle. How evolutionary crystal structure prediction works and why. *Accounts of Chemical Research*, 44(3):227–237, 2011.

[161] Chris J Pickard and RJ Needs. Ab initio random structure searching. *Journal of Physics: Condensed Matter*, 23(5):053201, 2011.

[162] Yanchao Wang, Jian Lv, Li Zhu, and Yanming Ma. CALYPSO: A method for crystal structure prediction. *Computer Physics Communications*, 183(10):2063–2070, 2012.

[163] Geoffroy Hautier, Chris Fischer, Virginie Ehrlacher, Anubhav Jain, and Gerbrand Ceder. Data mined ionic substitutions for the discovery of new compounds. *Inorganic Chemistry*, 50(2):656–663, 2011.

[164] Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Predicting stable crystalline compounds using chemical similarity. *npj Computational Materials*, 7(1):1–9, 2021.

[165] Dipendra Jha, Logan Ward, Arindam Paul, Wei-keng Liao, Alok Choudhary, Chris Wolverton, and Ankit Agrawal. ElemNet: Deep learning the chemistry of materials from only elemental composition. *Scientific Reports*, 8(1):17593, December 2018.

[166] Anthony Yu-Tung Wang, Steven K Kauwe, Ryan J Murdock, and Taylor D Sparks. Compositionally restricted attention-based network for materials property predictions. *npj Computational Materials*, 7(1):1–10, 2021.

[167] Austin D Sendek, Ekin D Cubuk, Evan R Antoniuk, Gowoon Cheon, Yi Cui, and Evan J Reed. Machine learning-assisted discovery of solid Li-ion conducting materials. *Chemistry of Materials*, 31(2):342–352, 2018.

[168] Aria Mansouri Tehrani, Anton O Oliynyk, Marcus Parry, Zeshan Rizvi, Samantha Couper, Feng Lin, Lowell Miyagi, Taylor D Sparks, and Jakoah Brgoch. Machine learning directed search for ultra-incompressible, superhard materials. *Journal of the American Chemical Society*, 140(31):9844–9853, 2018.

[169] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2224–2232, 2015.

[170] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 2017.

[171] Jingzhao Zhang, Kinfai Tse, Manhoi Wong, Yiou Zhang, and Junyi Zhu. A brief review of co-doping. *Frontiers of Physics*, 11(6):117405, 2016.

[172] Quan Zhou, Peizhe Tang, Shenxiu Liu, Jinbo Pan, Qimin Yan, and Shou-Cheng Zhang. Learning atoms for materials discovery. *Proceedings of the National Academy of Sciences*, 115(28):E6411–e6417, July 2018.

[173] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

[174] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[175] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[176] Tian Xie, Arthur France-Lanord, Yanming Wang, Yang Shao-Horn, and Jeffrey C Grossman. Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials. *Nature Communications*, 10(1):2667, 2019.

[177] K-R Müller, Michael Finke, Noboru Murata, Klaus Schulten, and Shun-ichi Amari. A numerical study on learning curves in stochastic multilayer feedforward networks. *Neural Computation*, 8(5):1085–1106, 1996.

[178] Michael W. Gaultois, Taylor D. Sparks, Christopher K. H. Borg, Ram Seshadri, William D. Bonificio, and David R. Clarke. Data-Driven Review of Thermoelectric Materials: Performance and Resource Considerations. *Chemistry of Materials*, 25(15):2911–2920, August 2013.

[179] Stéphane Gorsse, MH Nguyen, Oleg N Senkov, and Daniel B Miracle. Database on the mechanical properties of high entropy alloys and complex concentrated alloys. *Data in Brief*, 21:2664–2678, 2018.

[180] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

[181] Hao Song, Tom Diethe, Meelis Kull, and Peter A. Flach. Distribution calibration for regression. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5897–5906. PMLR, 2019.

[182] Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D. Sculley, Joshua V. Dillon, Jie Ren, and Zachary Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13969–13980, 2019.

[183] Lindsay Bassman, Pankaj Rajak, Rajiv K Kalia, Aiichiro Nakano, Fei Sha, Jifeng Sun, David J Singh, Muratahan Aykol, Patrick Huck, Kristin Persson, et al. Active learning for accelerated design of layered materials. *npj Computational Materials*, 4(1):1–9, 2018.

[184] Alexandros Solomou, Guang Zhao, Shahin Boluki, Jobin K Joy, Xiaoning Qian, Ibrahim Karaman, Raymundo Arróyave, and Dimitris C Lagoudas. Multi-objective Bayesian materials discovery: Application on the discovery of precipitation strengthened NiTi shape memory alloys through micromechanical modeling. *Materials & Design*, 160:810–827, 2018.

[185] Turab Lookman, Prasanna V. Balachandran, Dezhen Xue, and Ruihao Yuan. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials*, 5(1):21, February 2019.

[186] Apoorv Agnihotri and Nipun Batra. Exploring Bayesian optimization. *Distill 10.23915/distill.00026*, 2020.

[187] Florian Hase, Loïc M Roch, Christoph Kreisbeck, and Alán Aspuru-Guzik. Phoenics: A Bayesian optimizer for chemistry. *ACS Central Science*, 4(9):1134–1145, 2018.

[188] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning - ICANN 2018 - 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III*, volume 11141 of *Lecture Notes in Computer Science*, pages 270–279. Springer, 2018.

[189] Dipendra Jha, Kamal Choudhary, Francesca Tavazza, Wei-keng Liao, Alok Choudhary, Carelyn Campbell, and Ankit Agrawal. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nature Communications*, 10(1):1–12, 2019.

[190] Chi Chen and Shyue Ping Ong. AtomSets as a hierarchical transfer learning framework for small and large materials datasets. *npj Computational Materials*, 7(1):173, Oct 2021.

[191] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[192] Ryan J Murdock, Steven K Kauwe, Anthony Yu-Tung Wang, and Taylor D Sparks. Is domain knowledge necessary for machine learning materials properties? *Integrating Materials and Manufacturing Innovation*, 9(3):221–227, 2020.

[193] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[194] Yao Zhang and Alpha A. Lee. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chemical Science*, 10:8154–8163, 2019.

[195] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay S. Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[196] Anton O Oliynyk, Erin Antono, Taylor D Sparks, Leila Ghadbeigi, Michael W Gaultois, Bryce Meredig, and Arthur Mar. High-throughput machine-learning-driven synthesis of full-heusler compounds. *Chemistry of Materials*, 28(20):7324–7331, 2016.

[197] Mercouri G. Kanatzidis. Discovery-synthesis, design, and prediction of chalcogenide phases. *Inorganic Chemistry*, 56(6):3158–3173, 03 2017.

[198] Muratahan Aykol, Vinay I Hegde, Linda Hung, Santosh Suram, Patrick Herring, Chris Wolverton, and Jens S Hummelshøj. Network analysis of synthesizable materials discovery. *Nature Communications*, 10(1):1–7, 2019.

[199] Matthew J McDermott, Shyam S Dwaraknath, and Kristin A Persson. A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. *Nature Communications*, 12(1):1–12, 2021.

[200] Muratahan Aykol, Joseph H Montoya, and Jens Hummelshøj. Rational solid-state synthesis routes for inorganic materials. *Journal of the American Chemical Society*, 2021.

[201] Kirill Kovnir. Predictive synthesis. *Chemistry of Materials*, 33(13):4835–4841, 2021.

[202] Edward Kim, Zach Jensen, Alexander van Grootel, Kevin Huang, Matthew Staib, Sheshera Mysore, Haw-Shiuan Chang, Emma Strubell, Andrew McCallum, Stefanie Jegelka, and Elsa Olivetti. Inorganic materials synthesis planning with literature-trained neural networks. *Journal of Chemical Information and Modeling*, 60(3):1194–1201, 2020.

[203] Henning Glawe, Antonio Sanna, EKU Gross, and Miguel AL Marques. The optimal one dimensional periodic table: a modified pettifor chemical scale from data mining. *New Journal of Physics*, 18(9):093011, 2016.

[204] Cameron J Hargreaves, Matthew S Dyer, Michael W Gaultois, Vitaliy A Kurlin, and Matthew J Rosseinsky. The earth mover's distance as a metric for the space of inorganic compositions. *Chemistry of Materials*, 2020.

[205] Xiaojing Yang, Kazunori Takada, Masayuki Itose, Yasuo Ebina, Renzhi Ma, Katsutoshi Fukuda, and Takayoshi Sasaki. Highly swollen layered nickel oxide with a trilayer hydrate structure. *Chemistry of Materials*, 20(2):479–485, 2008.

[206] D. Prabhakaran, F.R. Wondre, and A.T. Boothroyd. Preparation of large single crystals of ANb_2O_6 (A=Ni, Co, Fe, Mn) by the floating-zone method. *Journal of Crystal Growth*, 250(1):72–76, 2003.

[207] H. Zhou, X.P. Zhang, B.T. Xie, Y.S. Xiao, C.X. Yang, Y.J. He, and Y.G. Zhao. Fabrication of Na_xCoO_2 thin films by pulsed laser deposition. *Thin Solid Films*, 497(1):338–340, 2006.

[208] Yoshinori Tokura, Akira Urushibara, Yutaka Moritomo, Takahisa Arima, Atsushi Asamitsu, Giyu Kido, and Nobuo Furukawa. Giant magnetotransport phenomena in filling-controlled Kondo lattice system: $La_{1-x}Sr_xMnO_3$. *Journal of the Physical Society of Japan*, 63(11):3931–3935, 1994.

[209] I Sfifir Debbebi, H Omrani, W Cheikhrouhou-Koubaa, and A Cheikhrouhou. A-site deficiency effects on the critical behavior of $La_{0.6}Ca_{0.15\cdot0.05}Ba_{0.2}MnO_3$. *Journal of Physics and Chemistry of Solids*, 113:67–73, 2018.

[210] Qiong Liu, Zhi Li, Ying Wang, Xin Su, Zhihua Yang, and Shilie Pan. $LiMCO_3$ (M=K,Rb,Cs): a series of mixed alkali carbonates with large birefringence. *Dalton Transactions*, 46(21):6894–6899, 2017.

[211] Hyeong Seok Lee and Jung Whan Yoo. Yellow phosphors coated with TiO_2 for the enhancement of photoluminescence and thermal stability. *Applied Surface Science*, 257(20):8355–8359, 2011.

[212] Franziska Thoss, Lars Giebeler, Steffen Oswald, Helmut Ehrenberg, and Jürgen Eckert. Study on the reversible Li-insertion of amorphous and partially crystalline al86ni8la6 and $Al_{86}Ni_8Y_6$ alloys as anode materials for Li-ion batteries. *Electrochimica Acta*, 60:85–94, 2012.

[213] Hongfei Sun, Xuewen Li, Peng Zhang, and Wenbin Fang. The microstructure and tensile properties of the Ti_2AlC reinforced TiAl composites fabricated by powder metallurgy. *Materials Science and Engineering: A*, 611:257–262, 2014.

[214] Edward Kim, Kevin Huang, Adam Saunders, Andrew McCallum, Gerbrand Ceder, and Elsa Olivetti. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials*, 29(21):9436–9444, 11 2017.

[215] Marco Cuturi. Sinkhorn Distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2292–2300, 2013.

[216] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477, 2019.

[217] Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012.

[218] Jean-Louis Reymond. The chemical space project. *Accounts of Chemical Research*, 48(3):722–730, 2015.

[219] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B*, 89(9):094104, March 2014.

[220] Jonathan Schmidt, Jingming Shi, Pedro Borlido, Liming Chen, Silvana Botti, and Miguel AL Marques. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chemistry of Materials*, 29(12):5090–5103, 2017.

[221] Haiying Liu, Jiucheng Cheng, Hongzhou Dong, Jianguang Feng, Beili Pang, Ziya Tian, Shuai Ma, Fengjin Xia, Chunkai Zhang, and Lifeng Dong. Screening stable and metastable ABO_3 perovskites using machine learning and the materials project. *Computational Materials Science*, 177:109614, 2020.

[222] Ankit Jain and Thomas Bligaard. Atomic-position independent descriptor for machine learning of material properties. *Physical Review B*, 98:214112, December 2018.

[223] Kyoungdoc Kim, Logan Ward, Jiangang He, Amar Krishna, Ankit Agrawal, and C Wolverton. Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary heusler compounds. *Physical Review Materials*, 2(12):123801, 2018.

[224] Felix A. Faber, Alexander Lindmaa, O. Anatole von Lilienfeld, and Rickard Armiento. Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals. *Physical Review Letters*, 117(13):135502, September 2016.

[225] Cheol Woo Park and Chris Wolverton. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Physical Review Materials*, 4(6):063801, 2020.

[226] Cheol Woo Park, Mordechai Kornbluth, Jonathan Vandermause, Chris Wolverton, Boris Kozinsky, and Jonathan P Mailoa. Accurate and scalable graph neural network force field and molecular dynamics with direct force architecture. *npj Computational Materials*, 7(1):1–9, 2021.

[227] Yury Lysogorskiy, Cas van der Oord, Anton Bochkarev, Sarath Menon, Matteo Rinaldi, Thomas Hammerschmidt, Matous Mrovec, Aidan Thompson, Gábor Csányi, Christoph Ortner, et al. Performant implementation of the atomic cluster expansion (PACE) and application to copper and silicon. *npj Computational Materials*, 7(1):1–12, 2021.

[228] Bingqing Cheng, Guglielmo Mazzola, Chris J Pickard, and Michele Ceriotti. Evidence for supercritical behaviour of high-pressure liquid hydrogen. *Nature*, 585(7824):217–220, 2020.

[229] Jeffrey M Blaney and J Scott Dixon. Distance geometry in molecular modeling. *Reviews in Computational Chemistry*, pages 299–335, 1994.

[230] Sereina Riniker and Gregory A Landrum. Better informed distance geometry: using what we know to improve conformation generation. *Journal of Chemical Information and Modeling*, 55(12):2562–2574, 2015.

[231] Anders S Christensen, Tomas Kubar, Qiang Cui, and Marcus Elstner. Semiempirical quantum mechanical methods for noncovalent interactions for chemical and biochemical applications. *Chemical Reviews*, 116(9):5301–5337, 2016.

[232] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The Δ-machine learning approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096, 2015.

[233] Zhuoran Qiao, Matthew Welborn, Animashree Anandkumar, Frederick R Manby, and Thomas F Miller III. OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *The Journal of Chemical Physics*, 153(12):124111, 2020.

[234] Ralph Walter Graystone Wyckoff. *The Analytical Expression of the Results of the Theory of Space-groups*, volume 318. Carnegie institution of Washington, 1922.

[235] Christopher Tholander, CBA Andersson, Rickard Armiento, Ferenc Tasnadi, and Björn Alling. Strong piezoelectric response in stable $TiZnN_2$, $ZrZnN_2$, and $HfZnN_2$ found by ab initio high-throughput approach. *Journal of Applied Physics*, 120(22):225102, 2016.

[236] Atsushi Togo and Isao Tanaka. Spglib: A software library for crystal symmetry search. *arXiv preprint arXiv:1808.01590*, 2018.

[237] David Hicks, Corey Oses, Eric Gossett, Geena Gomez, Richard H Taylor, Cormac Toher, Michael J Mehl, Ohad Levy, and Stefano Curtarolo. AFLOW-SYM: platform for the complete, automatic and self-consistent symmetry analysis of crystals. *Acta Crystallographica Section A: Foundations and Advances*, 74(3):184–203, 2018.

[238] Michael J Mehl, David Hicks, Cormac Toher, Ohad Levy, Robert M Hanson, Gus Hart, and Stefano Curtarolo. The aflow library of crystallographic prototypes: part 1. *Computational Materials Science*, 136:S1–s828, 2017.

[239] David Hicks, Michael J Mehl, Eric Gossett, Cormac Toher, Ohad Levy, Robert M Hanson, Gus Hart, and Stefano Curtarolo. The aflow library of crystallographic prototypes: part 2. *Computational Materials Science*, 161:S1–s1011, 2019.

[240] Mois Ilia Aroyo, Juan Manuel Perez-Mato, Cesar Capillas, Eli Kroumova, Svetoslav Ivantchev, Gotzon Madariaga, Asen Kirov, and Hans Wondratschek. Bilbao Crystallographic Server: I. databases and crystallographic computing programs. *Zeitschrift für Kristallographie-Crystalline Materials*, 221(1):15–27, 2006.

[241] LL Boyle and JE Lawrenson. The origin dependence of Wyckoff site description of a crystal structure. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 29(4):353–357, 1973.

[242] Logan Ward, Ruoqian Liu, Amar Krishna, Vinay I. Hegde, Ankit Agrawal, Alok Choudhary, and Chris Wolverton. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Physical Review B*, 96(2):024104, July 2017.

[243] Peter Bjørn Jørgensen, Estefanía Garijo del Río, Mikkel N Schmidt, and Karsten Wedel Jacobsen. Materials property prediction using symmetry-labeled graphs as atomic position independent descriptors. *Physical Review B*, 100(10):104114, 2019.

[244] Yunxing Zuo, Mingde Qin, Chi Chen, Weike Ye, Xiangguo Li, Jian Luo, and Shyue Ping Ong. Accelerating materials discovery with Bayesian optimization and graph deep learning. *Materials Today*, 2021.

[245] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5):e1603015, 2017.

[246] Anders S Christensen and O Anatole von Lilienfeld. On the role of gradients for machine learning of molecular energies and forces. *Machine Learning: Science and Technology*, 1(4):045018, 2020.

[247] Andrew J Morris, CP Grey, and Chris J Pickard. Thermodynamically stable lithium silicides and germanides from density functional theory calculations. *Physical Review B*, 90(5):054111, 2014.

[248] Angela F Harper, Matthew L Evans, and Andrew J Morris. Computational investigation of copper phosphides as conversion anodes for lithium-ion batteries. *Chemistry of Materials*, 32(15):6629–6639, 2020.

[249] Andreas Bender and Robert C Glen. A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. *Journal of Chemical Information and Modeling*, 45(5):1369–1375, 2005.

[250] Niu Huang, Brian K Shoichet, and John J Irwin. Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry*, 49(23):6789–6801, 2006.

[251] Paul Raccuglia, Katherine C. Elbert, Philip D. F. Adler, Casey Falk, Malia B. Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A. Friedler, Joshua Schrier, and Alexander J. Norquist. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601):73–76, 2016.

[252] Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):1–36, 2019.

[253] MK Horton, S Dwaraknath, and KA Persson. Promises and perils of computational materials databases. *Nature Computational Science*, 1(1):3–5, 2021.

[254] Joseph H Montoya, Kirsten T Winther, Raul A Flores, Thomas Bligaard, Jens S Hummelshøj, and Muratahan Aykol. Autonomous intelligent agents for accelerated materials discovery. *Chemical Science*, 11(32):8517–8532, 2020.

[255] Shidong Wang, Zhao Wang, Wahyu Setyawan, Natalio Mingo, and Stefano Curtarolo. Assessing the thermoelectric properties of sintered compounds via high-throughput ab-initio calculations. *Physical Review X*, 1(2):021012, 2011.

[256] Austin D Sendek, Qian Yang, Ekin D Cubuk, Karel-Alexander N Duerloo, Yi Cui, and Evan J Reed. Holistic computational structure screening of more than 12,000 candidates for solid lithium-ion conductor materials. *Energy & Environmental Science*, 10(1):306–320, 2017.

[257] Pieremanuele Canepa, Gopalakrishnan Sai Gautam, Daniel C Hannah, Rahul Malik, Miao Liu, Kevin G Gallagher, Kristin A Persson, and Gerbrand Ceder. Odyssey of multivalent cathode materials: open questions and future challenges. *Chemical Reviews*, 117(5):4287–4341, 2017.

[258] Nikolai A Zarkevich, Duane D Johnson, and Vitalij K Pecharsky. High-throughput search for caloric materials: the caloricool approach. *Journal of Physics D: Applied Physics*, 51(2):024002, 2017.

[259] David Hicks, Cormac Toher, Denise C Ford, Frisco Rose, Carlo De Santo, Ohad Levy, Michael J Mehl, and Stefano Curtarolo. AFLOW-XtalFinder: A reliable choice to identify crystalline prototype. *npj Computational Materials*, 7(1):1–20, 2021.

[260] R Hundt, JC Schön, and M Jansen. CMPZ: An algorithm for the efficient comparison of periodic structures. *Journal of Applied Crystallography*, 39(1):6–16, 2006.

[261] G Flor, Danel Orobengoa, Emre Tasci, Juan Manuel Perez-Mato, and Mois I Aroyo. Comparison of structures applying the tools available at the Bilbao crystallographic server. *Journal of Applied Crystallography*, 49(2):653–664, 2016.

[262] Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698):604–610, 2018.

[263] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. In *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016*, volume 1695 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[264] David Mrdjenovich, Matthew K Horton, Joseph H Montoya, Christian M Legaspi, Shyam Dwaraknath, Vahe Tshitoyan, Anubhav Jain, and Kristin A Persson. propnet: A knowledge graph for materials science. *Matter*, 2(2):464–480, 2020.

[265] Vinay I Hegde, Muratahan Aykol, Scott Kirklin, and Chris Wolverton. The phase stability network of all inorganic materials. *Science Advances*, 6(9):eaay5606, 2020.

[266] Haoyan Huo, Ziqin Rong, Olga Kononova, Wenhao Sun, Tiago Botari, Tanjin He, Vahe Tshitoyan, and Gerbrand Ceder. Semi-supervised machine-learning classification of materials synthesis procedures. *npj Computational Materials*, 5(1):1–7, 2019.

[267] Oren M Becker and Martin Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *The Journal of Chemical Physics*, 106(4):1495–1517, 1997.

[268] David J Wales, Mark A Miller, and Tiffany R Walsh. Archetypal energy landscapes. *Nature*, 394(6695):758–760, 1998.

[269] Stefan Chmiela, Huziel E. Sauceda, Klaus-Robert Müller, and Alexandre Tkatchenko. Towards exact molecular dynamics simulations with machine-learned force fields. *Nature Communications*, 9(1):3887, September 2018.

[270] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):1–10, 2020.

[271] Zhantao Chen, Nina Andrejevic, Tess Smidt, Zhiwei Ding, Qian Xu, Yen-Ting Chi, Quynh T Nguyen, Ahmet Alatas, Jing Kong, and Mingda Li. Direct prediction of phonon density of states with euclidean neural networks. *Advanced Science*, page 2004214, 2021.

[272] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[273] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, Kyle Chard, Mark Asta, Kristin Persson, G. Snyder, Ian Foster, and Anubhav Jain. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 09 2018.

[274] Shyue Ping Ong, Shreyas Cholia, Anubhav Jain, Miriam Brafman, Dan Gunter, Gerbrand Ceder, and Kristin A Persson. The materials application programming interface (API): A simple, flexible and efficient API for materials data based on representational state transfer (REST) principles. *Computational Materials Science*, 97:209–215, 2015.

[275] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[276] David Vanderbilt. Soft self-consistent pseudopotentials in a generalized eigenvalue formalism. *Physical Review B*, 41(11):7892–7895, April 1990.

[277] Giovanni Pizzi, Andrea Cepellotti, Riccardo Sabatini, Nicola Marzari, and Boris Kozinsky. AiiDA: Automated interactive infrastructure and database for computational science. *Computational Materials Science*, 111:218–230, January 2016.

[278] Georg Kresse and Jürgen Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science*, 6(1):15–50, 1996.

[279] Georg Kresse and Jürgen Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16):11169, 1996.

[280] Peter E Blöchl. Projector augmented-wave method. *Physical Review B*, 50(24):17953, 1994.

[281] Georg Kresse and Daniel Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B*, 59(3):1758, 1999.

[282] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

[283] Rickard Armiento. *Database-Driven High-Throughput Calculations and Machine Learning Models for Materials Design*. Springer International Publishing, Cham, 2020.