Accelerating Materials Discovery with Machine Learning

By **Luke P. J. Gilligan**

A thesis submitted for the degree of

Doctor of Philosophy

School of Physics/CRANN

Trinity College Dublin

2024

Declaration

I, Luke P.J. Gilligan, hereby declare that this thesis has not been submitted as an exercise for a degree at this or any other university.

It comprises work performed entirely by myself during the course of my Ph. D. studies at Trinity College Dublin. I was involved in a number of collaborations and, where it is appropriate, my collaborators are acknowledged for their contributions.

I agree to deposit this thesis in the University's open-access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish.

Luke P.J. Gilligan

Acknowledgements

It would be entirely impossible to summarise in a few pages the full contribution of everyone in my life that I credit with helping me get through this experience relatively unscathed.

First of all, I would like to extend my deep thanks to my supervisor, Professor Stefano Sanvito. Stefano gave me an opportunity to join his group as a summer intern in the third year of my undergraduate (despite me spelling his name wrong in my initial email). Since that time, I can credit most of my advancement in scientific research to his advice and guidance. I certainly wouldn't have been able to produce this work without his support and I am very grateful. I would also like to thank the group administrator, Stefania Negro, for her assistance throughout the last four years.

I would like to thank the Irish Research Council for their financial support for this Ph. D. (GOIPG/2019/3672) as part of the Government of Ireland Postgraduate Scholarship. I would also like to thank the Trinity Centre for High-Performance Computing (TCHPC), the Irish Centre for High-End Computing (ICHEC) and NVIDIA for the use of their computational resources and hardware.

Over the course of this work, the collaborations with a number of my colleagues and friends invigorated and refreshed my work and, indeed, my relationship with my work. When there was potential for burnout or stagnation of ideas, my collaborators were always on hand to offer a fresh perspective and advice. My two main collaborators in this respect were Matteo Cobelli and Rui Dong, both of whom are good friends as well as excellent scientists. I am very lucky to work closely with them and consistently learn a huge amount from them.

A key aspect of being in this research group is constantly being surrounded by individuals with amazing insight and a deep understanding of complex ideas. There have been countless discussions that have enhanced this work and challenged my preconceptions. I extend my thanks to my friends and colleagues, Akash Bajaj and Hugo Rossignol for their helpful and insightful discussions on the contents of some of the chapters of this thesis. It was very valuable and I believe it has enhanced this work. Hugo also had to live with me for several years so I am grateful to him for putting up with that burden on top of his own Ph. D. This is also a burden that my friends Eoin Monaghan and Laura Teefey currently have to face, for which I am also grateful.

I would also like to thank Mike Minotakis, Laura Gambini and Michelangelo Domina, who started this journey with me 4 years ago, and Willy Laborde for always being up for a laugh and a chat. I would also like to acknowledge Bruno Focassio, Rajarshi Tiwari and Urvesh Patil, who are great friends and always seem to already be experts at pretty much everything that I need to figure out. Also my thanks to Simone Koecher, Yanhui Zhang, Peng Jiang, Antik Sihi and Reena Gupta for having to share the office with Hugo, Mike, Bruno and me at various stages. I'd love to say we created a tranquil place to work, but I am not sure that would be entirely truthful.

I am very privileged to be surrounded by a group of very kind and supportive friends and family, who I am eternally grateful for. I have really felt everyone willing me on at every stage of this experience and I don't know where I'd be without that support. In particular, I'd like to thank my siblings, Maud, Eve and Shane, for being a mixture of both my friends and family; my grandparents, Joe and Anne Burke and Paddy and Monica Gilligan, who are a constant source of inspiration; and all of the rest of my extended family, who are too numerous to name individually.

I must give a specific thanks to my wonderful girlfriend Julia Appleby. She has been a constant through this whole experience and someone I doubt I could have completed this without. She is an inspiration and a joy, and it has been a pleasure and a privilege to grow alongside her over the years. I would also like to thank her family for their support, particularly her mother, Madeleine.

Finally, I would like to thank my own parents, Marie Burke and John Gilligan. This thesis represents the culmination of years of hard work and sacrifice on their part to ensure my education. I will always be grateful for that and can never thank them enough. I will always work to make them proud and I am very proud to be their son.

Ar scáth a chéile a mhaireann na daoine.

Abstract

Materials discovery has always been constrained by the classic approach to scientific discovery, often characterized by a combination of either human intuition or luck. Machine learning (ML) gives us the opportunity to turn this paradigm on its head. Computational techniques, based on ML algorithms, offer the potential to invert the discovery-to-design pipeline and target materials design to pre-defined properties, which are desirable for given applications.

This thesis developed new methods for executing the various stages of this inverse-design pipeline, by employing techniques that originate in several disparate fields within the domain of ML, ranging from regression techniques all the way to the newest generation of transformer networks, primarily used for natural language processing. Libraries of SNAP potential energy surfaces for two-dimensional materials were generated, with which the vibrational and thermal properties of composite heterojunctions could rapidly be computed. Such a step allows for the materials property space to be sampled for rapid property screening applications. These computations were performed and benchmarked against their first-principles equivalents and also experimental results, demonstrating very good agreement with both.

Further to this, a pipeline was constructed to isolate arbitrary compound-property relationships directly from scientific literature with minimal human intervention, in order to bypass any materials property calculations to construct property screening models. This step was executed by leveraging the superior natural language understanding of transformer networks. Models based on these networks were chained together to form an extraction pipeline that could be constructed using a few annotated examples, representing the totality of human intervention required. The resulting databases were demonstrated to be useful for rapid property screening, demonstrating the screening of high-Curie temperature compounds with a precision of 97%.

Finally, these same transformer networks were leveraged to construct materials representations for machine learning tasks, with context learned from literature embedded in the resulting representations. The resulting representations were subsequently demonstrated to show potential for improving the future ability of ML models to predict materials properties, a potential which exists due to the encoding of contextual information in the representation. The embedded contextual information can further inform ML model predictions by including a consideration of material properties that would otherwise be immensely difficult to include.

Keywords: *Materials Science, Inverse-Materials Design, Machine Learning, Thermal Properties, Natural Language Processing*

Contents

1 Introduction **1**

1.1 Rapid Property Screening 4

1.2 Screening for Stability 6

1.2.1 Energetic Stability 6

1.2.2 Dynamical Stability 7

1.3 First-Principles Analysis 7

1.4 Thesis Outline 8

2 Methods **11**

2.1 Machine Learning 11

2.1.1 Supervised Learning 12

2.1.2 Unsupervised Learning 23

2.1.3 Neural Networks 24

2.2 Density Functional Theory 29

2.2.1 Hohenberg-Kohn Theorems 31

2.2.2 Kohn-Sham Theory 31

2.2.3 Exchange-Correlation Energy 34

2.2.4 Pseudopotentials 36

2.2.5 Basis Sets 37

2.3 Summary 37

3 First-Principles Phonon Calculations **39**

3.1 Phonons 40

3.1.1 Frozen-Phonon/Finite Displacement Calculations 46

3.1.2 Density Functional Perturbation Theory 48

3.2 Niobium Disulphide (NbS_2) 52

3.2.1 Band Structure Calculations 53

3.2.2 Vibrational Dynamics 56

3.3 Summary & Conclusions 59

4 Machine-Learned Thermal Properties 61

4.1 Machine-Learned Interatomic Potentials 63

4.1.1 Feature Representation 64

4.1.2 Fitting Procedure 69

4.1.3 Training Dataset 69

4.1.4 Comparison with DFPT 75

4.1.5 Van der Waal's Interaction 78

4.2 Calculation of Thermal Properties 81

4.2.1 Thermal Conductivity 82

4.2.2 Interfacial Thermal Conductance 90

4.3 Summary & Conclusions 94

5 Databases from Language Models 97

5.1 Early Use of Language Models In Materials Science 98

5.2 Transformers 100

5.2.1 Self-Attention 101

5.2.2 Transformer Architecture 102

5.2.3 BERT 104

5.2.4 GPT 105

5.2.5 Transformer Models in Materials Science 106

5.3 Transformer Models for Database Construction 108

5.3.1 BERT-PSIE: Precise Scientific Information Extraction 109

5.4 Evaluation of Resulting Databases 116

5.4.1 Curie Temperature 117

5.4.2 Electronic band gap 130

5.5 GPT-Based Systems 136

5.6 Summary & Conclusions 139

6 Contextual Representations for Materials 141

6.1 Word Embeddings . . . 142

6.1.1 Word2Vec . . . 145

6.1.2 GloVe . . . 147

6.1.3 Contexualized Representations . . . 149

6.2 Word Embeddings in Materials Science . . . 150

6.2.1 Mat2Vec . . . 151

6.2.2 Word Embeddings for Materials Property Predictions . . . 153

6.3 Contextual Embeddings in Materials Science . . . 154

6.3.1 Impact of Domain-Specific Pre-Training . . . 156

6.3.2 Comparison of Contextual Embeddings . . . 163

6.4 Summary & Conclusions . . . 165

7 Conclusions and Future Work 167

A List of Publications 191

List of Tables

4.1 Evaluations metrics of the fit for the library of SNAP potentials for non-magnetic, hexagonal 2D monolayers. 74

4.2 Binding energy of MoS_2 as calculated with a variety of different flavours of VdW corrections. 79

4.3 Thermal conductivities of materials that were estimated using SNAP at 300 K, compared with DFT and experiment. 86

5.1 Performance of the three NLP modules developed for the Curie temperature extraction. 113

5.2 Performance of the three NLP modules developed for band-gap extraction.115

5.3 Comparison of the performances of the different Curie temperature databases on their query quality and their predictive quality. 121

5.4 Comparison of the performances of the different band-gap databases on their query quality and their predictive quality. 133

6.1 Sizes of the different databases for the evaluation of the static embeddings reduced from contextual representations. 155

List of Figures

1.1 The inverse materials-design pipeline. 2

1.2 Substituting first-principles calculations with surrogate methods. 5

2.1 A comparison of the outcomes of regression against classification methods in supervised machine learning. 12

2.2 The result of overfitting and underfitting a machine learning model. . . 14

2.3 The confusion matrix for classification tasks. 19

2.4 A simple feed-forward neural network. 25

2.5 Jacob's ladder of exchange-correlation functionals. 35

3.1 The procedure for calculating the matrix of force constants. 47

3.2 The atomic structure of monolayer 2H-NbS_2. 52

3.3 Energy band diagrams for ferromagnetic and non-magnetic monolayer 2H-NbS_2. 55

3.4 The phonon dispersion and phonon density of states for monolayer 2H-NbS_2, calculated using DFPT. 57

4.1 The general workflow for the construction of machine-learned interatomic potentials. 63

4.2 Plots describing the optimal displacement to create a dataset of distorted cells for training SNAP-MLIPs. 70

4.3 An example of a standard parity plot of the predicted values of energy against the DFT values for 2H-NbS_2 after a SNAP fit. 73

4.4 Comparison of the phonon dispersion calculated using first-principles DFPT and finite-difference using the SNAP-MLIP for monolayer NbS_2. 76

4.5 Comparison of the phonon dispersion calculated using first-principles DFPT and using the SNAP-MLIP for graphene and MoS_2. 77

4.6 A comparison of the phonon dispersion and density of states for bilayer graphene obtained using SNAP and DFPT. 80

4.7 A supercell of graphene encapsulated between two monolayers of hBN along with the phonon dispersion of the system. 81

4.8 Phonon lifetimes, calculated using SNAP-MLIPs for monolayer, bilayer, and trilayer hBN. 89

4.9 The cumulative thermal conductivity of monolayer hBN with the contributions of phonon modes of increasing frequencies. 90

4.10 A MoS_2 monolayer encapsulated with two monolayers of hBN. 92

4.11 An example of the time-dependent temperature of a MD simulation to calculate interfacial thermal conductance. 93

4.12 The linear fit of the temperature gradient between the hBN layers and the higher-temperature MoS_2 layer. 94

5.1 A flowchart of a typical transformer network architecture. 103

5.2 A diagram of the input representations of BERT models. 104

5.3 A diagram of the BERT-PSIE workflow for the automated extraction of scientific information from unstructured literature. 110

5.4 A diagram of the variant of the relationship classification strategy used in this work. 112

5.5 Comparison between the distributions of the different Curie temperature databases. 119

5.6 Violin plots comparing the T_C distribution of the compounds between the different databases. 120

5.7 Comparison between the T_C queried from the manual and BERT-PSIE dataset. 122

5.8 Parity plot for the predicted T_C from BERT-PSIE against the manually extracted T_C. 125

5.9 Violin plots showing the T_C distributions of the compounds screened using a RF model for different temperature brackets. 128

5.10 Comparison between the distributions of the different band-gap databases. 131

5.11 The distribution of band-gap values for the five most common chemical formulas found in the extracted database. 132

5.12 Comparison between the band gaps queried from the manual and BERT-PSIE dataset. 135

5.13 Comparison between the T_C queried directly from ChatGPT and the values contained in the manually curated dataset. 136

5.14 Flowchart of the parsing of compound/property mentions from a corpus of scientific literature, using GPT-based models. 138

6.1 A diagram depicting word offset vectors in word representations. 145

6.2 A diagram of the CBOW Word2Vec architecture. 147

6.3 A diagram of the skip-gram Word2Vec architecture. 148

6.4 A diagram depicting word offset vectors in word representations for materials science. 152

6.5 Evaluation metrics for the prediction of the formation energies of elpasolite compounds from various decontextualized and pooled BERT models. 157

6.6 Evaluation metrics for the prediction of the Curie temperature of ferromagnetic compounds from various decontextualized and pooled BERT models. 160

6.7 Evaluation metrics for the prediction of the bulk modulus compounds from various decontextualized and pooled BERT models. 162

6.8 A comparison of the performance of the best BERT representation, GPT-3 and `mat2vec` for various material property predictions. 164

Chapter 1

Introduction

"An luibh ná faightear is í a fhóireann."

Advancing our mastery over materials has been one of the main driving forces behind every technological leap forward that we have collectively achieved. Our eras are defined by it, from the stone age to the silicon age. Currently, we are faced with more challenges now than ever before. These challenges necessitate numerous, rapid technological advances. Historically, however, such technological advancement has largely been serendipitous in nature. Science has been reliant on experiments in labs, relying primarily on a combination of human intuition and luck, except in cases where a theory had already been established for the studied phenomenon.

As the field of machine learning (ML) has rapidly developed, opportunities have presented themselves to turn this convention on its head. We now have the capability to analyse research trends and physical phenomena at scale and identify promising research directions to take, which have the potential to drastically reduce the experiment-to-application pipeline. The aim of this research has been to utilise the toolkit of a machine learning practitioner in order to achieve this potential in the domain of materials science.

Fig. 1.1 outlines some of the key steps in the process of establishing a workflow that can help to cope with the vast compositional and structural space that constitutes every conceivable material, a space that encompasses an incomprehensible quantity of compounds. Such an amount of potential materials is a result of the combinatorial explosion that results from every possible compositional and structural configuration

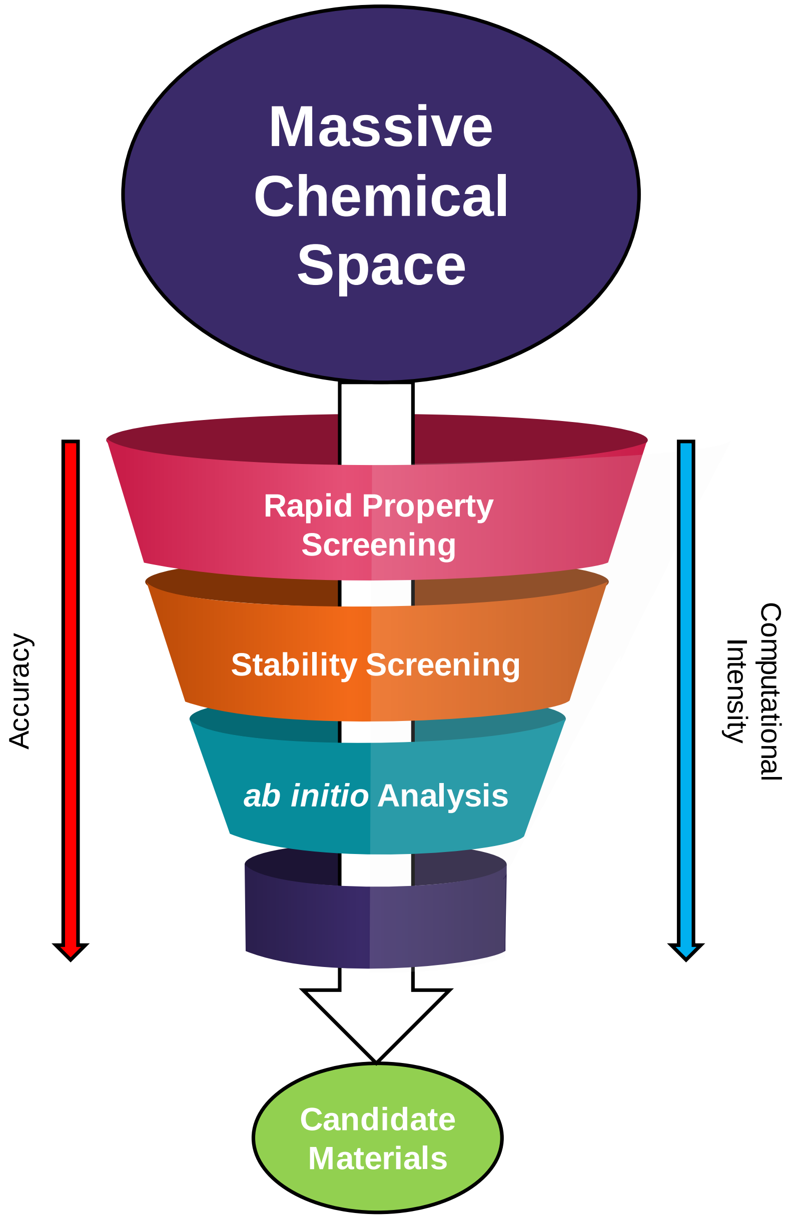

Figure 1.1: The steps required to efficiently identify promising candidates from the vast space of possible chemical structures and configurations such that they are optimised for a given targeted property. The steps are as follows: **1.** A coarse scan to rapidly compute the probable properties of a given structure or configuration. **2.** The funnelling of resulting candidate materials into a system which can rapidly predict how stable they are likely to be. **3.** Performing *ab initio* analyses on the materials which are likely to be stable, to identify the best candidates for synthesis and experiment.

of atoms. It is clear that trial and error is not a viable process to find regions of interest in this vast compositional space and, therefore, a sequence of simple and rapid tests to quickly narrow down the possible number of candidates from the original compositional space is the only strategy that has any potential for addressing this particular issue. A streamlined pipeline necessitates the initial assessment of a material's suitability for a specific application to be both lightweight and computationally efficient. This ensures comprehensive sampling of the entire chemical space, resulting in a subdomain that encompasses the optimal compounds for the desired property. This step should serve as a filter through which the initial massive configurational space can be narrowed down to a smaller quantity of candidates. This step has been effectively executed if the resulting subdomain of the initial configuration space is biased towards materials configurations that are high-performing for the target application. This target application could be any intrinsic material property for which there are valid computational tests.

After this initial filtering, a separate study should be performed such that the stability of the structures and compositions that are deemed reasonable candidates from the prior step can be assessed. This is also required to be lightweight as the initial step will not be sufficiently discriminative to ensure that the resulting subdomain is of a manageable size. Thus, this step will also act as a filter, further limiting the number of candidates to a more manageable size for more computationally demanding tasks. Further to this, the rapid property screening generally will have no way of knowing if a compound that is predicted to have the desired property is energetically favourable, or is of sufficient mechanical stability to exist.

The final step in this pipeline is that of the more computationally intensive, *ab initio* analysis of the resulting set of candidate materials after these prior, coarser steps. This final analysis serves to verify the results of the predictions of previous steps in the workflow, by analysing whether the structures are indeed energetically favourable and structurally stable. It would additionally determine whether the initial screening model's property prediction accuracy aligns with the desired property as per the true theoretical prediction derived from first principles. The final step of this workflow acts as a final barrier, which limits the total number of candidate materials further and, hopefully, would contain the optimal existing material for the target application. Thus, the need for an experimental, trial-and-error approach and the associated wastage of resources is eliminated.

The need for a sequence of progressively more computationally intensive, but more accurate steps, to identify the optimal candidate materials for a given task is clear. The next question that needs to be addressed, however, is how exactly can each of the steps in this computational workflow for materials discovery be constructed and executed in a high-throughput manner. Furthermore, we must ask how can these steps be executed in such a way that the methods involved are sufficiently efficient such that the full pipeline is, in fact, useful, practical and accessible for general applications.

This thesis aims to highlight some methods that may be useful in constructing such a pipeline and leverage the potential of machine learning techniques to improve our ability to rapidly identify candidate materials. Ultimately, the goal of this work is to contribute to practical solutions to aid in the creation of a fully data-driven pipeline, with the potential to reduce the theory-to-application bottleneck and, thereby, accelerate progress within the materials science community.

1.1 Rapid Property Screening

Two main challenges generally face the computational prediction of materials properties, particularly when attempting to do so in a high-throughput workflow. The first of these challenges is, as mentioned, the intensive nature of a lot of these calculations. The feasibility of using a computational method to rapidly sample the materials space and look for promising new candidate materials is very much contingent on the method's efficiency in performing that search. Thus, a key driver in considering new methods for property prediction is the computational expense of such a new method.

The second difficulty that is faced when attempting to conceive of a materials property prediction pipeline is that many useful properties are quite inaccessible to calculations at all. Many of these predictive techniques require the computation of the full electronic charge density, wavefunction, or the phonon wavefunction such that their dependent properties can be deduced by solving the corresponding Hamiltonian. Even after this, certain materials properties require the computation of a range of related phenomena and are inaccessible directly to calculation, meaning their use in a high-throughput pipeline is limited. To further compound this issue, many calculations fail to take into account a myriad of real-world sources of confusion and error due to no experimental information about these systems being readily available.

Therefore, to construct a valuable means of achieving a high-throughput screening for materials properties, we have an array of criteria that must be fulfilled. The first is

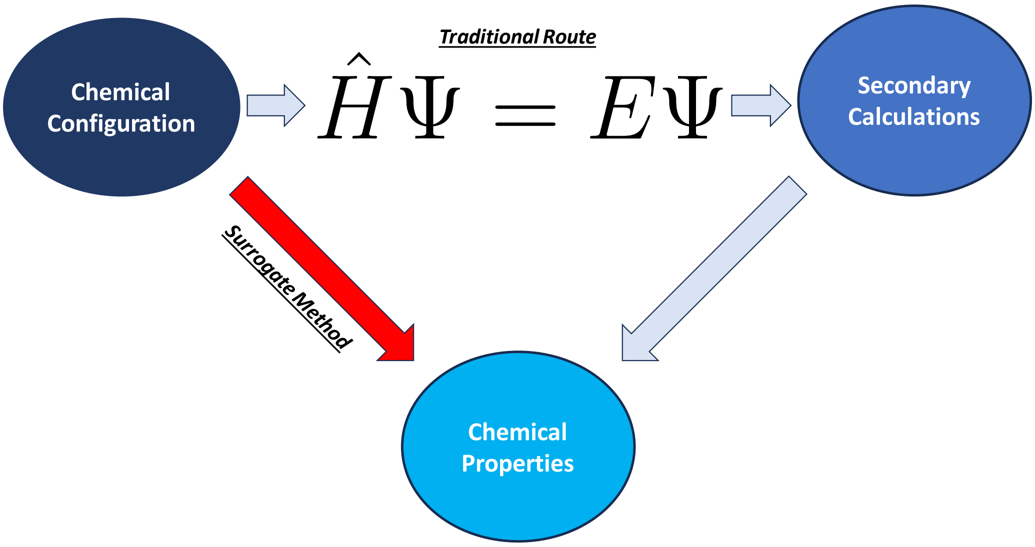

Figure 1.2: Using surrogate methods to map the chemical space directly to the property space can bypass much of the computational expense that is a feature of more traditional computational methods for property prediction, such as DFT.

that, where practically possible, we must conceive of approximations that simplify the computational complexity of the prediction itself. This may mean attempting to bypass the computation of more fundamental quantities from first principles (see Fig. 1.2) and, instead, attempting to come up with some function that maps the compositional and structural space directly to the property space. Further to this, it would be immensely valuable for practical applications if there was some way to incorporate information derived from experiments, in order to bias the predictions towards real-world outcomes, as opposed to theoretical ones, which may not represent the complete scenario.

It is a non-trivial issue to map a chemical space directly onto a property space. It is also quite likely that there is no analytical function that maps these two spaces to each other, and any function that performs this task is bound to be prohibitively high-dimensional. Therefore, it is reasonable to attempt to construct some form of approximate expression that can generally capture some relationship between the chemical space and the property space, given some features that are reasonably descriptive of the material in question. This can serve its purpose well as a coarse screening step to rapidly filter out the majority of materials that are unlikely to exhibit the desired behaviour, as described in the pipeline of Fig. 1.1.

Thus, a means of achieving a computationally efficient approximation of a complex, non-linear and high-dimensional function is necessary in order to execute the first step of the inverse-materials-design pipeline. Such a task is an ideal fit for the field of ML, which has recently been demonstrating its immense usefulness to basically every domain, as modern computers have become more powerful. There is, however, a vast array of potential applications that remain unexplored in the domain of materials science.

1.2 Screening for Stability

Once reasonable predictions have been given for compositions or structures that exhibit the targeted property, the next step is ensuring that the predicted material can exist. This can either involve checking that a predicted structure is stable, or if the rapid screening step was purely compositional in nature, it can involve determining stable structures with the predicted optimal composition. There are generally two considerations to make for the screening of material stability, that of the energetic stability and the dynamical stability. If these two criteria are not satisfied, any material yielded by the initial step is of no use as it is probably not synthesizable.

1.2.1 Energetic Stability

The first of these criteria is that of the energetic stability. That is, if the structure and composition is not the most thermodynamically favourable combination available, it will be more difficult to synthesise because it will likely spontaneously decompose into a lower-energy phase. Thermodynamic favourability involves the minimization of the Gibbs free energy G of the system which is a combination of the enthalpy $H = U + PV$ and entropy S of the system

$$G = H + TS, \tag{1.1}$$

where T is the temperature. When performing a high-throughput search, T is normally set to 0, meaning the entropy term can be dropped. Further to this, the pressure P is generally also set to 0, meaning a single total energy calculation is all that is needed to calculate the system enthalpy and, thereby, the Gibbs free energy. Once this quantity has been calculated, it must be compared with other material phases to ensure the system is in the lowest energy phase possible.

The energetic stability aspect of the inverse design workflow was not made a priority in this thesis as there has been an intensive focus on screening of this nature elsewhere. In fact, some of my colleagues have done extensive work in determining the optimal structure for a given composition, which is a key aspect of the workflow [1, 2].

1.2.2 Dynamical Stability

The second consideration of analysing the stability of a material is in evaluating the dynamical stability of the material. Dynamical stability is the tendency of a material to maintain a well-defined, stable crystal lattice structure in spite of the displacements of atoms and their oscillations within the crystal structure. Generally, the ability of a material to resist structural distortion and phase changes is encapsulated by the phonon dispersion and phonon density of states of the material, which generally involves intensive calculations to obtain. If there are negative phonon frequencies present in the density of states and dispersion, it generally means that the structure is dynamically unstable and will not remain in that configuration without undergoing some lattice distortion or some phase change, which could potentially change the physical properties of the system.

Thus, it is of immense value to rapidly assess whether or not it is dynamically feasible to synthesise a material without these potentially detrimental effects. Once again, ML offers a means of rapidly generating phonon dispersions for a large variety of materials, offering a massive increase in the efficiency of generating such analyses and furthering the ability to filter out candidate materials from the workflow that may not be dynamically stable.

1.3 First-Principles Analysis

Once the prior steps have been implemented in order to identify the optimal candidate materials for the desired application, a final test of viability is performed on the resulting, vastly reduced subspace of the materials space using *ab initio* methods. Such a step is necessary in order to validate the results of the prior steps and perform a more in-depth analysis to ensure that the predicted values and predicted stability of the material are consistent with the true theoretical predictions for the same quantities.

This negates the potential for the workflow to output non-viable materials due to any potential noise or effects that surrogate ML models might not feasibly take into account. Such an analysis would generally be performed using well-established computational techniques such as density functional theory (DFT) where it is reasonable to apply such methods. This final verification, however, cannot necessarily be implemented for screenings of every possible material property that this workflow could screen for as certain properties remain very difficult to calculate from theory, such as the melting point or the Curie temperature. This step can still be used to verify the synthesizability of any outputted compound and is, therefore, still a valuable step, even for difficult-to-calculate property screening.

1.4 Thesis Outline

In this thesis, I will present a range of results and workflows, which have the potential to enhance the aforementioned inverse-design pipeline. These results fall into two main branches, those that work to model the potential energy surfaces of classes of materials in order to bypass *ab initio* techniques, and those involving the use of natural language techniques in order to enhance the inverse-design workflow.

Every chapter in this thesis will involve the use of ML for these various applications and, therefore, I will describe ML in detail in Chapter 2. Also within this chapter, I will describe DFT, which constitutes the main method that would be used for the final step of the computational workflow outlined in Fig. 1.1. I will also employ DFT to generate the training data and benchmarks for the work in Chapter 3 and Chapter 4.

In the first of these two chapters, I will discuss phonons and various methods that can be used for their calculation, both with DFT and with ML methods in the case of finite displacement. I will also outline the theoretical background of phonons and give an example of a two-dimensional, monolayer system, NbS_2, which exhibits a variety of complex behaviours. This system is useful for highlighting the abilities and general weaknesses of ML approximations of potential energy surfaces.

In Chapter 4, I will describe the process of building potential energy surfaces using

ML. This approach aims to bypass the need for first-principles techniques, enabling the rapid generation of computationally intensive properties and facilitating the construction of property predictor models. I will demonstrate this for the cases of thermal conductivity and interfacial thermal conductance of layered two-dimensional materials. Alternatively, such models can be used to perform rapid analyses of the dynamical properties of such systems, through calculations of phonon properties of systems more efficiently. I will also demonstrate the execution of these calculations for system sizes that previously would have been infeasible using first-principles techniques.

Subsequently, in Chapter 5 I will introduce transformer networks and highlight their improved ability as next-generation language models. Furthermore, I explore their potential in isolating properties of interest for the automated construction of databases, through the creation of two databases of Curie temperatures and electronic band gaps of unique compounds from corpora of scientific papers. I will perform tests to examine the viability of constructing ML models from these databases to execute the first step of the workflow in Fig. 1.1. These tests will highlight the ability of the resulting models to rapidly narrow down the compositional space to compositions likely to exhibit the desired properties.

I will explore the potential of transformer networks to construct representations that can further enhance the rapid prediction of materials properties in Chapter 6. I will do this by constructing elemental representations using transformers with two separate approaches. The first approach will involve obtaining representations from pre-trained language models by inputting the elements themselves without context. The second strategy will involve pooling representations of elemental entities as they appeared in sentences. I will use these representations to perform materials property predictions for a variety of available property databases.

In the final chapter, I will draw several conclusions about the potential value of these various techniques to the overall aims of the workflow. As a final discussion, I will outline potential future directions in which these investigations could be taken.

Chapter 2

Methods

"I ndiaidh a chéile a thógtar na caisleáin."

With the overall aim of this work now established, I will outline some of the methods that feature prominently throughout this thesis. In particular, I will first focus on machine learning, a key topic in every part of this thesis. As a part of this discussion, I will give an overview of all of the general algorithms within the field that feature across several of the subsequent chapters along with a focus on the key considerations that are necessary in order to use it effectively as a predictive tool.

After this, I will describe density functional theory, the first-principles computational method that much of the foundations of Chapters 3 and 4 are based upon. I will outline the key concepts of this field, which has been a mainstay of the domain of computational physics for decades.

2.1 Machine Learning

Machine learning (ML) is generally described as being the ability of a computer to 'learn' how to solve a problem by iteratively discovering its own algorithm with which to solve it. More precisely, as defined in Ref. [3]:

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

ML has found a large array of applications from its conception, ranging from speech recognition [4] and medicine [5] to self-driving vehicles [6] and computer vision [7], along with too many more to be included. In this section, I will give an overview of the different classes of ML algorithms. I will also describe some of the main ML algorithms, that feature prominently in this work, in detail. Further to this, I will give an overview of the main considerations of these models and their training procedures.

2.1.1 Supervised Learning

The overall aim of supervised ML is to determine some function f, which can map a set of inputs x to a set of outputs, y

$$f : x \mapsto y. \tag{2.1}$$

The set of inputs x to a ML model is a collection of quantities, known as the set of features, which can be any of vectors, matrices or tensors. These features contain information that allows the prediction of the set of target quantities y, or as they are commonly known in the field of ML, targets. The nature of targets will similarly change depending on the task, which is broadly broken up into two branches in supervised learning. The difference between these two branches can be understood by looking at Fig. 2.1.

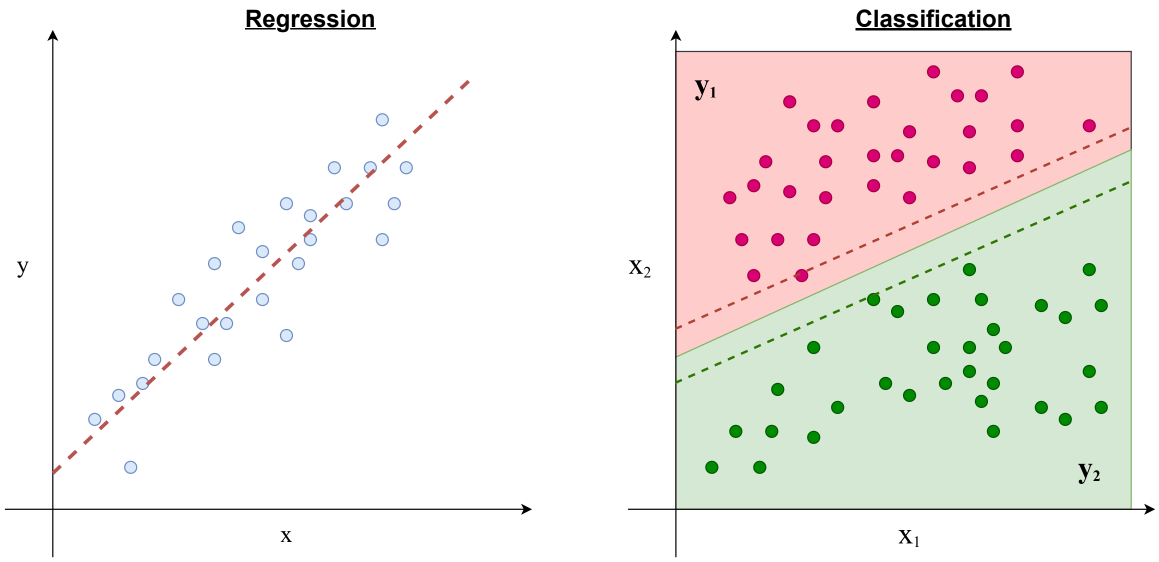

Figure 2.1: Regression is the task corresponding to predicting a numerical target variable y, given an input feature x. In contrast, a classification model will associate a class with a given input feature x by predicting the class label y as a target variable.

The first of these two branches is known as regression. In this case, the values contained in the targets y are quantitative in nature and will mostly be scalar or vector quantities. The second is classification, which is the case where the target variables are qualitative in nature. This generally means that the ML model will attempt to associate some class label to a given input feature. While this class label is a qualitative label, it is still represented numerically, often as a vector quantity, where each term in the vector corresponds to one of a number of classes that the feature can be categorised into. The feature is predicted as belonging to a class if the corresponding term in the target vector is predicted to be 1 by the ML model and 0 for the alternative case. In practice, however, the terms in the vector are likely to range between 0 and 1, with the value of each term of the target vector being the probability that the feature belongs to each class.

This type of learning is known as supervised in that there is some human intervention in the choice of training data that the model is exposed to. The training data should be as representative as possible of the data that the model is likely to encounter in its use case. Thus, the data should be free from any bias towards one outcome over another, unless the practitioner desires that bias to be inherent in the ML model conceived. This may mean the inclusion of equal numbers of training data points for different classes or reweighting the model to account for a lack of a certain class of data points.

The training data aggregated should also be partitioned into three distinct sets, which are commonly known as the training, validation and test sets. This split is needed in order to account for the risk of overfitting or underfitting the training data (see Fig. 2.2). The first of these issues, overfitting, results from the risk of finding a function that, while the model error on the training data is at a minimum, is unable to generalise as it has simply learned the training data distribution as opposed to the trend correlating the training data points. Underfitting is the case where the model has learned too simple a function to fully capture the trend relating the training data points.

There will always be a tradeoff between overfitting and underfitting when training a ML model. This tradeoff can be further understood by examining the breakdown of the error of an ML model. If we are attempting to model the true function $f(x)$ relating

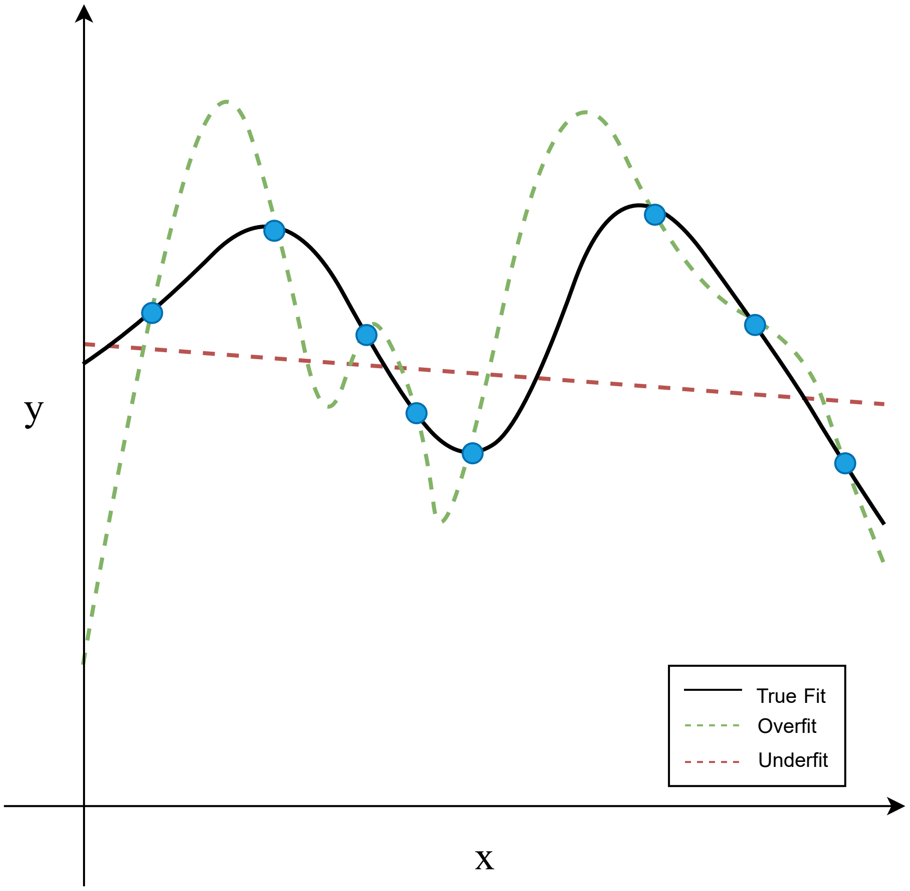

Figure 2.2: The true function *(solid line)* is improperly approximated when the model is either overfit *(green dashed line)* or underfit *(red dashed line)*. Overfitting occurs when the model learns each of the data points in the data set without capturing the simplest relationship describing them. In contrast, underfitting occurs when the function a model has learned is too simple to capture the relationship between data points.

our features x with our targets y, we create the approximation of this relationship with our ML model denoted as $\hat{f}(x)$. The expected square error of the model at x is

$$\mathrm{Err}(x) = E\left[(y - \hat{f}(x))^2\right]. \tag{2.2}$$

Here $E[f(x)]$ is the expected value of a function $f(x)$. This error can further be decomposed [8]

$$\begin{aligned}\mathrm{Err}(x) &= \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(E[\hat{f}(x)] - \hat{f}(x)\right)^2\right] + \sigma^2 \\ &= \left(\mathrm{Bias}\left[\hat{f}(x)\right]\right)^2 + \mathrm{Var}\left[\hat{f}(x)\right] + \sigma^2\end{aligned} \tag{2.3}$$

where σ^2 is the irreducible error, which is intrinsic noise contained within the dataset provided. The other two terms in this expression are known as the bias and variance of the model. The bias is the error which is associated with some erroneous assumptions made in the process of constructing the estimator. In contrast, the variance of a model is the error relating to small fluctuations in the training data of the model. In the case of underfitting, the model is likely to be consistently erroneous as a result of the assumption of too simple a model linking the features to the targets, meaning it will have a high bias. The model will also not have the capacity to learn more with the inclusion of more data and therefore, the variance of the model is likely to be low. In the case of overfitting, this will be reversed and the model will have a low bias but a high variance. An optimal model will exhibit low bias and low variance.

There are a number of steps one can take to limit the impact of underfitting, one is increasing the complexity of the ML model with the addition of more parameters or a more complex model structure. Another is to increase the number of training points. Overfitting is a more difficult issue to deal with and is dealt with by the aforementioned training, validation and test splits of the available data.

The training data is used to adjust the parameters θ of the model such that some measurement of error of the model, known as a loss function, is minimised. Overfitting is prevented by introducing some model hyperparameters, here denoted by λ, which restrict the ability of the model to overfit the training data. These λ values are adjusted by making predictions on the validation set, which the model has not seen during the training, and taking the values for λ that yield the optimal performance on this validation set. Finally, the model's ability to generalize to further unseen data is tested by performing predictions on the test set and analysing the error in these results. The validation set may not act as a test set because there may be information leakage from the validation set that results from the tuning of λ. In a limited-data regime, k-fold cross-validation is generally employed, where the training and validation data are combined and this combined dataset is split up into k subsets. The model reserves one of these subsets and performs the training on the other $(k-1)$ subsets. This step is performed until all subsets have acted as the validation set and the average of all of the validation errors is taken instead of the fixed set for performing the optimisation of λ.

Loss Functions

As was previously indicated, the loss function is essentially a score that indicates how optimal the parameters θ of a model are at predicting the values of the set of targets y, given a set of input features x. These loss functions must be minimised with respect to the model parameters in order to obtain the best model performance on the training data. This converts the problem of ML into an optimization problem for which there are a large variety of methods to solve, as optimization algorithms have been a staple domain of computer science for decades.

Loss functions are very closely related to the concept of error functions, however, loss functions do come in different flavours depending on the task at hand. By far and away the most common loss function used in regression problems is that of the mean square error (MSE)

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} (y - \hat{f}(x))^2}{n}. \tag{2.4}$$

The value of this particular metric is that predicted values that are far from the real values are penalised proportionally more heavily than those that are close to being correct as a result of squaring the variance. This is in contrast with another common loss function in regression tasks, mean absolute error (MAE)

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left| y - \hat{f}(x) \right|}{n}, \tag{2.5}$$

which penalises outliers far less severely than in the case of MSE.

In the case of classification problems, the most common loss function is known as the cross-entropy loss, which in the binary classification case can be expressed as

$$\mathrm{CrossEntropyLoss} = -y \log{(\hat{f}(x))} + (1 - y) \log{(1 - \hat{f}(x))}, \tag{2.6}$$

which will increase as the predicted probability of a feature belonging to a particular class diverges from the real class label. When the number of classes is greater than two, a separate cross-entropy loss can be calculated for each class label and the results can be summed to form a loss function accounting for every class in the data.

Evaluation Metrics

Establishing metrics to evaluate the overall performance of a given ML model is of immense importance to the usefulness of the resulting models. These metrics give a recognisable indication of how trustworthy the predictions that are outputted by a model are. They are primarily calculated for the results obtained on the test data. This is done in order to ensure that the model generalises well to unseen data.

For regression tasks, several common metrics can be used to evaluate the quality of the regressor. The first is the one encountered for use as a loss function, the MAE, Eq. (2.5). This essentially indicates the expected absolute deviation from the true result by the regressor. The second metric that is useful for regression is a variant of another loss function that we have previously encountered. The root-mean-square error (RMSE) is very simply calculated by taking the square root of the MSE

$$\text{RMSE} = \sqrt{\text{MSE}}. \tag{2.7}$$

A result of the MSE scaling with the square of the deviation from the true value of the regressor is that outliers will disproportionately affect this particular error metric. Therefore, the RMSE will also more strongly indicate the presence of outliers in the model predictions, which is useful information for indicating the quality of the predicted data.

The final, common metric that is used when assessing the quality of a regressor is the R^2 coefficient or the coefficient of determination. This serves to represent the proportion of variance in the target variables of the test data that the model is capable of predicting with its test features i.e.

$$R^2 = 1 - \frac{\sum \left(y - \hat{f}(x)\right)^2}{\sum \left(y - \bar{y}\right)^2}, \tag{2.8}$$

where $\bar{y}$ is the mean of the target variables

$$\bar{y} = \frac{1}{N} \sum y, \tag{2.9}$$

and N is the number of samples in the test data.

Due to the different nature of classification tasks in associating an absolute label to a data point as opposed to attempting to approximate it as closely as possible in geometric space, a different set of evaluation metrics is required. These metrics should gauge the efficacy of a classification model associating the class label to a given data point. When performing classification tasks, there are four possible outcomes to any prediction:

- True Positive (TP): A data point is predicted to belong to a class that it truly belongs to.

- True Negative (TN): A data point is predicted not to belong to a class that it truly does not belong to.

- False Positive (FP): A data point is predicted to belong to a class that it truly does not belong to.

- False Negative (FN): A data point is predicted not to belong to a class that it truly belongs to.

The predicted class labels of the test set can be compared to the true values and a construction known as a confusion matrix (see Fig. 2.3) can be constructed as a result, by counting the number of examples which fall into each case.

The first and most intuitive of the classification metrics that can be conceived is that of the accuracy of the classifier. This metric will yield the percentage of correct predictions for the test data. It can be calculated as

$$\text{accuracy} = \frac{TP + TN}{N}. \tag{2.10}$$

This is a very valuable metric but it is not sufficient on its own to adequately evaluate the model performance. For example, if there is a class label that only rarely occurs, a model can achieve a very high accuracy score if it simply classifies everything as being negative. Thus, a metric that specifically takes into account the labels that are predicted to be positive is required. This is the logic behind the use of the precision P, which is defined as being the ratio of true positive predictions against all of the

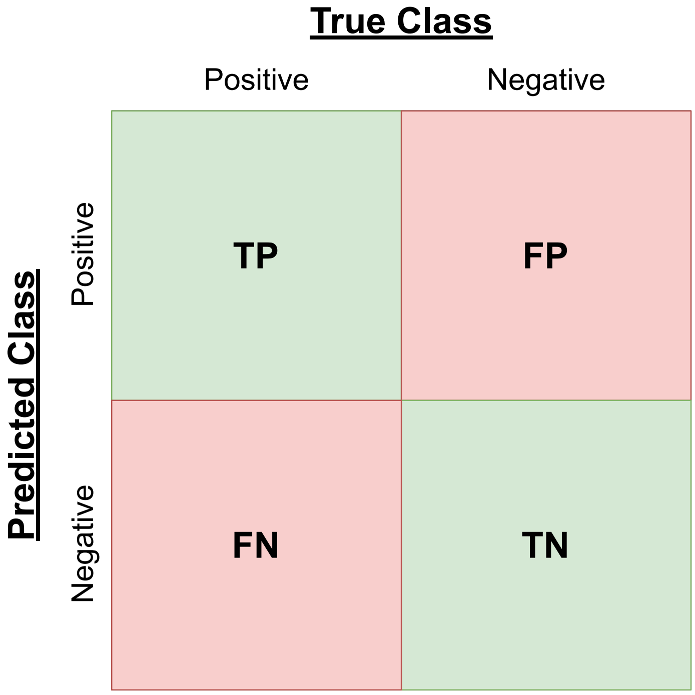

Figure 2.3: The confusion matrix is a tool for assessing the efficacy of a model at performing a classification task. A separate confusion matrix must be computed for each class label.

positive predictions the model made, i.e.

$$P = \frac{TP}{TP + FP}. \tag{2.11}$$

The precision is essentially an indication of the proportion of the data points predicted to belong to a class that is correct. The second metric constructed using this logic is the ratio of the true positive examples, compared with every data point in the data set which are truly positive. This is known as the recall R and can be computed as

$$R = \frac{TP}{TP + FN}. \tag{2.12}$$

The recall is a metric that captures the proportion of true members of a class in a dataset, which a model is able to correctly predict. Both of these metrics convey important information about the ability of the system to correctly classify data points, however, it is desirable to combine both of these metrics to most effectively indicate the effectiveness of the model. Thus, it is reasonable to take the harmonic mean of these two metrics to obtain the F_1 score

$$F_1 = \frac{2PR}{P+R} \tag{2.13}$$

which is a metric for the overall model performance. Just as both the precision and recall are values which range between 0 and 1, with 1 indicating optimal performance, so too is this the case for the F_1 score. This is generally the main metric for use in evaluating ML classifiers.

Ridge Regression

An example of a supervised ML model, which is used extensively in this work is that of ridge regression [9]. As the name suggests, this is a regression method that is primarily used in this work as a part of the SNAP method (see Section 4.1.1) to predict the total energy of a given compound. Ridge regression is an elaboration on the most simple regression model, that of linear regression, which aims to model a sequence of data points with their targets by assuming a linear relationship between the two. In the case of linear regression, given a set of n feature vectors $\boldsymbol{x}$ of rank p, forming a $(n \times p)$ matrix $\mathbf{X}$, the relationship between the set of target variables $\boldsymbol{y}$ and $\mathbf{X}$ is approximated as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.14}$$

where $\boldsymbol{\beta}$ is a set of linear coefficients, which must be determined and $\boldsymbol{\varepsilon}$ is the errors of the N data points in the data set. The loss function J, using the mean squared error can thus be expressed as [8]

$$J(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^{\intercal}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.15}$$

which gives an estimate of the linear coefficients as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\boldsymbol{y}. \tag{2.16}$$

This system of determining an estimate of the optimal linear coefficients for a predictor is known as ordinary least squares. There is the potential, however, that this setup can lead to the issue of overfitting as described previously and a way of coping with this is through the introduction of another term in the loss function that prevents the components of $\boldsymbol{\beta}$ from growing too large.

$$J(\boldsymbol{\beta}, \lambda) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^{\intercal}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\intercal}\boldsymbol{\beta}, \tag{2.17}$$

where λ is the regularization strength. This construction is known as L_2 regularization and it enables the approximation for the regularized $\boldsymbol{\beta}$ coefficients of

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{X}^{\intercal}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\intercal}\boldsymbol{y}, \tag{2.18}$$

thereby mitigating the risk of overfitting the linear model.

Random Forest Regression

Random forest (RF) regression [10] is employed as the regressor of choice in Chapters 5 and 6, a choice based on the work from Ref. [11], which identified it as being the best method tested for property prediction based on compositional features. Random forest models are based on the output of an ensemble of regression trees. A regression tree is a type of decision tree that outputs a continuous variable instead of a class label. A regression tree T will partition the feature space into M partitions $R_1, R_2, ..., R_M$, each of which attempts to model the target of a given feature by associating some constant c_m in each partition, i.e.

$$T(x) = \sum_{m=1}^{M} c_m I(x \in R_m), \tag{2.19}$$

where I is the indicator function, which maps the feature to the partitioned region of the feature space. It is simple to see that the value of c_m that will best approximate the target related to feature x that occurs in partition R_m is going to be the mean value of the targets that relate to features within that partition, i.e. for N data points where $i = 1, 2, ..., N$,

$$\hat{c}_m = \text{mean}(y_i | x_i \in R_m). \tag{2.20}$$

The difficulty now is in determining how the feature space should be partitioned in order to best approximate the relationship between x and y. This can be achieved using a greedy algorithm as a replacement for determining a computationally infeasible globally optimal solution using an algorithm akin to minimizing the sum of squares. First taking all of the data available, a splitting variable j and a split point s are conceived such that

$$R_1(j, s) = \{x | x_j \leq s\} \quad \text{and} \quad R_2(j, s) = \{x | x_j > s\}. \tag{2.21}$$

Once this partitioning is performed, the following function is minimized with respect to the choice of j and s

$$J(j, s) = \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2, \tag{2.22}$$

where, once again, the inner constants can be given by

$$\hat{c}_1 = \text{mean}(y_i | x_i \in R_1(j, s)) \quad \text{and} \quad \hat{c}_2 = \text{mean}(y_i | x_i \in R_2(j, s)). \tag{2.23}$$

Once this procedure has been executed for the first region, this same procedure can be executed iteratively on each of the subregions.

Eventually, the larger the tree grows, the more overfit the regression tree will become. There are a number of ways that this can be dealt with, through techniques like cost-complexity pruning and weakest-link pruning. Random forest regression solves the issue of overfitting by building an ensemble of identically distributed, de-correlated regression trees and averaging over them. It performs this step using a technique known as bagging or bootstrap aggregation.

Bagging is the process of averaging a large number of high-variance models, which are trained by randomly sampling the training data set with replacement (different training data sets may be constructed with a proportion of the same data point repeated). This step is what de-correlates the resulting trees. Given an ensemble of B different regression trees, a random forest model can be computed using each of the resulting trees as

$$\hat{f}^B_{\text{rf}} = \frac{1}{B}\sum_{b=1}^{B} T_b(x). \tag{2.24}$$

The bias of the averaged trees will be the same as any individual one as a result of an identical distribution of regression trees generated. The reduction in variance of the aggregated model can be understood by considering the averaging process of B trees. If each tree exhibits a variance of σ^2 and has a pairwise correlation factor of ρ. The expected variance for a prediction of the aggregated model is

$$\sigma^2_{\text{rf}} = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{2.25}$$

Thus, the expected variance of the aggregated trees is reduced for large B as $(1-\rho)/B$ will go to zero, provided the models are de-correlated, meaning $\rho < 1$.

2.1.2 Unsupervised Learning

Beyond supervised learning, in which there is an available set of labelled training data that the model is provided with for learning, there is another suite of techniques in ML in which patterns or hidden structures are learned from unlabelled data. This paradigm is known as unsupervised learning.

Unsupervised learning is primarily used for two applications, dimensionality reduction and the clustering of similar data points. In their essence, these two seemingly different tasks are quite similar in nature. As mentioned, clustering is the process of associating data points into subsets based on their proximity in the feature space. A space that represents information about the feature vector, i.e. two clustered data points will be determined to convey similar information. This can be achieved by associating data points with their nearest neighbours, or based on their proximity to an average of the other data points in the cluster.

In comparison, dimensionality reduction is a useful method of reducing the amount of redundancy contained in a feature representation by determining the similarity of information conveyed between different components in the feature representation itself. This is generally achieved by projecting a high-dimensional representation onto a lower-dimensional one such that the covariance between data points is maximized, thus preserving the maximal amount of information, while reducing redundancy and ensuring that the feature representation is more lightweight. This technique is known as principal component analysis or PCA.

Unsupervised learning is not a focus of this work and therefore will not be described in detail, however, many static term representations such as GloVe or Word2Vec, which are outlined in Chapter 6, may benefit from the reduction in redundancies in the representation that have been processed through dimensionality reduction techniques.

2.1.3 Neural Networks

Artificial neural networks (NN) are designed to approximate the mechanism through which the human brain works, using a simple model of nodes and connections between these nodes, called edges. These nodes are approximations of the units of the biological thinking process, neurons, hence the name neural network. In the field of ML, however, it is not actually of concern how well the model actually approximates the biological process of thought, as long as the predictions outputted by these models are of sufficient quality.

Fig. 2.4 gives an example of a very simple form of a NN, a feed-forward NN. This method is described as feed-forward based on the direction of information flow, which only ever goes forward through the NN (left to right in Fig. 2.4). As can be seen from the figure, the general architecture can be summarized as an input layer, an output layer and a number of hidden layers. The number of hidden layers is variable and is a hyperparameter of the model. Another model hyperparameter is the number of nodes per layer, which can vary from layer to layer.

For regression tasks, the output layer will normally be comprised of a single node, whereas for classification tasks, for K total classes, there will be a total of K different nodes in the output (regression means $K = 1$). By convention, the input layer is known as the zeroth layer giving $\boldsymbol{h}^{(0)} = \boldsymbol{x}$. For each of the hidden layers, the output of the

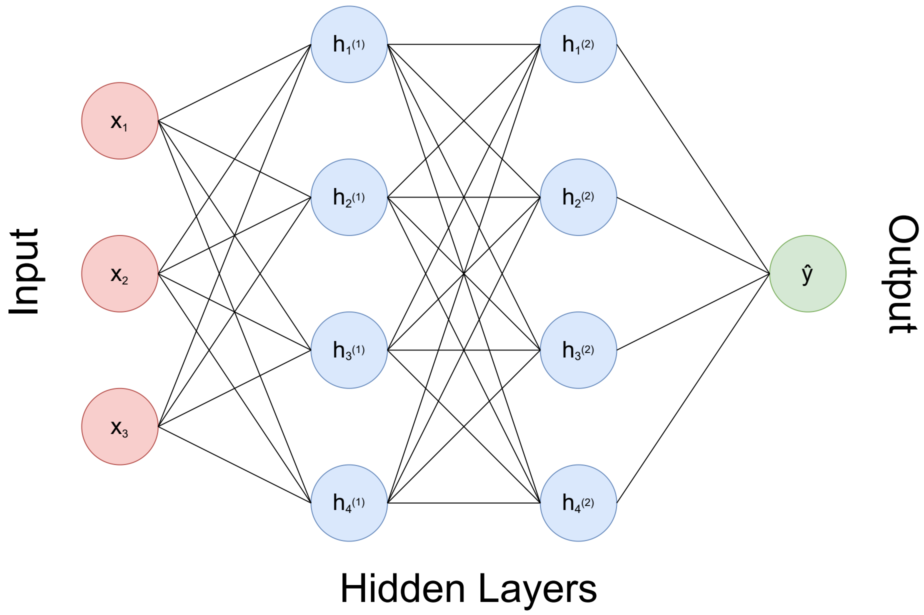

Figure 2.4: An example of a simple feed-forward neural network with two hidden layers, as well as the input and output layers. The bias term in each layer is omitted from this diagram. This neural network would generally act as a regressor due to the single output node.

layer will be given by

$$\boldsymbol{h}^{(i)} = g(\mathbf{w}_0^{(i-1)} + \mathbf{w}^{(i-1)\intercal}\boldsymbol{h}^{(i-1)}), \tag{2.26}$$

where $\mathbf{w}$ denotes the weights of the edges between the nodes and $\mathbf{w}_0$ is the bias term of the layer, which is an additional weight variable to tune in the training. Further to this, g is some activation function, which corresponds to the strength of activation of a given neuron in the NN.

There are numerous choices for the model activation. Originally, the sigmoid function $g(v) = 1/(1 + \exp(-v))$ was the default activation function for the nodes in a network. This is still commonly used, however, there are some difficulties with this particular activation. One is that the strength of activation does not scale with an increase in v, meaning the sigmoid will saturate, losing any information that would otherwise be propagated. The second is that it is more difficult to compute a sigmoid function for every node in a large NN than a simpler choice of activation. Thus, the ReLu or rectified linear units activation [12] became the activation function of choice with the advent of larger NNs. This activation function can be written as

$$\mathrm{ReLu}(v) = \max(0, v). \tag{2.27}$$

Therefore, not only does the output scale with increasing input indefinitely, it is also quick to compute, requiring only a single if statement, and it saturates at 0 with diminishing input, which indicates no node activation. These features allowed it to replace the sigmoid as the activation function of choice. There is, however, a whole toolkit of other activation functions, which may be useful for different tasks, such as the hyperbolic tangent or a modified version of ReLu that doesn't saturate at 0, known as leaky ReLu [13].

Once the data has been processed through each of the L hidden layers, the output is obtained as

$$\hat{y} = q(\mathbf{w}_0^{(L)} + \mathbf{w}^{(L)\intercal}\boldsymbol{h}^{(L)}), \tag{2.28}$$

where q is some function that captures the nature of the desired output. For regression tasks, where $K = 1$, this is generally chosen to be the identity function, $q(h) = h$, such that the output of the NN is just the value obtained when propagating the final hidden layer forward into the single output node. For an output classifying a data point into one of the K classes, the preferred output function is the softmax function,

$$q_k(h) = \frac{e^{h_k}}{\sum_{\ell=1}^{K} e^{h_\ell}}. \tag{2.29}$$

This function will produce positive estimates that sum to one, indicating the probability that the output belongs to class k.

Back-Propagation

With the NN architecture now established, we must conceive of a way of adapting the weights of the NN such that the output is as close to the targets as possible. We have already considered the forward propagation of information through the system, now we must consider the information propagating back through the system in order to retroactively update the model weights. This is, intuitively, known as back-propagation.

Back-propagation aims to minimize a loss function J with an adjustment of the weights and biases of the NN model. This requires the computation of the gradient of the loss function with respect to the model parameters θ. Taking the square error loss function from before,

$$J(x,\theta) = \frac{1}{2}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2, \tag{2.30}$$

the gradient of the loss function with respect to each of the individual model weights w_{ij}^{l}, for weight corresponding to node j in layer l going into node i can be computed as

$$\frac{\partial J(x,\theta)}{\partial w_{ij}^{l}} = \sum_{n=1}^{N}\frac{\partial J_n}{\partial w_{ij}^{l}}. \tag{2.31}$$

Thus, there is only a need to consider single training points at a time to obtain the full loss gradient, which can be computed with a simple summation once the individual contributions have been calculated. The n will be omitted for simplicity in future steps.

The derivatives of the loss function with respect to an individual weight can be obtained using the chain rule

$$\frac{\partial J}{\partial w_{ij}^{l}} = \frac{\partial J}{\partial a_i^l}\frac{\partial a_i^l}{\partial w_{ij}^{l}}, \tag{2.32}$$

where $a_i^l = w_{0i}^l + \sum_{j=1}^{M} w_{ij}^l h_j^{l-1}$ is the output of node i in layer l before activation where l contains M nodes. The first term in the product of Eq. (2.32) is normally called the error and is written as

$$\delta_j^l \equiv \frac{\partial J}{\partial a_i^l}. \tag{2.33}$$

Then the second term can be calculated as

$$\frac{\partial a_i^l}{\partial w_{ij}^{l}} = \frac{\partial}{\partial w_{ij}^{l}}\left(\sum_{m=0}^{M} w_{mj}^l h_m^{l-1}\right) = h_i^{l-1}. \tag{2.34}$$

Thus, the gradient of the loss with respect to a given model weight can be computed as

$$\frac{\partial J}{\partial w_{ij}^{l}} = \delta_j^l h_i^{l-1}. \tag{2.35}$$

Thus, the weights of the model can be iteratively updated in such a way that the model loss is minimized by updating the model weights at each iteration by calculating δ_j^l for every node, working backwards from the output. The model weights can then be adjusted by calculating the total loss function with respect to each weight and updating the weight in the direction of negative gradient, using an optimization hyperparameter, known as the learning rate α as

$$\Delta w_{ij}^{l} = -\alpha \frac{\partial J(x, \theta)}{\partial w_{ij}^{l}}. \tag{2.36}$$

Here the learning rate is a value, which controls how quickly the model will converge to a minimum. However, too large a learning rate will mean the model will struggle to converge to a solution, which minimizes the loss function adequately.

Regularization can also be applied to this minimization procedure to avoid overfitting, which is once again achieved by performing these steps, with the addition of a regularization term in the loss function, similar to the case of ridge regression. There are several other methods to avoid overfitting such as early stopping, which stops the training procedure once the validation error begins to rise, which is an indication of the model beginning to overfit. The other common example is the introduction of dropout [14], which will randomly set node outputs to 0, based on some probability that the user defines. This forces the model to learn different functions which map the inputs to outputs, thus improving the generalisability of the model.

Other Types of Neural Network

There is an entire zoo of different NN architectures that have been shown to be useful for a variety of different use-case scenarios. Some of the most prevalent of these other flavours of NN are the convolutional neural network (CNN) [15] and the recurrent neural network (RNN).

The CNN introduces a sequence of convolutional layers or filter layers through which the model can learn its own representation of the features. It couples these layers with conventional, fully connected feed-forward layers, through which the convolutions are passed. A major benefit of CNNs is that they are capable of dealing with a variable input size, meaning the input is more flexible than in the case of other NN flavours.

RNNs are a type of NN in which the information flow is not restricted to the forward direction alone. In this type of network, the output of a node can impact the future inputs to the same node. This bidirectionality means that these NNs can be useful for processing sequences in which the inputs are connected in some way, such as speech recognition [16] or other linguistic applications, such as named entity recognition [17].

2.2 Density Functional Theory

Density functional theory (DFT) [18] is a powerful computational technique in the field of computational physics and chemistry, which allows for the calculation of material and chemical properties from first principles of quantum mechanics. These techniques are used extensively in Chapters 3 and 4 of this thesis as a basis for benchmarking the performance of and training the ML models produced.

The fundamental assumption of DFT is that any property of a system of interacting particles can be viewed as a functional of its ground state density $n_0(\mathbf{r})$. This means that in order to calculate the quantum-mechanical properties of a system, it is not necessary to obtain the full wavefunction of a many-body system, for which there is no means of obtaining an analytical solution.

In order to understand the problem of calculating the properties of many-body systems using quantum mechanics, we must first consider the Schrödinger equation

$$\hat{H}\psi = E\psi, \tag{2.37}$$

where ψ is the many-body wavefunction, E is the system energy and $\hat{H}$ is the full Hamiltonian operator for a system of interacting electrons. This operator can be expressed as

$$\begin{aligned}\hat{H} = -\frac{\hbar}{2m_e}\sum_i \frac{\partial^2}{\partial \mathbf{r}_i^2} - \frac{\hbar}{2}\sum_I \frac{1}{M_I}\frac{\partial^2}{\partial \mathbf{R}_I^2} + \frac{e^2}{2}\sum_{i\neq j}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \\ - \sum_{iI}\frac{Z_I e^2}{|\mathbf{r}_i - \mathbf{R}_I|} + \frac{e^2}{2}\sum_{I\neq J}\frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|},\end{aligned} \tag{2.38}$$

where $\hbar$ is the Planck constant, $\mathbf{r}_i$ is the position of the i-th electron, $\mathbf{R}_I$ is the position of the I-th nucleus, e is the electronic charge, Z_I is the charge of the I-th nucleus, m_e is the electronic mass and M_I is the mass of the I-th nucleus. The initial two terms in this expression capture the kinetic energy of the system and the rest attempt to capture the electron-electron, electron-nucleus and nucleus-nucleus Coulombic contributions to the overall potential energy, respectively.

The Born-Oppenheimer approximation is useful here, which allows the wavefunctions for the electrons and nuclei to be decoupled due to the relative difference in velocity between the two, allowing the nuclei to be treated as fixed. This means that the contribution of the nuclei to the kinetic energy can be dropped, meaning that $\hat{H}$ may be rewritten as

$$\begin{aligned}\hat{H} = -\frac{\hbar}{2m_e}\sum_i \frac{\partial^2}{\partial \mathbf{r}_i^2} + \frac{e^2}{2}\sum_{i\neq j}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \\ - \sum_{iI}\frac{Z_I e^2}{|\mathbf{r}_i - \mathbf{R}_I|} + \frac{e^2}{2}\sum_{I\neq J}\frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|},\end{aligned} \tag{2.39}$$

which, on inspection, involves a term capturing the contribution to the kinetic energy of the individual electrons, a term for the electron-electron contribution to the potential energy, a term for the Coulombic interaction of the electron in a fixed field of nuclei and a constant potential energy contribution from the nucleus-nucleus interaction, i.e.

$$\hat{H} = \hat{T} + \hat{V}_{ee} + \sum_i v_{\text{ext}}(\mathbf{r}_i), \tag{2.40}$$

where $\hat{T}$ is the kinetic energy, V_{ee} is the electron-electron contribution to the potential and $\sum_i v_{ext}(\mathbf{r}_i)$ is the contribution of the external potential including the result of the electronic interactions with the nuclei.

This formulation of the Schrödinger equation is the basis for the solution to the many-particle equation proposed by Hohenberg and Kohn in 1964 [19], which represents the foundations of DFT.

2.2.1 Hohenberg-Kohn Theorems

In this work, Hohenberg and Kohn proposed a formulation of an exact solution to the Schrödinger equation, with the Hamiltonian as written in Eq. (2.39). The theory was based on two foundational theorems.

The first of these was that for every system of interacting particles in an external potential, there exists an external potential, which is determined uniquely, with the exception of a constant, by the ground state particle density $n_0(\mathbf{r})$. Thus, every system property can be determined only through the ground state charge density.

The second theorem that they outlined states that a universal functional for the energy $E[n]$, in terms of the density $n(\mathbf{r})$ can be defined, which is valid for any external potential $V_{ext}(\mathbf{r})$. For any $V_{ext}(\mathbf{r})$, the exact ground-state energy of this system is the global minimum of this functional, and the density $n(\mathbf{r})$ that defines this minimum is the ground-state density $n_0(\mathbf{r})$.

By combining these two theorems, it follows that when the Hohenberg-Kohn functional is available, the system properties calculable through the Schrödinger equation can be acquired by minimizing this universal functional while considering variations in the charge density. This process leads to the determination of the ground-state charge density.

2.2.2 Kohn-Sham Theory

In order to simplify the difficult problem of the fully interacting system of electrons, which is governed by Eq. (2.39), Kohn and Sham [20] proposed to replace this fully interacting system with a different, more tractable one. To that end, the Kohn-Sham (KS) approach replaces the original, fully-interacting system with that of a non-interacting system, existing in an effective potential. This effective potential incorporates all of the many-body effects through an approximated exchange-correlation functional.

The KS construction is founded on two assumptions. The first of these is that the exact ground-state density can be obtained by mapping the interacting system onto a non-interacting one. The second one is that there is an effective local potential $V_{eff}(\mathbf{r})$ acting on an electron at point $\mathbf{r}$. These assumptions allow for the construction of a new, auxiliary Hamiltonian, using Hartree atomic units where the constants are set to one

$$\hat{H} = -\frac{1}{2}\nabla^2 + V_{\text{eff}}(\mathbf{r}). \tag{2.41}$$

For a system of N independent electrons that obey Eq. (2.41), each electron will occupy one of the N resulting KS orbitals $\psi_i(\mathbf{r})$, which have the lowest energy eigenvalues ε_i. The density of the auxiliary system is then given by the summation of the absolute squares of the KS orbitals for the N lowest energy orbitals

$$n(\mathbf{r}) = \sum_{i=1}^{N} |\psi_i(\mathbf{r})|^2. \tag{2.42}$$

The Hohenberg-Kohn (HK) functional can, therefore, be constructed using this density. The independent particle kinetic energy functional T_s can be written as

$$T_s = -\frac{1}{2}\sum_{i=1}^{N}\int d^3r |\nabla\psi_i(\mathbf{r})|^2. \tag{2.43}$$

The Coulombic interaction of the electron density $n(\mathbf{r})$ with itself can be obtained as

$$E_{\text{Hartree}}[n] = \frac{1}{2}\int d^3r \int d^3r' \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}. \tag{2.44}$$

This is term is called the Hartree energy. The HK functional is therefore expressed as

$$E_{\text{KS}}[n] = T_s[n] + \int d\mathbf{r} V_{\text{ext}}(\mathbf{r})n(\mathbf{r}) + E_{\text{Hartree}}[n] + E_{II} + E_{\text{xc}}[n], \tag{2.45}$$

where E_{II} is the nucleus-nucleus interaction and $V_{\text{ext}}(\mathbf{r})$ is the external potential, which is caused by the nuclei and any other external fields.

As mentioned previously, all of the many-body effects of exchange and correlation are baked into the exchange-correlation energy functional $E_{\text{xc}}[n]$. If the true form of this functional was known, then the exact energy and ground-state density of the many-body system of electrons could be obtained. This, however, is intractable and, therefore, we must make do with approximations for this functional, some of which will be discussed in Section 2.2.3.

With a valid approximation for $E_{\mathrm{xc}}[n]$ and considering the non-interacting kinetic energy term is expressed in terms of the KS orbitals, while the other terms are functionals of the density, the chain rule can be employed to write the variational equation of the full functional as

$$\frac{\delta E_{\mathrm{KS}}}{\delta \psi_i^*(\mathbf{r})} = \frac{\delta T_s}{\delta \psi_i^*(\mathbf{r})} + \left[\frac{\delta E_{\mathrm{ext}}}{\delta n(\mathbf{r})} + \frac{\delta E_{\mathrm{Hartree}}}{\delta n(\mathbf{r})} + \frac{\delta E_{\mathrm{xc}}}{\delta n(\mathbf{r})}\right] \frac{\delta n(\mathbf{r})}{\delta \psi_i^*(\mathbf{r})} = 0. \tag{2.46}$$

Using equations Eq. (2.42) and Eq. (2.43) for $n(\mathbf{r})$ and T_s, respectively, yields

$$\frac{\delta n(\mathbf{r})}{\delta \psi_i^*(\mathbf{r})} = \psi_i(\mathbf{r}); \quad \frac{\delta T_s}{\delta \psi_i^*(\mathbf{r})} = -\frac{1}{2}\nabla^2 \psi_i(\mathbf{r}), \tag{2.47}$$

and the Lagrange multiplier method for handling the orthonormalization constraints $\langle \psi_i | \psi_j \rangle = \delta_{ij}$, gives the KS equations

$$(\hat{H}_{\mathrm{KS}} - \varepsilon_i)\psi_i(\mathbf{r}) = 0, \tag{2.48}$$

where ε_i are the energy eigenvalues and $\hat{H}_{\mathrm{KS}}$ is the effective KS Hamiltonian

$$\hat{H}_{\mathrm{KS}} = -\frac{1}{2}\nabla^2 + V_{\mathrm{KS}}(\mathbf{r}), \tag{2.49}$$

where

$$\begin{aligned} V_{\mathrm{KS}}(\mathbf{r}) =& V_{ext}(\mathbf{r}) + \frac{\delta E_{\mathrm{Hartree}}}{\delta n(\mathbf{r})} + \frac{\delta E_{\mathrm{xc}}}{\delta n(\mathbf{r})} \\ =& V_{\mathrm{ext}}(\mathbf{r}) + V_{\mathrm{Hartree}}(\mathbf{r}) + V_{\mathrm{xc}}(\mathbf{r}). \end{aligned} \tag{2.50}$$

Thus, the ground-state density can be obtained iteratively by first making a guess for the initial charge density. This guess is then used to calculate the effective potential and, thereby, solve the KS equation. The resultant KS orbitals are used to construct the charge density and a check is performed to see if the resulting density is self-consistent with the initial density, to within a pre-defined tolerance. If this is not the case, the cycle is repeated with the new density until self-consistency has been achieved.

Once the ground-state density has been obtained, all dependent quantities can also be obtained such as energy, forces, stresses etc.

2.2.3 Exchange-Correlation Energy

The single loose end remaining is how best to approximate the exchange-correlation (XC) functional, which accounts for the many-body effects of a group of N electrons, all interacting with each other. The exchange energy is a result of the antisymmetry of the electronic wavefunctions. The Pauli exclusion principle means that two electrons near each other must be antisymmetric with respect to an exchange in their positions and, therefore, contribute an exchange interaction energy to the overall system. The correlation energy is the term which accounts for any other electron-electron interaction. Correlation effects occur from the wide array of electronic interactions that involve the relative motions, distances and angles. There are entire families of approximations of the XC functional, which can ascend 'Jacob's Ladder' of density functional approximations [21], which increase in complexity and accuracy the higher the ladder is climbed as referenced in Fig. 2.5.

The first construction to consider when conceiving the best approximation for the XC functional is that of the local density approximation (LDA). The local density approximation models the XC energy as an integral over all space, where the energy density at each point is equivalent to that of a homogeneous electron gas with that density. This takes the energy density of a system which has an analytic form and assumes that the solid being modelled will be close to the homogeneous electron gas case. This assumption was first outlined in the original Kohn and Sham paper [20].

The second rung of the ladder is that of the generalized gradient approximation (GGA). This paradigm extends the idea of the LDA to incorporate the gradient of the density, as well as the value of the density at each point. The inclusion of the density gradient contributes to the GGA approximation being more accurate for large classes of compounds. The most commonly used functional in the community, as well as in this thesis, is that of the Perdew-Burke-Ernzerhof (PBE) functional [22].

The final rung of Jacob's Ladder which is pertinent to this thesis is that of the hybrid functionals. The hybrid functional used in this thesis is the HSE06 functional [23, 24], which is used in order to determine the ground-state phase of NbS_2 in Chapter 3.

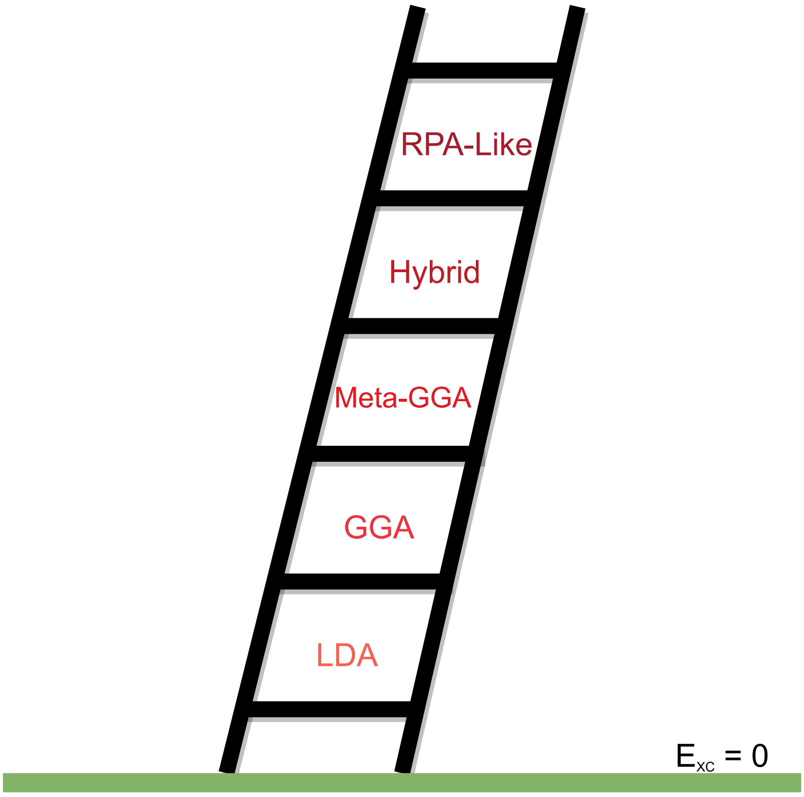

Figure 2.5: The 'Jacob's Ladder' of exchange-correlation functionals with increasingly complex but higher accuracy functionals as the ladder is ascended.

Hybrid functionals generally mix the contributions of Hartree-Fock theory [25, 26] and Kohn-Sham functionals to the exchange energy such that the functional is a mix of the two contributions according to some mixing parameter. The correlation term is taken from the GGA approximation. HSE06 is a form of hybrid functional, which employs the GGA-PBE functional for the GGA part of the functional and for the long-range exchange interaction, as the long-range and short-range interactions are separated. Thus, the hybrid nature of the functional is only present for the short-range exchange interaction. This hybrid functional introduces a screened Coulombic interaction, which is controlled by a screening parameter. The different HSE functionals differ only by this screening parameter.

The other two rungs on the ladder in Fig. 2.5 are meta-GGA [27] and random phase approximation [28] or RPA-type functionals. Neither of these functionals are used anywhere in this thesis so I will not go into any detail on these particular functional types. However, in brief, meta-GGA includes a consideration of the second derivative of the density, or of the KS kinetic energy density, as well as the gradient of the density and the density itself. RPA, in contrast, involves the estimation of the electron correlation effects based on the amplitude of fluctuations of the electron density.

2.2.4 Pseudopotentials

The properties of solids are generally dependent only on the interactions involving the outermost, valence electrons. The core electrons have little to no part to play in any interactions and, therefore, calculating their contributions to the overall system is unnecessary. A pseudopotential is a means of reducing the computational complexity resulting from the inclusion of these core electrons in the DFT calculation, which causes the wavefunction of the valence electrons to oscillate rapidly as it approaches the core.

These oscillations require increased computational cost to model, with little contribution to the physics of the system. Replacing the combined strong Coulomb potential and the screening effect of the tightly bound core electrons with a single effective potential will smooth out this wavefunction as it approaches the core. A process which creates a pseudo-wavefunction, which models the valence electrons well after some predefined cut-off distance from the ionic core. The valence electrons are, thereby, the only consideration necessary for the full DFT calculation, reducing cost by eliminating the rapid oscillations near the ion.

Similarly, in the case of the projector augmented wave (PAW) [29] approach, the wavefunctions of valence electrons, as they approach the cores, are transformed from the rapidly oscillating case, resulting from the core electrons to a smooth wavefunction. In contrast to the previous method, however, the all-electron properties can be obtained by a linear transformation of the resulting pseudo-wavefunction, using projectors, which map the pseudo-orbitals within some cutoff, onto the KS orbitals.

`VASP` [30–33] is the DFT library that is primarily used in this thesis and the calculations that are performed use the `VASP` library of potentials that employ the PAW method for their construction. The other library that is used is that of `FHI-AIMS`, which employs all-electron DFT and, therefore, does not use pseudopotentials.

2.2.5 Basis Sets

There are two main methods for solving the KS equation, corresponding to two types of basis sets that can be used to describe the KS orbitals. The first is the use of a basis set where the KS orbitals are approximated using some linear combination of functions, which are localized in space. These localized basis sets are generally comprised of Gaussian-type orbitals or Slater-type orbitals. These basis sets are not periodic and are, therefore, generally not used for periodic systems. However, they are useful for molecules and non-periodic systems as the vacuum surrounding these systems does not have to be modelled by the basis set in this case.

The one that is used throughout this thesis is the basis set that results from a linear combination of plane waves

$$\psi_i(\mathbf{r}) = \sum_k A_i(\mathbf{k}) \exp(i\mathbf{k} \cdot \mathbf{r}). \tag{2.51}$$

Plane-wave basis sets are very useful for modelling extended periodic systems, due to the periodicity naturally being built into the basis. Furthermore, the use of these basis sets means it is relatively simple to improve the overall accuracy of the calculation, as all that is required to do so is an increase in the number of plane waves representing the wavefunction. This is achieved by increasing the cut-off energy of the calculation. All plane waves with a kinetic energy less than the cut-off energy are included in the calculation. The more plane waves which are included improves the quality of the calculation, however, the increase in computational cost can be quite severe. Non-periodic systems, which require a large number of plane-wave basis functions to model the vacuum tend to be very expensive for the same reason and will generally require a very large cut-off energy.

2.3 Summary

This chapter first gave a description of machine learning (ML), including the overall inspiration for, and aims of the field. Further to this, a comprehensive description of supervised learning was outlined, including a discussion on the two main tasks involved in supervised learning, regression and classification. The concepts of over-

and underfitting were outlined, as well as a variety of strategies of accounting for these effects to improve the performance of ML models. In particular, the optimal composition and splitting of the training datasets into a train, validation and test split was outlined to account effectively for overfitting.

This led naturally to a discussion on various loss functions that can be employed for different supervised tasks, which must be minimized to optimize the performance of the ML model. Followed by a discussion on the various evaluation metrics that commonly appear in discussions on ML performance, which are obtained using the test dataset. The section on supervised ML was concluded by introducing both ridge regression and random forest, which are used throughout this thesis.

Once the discussion on supervised ML was complete, a brief discussion on unsupervised ML ensued that outlined the two main tasks it is suited for. Namely, those of clustering and dimensionality reduction. A description of neural networks (NNs) was subsequently given, giving an overview of both forward and back-propagation of information through NNs such that the whole network can be optimized for a given task.

Finally, the key concepts in density functional theory (DFT) were presented, including the Hohenberg-Kohn theorems, as well as the Kohn-Sham theory, both of which are foundational to the field. The relevance of the exchange-correlation term in the Kohn-Sham Hamiltonian was discussed and various methods of approximating it were given. Finally, a brief discussion of both pseudopotentials and basis sets for describing the Kohn-Sham orbitals was delivered.

This chapter has given an overview of the key concepts necessary to understand and execute the work that follows in this thesis, encompassing the foundations of two methodologies that consistently appear throughout. In the following chapters, these methods will be utilized both together and separately. This will be done along with a variety of other techniques, which will be outlined in the individual chapters, such that pipelines can be constructed to further the aims of inverse-materials design.

Chapter 3

First-Principles Phonon Calculations

"Breithnigh an abhainn sara dtéir ina cuilithe."

Before we are able to fully understand the capabilities of machine learning (ML) for bypassing first-principles calculations, we must formulate benchmarks or ground truths with which to compare the ML results. As Chapter 4 focuses mainly on using ML to bypass *ab initio* techniques for calculating vibrational and thermal properties of quasi-two-dimensional (2D) materials, methods to gain a ground-truth or benchmark of the vibrational behaviours of 2D materials must be considered. Thus, it is of great importance to fully understand the means of obtaining these benchmarks in this context.

Our focus is on a specific 2D material, NbS_2, known for its propensity to exhibit a diverse array of intricate and interlinked phenomena in its ground state. Remarkably, capturing these behaviours using traditional first-principles methods has proven challenging, let alone with the aid of ML techniques.

The calculations performed as part of this chapter introduce the methods that will be used subsequently to gauge the efficacy of any ML model to adequately replicate the first-principles dynamics and behaviour of complex 2D systems and their composites. Further to this, a full characterization of the complex interplays of behaviours of monolayer 2H-NbS_2 will be outlined. These results will prove in the next chapter to be useful for the discussion of the limitations of ML potentials. Thus, will form a large part of the understanding of such a model's capabilities and the intuition of their employment for valid use in ML pipelines for materials discovery, with a focus on pipelines for thermal property prediction.

In the first part of this chapter, I will outline in full the techniques for modelling phonons from first principles, laying the groundwork for subsequent analyses. In the latter part, I will present results outlining their use for the calculations of vibrational properties of NbS_2. This represents a holistic view of the phenomena underpinning the ground-state behaviour of the system, much of which is unique to 2D system dynamics.

3.1 Phonons

Much of the theory in this section has been adapted from a few diverse sources. Namely, the main body of the theory is adapted from a combination of Ref. [34] and Ref. [35], and some clarifications were taken from Ref. [36].

The first concept to introduce when establishing methods to capture the dynamics of materials and, thereby obtaining an understanding of their related quantities is that of phonons. While phonons play a vital role in our understanding of various thermal phenomena, mathematically interpreting heat transfer through a lattice as a result of individual atoms vibrating with excess energy presents a significant challenge. In this scenario, neighbouring atoms also vibrate, making it challenging to predict macroscopic properties like thermal conductivity. Thus, it is imperative that such an understanding be replaced with a theory that, instead of considering such individual behaviours of atoms at lattice sites, investigates the behaviour of a collective model.

This change in perspective directly follows the reasoning of the Einstein model for the specific heat, which considers all atoms in a lattice as being independent oscillators. This compares with the Debye model, which considers the specific heat of a solid as being derived from collective vibrations of lattice sites being considered as a gas of phonons in a box, where the box is the solid in question. In this model, the thermal energy of the system is distributed along the normal modes of the total crystal vibration as opposed to the individual contributions from individual lattice sites.

These collective lattice waves have the ability to transport thermal energy. However, a mechanism must be considered by which the resistance encountered by transporting such energy through the lattice can be effectively described. Phonon-phonon collisions, which are the scattering of phonons mediating the transport of this energy, are a valuable means of describing this diffusion of thermal energy as it travels through a medium. This does apparently contradict the definition of normal modes as being incapable of interacting with each other. However, this apparent contradiction is compensated for by eliminating an original assumption that the derivation of the model makes. This assumption is that the lattice sites oscillate in a purely quadratic potential. In reality, there are anharmonic terms also included in the potential, disrupting the initial assumptions. A discussion on the impacts of anharmonicity in phonon-phonon interactions is generally beyond the scope of this Chapter, but it will be touched upon in Chapter 4.

In order to model the properties of a solid-state body, some conventions must be established. The first of these is the definition of the crystal lattice, which is a regular array of sites in three-dimensional space, representing the positions of atoms in a crystal in its ground state. Formally, it is a set of translations with vectors satisfying the relationship

$$\boldsymbol{l} = l_1\boldsymbol{a_1} + l_2\boldsymbol{a_2} + l_3\boldsymbol{a_3}, \tag{3.1}$$

where $\boldsymbol{a_1}$, $\boldsymbol{a_2}$ and $\boldsymbol{a_3}$ are the primitive lattice vectors, and l_1, l_2 and l_3 are all integers, whose values lie inside a range defined by the size of the crystal as a whole.

From this fixed set of points in a grid, we can conceive our unit cell, which surrounds each grid point. This unit cell captures the fundamental unit from which the periodicity of our system is constructed. It is desirable to select the simplest possible choice for this unit cell, containing the minimum possible number of atoms. If there is a single atom in each cell, we have a Bravais lattice. If this is not the case, we require a basis, which is a set of vectors $\boldsymbol{b_1}, \boldsymbol{b_2}, ..., \boldsymbol{b_n}$ for n particles in the unit cell, defining the positions of the atoms in the unit cell, relative to an atom that we choose to be at the origin.

In order to discuss the motion of the lattice, it is necessary to establish the dynamical equations of the system in question. To obtain these equations, we must consider a Hamiltonian for the entire system. To aid in the formulation of our Hamiltonian, the adiabatic or Born-Oppenheimer approximation must be made such that the contribution to the potential energy from the dynamics of the electrons can be separated from that of the nuclear dynamics. This approximation operates on the assumption that the electrons move so much more rapidly than the atomic nuclei, that the potential energy, V, can be approximated purely as a function of atomic position $\boldsymbol{x_l}$, allowing the electronic contribution to be captured by an effective potential. The kinetic energy of the system, is also, simply, the sum of their individual momenta $\mathbf{p}_{\boldsymbol{l}}$. Thus the Hamiltonian corresponding to the nuclear dynamics, with $m_{\boldsymbol{l}}$ being the mass of the $\boldsymbol{l}$th atom, becomes

$$\hat{H} = \frac{1}{2}\sum_{\boldsymbol{l}} \frac{\mathbf{p}_{\boldsymbol{l}}^2}{m_{\boldsymbol{l}}} + V(\boldsymbol{x}_1, \boldsymbol{x}_2 ... \boldsymbol{x}_{\boldsymbol{l}} ...). \tag{3.2}$$

Considering the most general case of a lattice with a basis, each atom in the system is labelled using two symbols, $\boldsymbol{l}$ and $\boldsymbol{b}$, where $\boldsymbol{l}$ is defined as in Eq. (3.1) and $\boldsymbol{b}$ is the basis vector to the atom in the cell. Any configuration of the system in question may be given, therefore, using the coordinate vectors $\boldsymbol{x_{lb}}$ of all the atoms. A more natural choice of coordinate system, however, would be one in which the atomic positions are given relative to the equilibrium state, i.e.

$$\mathbf{u}_{\boldsymbol{lb}} = \boldsymbol{x_{lb}} - (\boldsymbol{l} + \boldsymbol{b}), \tag{3.3}$$

resulting in a significant simplification of our calculations as we now allow the equilibrium state to be a minimum in the potential energy of the system.

This convention allows us to expand the potential energy out into a Taylor series in the various vector displacements,

$$\begin{aligned} V(\mathbf{u}_{\boldsymbol{lb}\alpha}) &= V_0 + \frac{1}{2}\sum_{\boldsymbol{lb}\alpha;\boldsymbol{l'b'}\alpha'} \mathbf{u}_{\boldsymbol{lb}\alpha} \cdot \left.\frac{\partial^2 V}{\partial \mathbf{u}_{\boldsymbol{lb}\alpha}\partial \mathbf{u}_{\boldsymbol{l'b'}\alpha'}}\right|_{\mathbf{u}=\mathbf{0}} \cdot \mathbf{u}_{\boldsymbol{l'b'}\alpha'} \\ &= V_0 + \frac{1}{2}\sum_{\boldsymbol{lb}\alpha;\boldsymbol{l'b'}\alpha'} \mathbf{u}_{\boldsymbol{lb}\alpha} \cdot \boldsymbol{\Phi}_{\boldsymbol{lb}\alpha;\boldsymbol{l'b'}\alpha'} \cdot \mathbf{u}_{\boldsymbol{l'b'}\alpha'}, \end{aligned} \tag{3.4}$$

where $\alpha, \alpha' = (1, 2, 3)$, corresponding to the Cartesian direction of the vector component. The constant quantity V_0 is arbitrary and can be set to zero for simplicity. The significance of the construction $\boldsymbol{\Phi}_{\boldsymbol{lb};\boldsymbol{l'b'}} \cdot \mathbf{u}_{\boldsymbol{l'b'}}$ is the force acting on atom $(\boldsymbol{l}, \boldsymbol{b})$, when the atom occupying $(\boldsymbol{l'}, \boldsymbol{b'})$ is displaced by $\mathbf{u}_{\boldsymbol{l'b'}}$. Given that we have truncated the Taylor series expansion at the second order term, as part of the harmonic approximation of the potential energy, $\boldsymbol{\Phi}_{\boldsymbol{lb};\boldsymbol{l'b'}}$ is the matrix of second order force constants. If we had continued the expansion and included anharmonicity, we would have had to consider the force constants of the third-order, fourth-order etc.

Taking these conventions into account, the system is described by the Hamiltonian operator

$$\hat{H} = \frac{1}{2}\sum_{\boldsymbol{lb}\alpha}\left(\frac{1}{m_{\boldsymbol{b}}}\right)\mathbf{p}_{\boldsymbol{lb}\alpha}\cdot\mathbf{p}_{\boldsymbol{lb}\alpha} + \frac{1}{2}\sum_{\boldsymbol{lb}\alpha;\boldsymbol{l'b'}\alpha'}\mathbf{u}_{\boldsymbol{lb}\alpha}\cdot\boldsymbol{\Phi}_{\boldsymbol{lb}\alpha;\boldsymbol{l'b'}\alpha'}\cdot\mathbf{u}_{\boldsymbol{l'b'}\alpha'}, \tag{3.5}$$

whose eigenvalues and eigenfunctions we wish to obtain to ascertain the permitted vibrational modes of the system and their corresponding vibrational energy. To simplify this derivation further, we can rewrite Eq. (3.5) as

$$\hat{H} = \frac{1}{2}\sum_{\boldsymbol{l},\boldsymbol{b},\alpha}\frac{1}{m_{\boldsymbol{b}}}\left(\frac{\partial\mathbf{u}_{\boldsymbol{l},\boldsymbol{b},\alpha}}{\partial t}\right)^2 + \frac{1}{2}\sum_{\boldsymbol{lb}\alpha;\boldsymbol{l'b'}\alpha'}\mathbf{u}_{\boldsymbol{lb}\alpha}\cdot\boldsymbol{\Phi}_{\boldsymbol{lb}\alpha;\boldsymbol{l'b'}\alpha'}\cdot\mathbf{u}_{\boldsymbol{l'b'}\alpha'}, \tag{3.6}$$

This can be further simplified by taking the mass-normalized displacement by scaling the atomic displacement based on the mass of the atom in question, i.e. $\tilde{\mathbf{u}}_{\boldsymbol{lb}\alpha} = \sqrt{m_{\boldsymbol{b}}}\mathbf{u}_{\boldsymbol{lb}\alpha}$. With this definition, Eq. (3.6) can be rewritten as

$$\hat{H} = \frac{1}{2}\sum_{\boldsymbol{l},\boldsymbol{b},\alpha}\left(\frac{\partial\tilde{\mathbf{u}}_{\boldsymbol{l},\boldsymbol{b},\alpha}}{\partial t}\right)^2 + \frac{1}{2}\sum_{\boldsymbol{lb}\alpha;\boldsymbol{l'b'}\alpha'}\tilde{\mathbf{u}}_{\boldsymbol{lb}\alpha}\left(\frac{1}{\sqrt{m_{\boldsymbol{b}}}}\boldsymbol{\Phi}_{\boldsymbol{lb}\alpha;\boldsymbol{l'b'}\alpha'}\frac{1}{\sqrt{m_{\boldsymbol{b'}}}}\right)\tilde{\mathbf{u}}_{\boldsymbol{l'b'}\alpha'}. \tag{3.7}$$

The reduced force constants $\tilde{\Phi}_{\boldsymbol{lb}\alpha;\boldsymbol{l'b'}\alpha'} = \frac{1}{\sqrt{m_{\boldsymbol{b}}}}\boldsymbol{\Phi}_{\boldsymbol{lb}\alpha;\boldsymbol{l'b'}\alpha'}\frac{1}{\sqrt{m_{\boldsymbol{b'}}}}$ are a square, symmetric matrix, and can therefore be diagonalised using eigenvalue decomposition as

$$\tilde{\boldsymbol{\Phi}} = \mathrm{U}\Omega^2\mathrm{U}^\top, \tag{3.8}$$

where $\Omega^2 = diag(...,\omega_\xi^2,...)$, the diagonal matrix of eigenvalues, corresponding to the squared frequencies of each normal mode and U is an orthogonal matrix where each column is the eigenvector $w_{\boldsymbol{lb}\alpha}(\xi)$. Here, ξ is the index of each normal mode of the system. The Hamiltonian in Eq. (3.7) is written in a matrix notation with the introduction of the column matrix $\tilde{\mathbf{u}} = (\sqrt{m_{\boldsymbol{b}}} u_{\boldsymbol{lb}\alpha} ...)^{\top}$ as

$$\begin{aligned} \hat{H} &= \frac{1}{2}\dot{\tilde{\mathbf{u}}}^{\top}\dot{\tilde{\mathbf{u}}} + \frac{1}{2}\tilde{\mathbf{u}}^{\top}\tilde{\boldsymbol{\Phi}}\tilde{\mathbf{u}} \\ &= \frac{1}{2}(\mathrm{U}^{\top}\dot{\tilde{\mathbf{u}}})^{\top}(\mathrm{U}^{\top}\dot{\tilde{\mathbf{u}}}) + \frac{1}{2}(\mathrm{U}^{\top}\tilde{\mathbf{u}})^{\top}\Omega^2(\mathrm{U}^{\top}\tilde{\mathbf{u}}) \\ &= \frac{1}{2}\dot{\mathbf{Q}}^{\top}\dot{\mathbf{Q}} + \frac{1}{2}\mathbf{Q}^{\top}\Omega^2\mathbf{Q}, \end{aligned} \tag{3.9}$$

where $\mathbf{Q}$ gives the normal coordinates of the system, defined as $\tilde{\mathbf{u}} = \mathrm{U}\mathbf{Q}$. This gives a construction that relates the mass-normalised displacements $\tilde{\mathbf{u}}$ with the eigenvectors $w_{\boldsymbol{lb}\alpha}(\xi)$. The normal mode coordinates are a derived coordinate system, representing the contribution of a given phonon mode to the mass-normalized displacement of an atom. The dots above $\tilde{\mathbf{u}}$ and $\mathbf{Q}$ imply that these quantities are time derivatives. The eigenvalue problem shown in Eq. (3.8) can be rewritten explicitly as

$$\sum_{\boldsymbol{l'b'}\alpha'} \tilde{\boldsymbol{\Phi}}_{\boldsymbol{lb}\alpha;\boldsymbol{l'b'}\alpha'} w_{\boldsymbol{l'b'}\alpha'}(\xi) = \omega_{\xi}^2 w_{\boldsymbol{lb}\alpha}(\xi). \tag{3.10}$$

Taking our analysis from real space into reciprocal space greatly simplifies the calculation of the vibrational frequencies of the different normal modes. Taking advantage of the periodicity of our system, this is achieved by use of the Bloch theorem, allowing us to rewrite the eigenvector for the wave vector $\mathbf{q}$ as

$$w_{\boldsymbol{lb}\alpha,\mathbf{q}}(\xi) = \frac{1}{\sqrt{N}} W_{\boldsymbol{b}\alpha}(\xi) e^{i\mathbf{q}\cdot(\boldsymbol{l}+\boldsymbol{b})}, \tag{3.11}$$

where $W_{\boldsymbol{b}\alpha}(\xi)$ is a function capturing the periodicity of the crystal lattice and $\frac{1}{\sqrt{N}}$ is a normalizing factor, where N corresponds to the total number of unit cells in the solid. By combining Eq. (3.10) and Eq. (3.11), multiplying both sides by the factor $e^{-i\mathbf{q}\cdot(\boldsymbol{l}+\boldsymbol{b})}/\sqrt{N}$ and summing over $\boldsymbol{l}$ on both sides, we obtain

$$\sum_{\boldsymbol{b'}\alpha'} D_{\boldsymbol{b}\alpha;\boldsymbol{b'}\alpha'}(\mathbf{q}) W_{\boldsymbol{b'}\alpha'}(\mathbf{q}\nu) = \omega_{\mathbf{q}\nu}^2(\mathbf{q}) W_{\boldsymbol{b}\alpha}(\mathbf{q}\nu), \tag{3.12}$$

where now the index ξ is replaced with the index $\mathbf{q}\nu$, combining the wave vector $\mathbf{q}$ with the index of the phonon band, ν. This new eigenvalue equation introduces the dynamical matrix,

$$\begin{aligned}
D_{\boldsymbol{b}\alpha;\boldsymbol{b}'\alpha'}(\mathbf{q}) &= \sum_{\boldsymbol{l}\boldsymbol{l}'} \frac{e^{-i\mathbf{q}\cdot(\boldsymbol{l}+\boldsymbol{b})}}{\sqrt{N}} \tilde{\boldsymbol{\Phi}}_{\boldsymbol{l}\boldsymbol{b}\alpha;\boldsymbol{l}'\boldsymbol{b}'\alpha'} \frac{e^{i\mathbf{q}\cdot(\boldsymbol{l}'+\boldsymbol{b}')}}{\sqrt{N}} \\
&= \frac{1}{N\sqrt{m_{\boldsymbol{b}} m_{\boldsymbol{b}'}}} \sum_{\boldsymbol{l}\boldsymbol{l}'} \boldsymbol{\Phi}_{\boldsymbol{l}\boldsymbol{b}\alpha;\boldsymbol{l}'\boldsymbol{b}'\alpha'} e^{i\mathbf{q}\cdot(\boldsymbol{l}'+\boldsymbol{b}'-\boldsymbol{l}-\boldsymbol{b})} \\
&= \frac{1}{\sqrt{m_{\boldsymbol{b}} m_{\boldsymbol{b}'}}} \sum_{\boldsymbol{l}'} \boldsymbol{\Phi}_{\boldsymbol{0}\boldsymbol{b}\alpha;\boldsymbol{l}'\boldsymbol{b}'\alpha'} e^{i\mathbf{q}\cdot(\boldsymbol{l}'+\boldsymbol{b}'-\boldsymbol{b})},
\end{aligned} \tag{3.13}$$

where the final definition of $D_{\boldsymbol{b}\alpha;\boldsymbol{b}'\alpha'}(\mathbf{q})$ in Eq. (3.13) is achieved by considering the lattice translational symmetry of the force constant matrix. From this definition of the dynamical matrix, it can be understood to be the mass-reduced Fourier transform of the matrix of second-order force constants $\boldsymbol{\Phi}_{\boldsymbol{l}\boldsymbol{b}\alpha;\boldsymbol{l}'\boldsymbol{b}'\alpha'}$. Thus, in order to obtain the dynamical matrix, all that is required to compute is $\boldsymbol{\Phi}_{\boldsymbol{l}\boldsymbol{b}\alpha;\boldsymbol{l}'\boldsymbol{b}'\alpha'}$.

The construction $\tilde{\mathbf{u}} = \mathrm{U}\mathbf{Q}$ from Eq. (3.9) can be reformulated in explicit notation and combined with Eq. (3.11) to give a relationship between the displacement of an atom $\mathbf{u}_{\boldsymbol{l}\boldsymbol{b}\alpha}$ and the phonon coordinates or the normal-mode coordinate in reciprocal space $Q(\mathbf{q}\nu)$,

$$\mathbf{u}_{\boldsymbol{l}\boldsymbol{b}\alpha}(\mathbf{q}\nu) = \frac{1}{\sqrt{N m_{\boldsymbol{b}}}} \sum_{\mathbf{q}\nu} Q(\mathbf{q}\nu) W_{\boldsymbol{b}\alpha}(\mathbf{q}\nu) e^{i\mathbf{q}\cdot(\boldsymbol{l}+\boldsymbol{b})}. \tag{3.14}$$

Through the use of the definition of phonon creation $a^{\dagger}_{\mathbf{q}\nu}$ and annihilation $a_{\mathbf{q}\nu}$ operators, the solution to the quantum harmonic oscillator problem implies that $Q(\mathbf{q}\nu)$ can be written as [36]

$$Q(\mathbf{q}\nu) = \sqrt{\frac{\hbar}{2\omega_{\mathbf{q}\nu}}}(a_{\mathbf{q}\nu} + a^{\dagger}_{-\mathbf{q}\nu}), \tag{3.15}$$

where $\hbar$ is the Planck constant. This allows us to reconstruct the Hamiltonian for a harmonic system, using these creation and annihilation operators as a sum over all phonon modes

$$\begin{aligned}
\hat{H} &= \sum_{\mathbf{q}\nu} \hbar\omega_{\mathbf{q}\nu} \left(a^{\dagger}_{\mathbf{q}\nu} a_{\mathbf{q}\nu} + \frac{1}{2} \right) \\
&= \sum_{\mathbf{q}\nu} \hbar\omega_{\mathbf{q}\nu} \left(n_{\mathbf{q}\nu} + \frac{1}{2} \right),
\end{aligned} \tag{3.16}$$

where $n_{\mathbf{q}\nu} = a^{\dagger}_{\mathbf{q}\nu} a_{\mathbf{q}\nu}$ is the occupation number operator of the $\mathbf{q}\nu$ phonon mode. The occupation number is temperature dependent and its expectation value is given by the Bose-Einstein distribution, as phonons are bosonic quasiparticles

$$\langle n_{\mathbf{q}\nu} \rangle = \frac{1}{e^{\frac{\hbar\omega_{\mathbf{q}\nu}}{k_B T}} - 1}, \tag{3.17}$$

where k_B is the Boltzmann constant and T is the temperature of the system. This derivation highlights the analogous relationship between phonons and photons as being a means of mediating energy. The phonons quantize the vibrational energy into discrete amounts, each with an energy of $\hbar\omega_{\mathbf{q}\nu}$, much like the case of photons for electromagnetic radiation. Furthermore, when the temperature increases, the implication is that the thermal energy is distributed between a larger quantity of phonon modes in a greater and greater number of excited states.

3.1.1 Frozen-Phonon/Finite Displacement Calculations

Looking at the problem in a practical sense, the key to understanding which phonon modes are important to the contribution of this vibrational energy transfer is, therefore, the dynamical matrix, defined in Eq. (3.13). Diagonalizing this operator gives rise to each phonon frequency and eigenvector, which can then be used to surmise the phonon coordinates corresponding to the largest degree of energy transfer. The discussion must then turn to methods with which to obtain this dynamical matrix.

The first of these methods is arguably the most common and intuitive means of obtaining the dynamical matrix, known as the frozen-phonon or finite displacement method. This process is executed simply by obtaining the restoring forces an atom experiences with a series of different atomic displacements

$$-f_{\boldsymbol{l}\boldsymbol{b}\alpha} = \sum_{\boldsymbol{l'}\boldsymbol{b'}\alpha'} \boldsymbol{\Phi}_{\boldsymbol{l}\boldsymbol{b}\alpha;\boldsymbol{l'}\boldsymbol{b'}\alpha'} \cdot \mathbf{u}_{\boldsymbol{l'}\boldsymbol{b'}\alpha'}. \tag{3.18}$$

This sequence of simultaneous linear equations is solved for each atom by constructing a supercell model of the system. This supercell is comprised of a linear combination of unit-cell basis vectors. In this new, expanded system, an atom is displaced and the supercell periodicity is broken by this displacement as is represented in Fig. 3.1

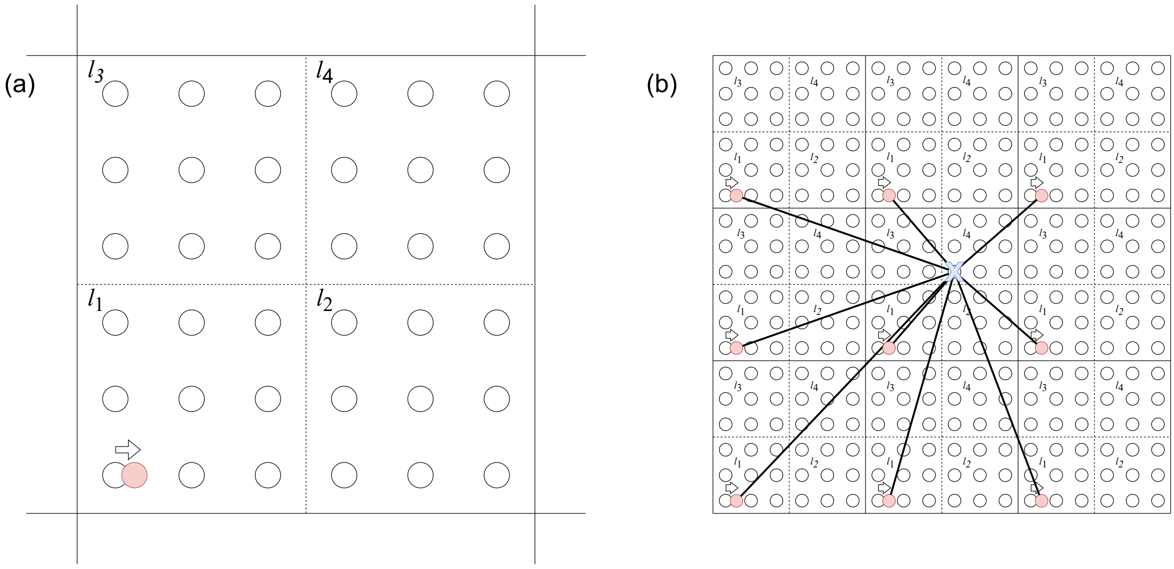

Figure 3.1: A schematic diagram of the procedure for calculating the matrix of force constants for a given system using the finite displacement method. **(a)** Depicts the expansion of the unit cell into a 2 × 2 supercell, where l_1, l_2, l_3 and l_4 are the indices of the constituent unit cells. The finite displacement of an atom in the supercell is depicted in red, breaking the symmetry within the supercell. **(b)** The periodic boundary conditions means the atom marked with the blue X will feel a resultant force due to the displacement of atoms in every mirror image of the supercell from the periodic boundary conditions. *This diagram was adapted from Ref. [35].*

(a). This leads to the supercell itself being the new unit of periodicity. Thus, this displacement is mirrored by all perceived mirror images of the supercell surrounding the one of interest as a result of the periodic boundary conditions of the system. The force acting on each atom in the supercell is calculated as a response of all of the displacements from every mirror image of the supercell Fig. 3.1 (b).

The number of unit cells in a given supercell is a parameter of the calculation that must be determined by converging the results of the calculations. If the forces calculated vary for a given atom in the supercell with increasing supercell size, the calculation is not yet converged. This is due to the fact that atomic interactions will decay with increasing distance. Therefore, if a supercell of sufficient magnitude is used, the dynamical matrix becomes a good approximation. Displacements must be applied on different atoms in the configuration in order to correctly model the forces acting between every atom and thereby identify the interatomic force constants directly. Various symmetries can be used for a given configuration that stands to reduce the computational complexity of the calculation by performing a single calculation for equivalent atomic sites.

The value of the finite-displacement method over other ways to obtain the dynamical matrix is that it is only dependent on the ability of a model to obtain the forces exerted on atoms by a less computationally intensive force calculator, instead of exclusively relying on first-principles methods, such as DFT. However, it is certainly reasonable to use such methods too. With this in mind, any quantum-accurate method for calculating interatomic forces should produce an extremely accurate picture of the quantized vibrational dynamics of a given system. The use of non-DFT force calculators is beneficial when there exists a system that would be very costly to model with *ab initio* methods due to the system size, or the amount of vacuum that would need to be considered.

3.1.2 Density Functional Perturbation Theory

The finite-displacement method is an intuitive and useful choice for obtaining the force constant matrix in a simple way. This method does suffer from some drawbacks when being used in conjunction with first-principles methods. This is particularly true when the method is employed for complex systems or unusual geometries. For large system sizes, for example, having a greater number of atoms in the primitive cell means that more calculations must be performed for the supercell. A fact that means finite-difference scales quite poorly with system size. Furthermore, if a system lacks symmetries, the reduction in the number of calculations due to the system symmetry can no longer be performed, further compounding this scaling issue. This is particularly true for few-layer, two-dimensional systems when employing DFT-based force calculators with a plane-wave basis set. As this basis set is intrinsically periodic in nature, there is a necessity to increase the distance between layers such that any inter-layer interactions effectively vanish. This requires a very large plane-wave cut-off in order to mimic the region devoid of charge density in the vacuum by the summation of a large number of plane-waves. The problem is avoided by using a local, or a hybrid basis, but here `VASP` was exclusively employed so that the resultant values are comparable to the standard calculations in the `AFLOW` repository [37, 38], which also employ `VASP`.

Coupling this concern with the likely large supercell that is necessary to capture the sometimes complex dynamics of quasi-2D monolayer means that such calculations may become prohibitively expensive for a sufficiently accurate picture of the system dynamics to be accessible. Thus, it is necessary that a method for obtaining the dynamical matrix of a given system is conceived, avoiding the need for this potentially expensive calculation, which involves the construction of a large number of large-scale supercells. Such a method is density functional perturbation theory (DFPT). This derivation will follow much of the logic of Ref. [39].

In order to understand DFPT, we must revisit the electronic Hamiltonian, moving in the field of fixed nuclei,

$$\hat{H}(\mathbf{R}) = -\frac{\hbar}{2m_e}\sum_i \frac{\partial^2}{\partial \mathbf{r}_i^2} + \frac{e^2}{2}\sum_{i\neq j}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} - \sum_{iI}\frac{Z_I e^2}{|\mathbf{r}_i - \mathbf{R}_I|} + E_N(\mathbf{R}), \tag{3.19}$$

where $\mathbf{R} = \boldsymbol{l} + \boldsymbol{b}$, m_e is the mass of the electron, Z_I is the charge of the I-th nucleus, $\mathbf{r}_i$ is the position of the i-th electron and E_N is the electrostatic interaction between different nuclei,

$$E_N(\mathbf{R}) = \frac{e^2}{2}\sum_{I\neq J}\frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}. \tag{3.20}$$

We can take a look at the Hellman-Feynman (HF) theorem [40] to determine a new means of calculating our force constants straight from our DFT functional. The HF theorem states that the first derivative of the eigenvalues of a Hamiltonian, dependent on a given parameter λ, $\hat{H}_\lambda$ can be given by the expectation value of the derivative of the Hamiltonian with respect to that parameter, namely,

$$\frac{\partial E_\lambda}{\partial \lambda} = \left\langle \Psi_\lambda \left| \frac{\partial \hat{H}_\lambda}{\partial \lambda} \right| \Psi_\lambda \right\rangle, \tag{3.21}$$

where Ψ_λ is the eigenfunction of $\hat{H}_\lambda$ corresponding to the E_λ eigenvalue, given by the eigenvalue equation $\hat{H}_\lambda \Psi_\lambda = E_\lambda \Psi_\lambda$. The Born-Oppenheimer approximation allows us to give the electronic Hamiltonian as a function of atomic coordinates, Eq. (3.19). According to the HF theorem, the force acting on the I-th atom can therefore be given by

$$\mathbf{F}_I = -\frac{\partial E(\mathbf{R})}{\partial \mathbf{R}_I} = -\left\langle \Psi(\mathbf{R}) \left| \frac{\partial \hat{H}(\mathbf{R})}{\partial \mathbf{R}_I} \right| \Psi(\mathbf{R}) \right\rangle, \tag{3.22}$$

where $\Psi(\mathbf{r},\mathbf{R})$ is the wave function of the ground state obtained from the electronic Hamiltonian Eq. (3.19). Within the Born-Oppenheimer approximation, the dependency of the Hamiltonian on $\mathbf{R}$ is only determined by the electron-ion interaction. Thus, the force acting on the nucleus is dependent only on the electron charge density. The HF theorem, therefore, states that

$$\mathbf{F}_I = -\int n_{\mathbf{R}}(\mathbf{r})\frac{\partial V_{\mathbf{R}}(\mathbf{r})}{\partial \mathbf{R}_I}d\mathbf{r} - \frac{\partial E_N(\mathbf{R})}{\partial \mathbf{R}_I}, \tag{3.23}$$

where $n_{\mathbf{R}}(\mathbf{r})$ is the ground-state electron charge density given by atomic configuration $\mathbf{R}$, and $V_{\mathbf{R}}(\mathbf{r})$ is the electron-nucleus interaction,

$$V_{\mathbf{R}}(\mathbf{r}) = -\sum_I \frac{Z_I e^2}{|\mathbf{r}-\mathbf{R}_I|}. \tag{3.24}$$

The Hessian matrix, which is the matrix of second-order partial derivatives of a scalar function, of the system energies can be obtained by taking the partial derivative of Eq. (3.23) with respect to nuclear coordinates, i.e.

$$\begin{aligned}\frac{\partial^2 E(\mathbf{R})}{\partial \mathbf{R}_I \partial \mathbf{R}_J} \equiv -\frac{\partial \mathbf{F}_I}{\partial \mathbf{R}_J} = \int \frac{\partial n_{\mathbf{R}}(\mathbf{r})}{\partial \mathbf{R}_J}\frac{\partial V_{\mathbf{R}}(\mathbf{r})}{\partial \mathbf{R}_I}d\mathbf{r} \\ + \int n_{\mathbf{R}}(\mathbf{r})\frac{\partial^2 V_{\mathbf{R}}(\mathbf{r})}{\partial \mathbf{R}_I \mathbf{R}_J}d\mathbf{r} + \frac{\partial^2 E_N(\mathbf{R})}{\partial \mathbf{R}_I \partial \mathbf{R}_J}.\end{aligned} \tag{3.25}$$

The Hessian matrix of the system energies with respect to a change in atomic configuration is the matrix of interatomic force constants $\mathbf{\Phi}_{IJ} = \partial^2 E(\mathbf{R})/\partial \mathbf{R}_I \partial \mathbf{R}_J$ necessary to construct the dynamical matrix as in Eq. (3.13). Thus, $\mathbf{\Phi}_{IJ}$ can be obtained by calculating the ground state of the electron charge density $n_{\mathbf{R}}(\mathbf{r})$ and its linear response to a change of nuclear geometry, $\partial n_{\mathbf{R}}(\mathbf{r})/\partial \mathbf{R}_I$. An expression for this response can be obtained by relating the density, $n(\mathbf{r})$, to the Kohn-Sham wavefunctions Eq. (2.42) from Section 2.2, $n(\mathbf{r}) = \sum_i |\psi_i(\mathbf{r})|^2$, and by taking this derivative with respect to the perturbation of atomic position $\mathbf{R}_I$

$$\frac{\partial n(\mathbf{r})}{\partial \mathbf{R}_I} = \sum_i \frac{\partial \psi_i^*(\mathbf{r})}{\partial \mathbf{R}_I}\psi_i + \psi_i^*\frac{\partial \psi_i(\mathbf{r})}{\partial \mathbf{R}_I}. \tag{3.26}$$

These wavefunctions obey the equation

$$\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial \mathbf{r}^2} + V_{\mathrm{KS}}(\mathbf{r})\right]\psi_i(\mathbf{r}) = \epsilon_i \psi_i(\mathbf{r}) \tag{3.27}$$

where V_{KS} is the same potential energy as defined in Eq. (2.50) and ϵ_i is the energy eigenvalue for Kohn-Sham orbital, ψ_i. These quantities can be expanded in a Taylor series for a small deviation of atomic configuration $\Delta\mathbf{R}_I$,

$$\begin{aligned} V_{\mathrm{KS}}(\mathbf{r}, \Delta\mathbf{R}_I) &= V_{\mathrm{KS}}(\mathbf{r}, 0) + \Delta\mathbf{R}_I \frac{\partial V_{\mathrm{KS}}(\mathbf{r})}{\partial \mathbf{R}_I} + ... \\ \psi_i(\mathbf{r}, \Delta\mathbf{R}_I) &= \psi_i(\mathbf{r}, 0) + \Delta\mathbf{R}_I \frac{\partial \psi_i(\mathbf{r})}{\partial \mathbf{R}_I} + ... \\ \epsilon_i(\Delta\mathbf{R}_I) &= \epsilon_i(0) + \Delta\mathbf{R}_I \frac{\partial \epsilon_i}{\partial \mathbf{R}_I} + ... \end{aligned} \tag{3.28}$$

Taking the first-order approximation in the expansions above and inserting them into Eq. (3.27), allows us to express this relationship as

$$\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial \mathbf{r}^2} + V_{\mathrm{KS}}(\mathbf{r}) - \epsilon_i\right]\frac{\partial \psi_i(\mathbf{r})}{\partial \mathbf{R}_I} = -\frac{\partial V_{\mathrm{KS}}(\mathbf{r})}{\partial \mathbf{R}_I}\psi_i(\mathbf{r}) + \frac{\partial \epsilon_i}{\partial \mathbf{R}_I}\psi_i(\mathbf{r}), \tag{3.29}$$

where

$$\frac{\partial V_{\mathrm{KS}}(\mathbf{r})}{\partial \mathbf{R}_I} = \frac{\partial V(\mathbf{r})}{\partial \mathbf{R}_I} + \int \frac{1}{|\mathbf{r}-\mathbf{r}'|}\frac{\partial n(\mathbf{r}')}{\partial \mathbf{R}_I} d\mathbf{r}' + \left.\frac{dV_{\mathrm{xc}}(n)}{dn}\right|_{n=n(\mathbf{r})}\frac{\partial n(\mathbf{r})}{\partial \mathbf{R}_I} \tag{3.30}$$

depends self-consistently on the charge density, induced after the introduction of the perturbation. An explicit evaluation of this equation would require a full knowledge of both the occupied and unoccupied states. However, only knowledge of the occupied states is required to evaluate the right-hand side of Eq. (3.29) [39]. Furthermore, the response of the charge density to a perturbation will only be dependent on components of the perturbation that couple occupied states to unoccupied ones. Thus, the introduction of a projector onto these empty states P_C allows Eq. (3.29) to be rewritten as [39]

$$\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial \mathbf{r}^2} + V_{\mathrm{KS}}(\mathbf{r}) - \epsilon_i\right] P_C \frac{\partial \psi_i(\mathbf{r})}{\partial \mathbf{R}_I} = -P_C \frac{\partial V_{\mathrm{KS}}(\mathbf{r})}{\partial \mathbf{R}_I}\psi_i(\mathbf{r}). \tag{3.31}$$

This gives us a means of calculating the linear response of the charge density with a change in atomic position, and therefore the matrix of interatomic force constants, $\mathbf{\Phi}_{IJ}$. This is due to the fact that Eq. (3.30) and Eq. (3.31) form a set of self-consistent equations, which can be solved analogously to the unperturbed case from Section 2.2.

3.2 Niobium Disulphide (NbS_2)

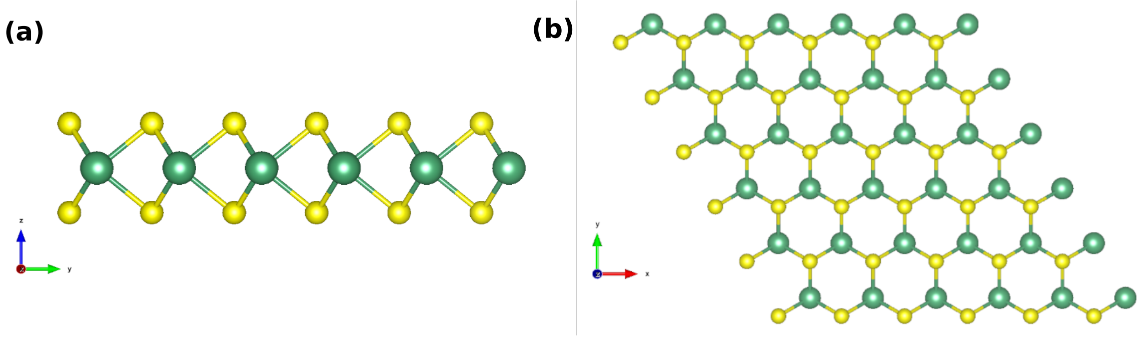

Figure 3.2: The atomic structure of monolayer 2H-NbS_2 (a) viewed along the x-axis (b) viewed along the z-axis. The larger niobium (Nb) atoms are depicted in green and the smaller sulphur (S) atoms are depicted in yellow.

Two-dimensional materials have been known to exhibit a large array of novel and potentially ground-breaking behaviours since the first isolation of monolayer graphene by Novoselov *et al.* [41] almost two decades ago. For instance, graphene exhibits a diverse range of fascinating behaviours, such as remarkable tensile strength [42], electrical conductivity [41] and thermal conductivity [43]. The group of layered materials from the family of transition metal dichalcogenides (TMDs), which is comprised of both metals and semiconductors, have also been shown to exhibit a wide variety of competing electronic phases such as charge-density waves [44], Mott-insulating phases [45] and superconducting phases [46], demonstrating a rich variety of potentially useful phenomena. If we are to use methods to predict the array of properties of these materials, we must begin by understanding the abilities of these models to capture such a diversity in behaviour, as well as their limitations in doing so.

An initial compound with which to identify strengths and weaknesses of various ML-based techniques for rapidly obtaining properties that result from vibrational properties could be niobium disulphide, NbS_2, a material that should prove to effectively demonstrate the bounds of performance of ML-based modelling techniques. NbS_2 exists in two main polytypes, which are rhombohedral (3R) and hexagonal (2H), each exhibiting unique electronic properties. 3R-NbS_2 is a metallic compound [47], whereas 2H-NbS_2 also exhibits metallicity, up until the onset of a superconducting phase below a critical temperature of 6 K [48]. Both polytypes are layered compounds and can be grown in bulk form by tuning the sulphur pressure to favour one polytype over the other [49]. In the few-layer limit, only 3R-NbS_2 has been grown, with mechanical exfoliation from bulk being the only way of isolating the more interesting, 2H few-layer phase of the compound [50].

NbS_2 is believed to be on the verge of a variety of different instabilities, with a variety of Coulombic and electron-phonon interactions leading to a unique combination of different charge-, spin- and superconducting phases, meaning that this material presents an immensely rich phase diagram [51]. The following study I will outline these competing phases and discuss how they may impact the vibrational dynamics of this system. In later chapters, I will further discuss these behaviours in the context of the capabilities of machine learning modelss in replicating the determined phenomena and the accuracy of these systems in adjusting their predictions in order to account for these complex dynamics.

3.2.1 Band Structure Calculations

An attempt was made to establish the lowest energy phase of monolayer 2H-NbS_2, and therefore the true system ground-state, upon which to perform the bulk of the subsequent analyses. All calculations in this section were performed using the `VASP` library [30–33] for DFT with a plane wave basis set. A high cut-off energy of 600 eV was necessary for this system in order to converge the calculations due to the 20 Å vacuum layer that needed to be included in order to accurately approximate the monolayer system, in spite of the inherent periodicity of the plane-wave basis set. The initial calculations on the system were performed using a GGA-PBE functional, with a Γ-centered k-mesh of $20 \times 20 \times 1$. The first step in this process is to relax the monolayer cell, with the initial spin configuration set to one of three possible magnetic phases; non-spin polarized, ferromagnetic and antiferromagnetic. The relaxation of the system forces were set to the very strict convergence criterion of 10^{-9} eV/Å, such that there was a sufficiently relaxed structure for the phonon calculations further down the pipeline, since this requires a strict convergence.

After this relaxation had been completed for each case, self-consistent field (SCF) calculations were performed in order to establish the lowest energy spin configuration of NbS_2 by comparing the resulting total system energies of the compound for each of the non-spin polarized, ferromagnetic and antiferromagnetic cases. Using the GGA-PBE functional, however, led to the total energies of each of these three systems being essentially identical, with a total of 1 meV separating the total energies of each of the three cases. Thus, it was decided that a hybrid functional may perform slightly better in the attempt at differentiating the ground state of the system. This suspicion was due to the stronger correlation effects in the functional potentially giving a stronger indication of the lower-energy spin state. Therefore, another SCF calculation was performed, this time using the HSE06 functional. As a result of the use of this functional, the antiferromagnetic phase was eliminated as a potential ground-state phase of the NbS_2 system, as this state was higher in energy by approximately 40 meV than that of the other two phases when the HSE06 hybrid functional was employed. The two remaining phases, however, were still degenerate in energy, providing evidence for competing magnetic and non-magnetic phases in the ground-state of NbS_2.

An understanding of the reasons for these competing phases can be obtained by studying the results of the band-structure calculations of Fig. 3.3, which demonstrates the difference in electronic band structure between the non-spin polarized state and the ferromagnetic state, as well as the corresponding difference in the density of states (DOS). In the non-spin polarized case, as seen in Fig. 3.3 (a), there is a peak in DOS present at the Fermi energy, which is evidence of the metallic nature of the material. The total energy of quasi-2D systems, however, can be reduced by the appearance of various types of long-range orderings that reduce the density of occupied states at the Fermi level. One such ordering is known as a charge-density wave (CDW) in which the total distribution of electron density across a low-dimensional system is periodically, spatially modulated. This periodic modulation of charge density, in turn, induces a periodic distortion in the atomic lattice itself. The increase in potential energy that it costs to have atoms in non-equilibrium positions is compensated for by the reduction in occupied states above the Fermi energy and, therefore, the system undergoes a phase transition. The transition from the metallic state to the insulating CDW state

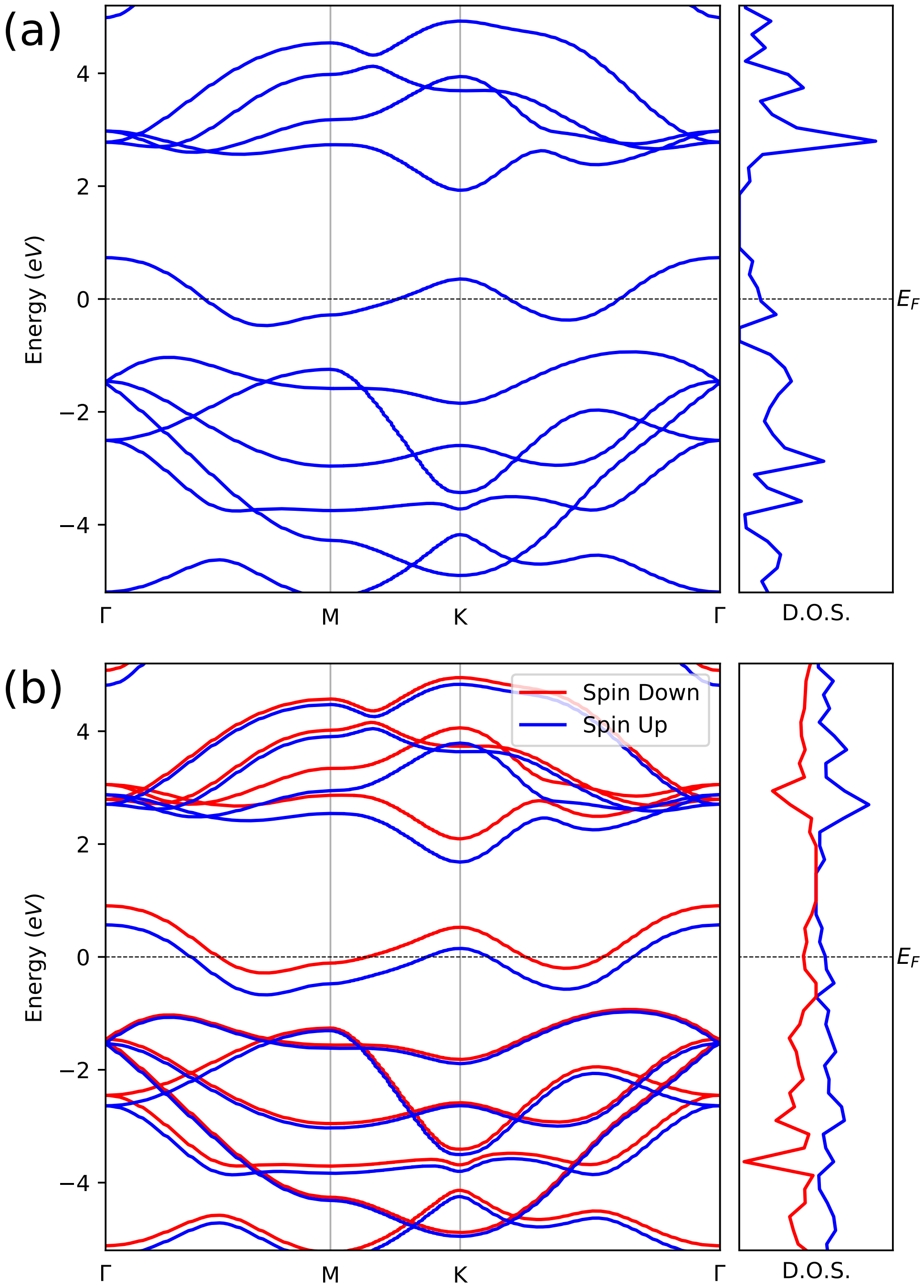

Figure 3.3: A comparison between the energy band structure diagram and density of states for monolayer 2H-NbS_2 in the (a) non-spin-polarized case and (b) ferromagnetic case. In the ferromagnetic case, the energy band splitting and the reduction of the density of states at the Fermi energy, which is represented by the dashed line, is evident upon transition from a non-spin-polarized to spin-polarized system. Both band structure diagrams were obtained using a GGA-PBE functional, with an energy cut-off of 600 eV and a Γ-centered k-mesh of $20 \times 20 \times 1$.

is known as a Peierls transition. NbS_2 is known not to exhibit a charge density wave, which makes it unusual among 2D TMD materials [52]. Studies have shown, however, that the material is in fact on the cusp of a CDW transition, which is mitigated by strong anharmonic effects affecting the electron-phonon coupling of the system [53].

Another way in which systems can lower their total energy is evident from the plot of the band structure visible in Fig. 3.3 (b). Considering spin-polarization can lead to a splitting in energies between the spin-up and spin-down electron energy levels, resulting in an overall reduction of total system energy due to the reduction in the occupied energy bands close to the Fermi energy.

Indeed, there is prior theoretical evidence for a tendency towards long-range magnetic ordering in monolayer NbS_2 [54], as well as the aforementioned tendency towards long-range electronic ordering, despite the fact that such electronic ordering has not been observed. In line with this, Ref. [53] suggests that the system is poised between two competing ordered phases. One of these phases is characterized by a charge density wave instability, while the other exhibits a modulation in the spin density, referred to as a spin density wave. This mirrors the findings of my study, which also indicates the presence of two competing ground-states: one featuring long-range magnetic ordering and the other devoid of such ordering, with long-range charge ordering instead.

3.2.2 Vibrational Dynamics

There are several notable phenomena that accompany these variances in electronic phase in 2D materials, be that a transition from a metallic to an insulating phase or a phase transition to superconductivity. As stated previously, another result of these complex phenomena in 2D materials is that the aforementioned modulation of charge density in CDWs, will cause a distortion in the unit cell of the compound which is periodic over a supercell of constituent unit cells. This will create the appearance of a dynamical instability in the structure, which manifests as negative phonon frequencies resulting from solving Eq. (3.12), corresponding to imaginary phonon modes.

In reality, the CDW instability in monolayer 2H-NbS_2 is, as stated, mitigated by the inclusion of anharmonic effects. However, the non-mitigated instability should appear when operating in the harmonic regime, a fact that enables us to explore the hypothetical boundary between these interesting states, even if in truth such a boundary is not a physical one. With this in mind, the phonon dispersion was obtained for monolayer 2H-NbS_2 in both the ferromagnetic and non-magnetic phases. Initially, the finite displacement method was employed, however, this led to a requirement for a very large supercell in order to converge the calculation, by causing an infeasible number of required force calculations.

To address this difficulty, DFPT was employed on the relaxed unit cell in order to directly calculate the matrix of first-order force constants, and thereby diagonalize the dynamical matrix of the system as described above in section 3.1.2, while avoiding the need for a great number of computationally demanding calculations on a wide number of supercell displacements for large supercells. Once again, a plane-wave basis set was used in conjunction with the implementation of DFPT as is available in the `VASP` library. A GGA-PBE functional, with a Γ-centered k-mesh of $20 \times 20 \times 1$ and an energy cutoff of 600 eV was, once again, employed in this study.

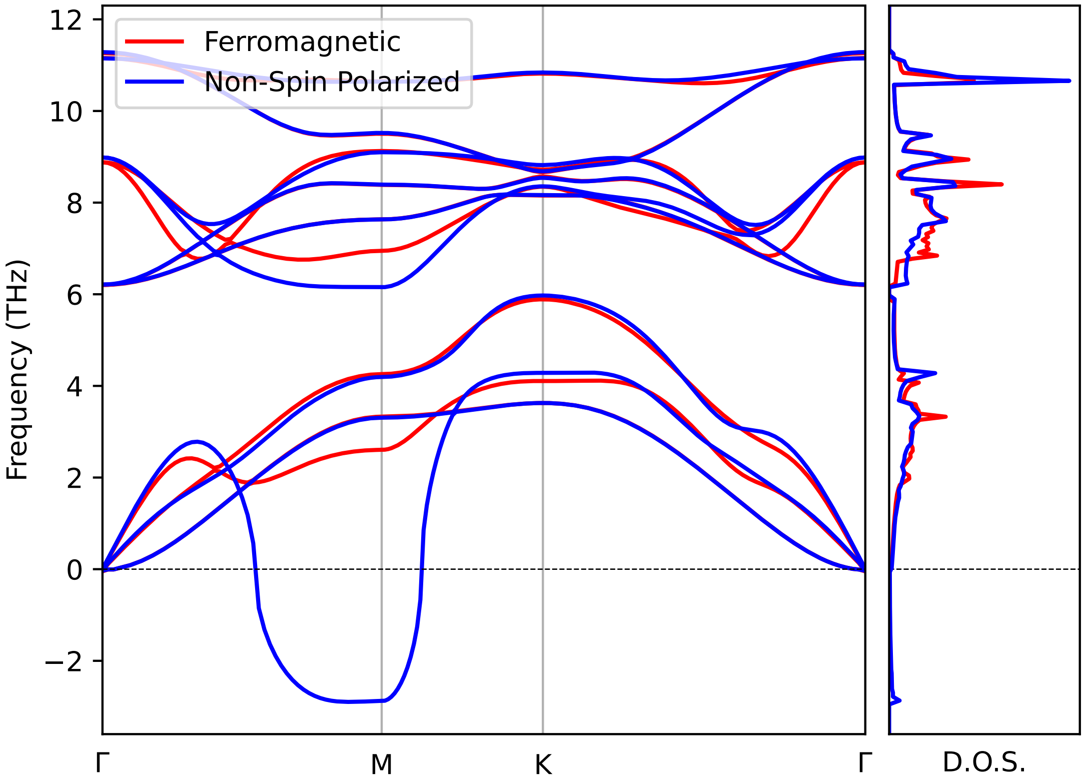

Figure 3.4: The phonon dispersion and phonon density of states for monolayer 2H-NbS_2, calculated using DFPT. The phonon dispersion for the system obtained in a non-spin polarized state is represented in blue and the ferromagnetic state is represented in red. The dynamical instability in the primitive cell for the non-spin polarized case can be clearly seen in the imaginary phonon modes at $\mathbf{q} \approx 2/3\Gamma M$. In the case of the ferromagnetic state, this apparent dynamical instability is not present, eliminating the imaginary phonon modes. Softening of the optical phonon modes can also be seen in the non-magnetic phase at the same point relative to the ferromagnetic phase.

The results of this analysis can be seen in Fig. 3.4, presenting a very stark comparison between the apparent unit cell distortion, resulting from the CDW formation in the harmonic approximation and the relaxed structure as it exists in the magnetic phase. This vastly differing dynamical behaviour of the system is a direct result of the electronic structure of the compound. It presents an ideal example of the vital importance of understanding the degree to which such factors can impact the resulting dynamics of the system. It also underscores the ability of novel methods of modelling these dynamics to replicate that behaviour.

The expected dynamical instability of the non-magnetic state is clearly visible in the phonon dispersion of Fig. 3.4 and appears at the phonon wavevector $\mathbf{q} \approx 2/3\Gamma M$, which agrees with the value reported for the harmonic approximation in Ref. [55]. This result implies that the system would undergo a 3×3 reconstruction. Further to the evident dynamical instability in the acoustic phonon modes, it should be noted that there is a softening of the optical phonon modes at the same wavevector as the instability in the non-magnetic case. This softening is not nearly as dramatic as the incipient instability. However, it is significant enough to warrant note in any future comparisons.

In the harmonic regime, this dynamical instability is completely eradicated by the phase transition to a ferromagnetic state. The tendency towards a CDW phase is offset in this case by the system energy reduction that the transition to a ferromagnetic phase yields. The fact that this magnetic phase and the non-magnetic phase are so similar in energy means that it is difficult to make a call on the exact nature of the ground state of the system. In fact, it is abundantly clear that there is a range of system dynamics that are directly in competition with each other, each presenting vastly different behaviour. Indeed, there has been some discussion on utilising this tendency towards multiple electronic phases in NbS_2 as a switch between conducting and insulating states, tuning the state with strain, which has been shown to mitigate the dampening of the CDW state resulting from anharmonicity [55].

The diversity of behaviours in NbS_2 and the presentation of multiple, competing ground-states that exhibit different dynamical behaviour means that NbS_2 is an ideal material with which to highlight the interplay between these states and also demonstrate the capabilities of less computationally expensive models than DFT to potentially capture this behaviour, or indeed, the limitations of such models in their ability to replicate this behaviour.

3.3 Summary & Conclusions

In this chapter, the methods for understanding the dynamical behaviour of systems at the quantum-level were first outlined. This began with a discussion of the nature of phonons as being a valid and useful means of translating the issue of thermal transport and vibrational dynamics of a system from an understanding based on individual atom dynamics to a model describing the collective behaviour of all the atoms in a system.

After gaining this understanding of the nature of phonons and their relationship to such macroscopic behaviours, methods with which the phononic properties of systems can be obtained were discussed. To that end, two methods were discussed: finite displacement, which aims to construct the dynamical matrix of the system by obtaining the force constants through the direct calculation of forces from atomic displacements in a supercell; and density functional perturbation theory, which attempts to calculate the same force constants through direct calculation of the Hessian matrix, using first-order perturbation theory.

A case study of calculating the behaviour of a two-dimensional system was presented after outlining the methods of calculating the desired dynamical behaviour. The system in question was chosen to be NbS_2 due to the wide array of diverse and complex behaviours that the monolayer, 2H polytype exhibits. Two ground-state phases were shown to be degenerate in energy, one of which was a ferromagnetic phase exhibiting long-range magnetic ordering and the other was a non-magnetic phase, on the cusp of entering into a CDW state. Both of these states exhibited vastly different dynamical behaviours on the calculation of the phonon dispersions for each state, with the non-magnetic phase presenting an instability in the harmonic regime, resulting from the incipient CDW state and the ferromagnetic state eliminating this instability entirely.

In order for the thermal behaviour of 2D materials and their heterostructures to be modelled effectively using ML techniques, an understanding of such model's ability to capture the dynamics resulting from the exotic behaviours of monolayers must be established. As is evident from the results of this chapter, NbS_2 offers us the ability to probe a large number of those behaviours and gain an insight into the potential, and the limitations of these ML models as we move beyond first-principles techniques and begin to attempt to bypass them.

Chapter 4

Machine-Learned Thermal Properties

"An rud a chíonn an leanbh is é an rud a níonn an leanbh."

In the previous chapter, methods for obtaining the dynamical properties of materials from first-principles calculations were discussed in the context of two-dimensional (2D) materials. Such methods are extremely valuable for our understanding of exotic phenomena in these materials and for expanding the theoretical horizons of our knowledge about those phenomena. Their practical use for the purpose of inverse materials design and, by extension, materials discovery, however, is extremely limited by the high computational cost of their use, particularly for low-dimensional materials requiring high energy cut-offs and large vacuum regions in the simulation cells.

Thus, a means of reducing the computational cost of energy and force calculations is required, such that there is little to no sacrifice in the fidelity or accuracy, relative to *ab initio* techniques. At first glance, this seems like an impossible task, due to the plethora of factors that influences the total energy/forces of a quantum mechanical system. However, this issue is greatly simplified by the progress in recent years in the capabilities of machine-learning (ML) methods, which are designed for and optimised for their ability to approximate high-dimensional, non-linear functions with relatively simple and computationally inexpensive algorithms. A fact that means that ML methods are an ideal candidate for approximating the potential energy surface (PES) of a complex molecular system. All one needs for such a task is an adequate means of representing the system and enough data for the model to be trained. With these criteria met, the total energy and forces of a system can easily be computed and with that, a range of system properties.

In this chapter, I will focus on the construction of a library of machine-learned interatomic potentials (MLIPs) for 2D materials, which can be arbitrarily extended to composite systems to facilitate the rapid computation of materials properties for such systems. The reason that this work targets 2D heterocomposite systems, in particular, is related to the fact that there exists a large family of 2D layered systems, with approximately 1,800 unique systems [56]. Due to the nature of the interlayer van der Waal's (VdW) binding in these layered systems, they can be arbitrarily stacked to form an infinite number of composite structures. The computational intensity of first-principles methods previously described in this thesis, would not be efficient enough to adequately sample the property space of this large family of potential materials. Consequently, a library of ML potentials is ideally suited for constructing a picture of this property space, which is helpful for targeting specific materials properties. This approach enables the rapid property screening step in an inverse-design workflow.

The main properties I targeted in this chapter are related to the thermal transport behaviours of 2D systems. 2D compounds are known to exhibit a wide range of thermal properties, with numerous examples exhibiting anomalously high- and low-thermal conductivities [57–59], and with certain compounds displaying highly efficient thermoelectric conversion [60]. This wide variance in the thermal properties of these materials implies that the 2D material class of compound represents an ideal playground for the synthesis of composite compounds with desired properties for any given thermal application.

In the first part of this chapter, I will discuss different methods for representing materials composition and structure such that they can be processed efficiently by ML algorithms to create MLIPs. I will then describe how the resulting potentials can be extended, to further enhance the capabilities of these MLIPs, by the inclusion of a consideration of the interlayer VdW interaction. The inclusion of this interaction facilitates the computation of energies and forces for composite heterostructures. Subsequently, our attention will be turned to the vibrational properties of these systems and to an in-depth analysis of the MLIPs' potential to replicate the phononic properties of monolayer and heterostructure materials. I will compare these analyses with some of the results from the previous section and other first-principles DFPT calculations.

Finally, I will turn the focus towards modelling the thermal properties of these systems, given their adequate performance on modelling the necessary phononic properties. Within this context, I will conduct an exploration of the methods that give rise to the determination of both the interlayer and the intralayer thermal conductivity of the studied materials. This presentation of a novel library of MLIPs for 2D materials should prove to be invaluable to the materials science community and in this chapter, I will give a full overview of the capabilities of these methods in replicating the thermal behaviour of such materials.

The work contained in this chapter was conducted in close collaboration with my colleague Rui Dong, with Rui broadly focused on the construction of the MLIPs, with several having been constructed by myself. However, my focus was primarily centred on the prediction of properties by utilising the resulting hybrid potentials.

4.1 Machine-Learned Interatomic Potentials

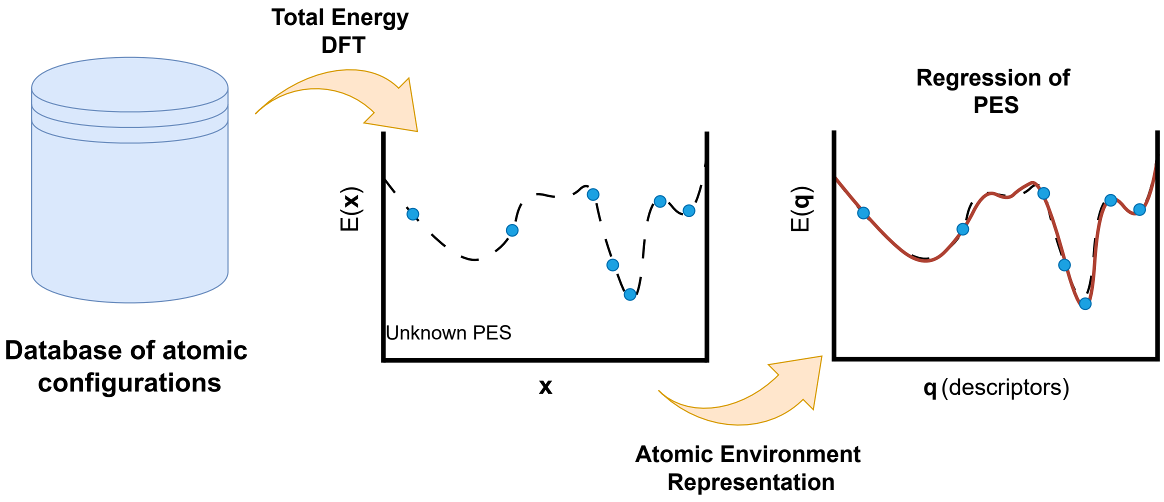

Figure 4.1: Diagram depicting the general workflow for the construction of machine-learned interatomic potentials. The descriptors **q** in the diagram are dependent on material properties and crystal structure, while **x** represents a generic geometry variable. *This diagram was adapted from Ref. [61].*

A general workflow for the construction of MLIPs is presented in Fig. 4.1. As stated, a regression task in ML is an ideal means of approximating a complex, high-dimensional, non-linear function with minimal computational expense, depending on the ML model chosen, the choice of feature representation and the size of the available data set. There are numerous choices for the construction of these interatomic potentials, some more suited to certain tasks than others. As with any model, there will always be some trade-offs between models in terms of how data-hungry or computationally efficient they are, or how well they are able to capture the complexity of the PES.

The first and arguably the most important choice when considering the model with which to construct an approximation of the PES is how much data is available for the training. Some models will converge with relatively little data available, whereas others require a vast amount of data collection before convergence becomes feasible. Generally speaking, the data required to train a model scales with the number of free parameters that are required to tune during the fitting procedure. There are two main philosophies involved when considering the encapsulation of the non-linear PES in a ML approximation. This non-linearity can be built in as part of your model, such as random forest models or neural networks, or the choice of feature representation through a non-linear kernel can translate the problem from a non-linear problem to a linear one, which facilitates the use of the far less computationally-demanding and more efficiently-trained ridge regression models. For this work, the latter option was chosen due to the relatively small number of data points required to train a satisfactory model. This means that fewer computationally intensive DFT calculations would be required to obtain a similar level of accuracy when compared with more complex, non-linear models. Thus, a valid choice of feature representation must be used to translate this problem to the linear space.

4.1.1 Feature Representation

There is a whole zoo of potential representations that have been conceived for the construction of PESs, from *ab-initio* training data. Due to the fact that away from select critical points, the behaviour of atoms is localized, meaning that the total energy for a given system can be obtained from a sum of atomic energies, namely

$$E = \sum_{i}^{N_{\text{atoms}}} \epsilon(\{\mathbf{r}_{ii'}\}), \tag{4.1}$$

where $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ is the relative distance between atoms i and j. Most of these descriptors involve a construction representing each local environment of the system, centred on each atom and encoding information about neighbours within the configuration. They can incorporate information about two- or three-body contributions or can alternatively extend to more involved many-body formalisms. There are some desirable properties that a valid representation must fulfil in order to vastly reduce the computational complexity of constructing a quantum-accurate PES with ML. Most of these criteria are based on the desired symmetry of the representation, such that equivalent systems will output the same value. Namely, the descriptor should exhibit permutational symmetry, in that the outcome should not change if two atoms of the same species are swapped; translational symmetry, in that the representation should yield the same result for any given translation of the lattice, which is already satisfied by the approximation in Eq. (4.1); and finally, rotational invariance, where the representation should give the same result regardless of any rotation of the system.

Several techniques have been proposed to fulfil these requirements for a good atomic environment descriptor. The initial approach, introduced in 2007 by Behler and Parinello [62], employed functions to represent these symmetries. This method entailed utilizing the difference between atomic positions instead of absolute positions for preserving translational symmetry. It also employed a radial symmetry function composed of the summation of Gaussian functions and a summation of angular symmetry functions to maintain rotational invariance. This representation also introduced the concept of a cutoff radius, such that the influence of a local environment vanishes at a pre-specified radius around the central atom, in order to reduce the computational intensity of the construction of these descriptors for little additional information. This work employed a neural network to perform the regression of the PES, requiring a relatively large number of first-principles DFT total-energy calculations to converge the model for a given system. The scaling of the symmetry function-based representation was also not particularly efficient, with the number of parameters of the descriptor exploding with an increase in system size.

This subsequently led to the conception of the Gaussian approximation potential (GAP) [63], which was designed in order to further reduce the computational cost of the regression of the PES by reducing the number of descriptor parameters and making the resulting descriptors amenable to regression by less computationally-demanding fitting algorithms. This work introduced the concept of the use of the signal correlation function between three points, the bispectrum, as an efficient means of representing an environment, that naturally encodes the desired invariants.

The relationship between the local environment and the bispectrum can be understood by first representing the local atomic density of the neighbours of atom i as

$$\rho_i(\mathbf{r}) = \delta(\mathbf{r}) + \sum_{i'} \delta(\mathbf{r} - \mathbf{r}_{ii'}) f_{\text{cut}}(|\mathbf{r}_{ii'}|), \tag{4.2}$$

where $f_{\text{cut}}(r) = 1/2 + \cos{(\pi r/r_{\text{cut}})}/2$ is the aforementioned cutoff function, that smoothly goes to zero at the cutoff radius r_{cut}, and δ is the usual Dirac delta function. This local density can then be projected onto the surface of a four-dimensional unit sphere, using the transformation

$$(\phi, \theta, \theta_0) = \left[\tan^{-1}\left(\frac{y}{x}\right),\ \cos^{-1}\left(\frac{z}{|\mathbf{r}|}\right),\ \frac{|\mathbf{r}|}{\mathbf{r}_0}\right], \tag{4.3}$$

where $r_0 > r_{\text{cut}}/\pi$. Thus, the resulting 4D surface naturally contains all the information of the 3D spherical region surrounding the central atom within the cutoff radius, including the radial dimension. This radial dimension is incorporated by the transformation outlined in Eq. (4.3). Thus, 4D hyperspherical harmonics, $U^j_{m'm}$, defined for $j = 0, \frac{1}{2}, 1, ...$ and $m, m' = -j, -j+1, ...j-1, j$ [64], form a natural complete basis for the interior of the 3D sphere within the cutoff, entirely eliminating the need for a radial basis function, as was introduced in Ref. [62]. This fact allows the projection of the atomic density onto the surface of the 4D sphere to be expanded in terms of the 4D spherical harmonics, using the coefficients

$$c^j_{m'm} = \left\langle U^j_{m'm} \,\middle|\, \rho \right\rangle, \tag{4.4}$$

where the index i is dropped for the sake of clarity. These coefficients allow the construction of the bispectrum, given by,

$$\begin{aligned} B_{j_1,j_2,j} = & \sum_{m_1,m'_1=-j_1}^{j_1} \sum_{m_2,m'_2=-j_2}^{j_2} \sum_{m,m'=-j}^{j} (c^{j}_{m,m'})^{*} C_{j_1 m_1 j_2 m_2}{}^{jm} \\ & \times C^{jm'}_{j_1 m'_1 j_2 m'_2} \cdot c^{j_1}_{m_1,m'_1} c^{j_2}_{m_2,m'_2}, \end{aligned} \tag{4.5}$$

where $C_{j_1 m_1 j_2 m_2}$ are the standard Clebsch-Gordan coefficients. In practice, in all MLIPs where the bispectrum is used, a truncated version of this construction is used, with $j, j_1, j_2 \leq J_{\text{max}}$. This simply limits the spatial resolution of the descriptors used to describe a given atomic environment. This function is then used to fit for the energy of the system by employing Gaussian process regression, in which a Gaussian kernel between bispectra of atomic environments is linearly fit. Thus, the coefficients of the regression are obtained by inverting the covariance, constructed from the Gaussian kernel, and multiplying the inversion by the expected energy values.

The ideas presented for the construction of GAP were later extended to account for some difficulties in representing systems with a large number of atoms contained in the local environments. This expansion led to the development of a method for directly creating a representation of similarity between atomic neighbourhoods, eliminating the need to calculate a similarity kernel based on two local atomic environment descriptors [65]. This new construction was called the smooth overlap of atomic potentials or SOAP. When implemented by replacing the Gaussian kernel in the GAP method with the new SOAP kernel (called the SOAP-GAP method), it led to superior performance on systems with arbitrary numbers of atoms within the cutoff sphere.

Thompson *et al.* [66] subsequently recognized that there was no necessity to compute the similarity kernel when effectively and accurately modelling a system's energy. This revelation meant that the local energy of a specific atomic environment could be accurately reproduced by creating a linear combination of the lowest-order bispectrum components, with linear coefficients that only depended on the chemical identity of the central atom

$$E^{i}_{\text{SNAP}}\left(\mathbf{B}^{i}\right) = \beta_0^{\alpha_i} + \sum_{k=1}^{K} \beta_k^{\alpha_i} B_k^{i} = \beta_0^{\alpha_i} + \boldsymbol{\beta}^{\alpha_i} \cdot \mathbf{B}^{i}, \tag{4.6}$$

where $\mathbf{B}^i = \{B_1^i, ..., B_K^i\}$ is the set of K bispectrum components representing the local environment of atom i, α_i is the chemical identity of the same atom and β_k^α are linear coefficients for atoms of type α. This linear form of the SNAP atomic contributions to the local energy is immensely useful for calculating properties that can be obtained directly from *ab-initio* DFT calculations. These formulae are specific to atoms of a single type for clarity, but can easily be extended to multiple chemical species. The contribution arising from the local environments to the total energy of a system with ionic positions $\mathbf{r}^N$ can be written in terms of the bispectrum components as

$$E_{\mathrm{SNAP}}\left(\mathbf{r}^N\right) = N\beta_0 + \boldsymbol{\beta} \cdot \sum_{i=1}^{N_{\mathrm{atoms}}} \mathbf{B}^i, \tag{4.7}$$

where $\boldsymbol{\beta}$ is the K-vector of SNAP coefficients, β_0 is the constant energy contribution for each atom and $\mathbf{B}^i$ is the K-vector of bispectrum components for atom i. Similarly, the forces acting on atom j can be computed by taking the derivative of the bispectrum components with respect to $\mathbf{r}_j$, which is the position vector of atom j

$$\mathbf{F}^j_{\mathrm{SNAP}} = -\nabla_j E_{\mathrm{SNAP}} = -\beta \cdot \sum_{i=1}^{N_{\mathrm{atoms}}} \frac{\partial \mathbf{B}^i}{\partial \mathbf{r}_j}, \tag{4.8}$$

where $\mathbf{F}^j_{\mathrm{SNAP}}$ is the force acting on atom j due to the SNAP energy. Finally, an expression for the contribution of this energy to the stress tensor can similarly be obtained as

$$\mathbf{W}_{\mathrm{SNAP}} = -\sum_{j=1}^{N_{\mathrm{atoms}}} \mathbf{r}_j \otimes \nabla_j E_{\mathrm{SNAP}} = -\boldsymbol{\beta} \cdot \sum_{j=1}^{N_{\mathrm{atoms}}} \mathbf{r}_j \otimes \sum_{i=1}^{N_{\mathrm{atoms}}} \frac{\partial \mathbf{B}^i}{\partial \mathbf{r}_j}, \tag{4.9}$$

where $\otimes$ is the Cartesian outer product operator and $\mathbf{W}_{\mathrm{SNAP}}$ is the contribution of the SNAP energy to the stress tensor.

All these expressions involve the vector of the SNAP coefficients $\boldsymbol{\beta}$ multiplying a vector of quantities calculating using the bispectrum components of the atoms in a given configuration. Thus, the structure of these resulting sets of equations facilitates the easy construction of a system of linear equations for each configuration in any given dataset, for the energy, forces or for the stress tensor. This fact allows us to optimise the values for $\boldsymbol{\beta}$, the vector of linear coefficients, using any of these three quantities.

4.1.2 Fitting Procedure

Given this SNAP representation, the question arises as to the optimal meethod of obtaining values for $\boldsymbol{\beta}$ that most accurately predict the desired quantity. Generally, this is achieved for SNAP potentials by solving a system of linear equations, for N_s atomic configurations in a given training dataset of DFT properties,

$$\begin{bmatrix} \vdots & \vdots \\ N_s & \sum_{i=1}^{N_s} \mathbf{B}_i \\ \vdots & \vdots \\ 0 & -\sum_{i=1}^{N_s} \frac{\partial \mathbf{B}_i}{\partial \mathbf{r}_j^{\alpha}} \\ \vdots & \vdots \\ 0 & -\sum_{j=1}^{N_s} \mathbf{r}_j^{\alpha} \sum_{i=1}^{N_s} \frac{\partial \mathbf{B}_i}{\partial \mathbf{r}_j^{\alpha}} \\ \vdots & \vdots \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \vdots \\ E_s^{DFT} \\ \vdots \\ \mathbf{F}_{j\alpha}^{DFT} \\ \vdots \\ \mathbf{W}_{\alpha\beta s}^{DFT} \\ \vdots \end{bmatrix}, \tag{4.10}$$

which is of type $\mathbf{A} \cdot \boldsymbol{\beta} = \mathbf{y}$, meaning it can be easily solved for the vector of coefficients $\boldsymbol{\beta}$. The best approximation of values for $\boldsymbol{\beta}$ can be obtained using the linear least squares method

$$\boldsymbol{\beta} = \underset{\beta}{\operatorname{argmin}} \parallel (\mathbf{A} \cdot \boldsymbol{\beta} - \mathbf{y}) \parallel^2 = \mathbf{A}^{-1} \cdot \mathbf{y}. \tag{4.11}$$

In practice, a full inversion of the $\mathbf{A}$ matrix is not performed, rather QR factorization[1] is used to solve the matrix equation for $\boldsymbol{\beta}$, leading to a very efficient and accurate means of obtaining the desired SNAP coefficients.

4.1.3 Training Dataset

In order to reduce the complexity of stacking heterostructures in computational simulations and have an intuitive understanding of the strain being induced by lattice mismatches and other complications, the focus of the construction of these MLIPs for monolayer compounds was on 2D materials with hexagonal lattices of the 1T and 2H polytype. In order to further reduce the complexity of the energy prediction from the model, this work was generally focused on non-magnetic materials. This will be

[1] QR factorization involves the decomposition of matrix $\mathbf{A}$ into a product $\mathbf{A} = \mathbf{QR}$, where $\mathbf{Q}$ is an orthonormal matrix and $\mathbf{R}$ is an upper triangular matrix.

discussed further in the context of NbS_2 in Section 4.1.4.

There were a total of 70 hexagonal, non-magnetic layered materials present in the Materials Cloud database [67] with a chemical composition of XY_2, where X and Y were different elements, of which 39 were non-metallic and 31 were metallic. All of these monolayers were either one of the 2H or 1T polytypes. Several materials from those remaining were excluded, mainly due to the inability to converge them in first-principles structural relaxation, prior to calculating the data for the training set, leaving a total of 69 hexagonal, non-magnetic monolayers for the library of MLIPs. Separate MLIPs were also trained for graphene and hBN due to their interest to the general community.

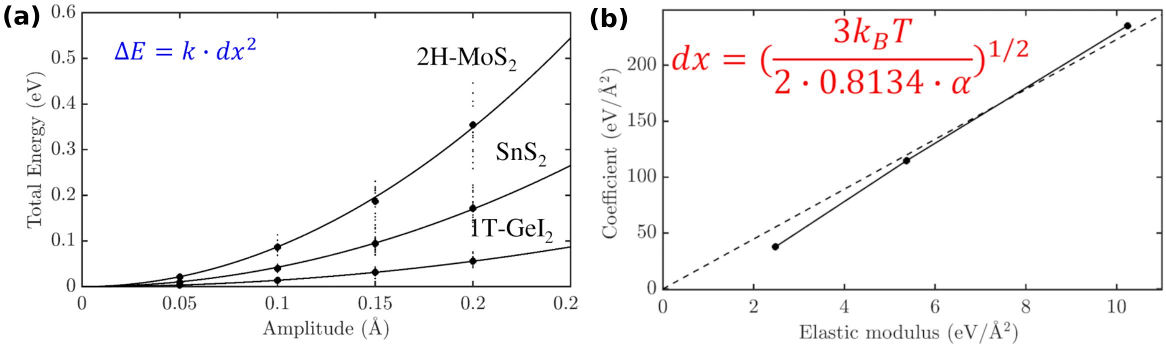

Figure 4.2: (a) Quadratic fit of the change in energy of three representative monolayer systems (2H-MoS_2, SnS_2 and 1T-GeI_2, respectively) against a series of atomic displacements. (b) Plot of the resulting coefficient for each of the quadratic fits vs. the elastic modulus of the monolayers. This relationship is subsequently linearly approximated (dashed line), yielding the optimal amplitude of displacement for the dataset.

In order to generate a valid training dataset for the construction of the MLIPs, the optimal displacement amplitude for the generation of the randomised displacements should be obtained. This is necessary to ensure that the training dataset will incorporate configurations that are representative of the real physical configurations that the system will likely be in for the desired molecular dynamics simulations or phonon calculations. Establishing this optimal value would be far too labour-intensive to perform for each monolayer in the dataset. Therefore, a small sample set of three monolayers, 2H-MoS_2, SnS_2 and 1T-GeI_2, were taken to establish some relationship between some readily available system property of the monolayers and the optimal displacement.

This goal was fulfilled by calculating the average change in potential energy per atom for 20 randomised configurations for a variety of values of small maximum displacements for each of these three monolayers. This is computed using `VASP` with a GGA-PBE potential, an energy cutoff of 500 eV and a Γ-centered k-mesh of $15 \times 15 \times 1$. The average change in the energy per atom was taken as being the mean value over all 20 configurations for each of these amplitudes. The change in resulting potential energy was fit quadratically with respect to the displacement amplitude, yielding a quadratic fit coefficient k. This quadratic fit is visible in Fig 4.2 (a). There is quadratic scaling of the change in the average potential energy per atom with an increase in average atomic displacement due to the assumption of a harmonic potential for each atom, implying the relationship

$$\Delta E_{pot} = k \cdot dx^2, \tag{4.12}$$

where dx is the maximal displacement. For a uniform distribution of randomized displacements, dx is double the mean displacement. Then there is a factor 4 between the mean displacement squared and the maximal squared displacement, which is absorbed by k, the coefficient of the quadratic fit. Assuming that the average change in potential energy per atom is fully converted into kinetic energy per atom at the temperature of interest allows for the following association

$$\Delta E_{\mathrm{pot}} = E_k = \frac{3}{2} k_B T, \tag{4.13}$$

where T is the temperature and k_B is the Boltzmann constant. This formula arises from the equipartition theorem. Combining Eq. (4.12) and Eq. (4.13) and solving for dx gives

$$dx = \left(\frac{3 k_B T}{2k} \right)^{\frac{1}{2}}. \tag{4.14}$$

Plotting the coefficient resulting from the quadratic fit against the elastic modulus α of each of the monolayers in question, a fit visible in Fig. 4.2 (b), allows us to linearly approximate the relationship between α and k. This gives a result of $k = 0.8134 \cdot \alpha$, giving us a simple means of associating the optimal maximum displacement for the SNAP training to the elasticity of the monolayer, which is an intrinsic property, for any temperature of interest. For example, using the established relationship, the optimal displacement to train a SNAP model for 2H-MoS$_2$ at 300 K, is 0.6 Å.

Once this relationship has been established, a strategy for constructing a distribution of data points for the training set must be established. These data points must incorporate a variety of strained lattices as well as the aforementioned displacements to ensure that strain is adequately represented in the training data. With this in mind, a variety of different configurations of training data sets were tested before an optimal distribution of lattice strains and sheared lattices was settled upon. The optimal distribution first involved 400 images with an unstrained unit cell, comprising configurations with randomized displacements up to a max of dx, the optimal maximum displacement derived for each monolayer above. This was initially done for 300 K. Taking the usual definition for the Cauchy strain tensor

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{bmatrix} = \begin{bmatrix} \frac{\partial u_x}{\partial x} & \frac{1}{2}\left(\frac{\partial u_x}{\partial y} + \frac{\partial u_y}{\partial x}\right) & \frac{1}{2}\left(\frac{\partial u_x}{\partial z} + \frac{\partial u_z}{\partial x}\right) \\ \frac{1}{2}\left(\frac{\partial u_y}{\partial x} + \frac{\partial u_x}{\partial y}\right) & \frac{\partial u_y}{\partial y} & \frac{1}{2}\left(\frac{\partial u_y}{\partial z} + \frac{\partial u_z}{\partial y}\right) \\ \frac{1}{2}\left(\frac{\partial u_z}{\partial x} + \frac{\partial u_x}{\partial z}\right) & \frac{1}{2}\left(\frac{\partial u_z}{\partial y} + \frac{\partial u_y}{\partial z}\right) & \frac{\partial u_z}{\partial z} \end{bmatrix}, \tag{4.15}$$

where u_α is the change in lattice vector in Cartesian direction α, the training set of normal lattice strains was obtained by taking every combination of $\varepsilon_{11} = \{\pm 0.03,\ \pm 0.05\}$ and $\varepsilon_{22} = \{\pm 0.03,\ \pm 0.05\}$ and obtaining 20 randomized configurations for each case, with an amplitude of dx with $T = 300$ K. Here, ε_{33} can be ignored for the 2D case. This yields a total of 240 strained configurations.

The final step necessary for a data set that adequately samples the possible configuration space is to adjust the shear components of the stress tensor $\boldsymbol{\varepsilon}$. This was sampled by, once again, taking a sample of 20 randomized configurations for each of the shear components $\varepsilon_{12} = \{\pm 0.005,\ \pm 0.01\}$, with an amplitude dx taken for $T = 300$ K. Once again, any shear component dependent on the z-direction can be neglected due to the 2D nature of the system. The final number of configurations in the fully representative training set for each of the monolayers was 720 configurations.

After having obtained a representative training data set, ridge regression is performed on the bispectrum components of the atomic environments of the system in question. Once this step has been executed, the usual mean absolute error (MAE) of the fit is consistently on the order of 1 meV/atom, indicating results that are within the error of DFT itself. Fig. 4.3 is an example of a parity plot of the test data against

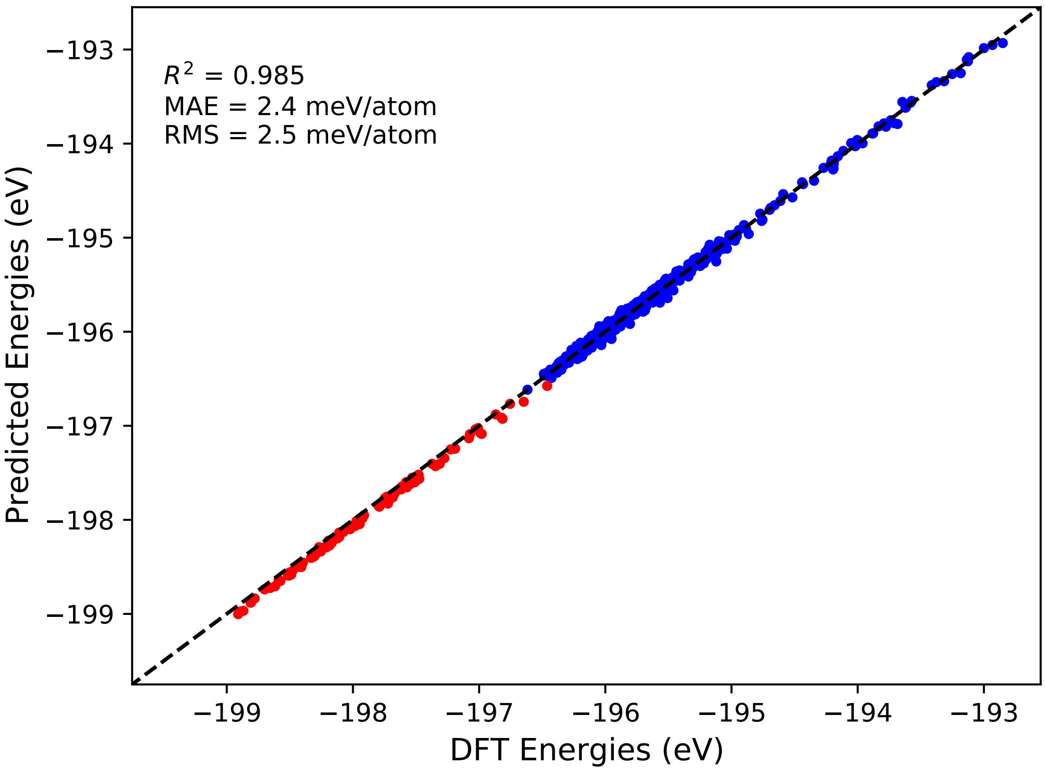

Figure 4.3: An example of a standard parity plot of the predicted values of energy against the DFT values for 2H-NbS$_2$ after the ridge regression was performed on the bispectrum components to obtain the total energy. This example had a MAE on the test data of 2.4 meV/atom and an RMS of 2.5 meV/atom. The training data is depicted in blue and the test data is depicted in red.

the predicted results for 2H-NbS$_2$, which had a MAE of 2.4 meV/atom on the test data after fitting. This test data for this system was obtained by running a molecular dynamics simulation, using the trained SNAP-MLIP and taking a snapshot of the system at 500 K every 1,000 time steps of 0.5 fs until 100 representative images of the system were obtained. The DFT energies of the resulting configurations were obtained by performing SCF calculations on the resulting configurations and the metrics mentioned above were obtained by comparing the SNAP energies outputted as part of the MD simulation, with the resulting DFT values.

The training MAE and RMSE for each of the rest of the non-magnetic, monolayer SNAP potentials in the full library of 71 SNAP potentials are presented in Table 4.1. As is evident from the presented results, the training error is consistently low for every case. The training errors are displayed as opposed to the test errors due

Table 4.1: The MAE and RMSE were calculated for the training of each of the 71 systems within the SNAP potentials library for non-magnetic, hexagonal 2D monolayers The magnitude of these errors in both cases are consistently on the order of 1 meV/atom, which is around the expected error of DFT calculations.

Monolayer	MAE	RMSE	Monolayer	MAE	RMSE
	(meV/atom)	(meV/atom)		(meV/atom)	(meV/atom)
1T-AuTe$_2$	2.942	4.335	1T-SnS$_2$	0.346	0.440
1T-BiTe$_2$	1.918	2.925	1T-SnSe$_2$	0.338	0.429
1T-CaI$_2$	0.313	0.401	1T-TaS$_2$	0.376	0.480
1T-CdBr$_2$	0.279	0.366	1T-TaSe$_2$	0.352	0.455
1T-CdCl$_2$	0.343	0.427	1T-TiBr$_2$	0.384	0.490
1T-CdI$_2$	0.249	0.319	1T-TiCl$_2$	0.331	0.426
1T-CoTe$_2$	0.477	0.598	1T-TiS$_2$	2.571	3.772
1T-FeBr$_2$	0.253	0.328	1T-TiSe$_2$	0.366	0.459
1T-GeBr$_2$	0.567	0.722	1T-TiTe$_2$	0.296	0.382
1T-GeI$_2$	0.430	0.533	1T-TmI$_2$	2.362	3.474
1T-HfS$_2$	0.364	0.476	1T-YbI$_2$	0.300	0.382
1T-HfSe$_2$	0.343	0.453	1T-YbSe$_2$	1.683	2.477
1T-HfTe$_2$	0.352	0.461	1T-ZnBr$_2$	0.235	0.302
1T-HgBr$_2$	2.361	3.246	1T-ZnCl$_2$	0.250	0.319
1T-IrTe$_2$	0.651	0.827	1T-ZnI$_2$	0.226	0.288
1T-MgBr$_2$	0.246	0.317	1T-ZrS$_2$	0.403	0.510
1T-MgCl$_2$	0.280	0.355	1T-ZrSe$_2$	0.369	0.470
1T-MgI$_2$	0.220	0.282	1T-ZrTe$_2$	0.381	0.494
1T-MoS$_2$	2.654	3.780	2H-CrSe$_2$	0.225	0.292
1T-Ba$_2$N	0.286	0.363	2H-GeI$_2$	0.369	0.474
1T-Ca$_2$N	0.204	0.256	2H-LaBr$_2$	2.121	3.130
1T-NbS$_2$	0.376	0.485	2H-MoS$_2$	0.231	0.305
1T-NbSe$_2$	0.346	0.441	2H-MoSe$_2$	0.233	0.302
1T-NbTe$_2$	0.512	0.641	2H-MoTe$_2$	0.255	0.328
1T-NiO$_2$	0.199	0.257	2H-NbS$_2$	0.591	0.756
1T-Tl$_2$O	0.433	0.548	2H-NbSe$_2$	0.545	0.683
1T-PSn$_2$	0.350	0.449	2H-ReSe$_2$	2.846	4.193
1T-PbI$_2$	0.696	0.879	2H-TaS$_2$	0.501	0.644
1T-PdTe$_2$	0.258	0.338	2H-TaSe$_2$	0.484	0.625
1T-PtO$_2$	0.264	0.343	2H-WS$_2$	0.234	0.307
1T-PtS$_2$	0.203	0.264	2H-WSe$_2$	0.247	0.312
1T-PtSe$_2$	0.201	0.260	2H-WTe$_2$	0.254	0.328
1T-PtTe$_2$	0.224	0.292	2H-ZrCl$_2$	0.201	0.257
1T-RhTe$_2$	3.264	4.683	Graphene	0.763	0.945
1T-Tl$_2$S	0.593	0.746	hBN	0.084	0.111
1T-SiTe$_2$	0.346	0.444			

to the infeasibility of constructing a representative test set for every system in the library due to the thousands of DFT calculations that would be required for such an endeavour. The training set error is, however, a reasonable indication of the quality of the resulting 71 potentials, which should prove to be of great value to the materials science community.

4.1.4 Comparison with DFPT

Taking the results from the first-principles calculation in Chapter 3, and comparing them with the results that are obtained using the 'quantum-accurate' SNAP MLIP, can offer valuable insight into the capabilities of these ML potentials, while also elucidating the reason for the exclusion of the magnetic materials from the library of MLIPs for 2D materials.

The workflow outlined in the previous sections was executed for the case of NbS_2, training the SNAP on total energy calculations in the non-magnetic case, the derivative of which was used to calculate the forces on each atom, as implemented in the `LAMMPS` molecular dynamics library [68]. The finite-displacement method described in section 3.1.1 was employed using the SNAP potential with a 3×3 supercell in order to calculate the phonon dispersion of the system. The parameter J_{max} was set to 4 and a cutoff radius of 5.1 Å was used for NbS_2 after having determined this as being the optimal cutoff. Studies performed on other materials were performed using a cutoff radius of 4.1 Å as that generally performed slightly better for the other cases tested.

This comparison between the SNAP-obtained phonon dispersion for the system and the one obtained using *ab initio* DFPT can be seen in Fig. 4.4. The expected dynamical instability, resulting from the apparent existence of the charge density wave (CDW) in the harmonic potential, disappears in the case of SNAP for NbS_2, as well as the softening of the optical phonon modes also being absent in the phonon dispersion obtained with the MLIP. This fact highlights the inability of the SNAP-based MLIP to capture the fundamental dynamics arising from the inclusion of spin dynamics or long-range electrostatic interactions, such as those which cause CDWs. This fact is related to the short-range nature of the descriptor and the assumption of SNAP that there only exists local contributions to the total energy. There have been some attempts to include spin dynamics into this family of computationally-efficient MLIPs [69], however, the inclusion of these additional features is currently beyond the scope of this work.

Even beyond the expected dynamical instability near the M-point, the overall qual-

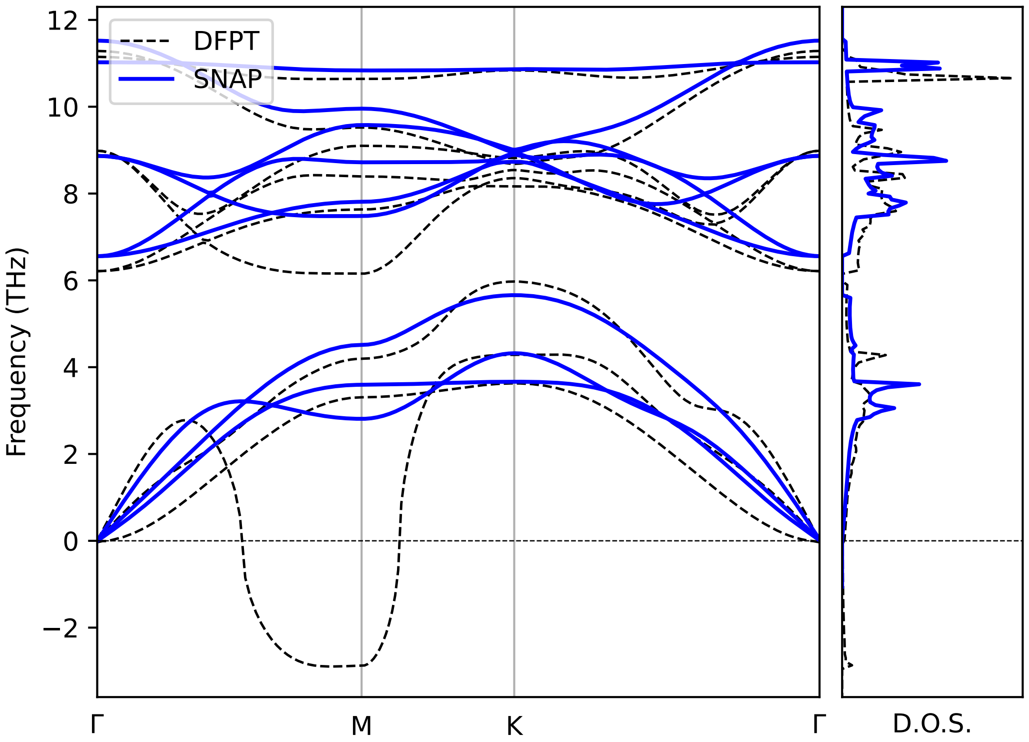

Figure 4.4: Comparison of the phonon dispersion for monolayer 2H-NbS$_2$ calculated using first-principles DFPT *(dashed line)* and finite-difference using the SNAP-MLIP *(blue line)*. The apparent CDW from first principles calculation in the harmonic approximation does not appear near the M-point for the dispersion obtained by using the locally-resolved SNAP-MLIP.

ity of the fit of the dispersion to the DFPT example is relatively poor for the SNAP potential. This is likely due to the model attempting to fit the long-range contributions to the energy resulting from the CDW state into the energy contributions from the local environments, further evidencing the issues arising from the inclusion of systems with complex dynamics into the library of MLIPs, without first resolving interactions of this nature in the workflow.

The inclusion of long-range electronic interactions can be achieved with the addition of another term in the energy, beyond the locally-resolved SNAP potential [66]. These long-range electrostatic contributions are generally tuned to each individual monolayer or represent coarse approximations of this long-range potential, and therefore, it was not deemed prudent to include such long-range interactions for the initial construction of our library of potentials. This is particularly valid considering that these long-range interactions are not present in every system, thus increasing the specificity of the requirement to certain systems. However, this may be explored in future work in order to adequately capture all of the rich dynamics of 2D materials emerging from their exotic phenomena. This currently, however, is a potential limiting factor in the ability to accurately portray the dynamics of the systems.

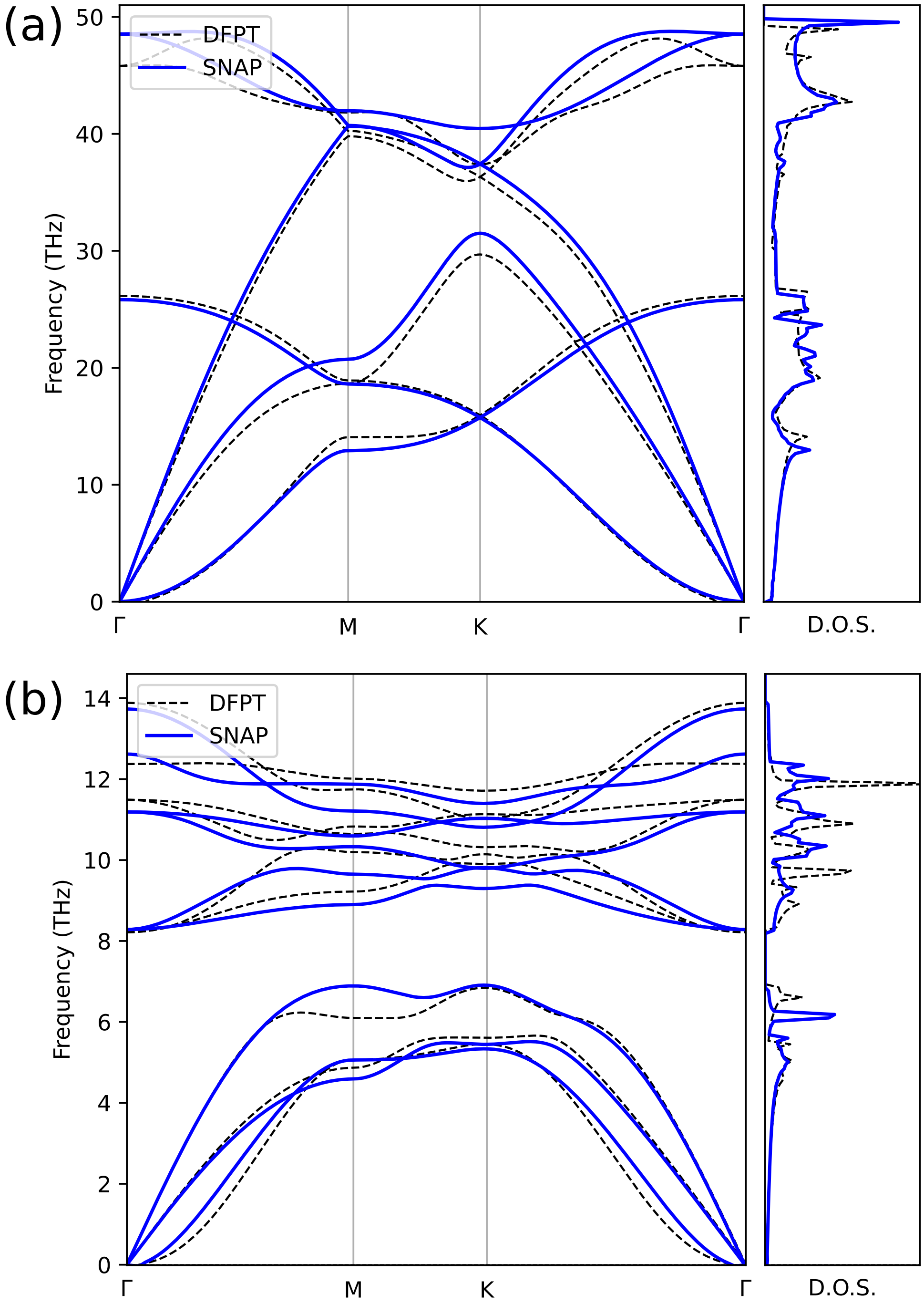

Figure 4.5: Comparison of the phonon dispersion calculated using first-principles DFPT *(dashed line)* and finite-difference using the SNAP-MLIP *(blue line)* for a (a) graphene monolayer and (b) MoS_2 monolayer.

The lack of ability to model the spin state is very simply dealt with by excluding monolayers that have already been established as being magnetic according to Materials Cloud [67] and assessing the quality of the resulting potentials without the concern of non-accounted-for phenomena impacting the faithfulness of the system dynamics. This is evident from the relative improvement in the overall fit of the phonon dispersions, examples of which can be seen for the sample cases of graphene in Fig. 4.5 (a) and MoS_2 in Fig. 4.5 (b). These were fit with a cutoff radius of 4.1 Å and a $J_{\mathrm{max}} = 4$.

These phonon dispersions generally show excellent agreement with DFPT data, particularly in the case of graphene, which only presents some slight deviations from the dynamical behaviour exhibited in the optical branches near the Γ-point. These deviations are very minor, however, and do not impact the remarkable accuracy of the reproduction of the phonon dispersion, using a relatively simple linear model. The case of MoS_2 also presents generally excellent agreement with DFPT, despite a slight softening of the longitudinal acoustic mode at the M-point, a fact that is not physical. This is likely due to the enhanced complexity of modelling the dynamics between multiple atomic species and could potentially be resolved by introducing different weighting between species as part of the fitting procedure. This is a relatively minor effect, however, and would add unnecessary complexity to the workflow of MLIP construction by requiring an additional tuning step for every MLIP produced.

Overall, the phonon dispersions obtained for non-magnetic monolayers exhibit excellent agreement with the results expected according to *ab initio* DFPT calculations. This fact implies that these MLIPs have the potential to be a useful tool for obtaining valid estimates of material properties that are dependent on phonon transport and vibrational dynamics.

4.1.5 Van der Waal's Interaction

In order to extend these potentials to arbitrary heterostructure stacks of 2D materials, such that the materials-property space can be fully explored, it is necessary to conceive of a means of adequately representing the interlayer coupling. This can simply be facilitated by parametrically fitting the Lennard-Jones (LJ) potential [70],

$$E_{\mathrm{LJ}} = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - 2\left(\frac{\sigma}{r}\right)^{6} \right], \quad r < r_{\mathrm{cutoff}} \tag{4.16}$$

where $r_0 = 2^{\frac{1}{6}}\sigma$ is the distance at which the energy is a minimum and ϵ is the depth of the potential well. Both parameters need to be fit for each atom type in each of the different systems. The LJ potential is immensely valuable for use in molecular dynamics due to the ability to implement this potential in conjunction with others (like SNAP) through the `LAMMPS` library [68]. The difficulty, however, is in the optimal means of obtaining adequate data with which to obtain the parameters for the fitting of this interlayer potential.

There are three main adjustments to the charge density in DFT that can be applied in order to mimic the VdW interaction using first-principles calculations, which are the DFT-D2 [71], DFT-D3 [72] and the Tkatchenko-Scheffler [73] (TS) methods. The TS method is implemented alongside iterative Hirshfeld partitioning to determine the interaction between atoms in a solid and modify the TS parameters accordingly. This iterative procedure does not work well with plane-wave basis sets and, therefore, `FHI-AIMS` [74] was employed in the place of `VASP` to obtain the VdW parameters for this comparison. In order to obtain the best of these algorithms for use in fitting the VdW parameters for the LJ potential, different flavours of VdW corrections were used for different interlayer distances in bilayer MoS_2. The interlayer binding energy was calculated by taking the difference between the total energy of the individual layers and the interacting layers, with varying interlayer distances. The binding energy is, as a result, taken to be the depth of the resulting potential well. The approximated binding energy resulting from these calculations was compared with the experimental result for MoS_2, which was 34.3 meV/Å^2 [75], as seen in Table 4.2.

Table 4.2: Binding energy of MoS_2 as calculated with a variety of different flavours of VdW corrections. These values were calculated using `FHI-AIMS` with varying interlayer distance and taking the difference between the individual layers and the interacting ones. The comparable value of binding energy determined from experiment [75] is presented in grey.

	D2	D3	TS	Exp.
E_b (meV/Å^2)	17.2	27.2	29.6	34.3

This same test was performed with variable cutoff distances for the potential. However, this variable was not shown to improve the results much after tuning, exhibiting very similar performance on the fit. Thus, it was determined that using the TS method, with a cutoff of 12 Å was the optimal and most efficient choice for parameterising the

VdW interaction, capturing the most realistic interlayer binding behaviour in comparison with experimental results.

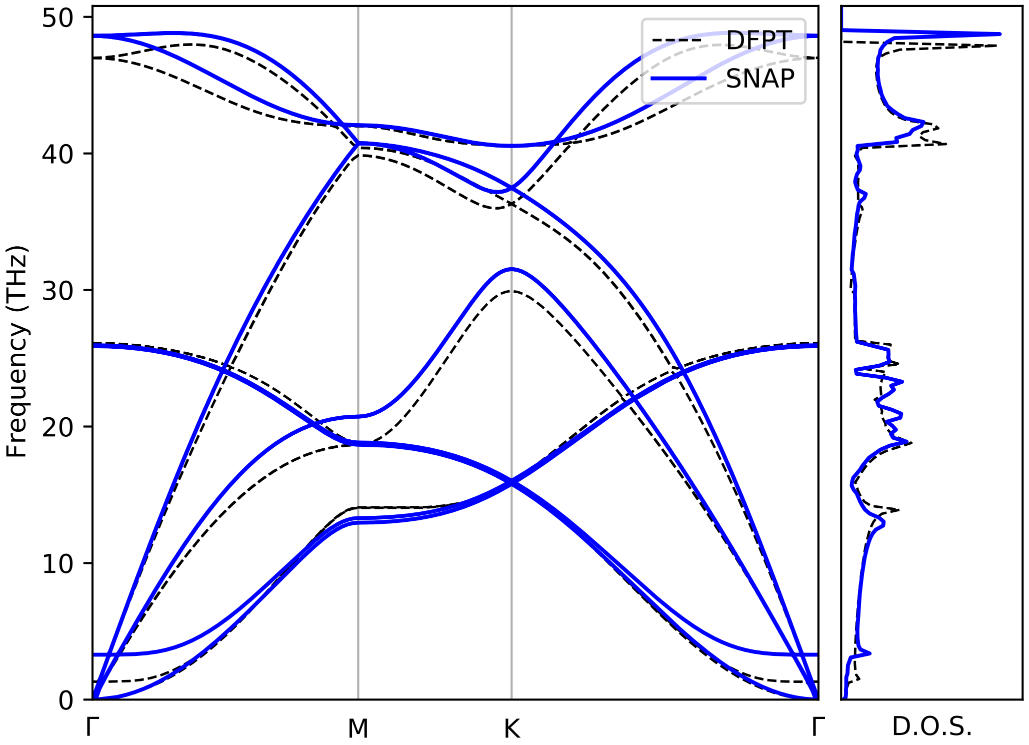

Figure 4.6: A comparison of the phonon dispersion and density of states for bilayer graphene obtained using SNAP *(solid blue line)* and DFPT *(dashed line)*. The interlayer interaction has introduced breathing and shear vibrational modes near the Γ-point of the dispersion.

This fact is evidenced by the comparison of dispersions obtained for bilayer graphene using this hybrid SNAP/LJ method with a 5 × 5 supercell, and the one obtained using DFPT in VASP with an equivalent supercell, as seen in Fig. 4.6. The SNAP/LJ manages to capture the dynamics of the bilayer system very well, with the emergence of the interlayer breathing and shear vibrational modes in the acoustic branches of the phonon dispersion near the Γ-point. The breathing mode frequency is slightly overestimated, implying a tighter interlayer binding than would be expected with DFPT. However, the system dynamics are still very well represented. With this replication of the expected results and having obtained results comparable to those *ab-initio* calculations, the door is open for these methods to be extended to systems of arbitrary size.

Evidence of this benefit can be seen in Fig. 4.7, which presents the phonon dispersion of a graphene system, which has been encapsulated by hBN monolayers, calculated using SNAP-MLIPs with a 5 × 5 supercell. This took a time of the order of several

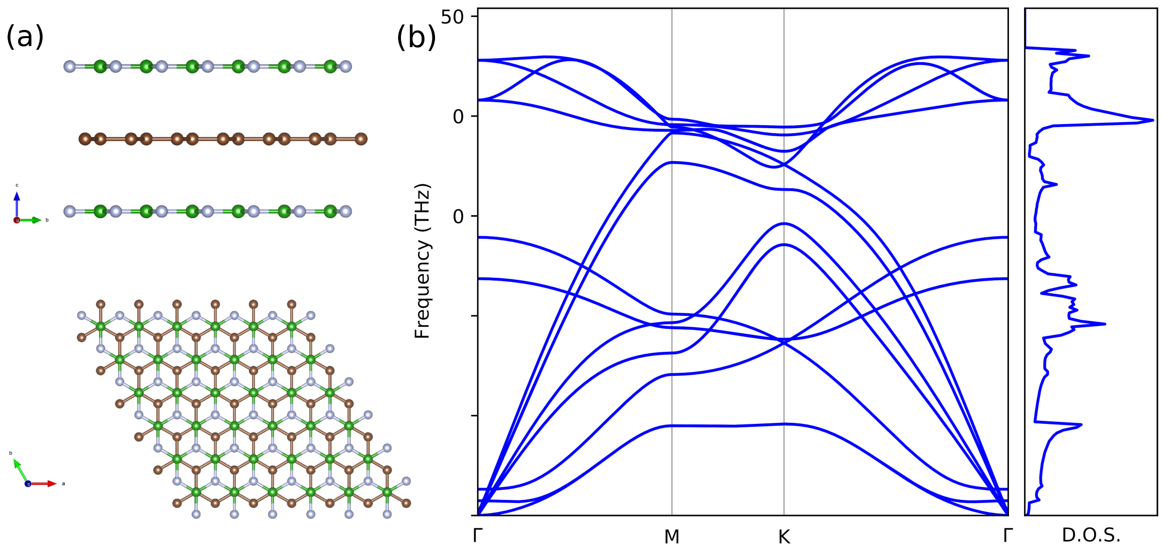

Figure 4.7: (a) The side and top view of the structure of a 6×6 supercell of a graphene monolayer *(brown)*, encapsulated between two monolayers of hBN *(green and silver for B and N, respectively)*. (b) An example of a phonon dispersion and density of states for the same system, calculated with a combination of SNAP-MLIPs and interlayer LJ potentials.

minutes, using the CPU of a laptop, as opposed to an *ab-initio* calculation for an equivalent system, which would either be completely inaccessible to first-principles calculations or take time on the order of hours or days depending on the method of calculation and the hardware accessible to a user. This fact indicates the massive potential of such MLIPs in simplifying the complex issue of scanning materials space for desirable composite materials for a given application, with minimal computational cost. This example also provides evidence that the strategy of extending the potential for multilayer systems beyond the bilayer, homogeneous example given above in Fig. 4.6, is sound. It also evidences that the potential can be employed for arbitrary combinations of heterogeneous layers.

4.2 Calculation of Thermal Properties

Once it was clear that the MLIPs, extended using parameterized interlayer interactions were sufficient to model the microscopic properties of composite, heterogeneous materials, all of the ingredients required to consider their performance on the task of computing macroscopic properties of interest were available. Thus, means of efficiently calculating properties of interest must be considered.

A feature of these methods that could also be of further interest is analysing the types of insight one can gain from these macroscopic properties using MLIPs that are able to effectively model quantum-accurate interactions between atoms. Such a fact means that these methods could give a means of achieving insights into how to best optimise a material for a given application by analysing by-products of these calculations, produced at no additional computational cost, and tuning the resulting material based on the insights offered.

This next section will discuss various ways of performing thermal transport calculations, using SNAP-based MLIPs and will demonstrate their use in a variety of circumstances. Comparisons will be done with experimental tests for the same properties and the potential of the use of these methods for massive systems will also be demonstrated. This comprehensive, but non-exhaustive discussion will evidence the use of this library of potentials for diverse physical and technological applications.

4.2.1 Thermal Conductivity

One of the most fundamental thermal properties of a solid-state system is the thermal conductivity κ, which corresponds to the amount of heat energy per unit time flowing through a given system. The origin of this quantity can clearly be understood through the classical view of thermal transport, Fourier's law as defined by Joseph Fourier in 1822

$$\mathbf{J}(t) = -\kappa \boldsymbol{\nabla} T(t), \tag{4.17}$$

where $\mathbf{J}$ corresponds to the heat flux density and $\boldsymbol{\nabla} T$ is the temperature gradient of the system. It is evident from Eq. (4.17), thermal conductivity is the key quantity describing the efficiency of the displacement of heat energy, when a thermal gradient is applied to a material. This quantity is obviously of vital importance for a huge variety of applications, be they displacing waste heat to ensure the efficiency of computational processes or transferring thermal energy from one point to another. The thermal conductivity is also of key importance for determining the efficiency of conversion from thermal energy into electrical energy in thermoelectric materials, with the thermoelectric figure of merit zT, being defined as

$$zT = \frac{S^2\sigma}{\kappa}T, \tag{4.18}$$

where S is the Seebeck coefficient, σ is the electrical conductivity and T is the absolute temperature. Thus, it is clear that, in order to have a pipeline that screens materials for their potential thermal applications, having a method of obtaining κ is of utmost importance. With a highly efficient means of calculating phonon dynamics for 2D materials and their heterostructure, the door is opened to calculate the macroscopic thermal conductivity by rapidly calculating interactions between phonons and, thereby, measuring the efficiency of thermal energy transfer in 2D materials and their heterostructures.

The thermal conductivity of a system is mediated by two main contributions, the electronic part κ_e and the lattice thermal conductivity, κ_L, such that $\kappa = \kappa_e + \kappa_L$. Generally speaking, in non-metallic systems, κ_L is the main contribution to κ. It metallic systems, κ_e increases in importance. However, κ_L remains an important contribution. This quantity can be described entirely in terms of phonons. Thus, here we focus on calculating κ_L and, hereafter, the term 'thermal conductivity' refers exclusively to the lattice contribution. Methods for calculating the thermal conductivity from phonon interactions fall into two main categories, those that are solved using the Green-Kubo theory of linear response [76–79], and those that are solved using the Boltzmann transport equation (BTE) for phonons. The former method tends to perform better for non-crystalline systems, whereas for extended, periodic systems, the latter BTE method tends to be perfectly adequate and, therefore, will be the focus of our attention. All calculations performed to obtain the thermal conductivity were performed by adapting the `phono3py` [35, 80] pipeline to be used with SNAP-MLIPs. This is achieved by using the `LAMMPS` calculator from the Atomic Simulation Environment (`ASE`) [81] library to calculate the forces for the atomic displacements provided by `phono3py`. The BTE can be written as [34]

$$\frac{\partial n^{(1)}_{\mathbf{q}\nu}}{\partial t} + \frac{\partial n^{(0)}_{\mathbf{q}\nu}}{\partial T}\frac{\partial T}{\partial \mathbf{r}} \cdot \mathbf{v}_{\mathbf{q}\nu} = C(\mathbf{q}\nu; n^{(1)}_{\mathbf{q}\nu}) + \frac{1}{2}D(\mathbf{q}\nu; n^{(1)}_{\mathbf{q}\nu}). \tag{4.19}$$

The phonon scattering contribution to the transport appears on the right-hand side of the equation through the collision processes term, $C(\mathbf{q}\nu)$, and the decay processes term, $D(\mathbf{q}\nu)$, where $\mathbf{q}\nu$ holds the same meaning as in Chapter 3 for the phonon wavevector $\mathbf{q}$ and phonon branch number ν. The term $n_{\mathbf{q}\nu}$ is the occupation function for the $\mathbf{q}\nu$-phonon mode. At equilibrium, the occupation function would follow Bose-Einstein statistics as in Eq. (3.17). This equilibrium occupation function is depicted as $n^{(0)}_{\mathbf{q}\nu}$, and $n^{(1)}_{\mathbf{q}\nu}$ is the first-order deviation from the equilibrium such that $n_{\mathbf{q}\nu} \approx n^{(0)}_{\mathbf{q}\nu} + n^{(1)}_{\mathbf{q}\nu}$. Finally, $\mathbf{v}_{\mathbf{q}\nu}$ is the velocity of a phonon occupying the $\mathbf{q}\nu$ state .

This expression for the phonon BTE can be rearranged [82] in order to make the relationship between the collision and the decay processes and the phonon-phonon interactions more explicit

$$C(\mathbf{q}\nu; n^{(1)}_{\mathbf{q}\nu}) + \frac{1}{2}D(\mathbf{q}\nu; n^{(1)}_{\mathbf{q}\nu}) = -\sum_{\mathbf{q}'\nu'} \mathbf{\Omega}_{\mathbf{q}\nu;\mathbf{q}'\nu'} n^{(1)}_{\mathbf{q}'\nu'} \frac{\sinh\left(\frac{\hbar\omega_{\mathbf{q}'\nu'}}{2k_BT}\right)}{\sinh\left(\frac{\hbar\omega_{\mathbf{q}\nu}}{2k_BT}\right)}, \tag{4.20}$$

which introduces the collision matrix $\mathbf{\Omega}_{\mathbf{q}\nu;\mathbf{q}'\nu'}$

$$\begin{aligned}\mathbf{\Omega}_{\mathbf{q}\nu;\mathbf{q}'\nu'} =& \delta_{\mathbf{q}\nu;\mathbf{q}'\nu'} \frac{1}{\tau_{\mathbf{q}\nu}} + \frac{36\pi}{\hbar^2} \sum_{\mathbf{q}''\nu''} |\mathbf{\Phi}_{\mathbf{q}\nu;\mathbf{q}'\nu';\mathbf{q}''\nu''}|^2 \frac{1}{\sinh\left(\frac{\hbar\omega_{\mathbf{q}''\nu''}}{2k_BT}\right)} \\ & \times [\delta(\omega_{\mathbf{q}\nu} - \omega_{\mathbf{q}'\nu'} - \omega_{\mathbf{q}''\nu''}) \\ & + \delta(\omega_{\mathbf{q}\nu} + \omega_{\mathbf{q}'\nu'} - \omega_{\mathbf{q}''\nu''}) + \delta(\omega_{\mathbf{q}\nu} - \omega_{\mathbf{q}'\nu'} + \omega_{\mathbf{q}''\nu''})],\end{aligned} \tag{4.21}$$

where $\tau_{\mathbf{q}\nu}$ is the phonon lifetime of a phonon occupying the $\mathbf{q}\nu$ state. The expression $\mathbf{\Phi}_{\mathbf{q}\nu;\mathbf{q}'\nu';\mathbf{q}''\nu''}$ is the matrix of the third-order force constants, which represents our first foray into anharmonic lattice dynamics. This term is calculated by extending the Taylor series expansion in Eq. (3.4), by another term in order to consider three-phonon processes beyond the harmonic regime. Given that the SNAP-MLIP is a highly accurate approximation of the potential energy surface, and has no assumptions of harmonicity, the anharmonic lattice dynamics should be accessible to calculation via SNAP-MLIPs. The calculation for the third-order force constants can be performed using the finite displacement method as described in Section 3.1.1, extended for the third-order anharmonic term as implemented in `phono3py`. This matrix can also be used to obtain the phonon lifetimes $\tau_{\mathbf{q}\nu}$.

The collision matrix $\mathbf{\Omega}_{\mathbf{q}\nu;\mathbf{q}'\nu'}$ is a quantity that can become very large very fast, with increasing **q**-point sampling density in reciprocal space. Thus, in order to reduce the size of this matrix, the crystal symmetries can be exploited and the velocity field can be restricted to the irreducible part of the Brillouin zone (BZ). Thus, the collision matrix can be transformed into a more compact, memory-efficient form,

$$\tilde{\mathbf{\Omega}}_{\mathbf{q}\nu;\mathbf{q}'\nu'} = \frac{1}{\sqrt{g_{\tilde{\mathbf{q}}}g_{\tilde{\mathbf{q}}'}}}\sum_{R\in P} R_{\alpha\alpha'}\mathbf{\Omega}_{\tilde{\mathbf{q}}\nu;R_{\tilde{\mathbf{q}}'\nu'}}, \tag{4.22}$$

where $\tilde{\mathbf{q}}$ is the phonon wavevector in the irreducible part of the BZ, R is the rotational operation in the point group P, α is the Cartesian coordinate index and $g_{\tilde{\mathbf{q}}}$ is the order of the point group at wavevector $\tilde{\mathbf{q}}$.

Ref. [82] demonstrated that the thermal conductivity tensor could be written as

$$\kappa_{\alpha\alpha'} = \frac{1}{4k_BT^2NV_c}\sum_{\mathbf{q}\nu;\mathbf{q}'\nu'}\frac{\hbar\omega_{\mathbf{q}\nu}v_{\mathbf{q}\nu\alpha}}{\sinh\left(\frac{\hbar\omega_{\mathbf{q}\nu}}{2k_BT}\right)}(\Omega^{\sim1})_{\mathbf{q}\nu;\mathbf{q}'\nu'}\frac{\hbar\omega_{\mathbf{q}'\nu'}v_{\mathbf{q}'\nu'\alpha'}}{\sinh\left(\frac{\hbar\omega_{\mathbf{q}'\nu'}}{2k_BT}\right)}, \tag{4.23}$$

where $\Omega^{\sim1}$ is the Moore-Penrose pseudoinversion of the collision matrix, V_c is the unit cell volume and N is the number of atoms. Rewriting this in terms of the compact collision matrix given in Eq. (4.22) gives

$$\kappa_{\beta\beta'} = \frac{1}{4k_BT^2NV_c}\sum_{\mathbf{q}\nu;\mathbf{q}'\nu'}\frac{\hbar\omega_{\mathbf{q}\nu}v_{\mathbf{q}\nu\alpha}}{\sinh\left(\frac{\hbar\omega_{\mathbf{q}\nu}}{2k_BT}\right)}\frac{[\Omega^{\sim1}I(\beta,\beta')]_{\tilde{\mathbf{q}}\nu\alpha;\tilde{\mathbf{q}}'\nu'\alpha'}}{\sqrt{g_{\tilde{\mathbf{q}}}g_{\tilde{\mathbf{q}}'}}}\frac{\hbar\omega_{\mathbf{q}'\nu'}v_{\mathbf{q}'\nu'\alpha'}}{\sinh\left(\frac{\hbar\omega_{\mathbf{q}'\nu'}}{2k_BT}\right)}, \tag{4.24}$$

here $I(\beta,\beta')$ is a matrix that allows the vectors in the irreducible BZ to map to the ones in the full BZ as

$$I_{\tilde{\mathbf{q}}\nu\alpha;\tilde{\mathbf{q}}'\nu'\alpha'}(\beta,\beta') = \delta_{\mathbf{q},\mathbf{q}'}\delta_{\nu,\nu'}\sum_{R\in P}\frac{R_{\beta\alpha}R_{\beta'\alpha'}+R_{\beta'\alpha}R_{\beta\alpha'}}{2}. \tag{4.25}$$

Combining all of these elements means that there is a means of obtaining the lattice thermal conductivity of a system by solving the linearized Boltzmann transport equation directly, from the direct computation of the third-order force constant matrix alone. This is a calculation, which is achievable at a vastly reduced computational cost as a result of the construction of the SNAP-MLIP potentials. Thus, SNAP-MLIPs could prove to be an invaluable tool for rapidly calculating the lattice thermal conductivity and its related quantities for arbitrary heterostructures.

Table 4.3: Thermal conductivities of materials that were estimated using SNAP at 300 K with a sampling mesh in the irreducible Brillouin zone of $71 \times 71 \times 1^1$, compared with those calculated from DFT-based studies and results taken from experiment. The materials in this representative sample of the 69 systems in the SNAP library were chosen due to the research focus the community has placed on them and the access to relevant information.

	κ_{SNAP} (W/mK)	κ_{DFT} (W/mK)	$\kappa_{Exp.}$ (W/mK)
MoS_2			
1-layer	87.5	104 [83]	84 ± 17 [84]
$MoSe_2$			
1-layer	32	54 [83]	59 ± 18 [84]
WS_2			
1-layer	155	142 [83]	32-63 [85–87]
WS_2/MoS_2			
Heterostructure	99.5	70 [88]	-
Graphene			
1-layer	3182	2000-5000 [89]	3000-5000 [90–92]
2-layer	1934	2200 [93]	1896 ± 410 [94]
3-layer	1837	-	1495 ± 150 [95]
hBN			
1-layer	456	650 [96]	751 ± 350 [57]
2-layer	296	-	484 +141/-24 [97]
3-layer	296	-	-

It is quite difficult to obtain data for thermal conductivity measurements from theory or experiment due to anomalous effects that may skew the theoretical result away from the experimental result, or the deviation due to equipment details on these generally very sensitive measurements. Despite this, Table 4.3 presents a select example of results from the calculations of thermal conductivity at 300 K resulting from the SNAP-MLIPs with irreducible BZ sampling of $71 \times 71 \times 1^1$, coupled with the direct solution of the linearized BTE. Broadly speaking, the results from the SNAP-MLIP do capture the trend of results expected from either experiment or obtained by using DFT-based methods.

For the case of the transition metal dichalcogenides (TMDs), the expected results are very well replicated for the case of MoS_2 and reasonably well replicated for $MoSe_2$. Both the SNAP-MLIP and DFT-based methods overestimate the thermal conductivity of single-layer WS_2. This, apparently, is due to the difficulty that the *ab initio*

[1] Both trilayer systems for hBN and graphene were sampled with a 41 × 41 × 1 sampling mesh because a grid this large was prohibitively expensive for such large systems.

effects have on capturing a range of competing effects relating to a large acoustic optical frequency gap due to the large mass difference between W and S [83]. It makes very intuitive sense that the SNAP-based method would lead to the same divergence in results of thermal conductivity as the DFT-based methods, considering that the SNAP-MLIP is purely trained using DFT data and, therefore, does not contain any additional information. Despite this, the SNAP-based approach exhibits only minor deviations from first-principles results when applied to TMDs. It consistently provides predictions for thermal conductivity that are within the same order of magnitude as those obtained through DFT. Further to this, a prediction of thermal conductivity for a composite heterostructure, WS_2/MoS_2, was achieved using the hybrid-SNAP/LJ method of constructing the potential energy surface. There are, unfortunately, no prior experimental results for this heterostructure for comparison, however, the MLIP-calculated value of 99.5 W/mK compares quite favourably with the DFT-based result of 70 W/mK [88], implying that the MLIP-based system gives reasonable thermal conductivity predictions for hetero-composite materials, and can give a reasonably good estimation of the impact of boundary scattering of phonons as a result of interlayer interactions between different monolayers.

Moving beyond the TMD systems and their heterostructures, similar calculations were performed for the thermal conductivities of materials known for their anomalously high thermal conductivities, namely graphene [90–92] and hBN [57]. Further study was also performed on the few-layer form of these layered structures as there is an abundance of experimental and DFT studies for comparison for these particular materials due to the focus of the materials science community on them. Once again, the results for the monolayer cases of both graphene and hBN in Table 4.3 show very good agreement with both DFT-based data and experiments, with both results falling within the error margin of the experimental measurements. This agreement continues for the multilayer graphene structures, with the expected trend of a reduction of thermal conductivity with an increase in layer number. While there is no explicit DFT measurement for the trilayer case of graphene, it is expected that the thermal conductivity converges to a value in the vicinity of 2000 W/mK on transition from monolayer to bulk [98]. Therefore, this result is fairly consistent with the DFT expectation, implying that the deleterious impact of increased scattering due to phonons on the thermal conductivity is well represented by the LJ estimation of the interlayer coupling.

While the predicted value for the thermal conductivity of monolayer hBN obtained with the SNAP-MLIP is within the experimental error, it does somewhat underestimate the thermal conductivity for the monolayer and bilayer case. While there are no available results for DFT or experiment describing the trilayer case, it is likely safe to assume that this value is also underestimated. This divergence is possibly related to a region of the irreducible BZ, contributing to the thermal conductivity, being slightly undersampled due to the distribution of the sampling points. This can sometimes occur even if a large grid is chosen in the irreducible BZ and it is difficult to achieve convergence. However, this underestimation is not severe for the two cases for which there is data and is still, therefore, useful as a reasonable prediction of the order of magnitude of the thermal conductivity of the multilayer system. Further to this, the expected trend on transition from monolayer to multilayer system is observed with the interlayer interaction inducing phonon boundary scattering, leading to a reduction in thermal conductivity.

Upon demonstrating that these MLIPs are capable of giving a good estimation of the thermal conductivity of multilayer systems, further use can be demonstrated through the analysis of the results obtained during the thermal conductivity calculations for these systems. A variety of analyses can be performed on the results that give key insights into why κ scales as it does for a variety of systems and the contributions to thermal conductivity measurements can be broken down by a variety of means such that the phonon modes that are most important to the thermal transport can be identified.

The phonon lifetimes are a valuable source of information about the relative contribution of phonons of a given frequency to thermal transport. Phonons with longer lifetimes generally have the ability to transfer energy over a longer distance due to the enhanced distance that they can travel before a scattering event. Fig. 4.8 is one such example of insights that can be drawn from this quantity. It depicts the phonon lifetimes of monolayer, bilayer and trilayer hBN. This phononic property emerges naturally from an inversion of the phonon scattering matrix defined in Eq. (4.22). As the number of layers increases, increased boundary scattering leads to a general reduction in the phonon lifetimes in the low-frequency region. This effect is significant in the transition from monolayer system to bilayer. However, this effect is reduced in the

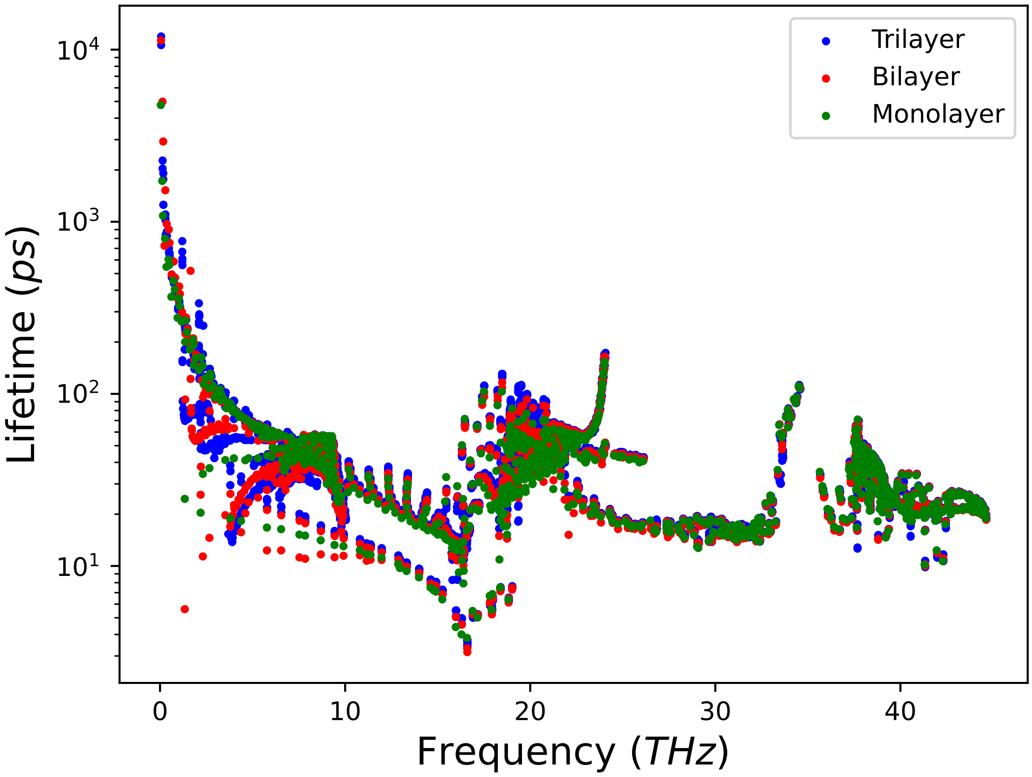

Figure 4.8: Phonon lifetimes, calculated using SNAP-MLIPs for monolayer *(green)*, bilayer *(red)*, and trilayer *(blue)* hBN. A reduction in lifetimes as a resut of increases in boundary scattering can be seen for the multilayer system in the low frequency region of the plot.

transition from bilayers to trilayers. Such a suppression of phonon lifetimes is still somewhat present on transitioning from bilayer to trilayer, while less pronounced. The result of this is that there is a more severe deterioration of the thermal conductivity on the transition from monolayer to bilayer than bilayer to trilayer, as evidenced by the values of κ for hBN presented in Table 4.3.

This can be further understood by analysing the cumulative contribution from each phonon mode to the total thermal conductivity of monolayer hBN, as seen in Fig. 4.9. The low-frequency, acoustic phonons below 10 THz are by far the most dominant in contributing to the overall thermal conductivity of the system. Thus, the suppression of phonons in that frequency range is going to have a detrimental effect on the overall thermal conductivity of the system, as was observed.

Thus, the combination of SNAP-MLIPs and LJ potentials to form a hybrid potential is capable of giving a consistently reasonable prediction for the thermal conductivity

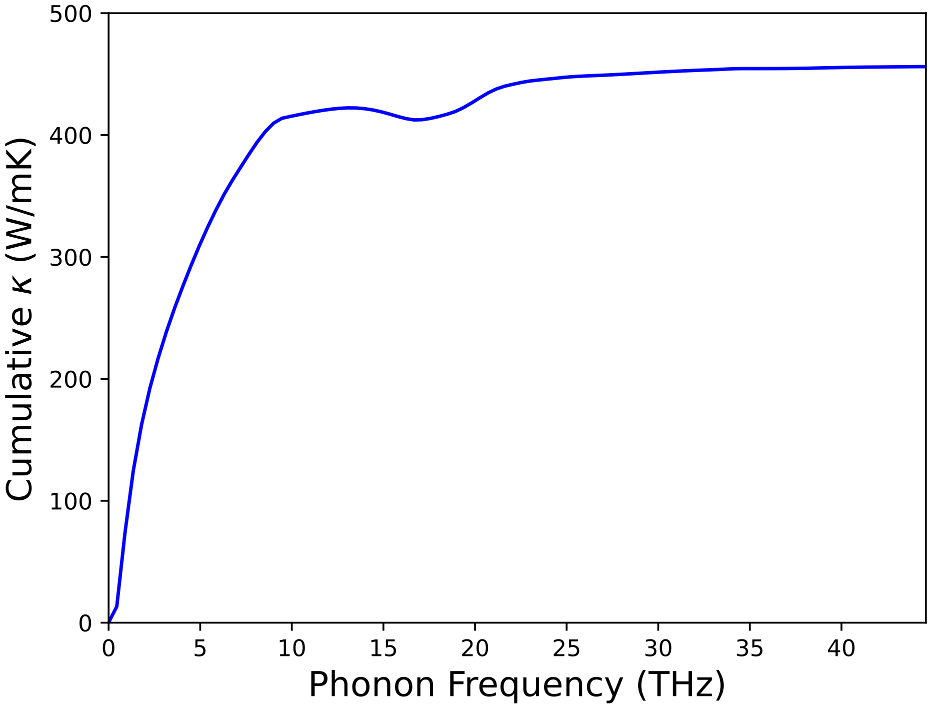

Figure 4.9: The cumulative thermal conductivity of monolayer hBN with the contributions of phonon modes of increasing frequencies. The dominant contribution to the overall thermal conductivity is in the low-frequency, acoustic phonon region, below 10 THz.

of layered materials and heterostructures, on par with the predictions resulting from DFT. This fact unlocks a massive amount of potential for performing calculations of these properties on systems that previously would have been very difficult to model or, indeed, impossible to model. Furthermore, these potentials demonstrate their use, not only in the prediction of the thermal conductivity but also to further enhance our understanding of the behaviour of thermal transport phenomena through the insights gained from auxiliary properties such as phonon lifetimes or the decomposition of the thermal conductivity into its contributions from different phonon modes.

4.2.2 Interfacial Thermal Conductance

In the former section, the SNAP-MLIP was used to perform a computationally efficient estimation of the in-plane thermal conductivity. The potential of these MLIPs, however, is not confined to analysing how interlayer phenomena impact in-plane properties. In fact, these potentials can be applied to study interlayer phenomena and properties. Monolayers have been shown to have potential applications in thermal management systems of electronics, and can potentially be used as a way to dissipate heat directly from electronic components. This behaviour would be mediated through the out-of-plane interlayer interaction between the monolayer and the component or active material in question.

In order to directly compare the ability of the hybrid MLIP to capture these effects, a system for which there is an experimental measurement of the interfacial thermal conductance must be identified. One such system is that of a monolayer of MoS_2, sandwiched between two monolayers of hBN. This is a structure of potential interest due to the direct semiconducting nature of the MoS_2 monolayer, meaning it is an ideal candidate for a number of potential electronic and optoelectronic applications. The thermal conductivity of MoS_2, however, is not extremely high and, therefore, this particular system may not be optimal for the dissipation of heat to improve the performance of devices, which are based on this semiconducting behaviour. Encapsulating the MoS_2 in the higher-κ hBN monolayers could be a valuable means of more effectively dissipating the thermal energy, thereby improving device performance. In order to gauge the efficacy of this transfer of thermal energy, the hybrid SNAP/LJ potential can be employed as a part of a molecular dynamics (MD) simulation, performed using `LAMMPS` to determine the interfacial thermal conductance between hBN and MoS_2, for which there is an experimental measurement [99].

In order to execute this step, a large system must be constructed to adequately capture the energetics of the system, while minimising the potential noise in the simulation. Therefore, a hBN/MoS_2/hBN heterostructure was constructed (see Fig. 4.10), with a total of 11,200 atoms, with 3,800 atoms in each of the hBN layers and 3,600 atoms in the MoS_2 layer. The estimation of the thermal conductance was executed by equilibrating the system such that a temperature difference is established between the hBN layers and the encapsulated MoS_2 layer. A separate temperature difference was maintained, ranging between 50 K and 100 K across the layers for a total of 10 MD simulations. This was achieved by allowing the system to equilibrate using a separate Nosé-Hoover thermostat, a constant temperature, constant volume ensemble acting essentially as a heat bath, for each layer in the system for a total of 1 ns with a time step of 1 fs. Once this step had been executed and each layer had equilibrated to

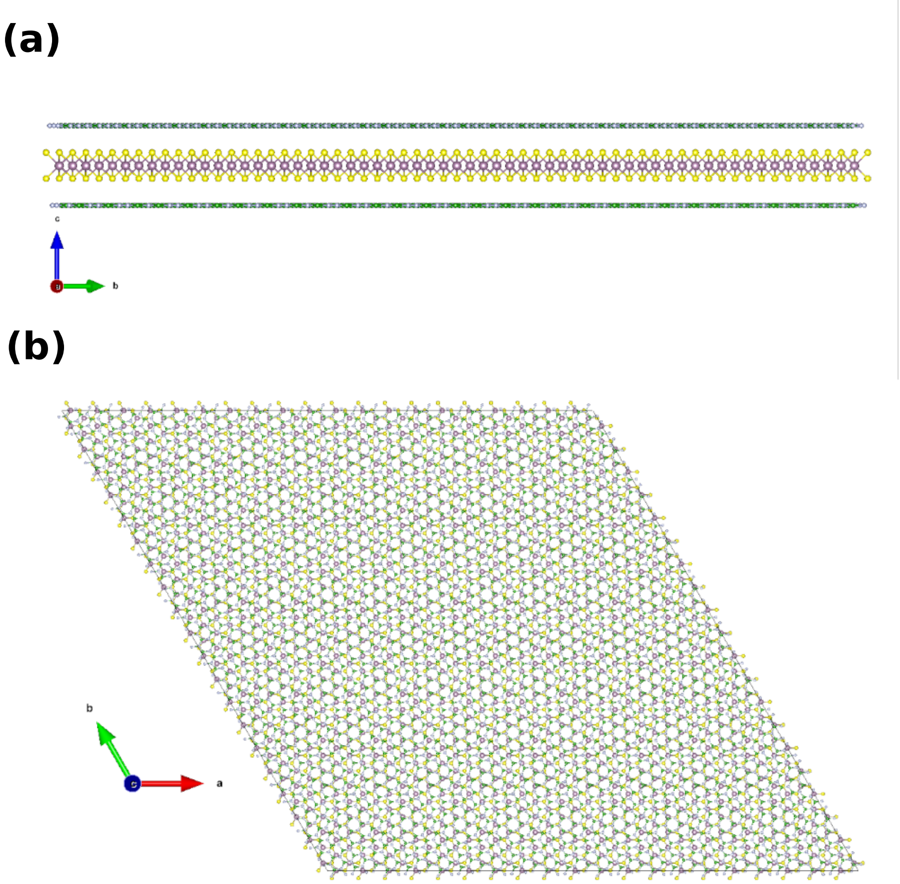

Figure 4.10: A MoS_2 monolayer encapsulated with two monolayers of hBN as seen from the (a) x-axis and (b) z-axis. This system is composed of a total of 11,200 atoms and would be inaccessible to calculations from most *ab initio* techniques.

their respective temperatures, each layer was simultaneously disconnected from their respective thermostats and, was instead allowed to evolve according to the plain-time integration from a standard microcanonical, constant energy ensemble, a plot of the temperature evolution of the system after this step is presented in Fig. 4.11. As the system is isolated, the only pathway for the system to equilibrate is for the excess thermal energy in the MoS_2 layer to be transferred to both of the hBN monolayers. Thus the interfacial thermal conductance can be estimated by tracking the time evolution of the energy of the MoS_2 layer as a function of the temperature difference between the layers, expressed as

$$\frac{dE}{dt} = AG\Delta T, \tag{4.26}$$

where ΔT is the difference in temperature between the hBN layers and the MoS_2, A is

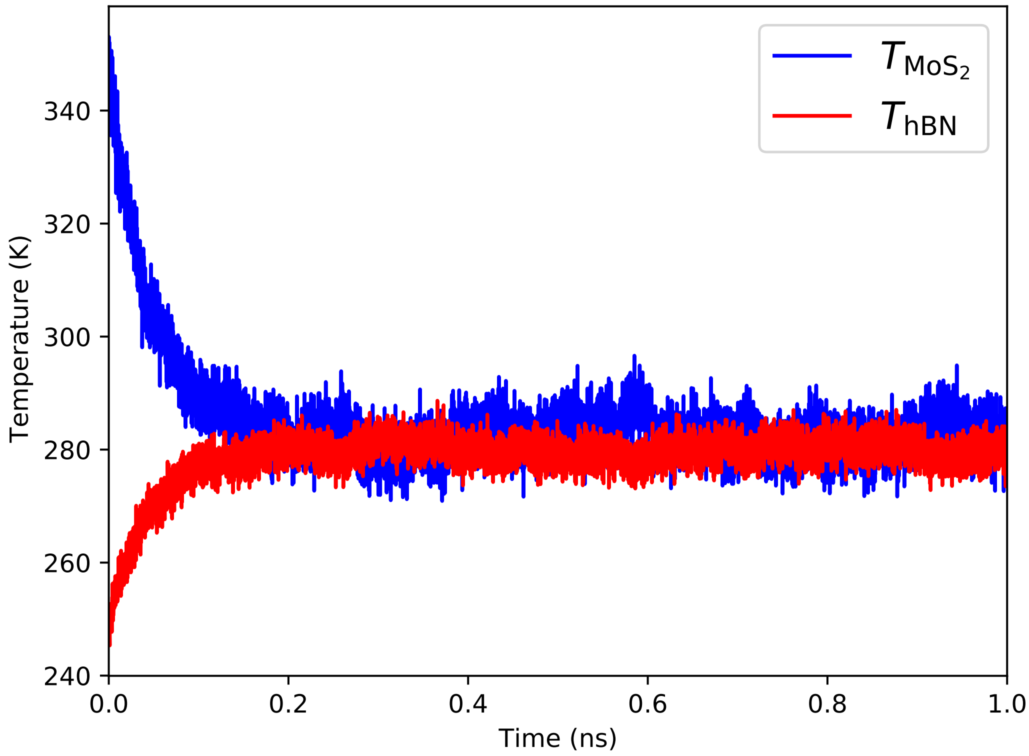

Figure 4.11: An example of the time-dependent temperature of a MD simulation to calculate the interfacial thermal conductance of the hBN/MoS_2/hBN heterostructure. The hBN temperature is represented in red and the MoS_2 is represented in blue. In this example, prior to this time evolution of the temperature with an NVE ensemble, the monolayers had been held at a temperature differential of 100 K, using an NVT thermostat.

the total interfacial area between the layers, and G is the coefficient of interfacial thermal conductance. In order to somewhat mitigate the impact of noise on any estimate of the time derivative of the energy, a running average of the change in energy for every 100 time steps was taken. Thus, by obtaining the best estimate of a linear fit between the averaged time-derivative of the energy, and the difference in temperature between the monolayers, the slope of the resulting linear fit will allow for an estimation of the interfacial thermal conductance between MoS_2 and hBN.

A linear fit of the model was obtained (see Fig. 4.12), giving an R^2 coefficient of 0.67. The resulting fit leads to an estimated interfacial thermal conductance of (11.2 ± 0.3) MW/m^{2}K between MoS_2 and hBN. Contrasting this result with that of the experimental value of (17.0 ± 0.4) MW/m^{2}K [99], implies that this method has the capability of giving a very reasonable prediction for the interfacial thermal conductivity of composite heterojunctions.

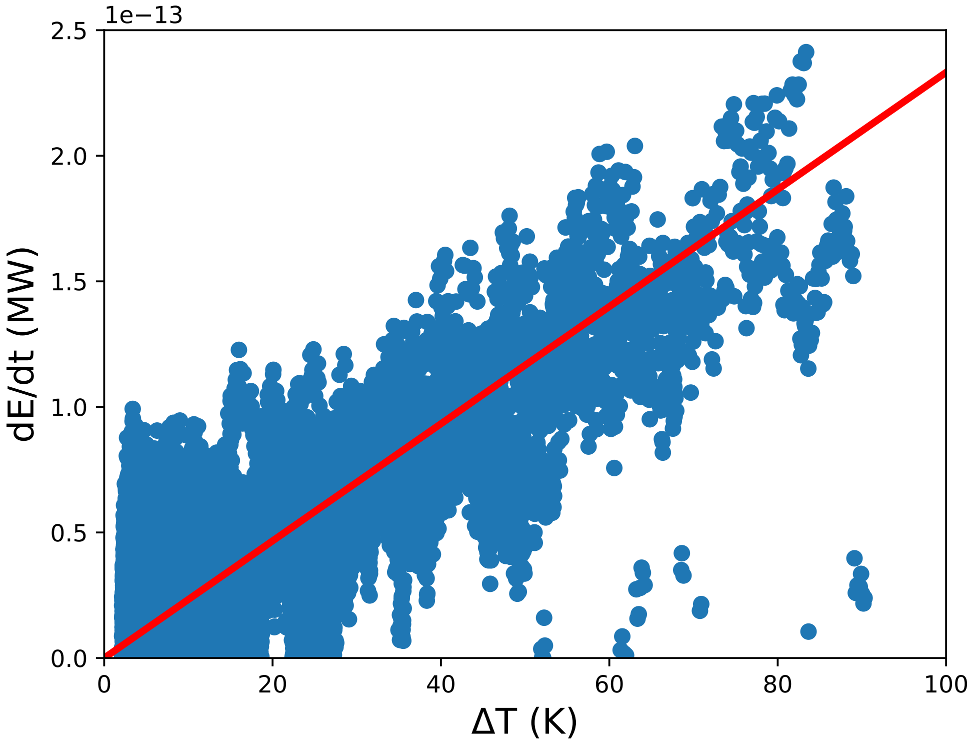

Figure 4.12: The linear fit *(red)* of the temperature gradient between the hBN layers and the higher-temperature MoS_2 layer, against the time derivative of the energy, taken as a running average over every 100 time steps in the MD simulation. The fit gave an R^2 coefficient of 0.67 and resulted in a standard error of the prediction of 0.3 MW/m^2K.

Thus, the use of this hybrid strategy for the construction of SNAP-MLIPs coupled with parameterized LJ interactions has been shown to be useful for the estimation of energy transfer between layers as well as for performing in-plane estimations. This result for the interfacial thermal conductance can be obtained with relatively little computational cost while showing reasonable agreement with experimental results. Any calculation of this scale would generally be entirely prohibitive for first-principles methods due to the computational costs incurred, highlighting a key advantage of the use of these potentials as a surrogate for *ab-initio* calculations, with comparable levels of accuracy, at a fraction of the computational cost.

4.3 Summary & Conclusions

It is evident after this analysis that machine-learned potentials exhibit strong promise for the acceleration of material property predictions. The genesis of machine-learned potentials was described at the start of this chapter, followed by an outline of the key developments, which improved the efficiency and accuracy of these methods. This culminated in a description of the SNAP method for constructing an interatomic potential energy surface, which can achieve near-DFT-level accuracy on energy predictions of compounds with minimal computational cost.

A library of 71 SNAP potentials was constructed for the case of non-magnetic, hexagonal 2D materials and these potentials were extended to account for interlayer interactions by including a parameterized Lennard-Jones interaction, describing the interlayer van der Waal's binding. Methods for calculating the lattice thermal conductivity of systems using these potentials were described, specifically the direct solution of the linearized phonon Boltzmann transport equation, which can be solved for the system through the construction of the third-order force constant matrix. This calculation is easily and efficiently achievable through the use of the hybrid SNAP/Lennard-Jones potential with the finite displacement method for constructing the force constant matrix for both the monolayer and multilayer cases. The resulting calculations for thermal conductivity demonstrated good agreement with those expected based on both experiments and other theoretical methods. Further to this, very useful insights for analysing the trends were obtained through the study of the distribution of phonon lifetimes and cumulative thermal conductivity with increasing phonon frequencies.

The value of the SNAP-MLIPs with the interlayer interaction was further demonstrated by calculating the interfacial thermal conductance between hBN and MoS_2. This calculation required a massive system of 11,200 atoms to converge, a number which is easily within reach for the lightweight MLIP but entirely prohibitive to first-principles methods. The resulting value for the thermal conductance was in good agreement with the value obtained through experimental measurement.

This chapter detailed a significant but non-exhaustive overview of the potential of the library of machine-learned interatomic potentials generated for 2D materials, as well as some use-case scenarios for the rapid prediction of thermal properties using them. Thereby paving another step on the way to mapping out the materials' property space to eventually achieve the goal of inverse-materials design via rapid property prediction.

Chapter 5

Databases from Language Models

"Bailíonn brobh beart."

Natural language processing (NLP) is an immensely rapidly developing field within the domain of artificial intelligence (AI). The state-of-the-art techniques within the field are evolving quickly, with much of the focus of AI research being placed squarely on this domain. Indeed, even since this work was started, there has been several paradigm shifts within this subdomain of AI, all worthy of *attention*.

The first section of this chapter will outline the progress that has been made in applying these NLP techniques to the domain of materials science. It will delve into some of the historical developments of NLP methods and the challenges they were used to address in materials science, first in the pre-transformer era and, later, will discuss how the advent of transformers has impacted the field.

In the subsequent sections, the chapter will explore the efficacy of more modern techniques in rapidly constructing large-scale materials science databases of materials properties. It will highlight automated approaches that leverage contextual language models to extract and organize relevant information from diverse sources. These advancements have changed the landscape of the way materials databases are created entirely, enabling researchers to access comprehensive datasets efficiently, while eliminating the need for researchers to possess extensive knowledge of ultra-specific rules-based techniques in natural language processing.

The chapter will also discuss novel ways of evaluating the quality of these resultant databases alongside their ability to predict materials properties. It will explore metrics and methodologies used to assess the reliability and accuracy of predictions generated by NLP-driven materials databases. This section will shed light on the challenges and advancements in ensuring the robustness of these databases.

Following this discussion, the chapter will shift its focus to the use of large language models that have gained significant attention in recent years. It will examine how these models, such as GPT-3 and its successors, have the potential to revolutionize materials science research. Much of this chapter draws from a publication derived from this work, which can be found in Ref. [100]. It was also performed in close collaboration with my colleague Matteo Cobelli.

5.1 Early Use of Language Models In Materials Science

The automated construction of databases in physics and chemistry has long been a desired goal of the field of NLP for materials science. Early attempts at doing this were mainly based on grammar rules or dictionary approaches. Such approaches relied quite heavily upon the assumption that scientific information in journals, reports and patents was generally highly procedural and rarely deviated from very rigid grammatical and syntactic structures. These tools, such as ChemicalTagger [101], OPSIN [102] and early versions of OSCAR [103], were applied specifically to chemical synthesis procedures. These methods utilized regular expressions, a string-searching algorithm. Their purpose was to discover patterns within a text, which would then be compared to strings adhering to user-defined rules. These rules were composed in a representative language specifically designed for standardized rule representation. There was some use of simple classifier models for chemical named entity recognition (NER) at the time [104, 105], however, the main extraction procedures relied mainly on dictionary methods.

These methods were powerful in the field of synthetic chemistry due to the prevalence of International Union of Pure and Applied Chemistry (IUPAC) standards in the field, which meant that chemicals within synthesis procedures are intentionally written in an international standard to facilitate their reproduction and synthesis. This, in turn, means that such rules are far more difficult to apply beyond scientific regimes that do not employ such rigid standards in naming conventions, such as in the fields of inorganic chemistry and materials science, for example. A change in strategy was needed in order to account for the increased variance in the reporting of such compounds.

Various ML algorithms were explored to this end and success was found in utilising hidden Markov models (HMMs) [106], maximum entropy Markov models (MEMMs) [107] and conditional random fields (CRFs) [108, 109] in a hybrid approach with rules-based methods in order to maximise the ability of chemical-literature-extraction models to extract chemical entities reliably with a higher fidelity than before [103, 110]. Both the HMM and MEMM algorithms work on the assumption that the labels associated to a word or token in a sentence are a Markov chain, meaning that the label assigned to a token depends on the previous token. HMMs learn the transition probability between each label in the sequence, learned from previous tokens in a given training data set. In contrast, the MEMM algorithm is based on a multinomial logistic regression classifier that attempts to assign a label to each token in a sentence. This simple model is extended in the case of MEMMs by discarding the assumption that the labels for the tokens are conditionally independent of each other, but rather, making the assumption that they constitute a Markov chain.

CRFs also take this idea of associations between labels obeying a Markovian relationship i.e. the probability of a certain label being associated with a word is dependent only on its neighbours. However, in this case, the semantic structure is represented as an undirected graph. Usually, when applying this method to natural language processing, linear-chain CRF is used in which the graph of labels is structured into a chain, with each label representing a single, associated token.

These hybrid methods became state-of-the-art and databases that were constructed using chemical NER became more widespread, with ChemDataExtractor becoming the most dominant tool in this space [111]. Automated databases were constructed using CRFs and dictionary-based methods that spanned the topics of synthetic chemistry [112, 113], magnetic materials [114] and battery materials [115]. The extraction of properties associated with these extracted chemical entities, however, involved the tedious definition of dictionaries and rules that attempted to encapsulate every possible combination of grammatical and syntactic structures in which a reported value could appear in a text. Furthermore, this step had to be repeated for each individual property under examination. As a result, these methods did not become widely utilised and their use was focused to a select few research groups.

Further to this, the databases resulting from these works were seldom used outside of the work they were published in. This could be related to the difficulty in evaluating the quality of the resulting databases. Even with established metrics for evaluating machine learning performance, such as F_1 score, precision and recall, it remains very difficult to directly evaluate the databases resulting from automated curation using machine learning. Most of these works focused on evaluating the efficacy of the model using a small, handpicked, sample of the texts from which values had been extracted. The evaluation was performed by verifying the information extracted from this small sample of texts. Such an analysis clearly is insufficient to draw any robust conclusions for the totality of the extracted values with any statistical significance.

Despite these shortcomings, there was a clear indication from the few works that involved the automated construction of databases from hybrid language models that they had valuable use for materials discovery applications, once a critical mass of values had been collected.

Given the immense potential of these methods, this chapter focuses on attempts at addressing the main concerns with the established methods of automatically constructing large-scale materials science databases for machine-learning applications in materials informatics.

5.2 Transformers

Transformer architectures were initially proposed by Vaswani *et al.* in 2017 [116]. This architecture represented a development over recurrent and convolutional neural networks for natural language applications. The best performing of these generally incorporates what is known as an attention function [117–119]. The transformer architecture advanced on these methods by entirely dropping recurrence and convolutions (described in Section 2.1.3) and simply applying attention on its own to these applications.

5.2.1 Self-Attention

Self-attention is the key concept of transformer architectures and is a means of associating the relative importance a word has with other words in a sentence when applied to natural language applications. It was originally conceived as an attempt to mirror the cognitive function of attention to certain terms in an expression over others, a natural way in which to identify the importance of certain terms to the interpretation of a sentence.

In a more general sense, the attention function will take three vector quantities, a query and a key-value pair. These vectors are computed for each token according to a weight matrix, which is obtained during the training of the full transformer network for these three vectors. Each input token to a transformer will produce all of these three vectors, by multiplying the input embedding by the universal query, key and value weight matrices. The attention function will then map these quantities to an output, which is a weighted sum of the value, where each term of the value vector is weighted according to a compatibility function comparing the query with the key. For practical purposes, to aid efficiency in the computation of the outputs, these vectors are aggregated into matrices of queries $\mathbf{Q}$, keys $\mathbf{K}$ and values $\mathbf{V}$ such that the outputs for each combination can be computed simultaneously. In the original paper, the authors applied an attention function known as the scaled dot-product attention,

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\intercal}}{\sqrt{d_{\mathbf{k}}}}\right)\mathbf{V}, \tag{5.1}$$

where $d_{\mathbf{k}}$ is the dimensionality of the key and query vectors. The reciprocal of its squared root is used as a scaling factor in Eq. (5.1) in order to avoid potentially negative impacts for the model convergence that would result from using the dot-product attention without such scaling. The softmax function is the same as was defined in Eq. (2.29).

This attention function is computed in parallel with different dimensionalities for each of the query, key and value vectors, the outputs of which are concatenated and fed through a single, fully connected, feed-forward, neural network layer. Such a system allows the model to learn information from different representations in parallel, which can all be incorporated into the model simultaneously. This construction is known as a multi-head attention.

The potential of systems based entirely on self-attention, as opposed to ones that involved recurrence or convolutions, was realised by comparing several aspects of purely attention-based architectures to the existing alternatives. The first of the criteria of comparison is the computational efficiency of the operations and the ability for such methods to be parallelised. The computation of attention is based entirely on matrix operations and different attention heads can be computed independently from each other. This means they can be trained and deployed using parallelised GPUs, vastly improving their computational efficiency.

Another potential advantage of the use of self-attention in a transformer architecture is the interpretability of the resulting representation. Similar to a convolutional-style construction, different attention heads learn to perform different tasks and, similarly the attention weights of sentences inputted into a transformer model often capture and reflect the semantic structure of the sentences. This is a key advantage in systems that require a nuanced knowledge of the syntactic structure of a sentence in order to execute their task.

5.2.2 Transformer Architecture

The architecture for the original transformer network can be found in Fig. 5.1. Transformers expand upon this idea of multi-head self-attention by stacking multi-head attention layers on top of each other. These layers combine a multi-head self-attention layer with a fully connected feed-forwards network layer. Both constituent sublayers within the layer employ a residual connection in which the input to each sublayer is added to the output and the resulting vector is normalised. This step ensures that the positional encoding is preserved throughout the network.

For the original architecture, this structure is used as a component of two separate stacks, as seen in Fig. 5.1, constituting an encoder-decoder architecture. This is an ideal choice for a large variety of NLP applications, such as sequence-to-sequence modelling or classification. Each of these stacks is composed of N repeated attention/feed-forward layers ($N = 6$ in the original paper). The left-hand stack in Fig. 5.1 is the encoder and on the right-hand side is the decoder.

The encoder functions as a means of creating a representation of the input sequence

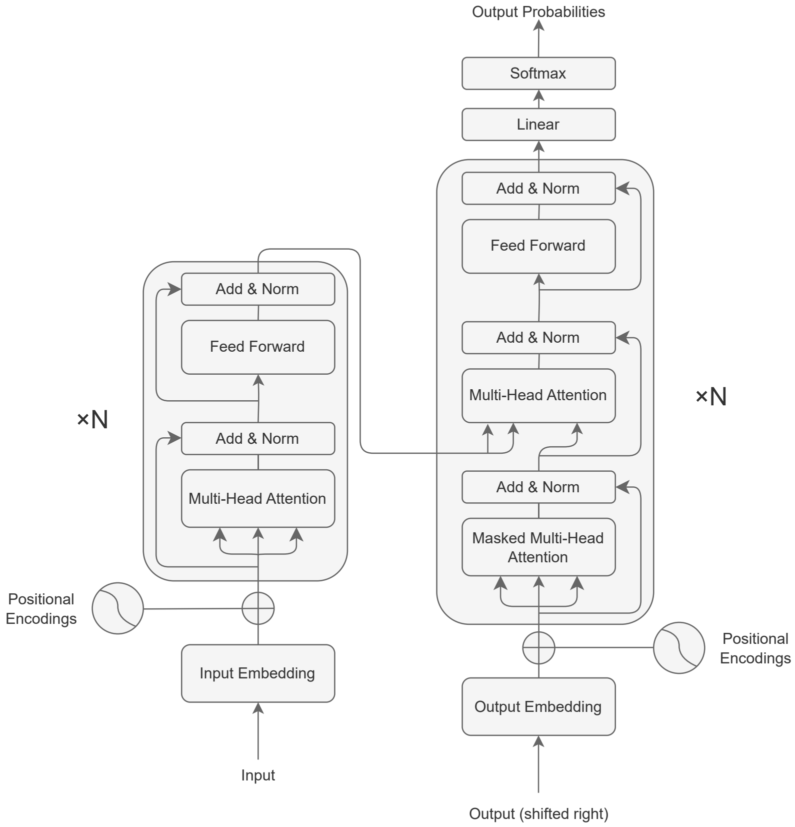

Figure 5.1: A flowchart of a typical transformer network architecture. On the left is the encoder stack and on the right is the decoder stack. The decoder takes as input the output from the previous pass to generate the output probabilities. *Adapted from Ref. [116].*

based on the stacking of the aforementioned attention/feed-forward sublayers. This creates a latent representation of the input that the decoder receives in combination with the decoder output of the previous pass in order to generate the output sequence. The decoder employs a further elaboration on the concept of self-attention, namely, masked multi-head attention.

Masked multi-head attention involves the modification of the self-attention mechanism in the sublayer to only attend to the preceding terms in the sequence as opposed to the totality of words that the encoder will attend to simultaneously. It does this by the introduction of a step called masking, in which all values after the specified position in the sequence are set to $-\infty$ before being input into the softmax function of Eq. (2.29) when computing the self-attention as in Eq. (5.1). Such a step enables the decoder layer to only consider the preceding words when making a prediction for a token. This is in order to ensure that the decoder is autoregressive, which means that the model output is dependent on its previous outputs.

5.2.3 BERT

For certain cases, the autoregressive feature of the decoder of transformer networks might not be optimal. In 2019, Devlin *et al.* [120] proposed a new language representation model known as BERT, standing for bidirectional encoder representations from transformers. This enabled the creation of natural language representations that were dependent on both left and right contexts throughout the architecture. BERT was proven to exhibit state-of-the-art performance on a wide range of tasks and quickly became the dominant model in the field upon its conception.

It considers both left and right contexts in this representation by changing the original masking paradigm described in section 5.2.2 by instead employing a masked language model pre-training objective. This strategy involves discarding the masking objective that enforces unidirectionality in the decoder proposed by Vaswani *et al.* [116] by masking all subsequent tokens in the calculation of the self-attention and, instead, randomly masks a selection of tokens from the input, which the model then tries to predict. This is the only deviation from the training of the original transformer architecture as shown in Fig. 5.1.

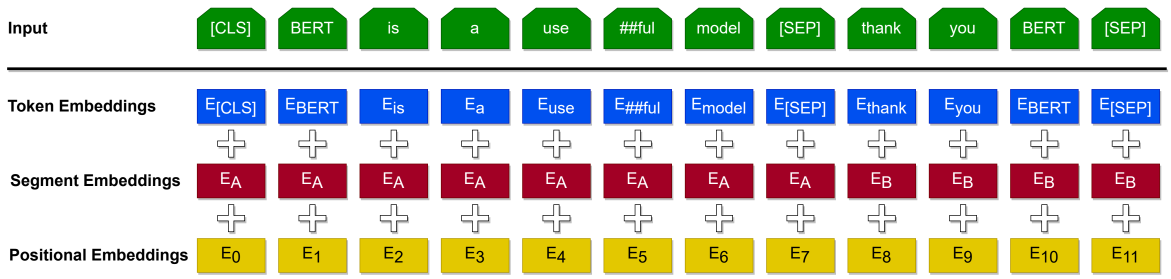

Figure 5.2: A diagram of the input representations of BERT models, which are composed of a summation of the token embedding, segment embeddings and positional embeddings. *Adapted from Ref. [120].*

The input representations for the BERT architecture (see Fig. 5.2) include some interesting innovations over the original transformer representations. Namely, the addition of two special tokens for the input embedding, the `[CLS]` and `[SEP]` tokens. The `[CLS]` token is always the first token in every sequence and corresponds to a classification token. The final hidden state of this token after being processed by the network is the aggregate representation of the whole sequence for sequence classification tasks. In contrast to this, the [SEP] token corresponds to a separator that delineates the point at which one sentence ends and the next begins. This is used extensively for tasks involving sequences of sentences such as next-sentence prediction and natural language inference.

5.2.4 GPT

Generative pre-trained transformer (GPT) models are the main alternative form of language models and are currently the main focus of development in the domain of general NLP techniques. They are commonly seen as being the key to general and conversational AI.

The original GPT model or GPT-1 [121], employed only the decoder stack of the transformer architecture (Fig. 5.1), using $N = 12$ for the number of stacked sublayers. The research behind this model revealed that training an autoregressive model to predict the next sentence based on a vast text database resulted in a pre-trained model that enhances predictive capabilities in discriminative tasks such as question answering and sentiment analysis.

Soon after this work, GPT-2 [122] was trained as a larger version of the original architecture, now with 1.5 billion tunable parameters, in contrast with the 117 million parameters of GPT-1 and the 340 million parameters of the largest BERT model [120], some small modifications to the sublayer structure and an increase in the size of the model vocabulary. This larger GPT model was proven to be proficient at zero-shot task transfer. Zero-shot task transfer concerns the ability of a model with sufficient capacity to learn and infer how to perform tasks in natural language, by learning how to predict the most likely next sentence in a vast corpus of natural text. This work by Radford *et al.* [122] demonstrated that the higher-capacity GPT model could learn how to perform a large variety of tasks that it was not trained for in an unsupervised manner simply from training the large model autoregressively on next sentence prediction, a key step on the road to general AI.

The next GPT version, GPT-3 [123] scaled this effort up to 175 billion tunable parameters and, again, made some minor adjustments to the architecture by refining the self-attention mechanism. This model performed very well across the board at adapting to generalised tasks using the few-shot learning paradigm. Few-shot learning, in contrast to zero-shot, involves prompting the model with few samples of the cases similar to the desired task as the model input, thus allowing the model to learn the desired output from the select samples in a carefully chosen input, as opposed to a full model fine-tuning or pre-training. There were tasks for which this system matched the performance of the state-of-the-art of fine-tuned systems when employing this few-shot technique for training.

There are further developments in this field with the advent of GPT-3.5 and GPT-4 [124], however, these models are proprietary in nature and therefore their architectures and robustly benchmarked performances are not available.

One crucial aspect to note regarding the utilization of GPT-based, autoregressive, large language models (LLMs) is that their primary limitation stems from their focus on optimizing predictions for the most probable token following a sequence of tokens. Consequently, these models exhibit a significant and concerning tendency to generate hallucinations in their output. For LLMs, hallucination is the tendency of these models to make up information based on the likeliest syntactic structure of the continuation of the given prompt. While these models do demonstrate a remarkable performance in many more specialised domains, such as science and medicine, outside of the general data that they are trained on, the information produced by them is often inaccurate [125, 126].

5.2.5 Transformer Models in Materials Science

With the growth of interest in transformer models, there has been an accompanied focus on utilising their superior natural-language performance for every domain. Materials science was no exception to this trend. Various attempts have been made since Ref. [116] to incorporate transformer-based systems into the toolkit of materials science.

The first of these attempts came in the form of BatteryBERT [115], a pipeline, which facilitates the automated extraction of compound-property relationships from unstructured battery-specific scientific texts. A total of six BERT models were pre-trained on a corpus of approximately 400,000 papers relating to battery research. These models were employed to enhance an existing rule-based pipeline for property extraction based on ChemDataExtractor by fine-tuning them for paper classification and question-answering. The paper classification step was used to improve on a filtering step to isolate relevant documents, that previously was performed with a combination of a term frequency-inverse document frequency (TF-IDF) document representation and logistic regression. The model was also fine-tuned to perform extractive question answering through the addition of a span classification layer on top of the BatteryBERT model. This effectively allows the model to supplement the records extracted using the rules-based technique, with the device function of each extracted material, namely the cathode material, the anode material and the electrolyte.

Another of the first explorations of the use of transformer models in the materials science domain is the work of Trewartha *et al.* [127] in which the advantage of domain-specific pre-training for transformer models was quantified for the first time. Interestingly, this paper found that the general BERT model was outperformed by every domain-specific, fine-tuned model on domain-specific tasks. Furthermore, it was found that a simpler system using a bidirectional long-short-term-model (BiLSTM) also outperformed the original BERT model on domain-specific tasks, clearly demonstrating the importance of using domain-specific pre-trained models for high-fidelity performance in non-conventional domains. This boost in performance was also demonstrated by using progressively more refined models in the materials science domain, with a model pre-trained for general science applications (SciBERT [128]) outperforming the BERT-base model and a model pre-trained for materials science (MatBERT [127]) further improving on this leap in performance.

Further to this work, another BERT model was constructed for materials science by taking the pre-trained weights of SciBERT and continuing this training, using a materials science-specific corpus of 150,000 documents, which were chosen from a variety of sub-domains of the field of materials science [129]. It also utilised the training strategy of RoBERTa [130], which applied a 15% dynamic whole word masking during the training. This new model was called MatSciBERT and demonstrated, once again, a superior ability to both SciBERT and BERT-base on downstream tasks specific to materials science.

A separate model was fine-tuned for the NER of polymer information by continuation of the pre-training of the weights of PubMedBERT, a BERT model pre-trained from scratch from literature taken from the PubMed database [131]. This model, named MaterialsBERT was then used to extract a database of approximately 300,000 polymer property records from a corpus of 650,000 abstracts. An attempt at benchmarking the quality of the extracted database was performed by analysing co-relation trends between key properties, which generally followed the expected trends.

The final attempt at integrating transformer models into a materials design or property extraction pipeline as of writing is the first attempt at exploring the use of fine-tuning a GPT-based LLM for materials property extraction [132]. The model produced was trained on three tasks, namely, linking dopants with their associated host materials, cataloguing metal-organic frameworks and general chemistry information extraction. There was an attempt at evaluating the produced output, however, generative models generally struggle to produce a systematic, consistent output and therefore, the authors of this paper defaulted to evaluating the results using the oversight of domain experts.

5.3 Transformer Models for Database Construction

There is a wide array of potential applications for databases of materials properties, some of which have already been demonstrated. Indeed, such databases have been used already to construct models, which are able to identify regions of materials composition space favourable to superconductivity [133], to design high-entropy alloys [134], to predict the existence of novel magnets [135] and to predict the zT thermoelectric figure of merit in inorganic materials [136], to name a few examples.

The vast majority of experimental data for materials science remains entirely locked away in literature. Most databases in this field are limited to calculated materials properties [37, 67, 137, 138], either introducing a systematic error in data in many cases or limiting the potential quantities contained within these databases to quantities amenable to rapid calculations. There is also the possibility that such calculations may also not reflect reality. This is, therefore, clearly sub-optimal. Most attempts at constructing experimental databases are proprietary in nature due to the involved labour costs [139, 140], despite some attempts at database construction by some open-access initiatives [141, 142]. This leaves an opportunity to delve into the development of workflows aimed at enhancing the accessibility and performance of models for the automated construction of materials databases from literature, making it a fertile area for exploration.

Previous attempts at utilising natural language models and transformers to extract information from unstructured literature, which are described in sections 5.1 and 5.2, highlight the strengths and weaknesses of such models. These must be addressed in order to construct a robust, high-fidelity and efficient information extraction system.

It is clear that transformers represent the current state-of-the-art with regard to general performance in natural language tasks, largely deprecating previous architectures. They also largely bypass the need for intricate syntactic and grammar rules that reduce the accessibility of such NLP systems to a large community. Domain-specific pre-training and fine-tuning is clearly demonstrated to enhance performance in domain-specific tasks [127] and, therefore, any pipeline constructed using transformers should rely on models optimised for the relevant domain. Furthermore, GPT-based systems have a tendency to generate inaccurate information, making output consistency a challenge. The proprietary nature of the most powerful systems also leads to a lack of transparency regarding privacy and model architecture. Additionally, performing domain-specific fine-tuning can be difficult. All these factors imply that there is much work to be done in order to establish a high-fidelity workflow for implementing these models. In Section 5.5, we will delve into the efforts aimed at addressing these challenges. This leaves a workflow based on fine-tuned BERT models the clear choice for executing a rule-definition-free workflow for precise information extraction from scientific literature. The remaining part of this chapter will focus on the work conducted as part of this thesis to implement such a workflow, which we call BERT for Precise Scientific Information Extraction (BERT-PSIE), along with a full evaluation of the resulting databases.

5.3.1 BERT-PSIE: Precise Scientific Information Extraction

Fig. 5.3 displays the workflow of our BERT pipeline for precise information extraction from scientific literature. This workflow aims to systematically construct databases of materials-properties relationships in an automated fashion from unstructured scientific literature, using a sequence of BERT models fine-tuned on downstream tasks. All BERT models used in the pipeline are derived from the weights of the MatSciBERT pre-trained model.

This workflow is shown to work for binary relationships and, in theory, there is no

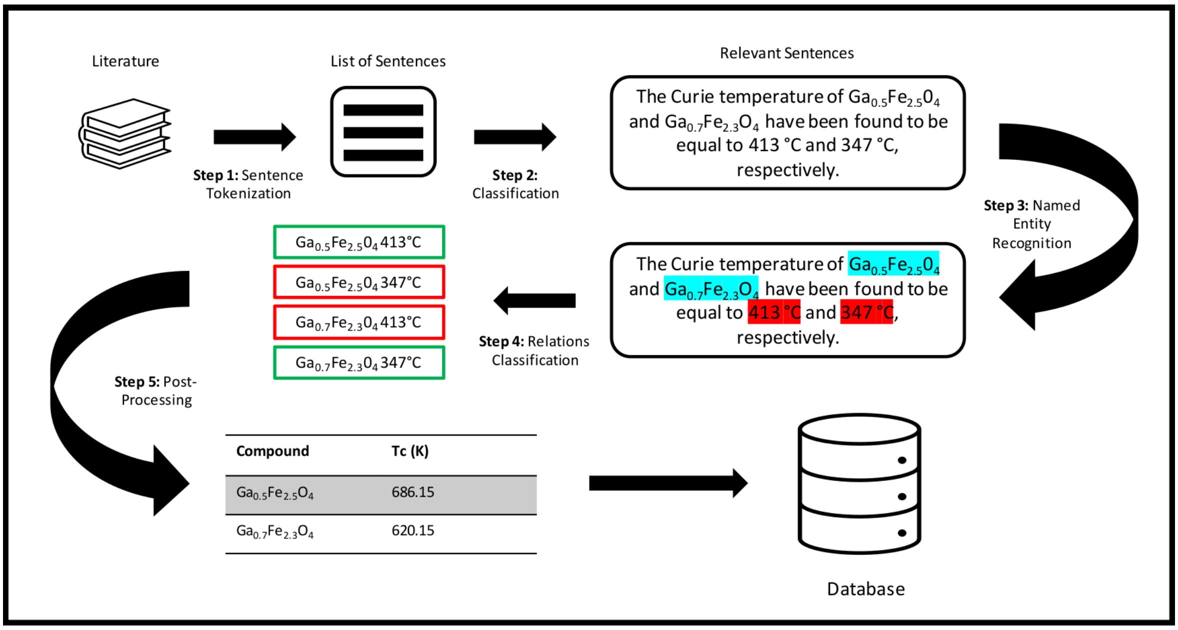
Literature
List of Sentences
Relevant Sentences
Step 1: Sentence Tokenization
Step 2: Classification
The Curie temperature of $Ga_{0.5}Fe_{2.5}0_4$ and $Ga_{0.7}Fe_{2.3}O_4$ have been found to be equal to 413 °C and 347 °C, respectively.
Step 3: Named Entity Recognition
The Curie temperature of $Ga_{0.5}Fe_{2.5}0_4$ and $Ga_{0.7}Fe_{2.3}O_4$ have been found to be equal to 413 °C and 347 °C, respectively.
Step 4: Relations Classification
$Ga_{0.5}Fe_{2.5}0_4$ 413°C
$Ga_{0.5}Fe_{2.5}0_4$ 347°C
$Ga_{0.7}Fe_{2.3}0_4$ 413°C
$Ga_{0.7}Fe_{2.3}0_4$ 347°C
Step 5: Post-Processing
Compound
Tc (K)
$Ga_{0.5}Fe_{2.5}O_4$
686.15
$Ga_{0.7}Fe_{2.3}O_4$
620.15
Database

Figure 5.3: A diagram of the BERT-PSIE workflow for the automated extraction of scientific information from unstructured literature, developed in this thesis. The workflow relies on a combination of BERT models trained on downstream tasks such as sentence classification, named entity recognition and relationship classification.

limit to the number of inter-related properties that it can be extended to.

Gathering a corpus of scientific papers for the training step is a necessary first step in the extraction pipeline. For the case of the construction of a database of Curie temperature values, the Crossref REST API is employed to execute a keyword search for the term ‘Curie temperature’ over all open-access literature published by Elsevier. This returns the metadata for these papers with the full text available, yielding approximately 180,000 papers. Elsevier is chosen to extract from for the abundance of literature available in an HTML format as opposed to other publishers, which offer access principally to PDF documents. PDF documents are generally less efficient and less consistent after parsing, whereas the consistency of the format of standard HTML documents makes them ideal candidates for data scraping. Of the returned metadata, 800 abstracts are manually annotated by employing the sentence tokenizer of the natural language toolkit (NLTK) [143], which separates these abstracts into their constituent sentences. The resulting, separated sentences are split into two categories, sentences that contain a Curie temperature and those that do not. This step yields a database of approximately 4,000 sentences, of which 189 are deemed relevant. All positions of compound and Curie temperature mentions in this collection of sentences are also labelled. This dataset is subsequently employed to fine-tune a BERT classifier model to find relevant sentences in a full corpus of papers extracted using the full-text links from the metadata from the CrossRef API.

A similar workflow is employed for the extraction of relevant sentences to the electronic band gap, with the difference being the source of the abstracts for training. In the case of the band gap, the metadata for the Cornell University arXiv pre-print repository is downloaded in its entirety from the Kaggle dataset [144]. A selection of 1,000 abstracts is taken from this dataset by searching the abstracts for the terms 'band gap,' 'band-gap,' or 'bandgap.' Abstracts containing a band-gap reference are tokenized into their constituent sentences, facilitating a database of 672 sentences, with 171 sentences within this database considered to be relevant.

The positions of all mentions of compounds and properties are labelled in a selection of 200 sentences deemed relevant by the classifier, as above. The new sentences are combined with the annotated abstracts, and this collection of annotated examples is used to train a BERT model fine-tuned to be a token-level named entity recognition (NER) system. Such a system is essentially a token level classifier, which classifies each input token as being either a `CHEM`, `TEMP`, or neither, with the `CHEM` token corresponding to a chemical entity mention and the `TEMP` token corresponding to a Curie temperature mention.

The CrossRef API is once again employed to download a corpus of papers for extraction based on a keyword search for the terms 'magnetic' and 'electronic' to construct a corpus of papers likely to contain a Curie temperature and electronic band gap, respectively. These corpora are comprised of approximately 180,000 papers for the magnetic database and approximately 77,000 papers for the electronic corpus. After the sentence classifier is run on these sets, a database of sentences is yielded for each quantity being extracted. These relevant sentence databases contain 55,000 and 126,000 sentences for the Curie temperature and electronic band gap, respectively.

The final task in the sequence of BERT models in the BERT-PSIE architecture is that of relationship classification. Firstly, the databases of sentences are run through the NER model to identify mentions of compounds and their associated properties. If a singular property and a singular compound are mentioned in a sentence, these are assumed to be related quantities and are added to the final database (step 5 of

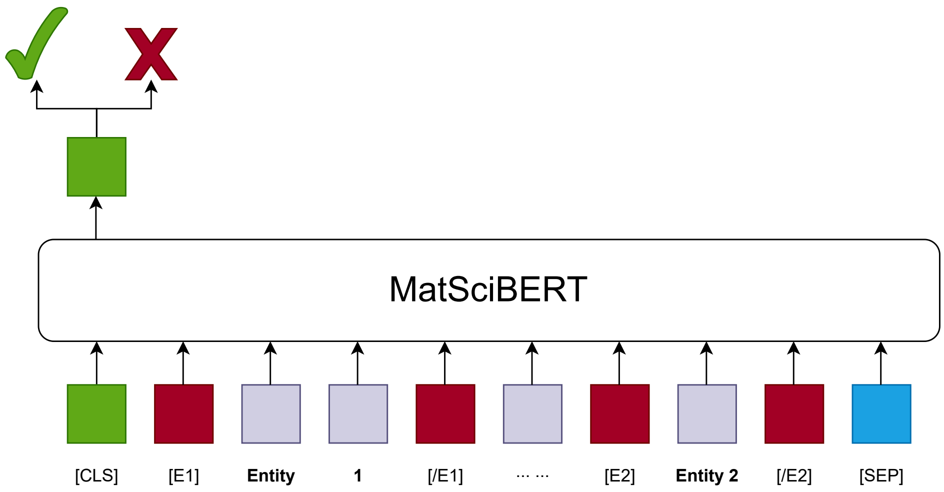

Figure 5.4: A diagram of the variant of the relationship classification strategy used in this work, derived from the work of Soares *et al.* [145]. Compound and property entities in a sentence are marked with `[E1]` and `[E2]` tags before being converted into MatSciBERT representations for the binary classification, which determines if the tagged entities are the related ones or not, a classification depicted here with the check mark or the X.

Fig. 5.3). For the more complex cases where multiple compounds and/or properties appear in the same sentence, there is an ambiguity due to any one of all of the possible compound/property associations. This ambiguity is certainly simple to deal with for a human reader; however, it is not a trivial problem to solve in an NLP pipeline. Here, the resolution of these cases is dealt with by adopting a simple classification scheme developed by Soares *et al.* [145]. More specifically, a BERT architecture is trained to classify whether sentences have a property relationship or not, after the addition of entity marker tags.

This scheme is visualised in Fig. 5.4. To take the example seen in step 4 of Fig. 5.3, the sentence: 'The Curie temperature of $Ga_{0.5}Fe_{2.5}O_4$ and $Ga_{0.7}Fe_{2.3}O_4$ have been found to be equal to 413°C and 347°C, respectively.' contains the pre-described ambiguity. From this sentence, four associations are made as follows: 1) $Ga_{0.5}Fe_{2.5}O_4$ and 413°C, 2) $Ga_{0.5}Fe_{2.5}O_4$ and 347°C, 3) $Ga_{0.7}Fe_{2.3}O_4$ and 413°C, 4) $Ga_{0.7}Fe_{2.3}O_4$ and 347°C. Each of these combinations is marked within the sentence using the tags `[E1`$_{\text{start}}$`]`, `[E1`$_{\text{end}}$`]` to identify the compound with `[E2`$_{\text{start}}$`]` and `[E2`$_{\text{end}}$`]` delineating the start and end point of the property mentions. After this step, the marked sentences are classified as positive or negative depending on whether the `[E1]` and `[E2]` tags are correctly associating the compound/property relationships.

Model Evaluations

Table 5.1: Performance of the three modules developed for the Curie temperature extraction: the sentence-level relevancy classifier, the NER and the relation classifier. Results are presented for the test sets. Here we report: precision, P, recall, R, and F_1 score. The size of the test (TeS) and training (TrS) sets are also given (number of sentences used). For the case of NER, we report results for both chemical entities (Chem) and T_C, as well as the support.

Model	Entity	P	R	F_1	Support	TrS	TeS
Classifier		0.77	0.82	0.79		3941	801
NER	Chem	0.92	0.86	0.89	754	1,769	168
	T_C	0.97	0.81	0.88	42		
Relation		0.72	0.64	0.68		200	50

In this section, the results for the constituent models of the BERT-PSIE pipeline are presented. Considering first the Curie temperature pipeline, the results for each of the constituent models are visible in Table 5.1.

The first row in Table 5.1 presents the standard classification evaluation metrics for the sentence relevancy classification, the precision P, recall R and the F_1 score, as defined in Section 2.1.1. Each of these metrics is approximately 0.8, indicating a high-level model performance on the test data. Such high values are compounded by the model accuracy of 98.83%. This would potentially be higher, but unfortunately, there is a certain amount of ambiguity in the classification of Curie temperature sentences. In fact, there are a large number of critical phase transition descriptions that use very similar syntactic structures to report the temperature at which they occur. This is, unfortunately, impossible to address meaningfully when operating at a sentence level. An example of such a sentence structure that may present this difficulty could be ‘The critical temperature of Nb is 9.2 K.’ In this instance, the Curie temperature is unlikely to be the value that the ‘critical temperature’ refers to. The sentence-level classifier, however, is quite likely to select such a sentence as being likely to be referencing a Curie temperature and, therefore, introduce noise into the resulting database. An attempt at mitigating the effect of this issue is obtained by selectively scraping data from texts exclusively relevant to the subdomain of magnetism.

The second row of Table 5.1 concerns the performance of the NER system on the test set, constituting step 3 of the pipeline described in Fig. 5.3. The precision, recall and F_1 score are all consistently high for both token-level labelled entities, namely, the Curie temperature (TEMP) and the chemical compound (CHEM). The high performance of both classifiers implies that these two entities can be recognised and extracted from plain text in one step, using this model. It does, however, suffer from the same drawback mentioned above. Specifically, the NER system will also struggle slightly at differentiating different types of critical temperatures due to the similar syntactic and grammatical contexts in which they appear. Despite this, the NER system can, in fact, differentiate between critical temperature mentions and mentions of other temperatures that do not involve a critical phase transition temperature. These factors demonstrate its excellent performance in identifying and extracting the desired quantity from relevant sentences in relevant literature.

The final row in Table 5.1 indicates the evaluation metrics for the case of the relationship extraction step in the workflow (step 4 in Fig. 5.3). As is clear from the evaluation metrics, relationship classification presented itself as the most challenging task of the workflow. This issue, however, is not unique to the case of relationship resolution based entirely on ML methods. The ambiguity in relationship resolution for the equivalent rule-based methods is equally, if not more, challenging. The total syntactic variance of every construction that includes multiple compound/property mentions must be captured by the rules defined by your rules-based method and, therefore, performance is highly dependent on the user's ability to define the rules. This is contrasted with purely ML-based methods, which allow for some syntactic variance in such relationships naturally due to their superior contextual awareness from the self-attention mechanism.

Indeed, the relationship classifier exhibits the lowest scores of the workflow, however, this is not entirely unexpected. A fact which is the result of the potential ambiguity of compound/property relationships, with some relationships being misleading, e.g. 'The compound exhibits a Curie temperature of 1000 K, similar to that of iron.' In this example, the Curie temperature does not actually refer to the compound which appears in the sentence, introducing a source of ambiguity. This issue is compounded with that of dopant and other material mentions unrelated to the binary relationship between the compound and the property. In spite of the complexity of this task, however, the classifier performance is still very good and adequate for enhancing an information extraction pipeline, thereby improving the final database.

Given that all constituent models in the workflow are very high-performing, a pipeline was constructed in order to automate the extraction of a database of materials properties and their associated compounds. Approximately 180,000 unstructured scientific texts were downloaded. After running the full-text papers through the sentence relevancy classification step, a database of 55,000 relevant sentences was obtained. This list of sentences was run through the NER model. Single mentions of Curie temperature and compounds were associated together when they appeared in sentences together. When the sentence contains multiple entities, a list of sentences with all possible constructions of pairs of compound and property entities is created. These resulting lists of sentences, with the tagged combinations of entities, are processed by the relationship classification model. The resulting, associated entities, predicted as being correct are added to the database. After post-processing, involving converting temperature units to Kelvin and scaling the chemical compositions to have reduced integer coefficients, e.g. $Ga_{0.5}Fe_{2.5}O_4$ became $GaFe_5O_8$, the database contained 3,518 distinct compound-property entries with their digital object identifiers (DOI).

Table 5.2: Performance of the three modules developed for the band-gap extraction: the sentence-level relevancy classifier, the NER and the relation classifier. Results are presented for the test sets. Here we report: precision, P, recall, R, and F_1 score. The size of the test (TeS) and training (TrS) sets are also given (number of sentences used). For the case of NER we report results for both chemical entities (Chem) and band gap, as well as the support.

Model	Entity	P	R	F_1	Support	TrS	TeS
Classifier		0.95	1.00	0.97		404	134
NER	Chem	0.80	0.96	0.87	1166	4000	1000
	band gap	0.78	0.97	0.87	119		
Relation		0.88	0.88	0.88		300	80

This same pipeline was executed once again for the case of the electronic band gap. Firstly, the performance metrics of the sentence-level relevancy classifier are presented in the first row of Table 5.2. This classifier significantly outperforms the classifier for the Curie temperature, with a perfect recall and a slightly lower precision of 0.95. These metrics indicate almost perfect performance on the test set in the ability to differentiate between sentences, which contain and do not contain information about the electronic band gap. This is most likely attributed to the reduction in ambiguity between sentences that contain band gaps and those that do not. In the case of the

Curie temperature, the model appears to have learned the concept of critical temperature in its differentiation of sentences, however, in the case of the band gap, there are far fewer constructions that share the same sort of syntactic similarity. Similarly, the NER model remains very performant on the test set data with only minor deviations from the range of values presented in Table 5.1.

In the case of the relationship classification task, it was found that the model outperformed the Curie temperature relationship model. This is believed to be related, once again to the relatively more consistent nature of reporting band-gap values when compared to the variability of descriptions of ferromagnetic phase transitions. This is also probably related to the aforementioned ambiguity in syntactic structure that accompanies Curie temperatures and other critical transition temperatures, which is not a feature of descriptions of electronic band-gap values. Such descriptions may have a tendency to be more formulaic and this feature may explain the improvement in model performance between different quantities.

Approximately 77,000 papers were downloaded for the band-gap extraction and tokenized into lists of sentences. Using the high-performing sentence-level relevancy classifier, a database of approximately 126,000 sentences deemed likely to contain a band-gap value was constructed. After the NER and post-processing steps were performed, once again normalizing the compound formulae and scaling all units to eV, a database of 2,090 unique compound/property relationships was obtained.

5.4 Evaluation of Resulting Databases

Evaluation metrics in machine learning such as precision, recall and F_1 score are often a very effective way of indicating how performant ML models are. However, there are some weaknesses of such methods that inhibit them from being useful when evaluating the efficacy of a full workflow of concatenated models. Thus, the conventional metrics on model performance only give a limited indication of workflow performance and, therefore, database quality. Furthermore, they only reveal the performance of individual, constituent processes in a data-mining workflow, without giving information about the efficacy of the entire sequence.

A better test would be a direct comparison of the resulting database quality with some form of an expected result, thereby confirming that the data in the automated database is of a similar quality to that of a database that could be considered a ground truth. With this in mind, in this section, the databases extracted using the workflow described in section 5.3.1 are compared with databases for the same compound/property relationships that have been curated manually. As previously discussed, the BERT-PSIE pipeline can, in theory, be extended to dependencies that stretch beyond the binary relationships, such as ternary or quaternary relationships between materials properties. However, there is a relative dearth of available manual databases containing such dependencies and, therefore, this evaluation is limited to binary relationships. As a result, the two properties that will be the focus of this comparison and evaluation are the Curie temperature (Section 5.4.1) and the electronic band gap (Section 5.4.2). An analysis will also be performed by comparing and contrasting BERT-PSIE with the state-of-the-art, rules-based methods for information extraction, namely, ChemDataExtractor.

5.4.1 Curie Temperature

For the evaluation of the Curie temperature database that resulted from the use of the BERT-PSIE pipeline, a combination of data from various sources was utilised. One of the primary sources of manually curated data was from the work of Nelson *et al.* [11], which was created by aggregating the *AtomWork* database [146], *Springer Materials* [147], the *Handbook of Magnetic Materials* [148], as well as the book, *Magnetism and Magnetic Materials* [149]. This database was subsequently combined with the database aggregated as a result of the work of Byland *et al.* [150], a dataset primarily focused on Co-containing compounds. The combined database amounted to a total of 3,638 unique ferromagnetic compounds and their associated Curie temperatures. This is taken as our ground truth for a Curie temperature database.

The results of the BERT-PSIE pipeline are also compared with those obtained using the semi-supervised Snowball algorithm coupled with a rules-based scheme [114]. The BERT-PSIE database and the ChemDataExtractor database both extract data from very similar sources in that they both employ the CrossRef metadata search and both were extracted quite recently relative to each other. In contrast, the aggregated manual dataset is largely based on data extracted from dense tables and contains relatively older data, with some results having gone back as far as the 1950s. While the two automatically extracted datasets arise mostly from similar sources, and they both contain several thousand data points each, there is a remarkably limited overlap between the two of them of merely 694 compounds. Both automatically extracted datasets have a similar magnitude of overlap with the manually curated cases, i.e. 687 for BERT-PSIE vs. manually curated and 595 for ChemDataExtractor vs. manually curated. Between all three datasets, there is only an overlap of 262 compounds. All comparisons done in this section are performed by exclusively taking the median value of the Curie temperature for compounds that have multiple associated Curie temperatures.

It should be noted that a significant source of error could arise in this analysis from the NER model of the workflow due to the difficulty that it may have in differentiating between elemental compounds and the same element when used as a dopant (e.g. bulk Mn vs. Mn-doped GaAs). Indeed, for the case of assigning a Curie temperature to an elemental compound using the BERT-PSIE workflow, it was observed that the error increased significantly. This is because dopants can appear in a multitude of concentrations and in a wide array of hosts, meaning these can be assigned to a vast multitude of different Curie temperature values. Despite this, as can be seen in the top panel of Fig. 5.5, there is generally an excellent agreement between the expected distribution of Curie temperatures, with the most significant peak in every distribution being expectedly close to absolute zero as a result of most being non-magnetic in nature.

Notably, there is a particularly strong agreement in the structure of the Curie temperature distribution between the two automatically extracted databases resulting from BERT-PSIE and ChemDataExtractor, with both exhibiting a strong peak in the distribution around room temperature, a feature which is not present in the manually curated dataset. There are two potential reasons for this artefact: 1) there is either a bias in the most recent literature in favour of critical temperatures close to the room temperature value of 300 K or 2) model errors aggregate T_{C} values close to ambient temperature. The latter hypothesis makes sense as references to ambient temperatures may feature heavily in sentences that contain the targeted property. For example, the sentence ‘The magnetization curve at 300 K was obtained and the Curie temperature was determined by TGA under a magnetic field, yielding a Curie temperature of 1043 K for Fe.’ mentions an experimental measurement at ambient temperature as well as the

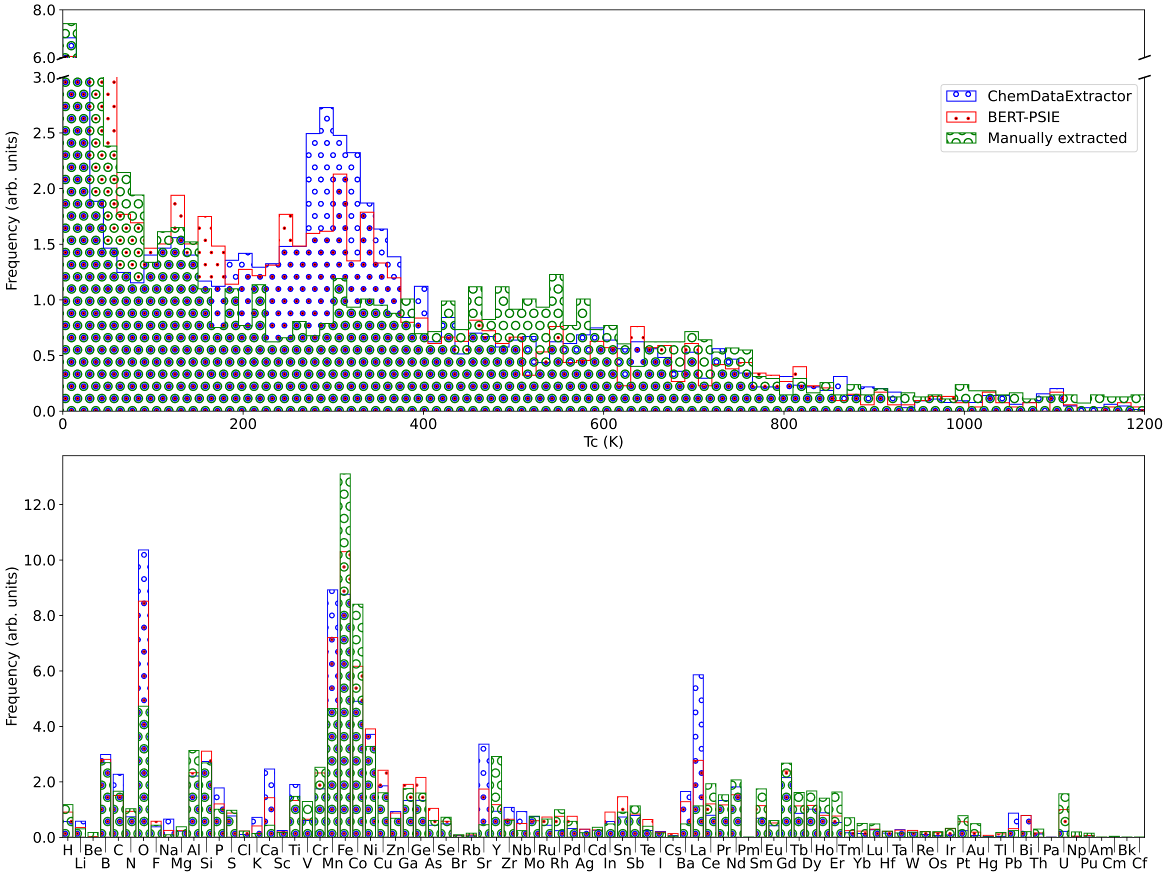

Figure 5.5: Comparison between the content of the different databases: *(red box)* BERT-PSIE, *(blue box)* ChemDataExtractor and *(green box)* the manually extracted database of Ref. [11] and Ref. [150]. Top panel: Normalized distribution of the Curie temperatures extracted. A peak is visible in the distribution around 300 K in both the autonomously extracted databases, which is not seen in the manually extracted one. Bottom panel: Relative elemental abundance across the compounds present in a database. Although there is general agreement among the three databases, additional peaks are observed for various elements in the case of automatically extracted data, which are not present in the manually curated dataset. The most severe of these discrepancies is in the relative abundance of Mn- and O-containing compounds. Note that the automatically extracted datasets and the manually curated one are based on different literature libraries, with the manually curated case arising from older reference books, Ref. [146–149], whereas the BERT-PSIE-extracted case came from scraping the Elsevier API.

desired Curie temperature. While the high performance of the relationship resolution and NER steps of BERT-PSIE may mitigate much of the influence of such noise in the database, it is almost impossible to entirely negate the effect of such cases.

Despite the aforementioned drawbacks, it is clear from the top panel of Fig. 5.5 that the automatically extracted database from BERT-PSIE adequately captures the relative abundance of high- and low-temperature ferromagnetic materials when compared to the expected distribution from the manually curated dataset.

A further understanding of the structure of the resulting database can be obtained by looking at the bottom panel of Fig. 5.5. This plot displays the relative elemental abundance of each unique compound in our database, i.e. the frequency at which a certain element in a compound appears in the database. As I expected, the largest abundances within all databases are found for the magnetic transition metals (Fe, Ni etc.), certain rare earths and oxygen. Interestingly, the automated databases have a tendency to overestimate the abundance of Mn and O, beyond the expected amount from the manually curated dataset, as well as that of di- and tri-valent alkali metals (Ca, Ba, Sr and La). This overestimation with respect to the manually extracted dataset is significantly more pronounced in the ChemDataExtractor dataset than for the data obtained using BERT-PSIE. This overestimation is likely due to the primary data sources used for the extraction. The data sources used for both BERT-PSIE and ChemDataExtractor are significantly more recent than the sources used for the manual curation and therefore this change in abundance is likely attributable to the shift in research focus over the years to novel structures and exotic magnetism. In particular, the overabundance of Ca-, Ba-, Sr- and La-containing compounds may be due to the significant focus being placed on perovskites, such as manganites, currently within the field.

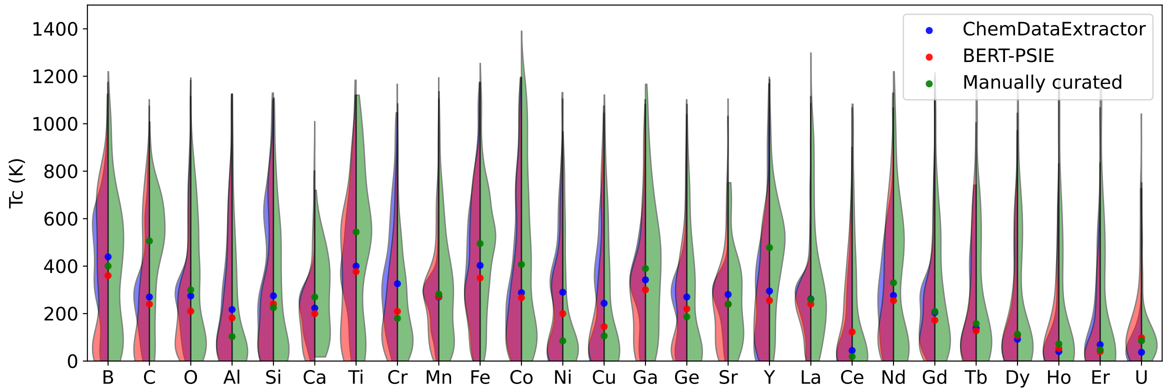

Figure 5.6: Violin plots comparing the T_C distribution of the compounds containing specific elements in the dataset automatically generated with BERT-PSIE *(red)* and ChemDataExtractor *(blue)*, and in the manually curated ground-truth *(green)*. Only the most common elements appearing in the datasets are displayed here. The dots show the median of each distribution.

The influence of primary data source on the distribution of data in the final dataset is further confirmed by looking at the T_C distributions of compounds containing the 25 most common elements in each of the three datasets, which are presented in Fig. 5.6. There is generally a very good agreement between the distributions and, in particular, between the two automatically extracted distributions, containing entries that were extracted from similar sources. BERT-PSIE does generally capture a similar distribution to the expected one from the manually extracted values. However, there are some discrepancies for certain elements. This is likely indicative of the aforementioned historical variance in research focus between the sources that were used for the ground truth when compared with the more modern, automatically extracted cases.

Table 5.3: Performance comparison between the different datasets against the manually curated one from Ref. [11] and Ref. [150]. The left-hand side of the table refers to the query test, while the right-hand side refers to the RF T_C predictor. Together with the BERT-PSIE and ChemDataExtractor databases we also consider different BERT-assembled datasets obtained by using different relation-classification strategies (see details in the text). The query benchmark is done over the 262 compounds that are shared by all the datasets, while the RF obtained is done over 2,623 compounds that are not present in any of the automatically collated datasets. Values for the best-performing datasets are in bold.

	Entries	Query			RF Predictions		
		R^2	MAE (K)	RMSE (K)	R^2	MAE (K)	RMSE (K)
ChemDataExtractor	4,289	0.78	**48**	137	0.65	**123**	176
This work							
Single Mentions	1,858	0.77	51	139	**0.66**	128	**174**
Order of Appearance	2,682	0.77	51	141	0.65	126	176
All Combinations	4,308	**0.81**	52	127	0.61	134	184
BERT-PSIE	3,518	**0.81**	50	**126**	0.65	126	**174**
BERT-PSIE + ChemDataExtractor	7,052	0.86	38	109	0.69	118	165

After confirming the capability of capturing the data distribution of a manually curated database with automated methods, using methods that do not rely on the labour-intensive definition of grammar rules, it is useful to measure the performance of using an automated database on downstream tasks to aid in the materials-design pipeline. Ultimately, the quality of a materials database can only be measured by its predictive power and its capability to return an adequate estimate of the correct value

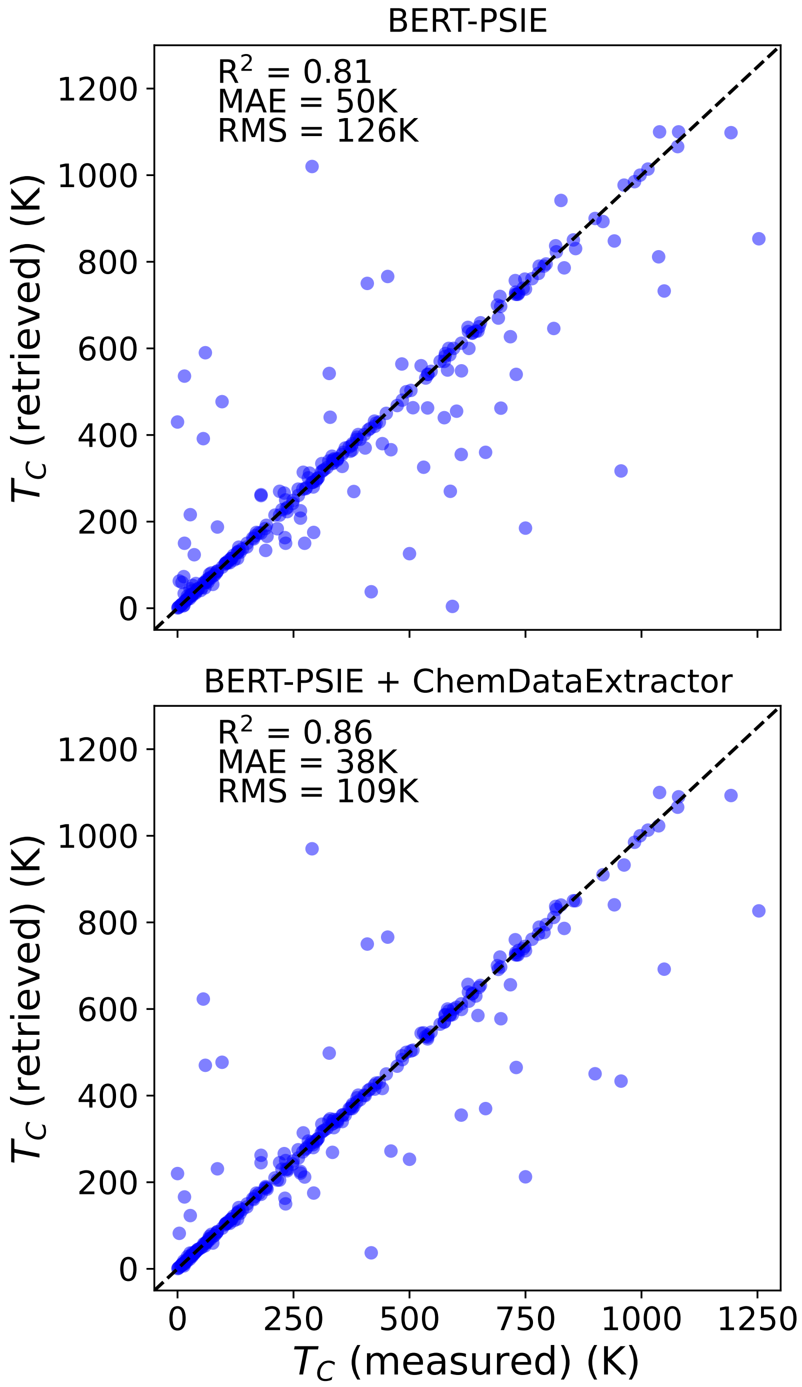

Figure 5.7: Comparison between the T_C queried in the dataset automatically generated by BERT-PSIE and the values contained in the manually curated dataset (top panel). The comparison is performed over the 262 compounds that are shared by all datasets examined in this work. The median value is returned whenever multiple T_C values are collected for a given compound. The same comparison is performed on the dataset resulting from the combination of the one generated by BERT-PSIE and the one generated by ChemDataExtractor (bottom panel).

of the Curie temperature when the database is queried. Firstly, in order to quantify the ability of the database to adequately return the correct value, a test was devised by simply directly comparing the automatically extracted Curie temperatures with the ones that were present in the manually curated dataset. There were also several simple rules-based strategies that were tested against the relationship classification strategy depicted in Fig. 5.4. The metrics were all calculated using the overlap shared by all of the datasets being compared, with the manually curated dataset, which constituted 262 compounds. The comparison between BERT-PSIE and the aggregated overlap from the manually curated dataset can be seen in the top panel of Fig. 5.7 and the evaluation metrics for all strategies in the query test are presented on the left-hand side of Table 5.3.

As was previously discussed (Section 5.3.1), the most challenging aspect of the extraction workflow is the relationship classification step. In order to ensure that there is value in adopting the unique model instead of a simpler, rule-based method, these methods were adopted as a part of the workflow and compared with the use of the relationship classifier and the performance of each was tested on the overlap of all datasets with the manually curated dataset. The first of these strategies involved only taking the values for which there was a single mention of a compound and a property in a sentence. This entirely voided the need for a relationship classification step as it was assumed that every solo mention of a property and compound that appeared in a sentence was related to each other ('Single Mentions' in Table 5.3). The second strategy worked under the assumption that it was generally true that when there are multiple mentions of compounds and properties, they can generally be associated based on the order that they appear in the sentence in question ('Order of Appearance' in Table 5.3). Finally, this was compared with the case where every possible combination of compounds and properties was added to the database. This step was taken as a way to map the improvement of the database with the inclusion of the relationship classification against randomised associations ('All Combinations' in Table 5.3). This comparison is identical to the situation in which the relationship classification model always outputs a positive classification for every relationship. Table 5.3 also presents the results obtained with the full BERT-PSIE workflow ('BERT-PSIE'), the data contained in the ChemDataExtractor database and by a combination of those two automated databases ('BERT-PSIE + ChemDataExtractor'). The query test parity plot for the aggregated dataset can also be seen in the bottom panel of

Fig. 5.7.

The results presented in Table 5.3 quite clearly indicate that every database compiled with the rule-free pipeline demonstrates an ability to query the manually curated database with a comparable ability to those of the rule-based ChemDataExtractor method. Further to this, BERT-PSIE appears to perform best in almost every query-test evaluation metric. In particular, the BERT-PSIE pipeline returns the best R^2 coefficient of 0.81 and RMSE of 126 K. Interestingly, BERT-PSIE does give a very slightly larger MAE than the one obtained from the ChemDataExtractor system. These metrics indicate that both manually curated databases extract databases of comparable quality, however, BERT-PSIE is less likely to display outliers as significant as ChemDataExtractor. The conventional notion of an outlier does not necessarily apply in the case of gauging the efficacy of a workflow in generating a database capable of querying an accurate temperature, however. As is evident from Fig. 5.7, the workflow will either extract the correct value, where entries appear on the parity line, or they will be incorrect and will appear as an outlier in the parity plot. There is no real correlation between the degree to which an outlier is incorrect and the accuracy of the workflow. This can more precisely be considered as a binary classification for each data point of either exact extraction or erroneous extraction.

Beyond this point, it is also evident from Table 5.3 that BERT-PSIE as a complete workflow outperforms every other BERT-based model that relies on alternative ways of associating the compounds and their properties. Such an increase in performance relative to these less complex strategies for relationship resolution between multiple potential compound-property pairs indicates the advantageous nature of the use of relationship classification as part of the BERT-PSIE pipeline. Admittedly, the performance of the relationship classification step does not yield a massively significant boost in performance over the inclusion of every possible combination of compound-value pairs, however, it still aids the workflow in producing a better-quality automated dataset.

The second test to evaluate the extraction workflow performance probes the ability of any of the data-extraction strategies to create databases of adequate quality to construct a predictive ML model. Determining whether the data aggregated from automated extraction can be a platform for the construction of models that have the

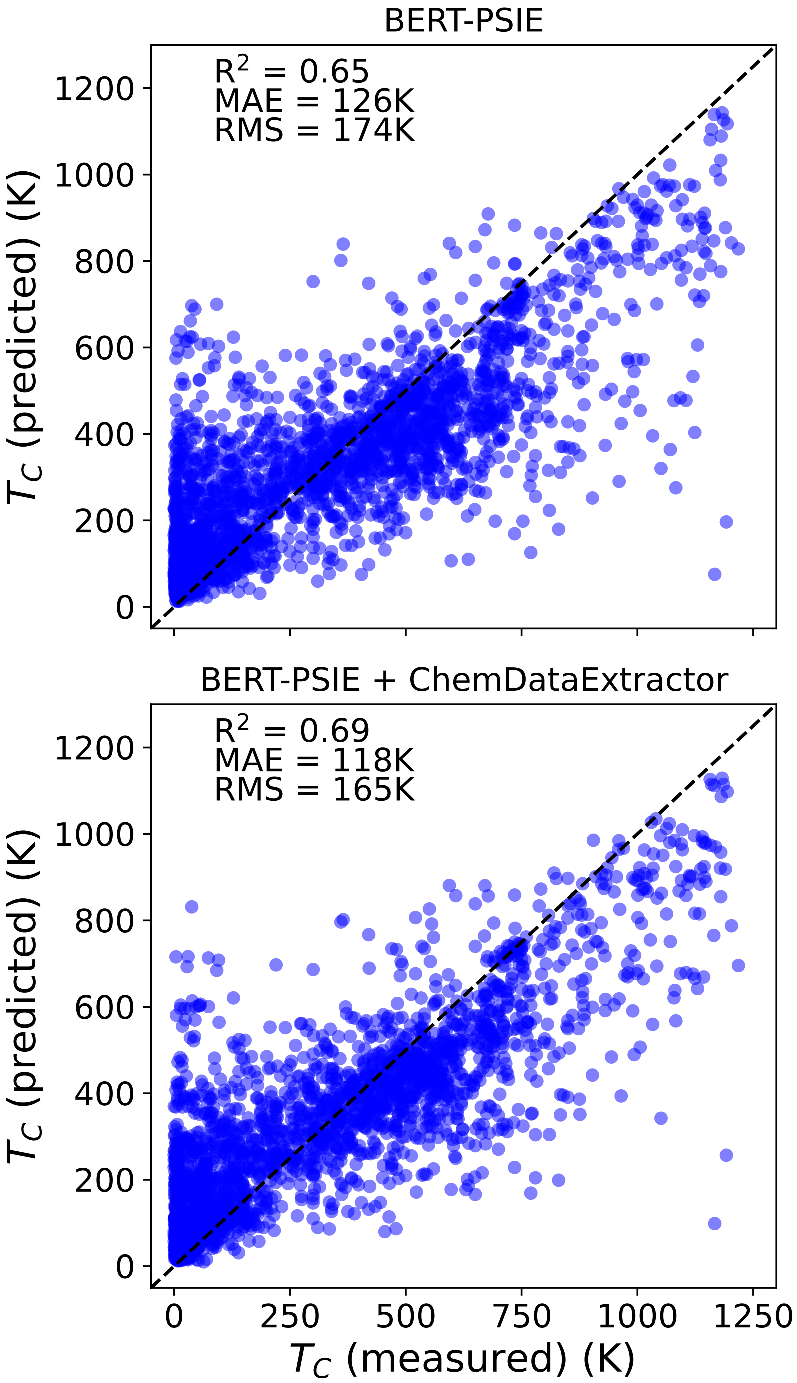

Figure 5.8: Parity plot (predicted T_C vs. manually extracted T_C) for the best RF compositional model constructed on the BERT-PSIE dataset (top panel) and on the combined BERT-PSIE and ChemDataExtractor dataset (bottom panel). The test set consists of the 2,623 compounds that are not present in any of the automatically generated datasets considered in this work, but for which we have a T_C manually extracted from the scientific literature.

ability to screen for the T_C of unseen compounds is a key step in gauging automated workflows' potential to be useful to the general field of materials science. Models that are constructed using representations based entirely on the fractional composition of elements contained within the compound have previously been demonstrated to produce reasonably good results when trained on manually curated data [11] with the caveat that the model produced by Ref. [11] does not describe non-ferromagnetic phases, e.g. antiferromagnetic structures. Further to this, such compositional representations can be constructed from information directly accessible from the extracted data, without the need for any additional extraction steps. In order to test the ability of a dataset to function as a reliable source with which to train useful ML models for property screening, a random forest (RF) model is trained on each automatically generated dataset. This takes compositional features as model input, as described in Ref. [11] and Ref. [151]. Each of the RF models trained employs the same input features as there has not been any observed improvement in model performance with the addition of more features. Once the training of the model was completed, the model predictions were subsequently compared with the manually extracted values that were not present in the training set of the model. For the Curie temperature dataset, the test set constituted 2,623 compounds for which there is a manually extracted T_C but do not appear in any of the automatically generated datasets. The use of this test set ensures that all tests are performed on the same compounds for every model considered. For compounds with multiple associated, extracted values, the median of the results is taken as being the associated Curie temperature of the compound, as in Ref. [11]. The mean and the mode were also tested, however, this did not contribute to any meaningful change in the results of the test set evaluation.

The results of this predictive-power test can be seen on the right-hand side of Table 5.3. The BERT-based extraction workflow, once again performs to a similar quality as the established, rule-based method. In particular, the full workflow, BERT-PSIE has an identical R^2 value, with a better RMSE and a slightly worse MAE once again, very similar to the query test results. Most interestingly, it was found that the inclusion of data extracted using the full BERT-PSIE workflow, including the relationship classification step does not improve the quality of the database for prediction. Single mentions only appear to be a better strategy for constructing a predictor with a better R^2 value of 0.66 and a RMSE of 174 K, while BERT-PSIE gives a slightly degraded R^2 of 0.65 and an identical RMSE, although it does improve the MAE by about 2 K. This is likely due to the inclusion of entries with multiple mentions and the error of the relationship classifier model will inherently add noise to the database, which will not facilitate an improvement in predictor performance. Therefore, no significant model improvement was detected, despite the fact that the full BERT-PSIE pipeline facilitates a much larger dataset.

The parity plot of the optimal RF model trained on the full BERT-PSIE workflow database is visible in the top panel of Fig. 5.8. The trend for the T_C is generally captured by the model. However, the model significantly underperforms relative to the one constructed on a manually curated dataset presented in Ref. [11], which reports an MAE of 57 K. This is approximately a factor of two smaller than the 126 K obtained for the RF model trained on data extracted with the full BERT-PSIE workflow. Such a result may partially be attributable to noise inherent in manually extracted datasets. It is likely, for example, that there are antiferromagnetic critical temperatures in the BERT-PSIE database. In contrast, the data used for the model in Ref. [11] was highly curated, even after the collection was complete. As a concrete example of this, additional data on paramagnets was included in order to improve the low-temperature part of the data distribution, while other data manipulation was employed to redistribute concentrations of metallic alloys, to better balance the chemical distribution of the dataset. Naturally, such steps were not taken as an aspect of the training of RF models based on the automated curation of data, since the task is to gauge the quality of the workflow in creating automatic databases without such manipulation. The expectation for these systems is that they should aggregate sufficient data so that such manipulation becomes redundant.

Further study was done in order to assess the impact of additional data on the quality of the database. It should be expected that the statistical significance of the choice of associating compounds with a Curie temperature, based on the median value, would increase significantly with a significant increase in available data. As the data extracted in the ChemDataExtractor pipeline and the BERT-PSIE one is of similar quality, and the overlap between the two datasets is so limited, an additional dataset was constructed by combining the entries for each of the two. This combined database contains 7,052 distinct entries and performs significantly better on every metric evaluated in each test devised (bottom row of Table 5.3 and the bottom plots in Fig. 5.7

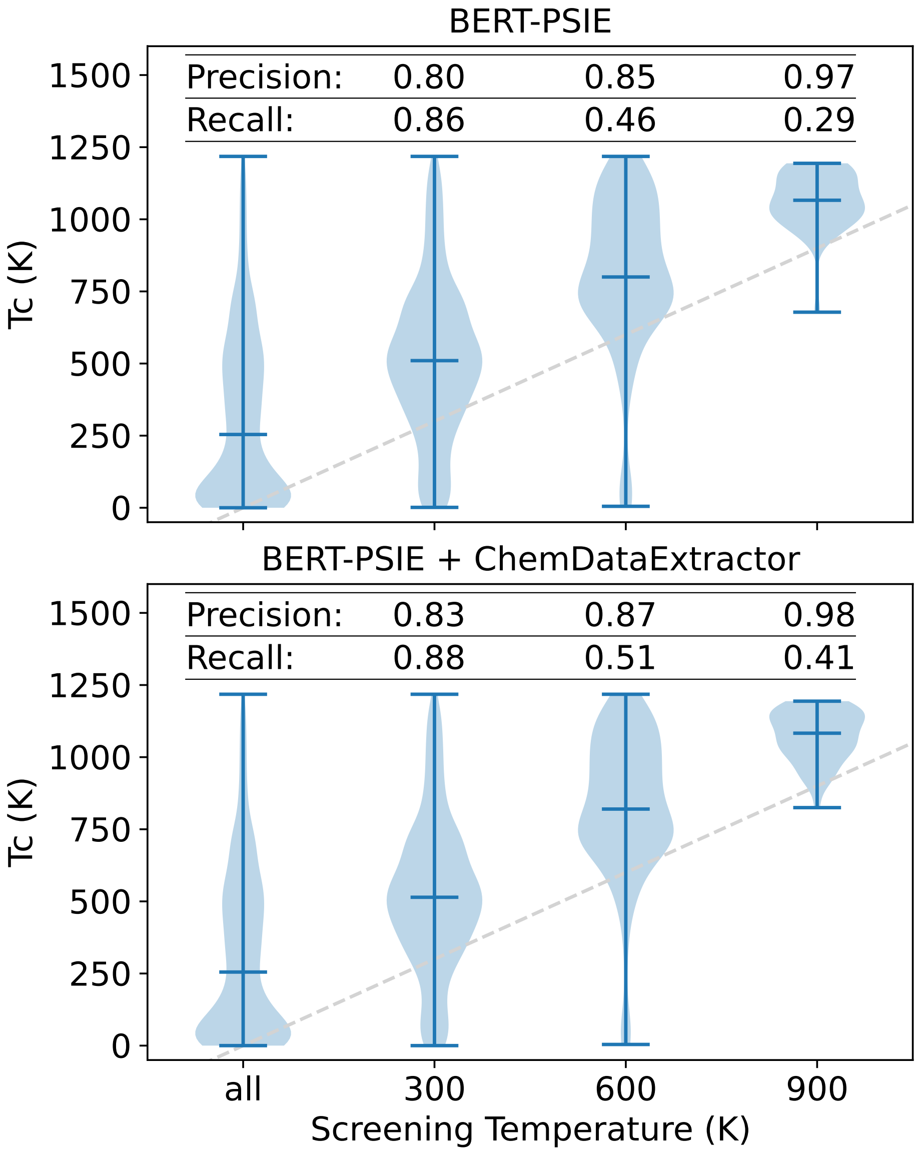

Figure 5.9: Violin plots showing the T_C distributions of the compounds screened using a RF model trained on the BERT-PSIE data and compared with the manually extracted values (top panel). The dashed line is the parity line highlighting how the median of the screened distribution increases as the screening threshold increases. Despite a low recall, the precision is high enough to select compounds likely to have a T_C higher than a given threshold. The screening is done on compounds not present in the training set of the RF. The same test is performed by training a RF model on the combination of the BERT-PSIE and ChemDataExtractor datasets (bottom panel).

and Fig. 5.8). This leap in performance is likely attributable to the far larger size of the combined database relative to the other two automatically constructed constituent ones. Indeed, the combined dataset is roughly double the size of either of the original ones. As expected, this larger size seems to have reduced the noise present in the median values. For the RF model, the larger number of entries means that the chemical space is far better sampled. The combined database represents the best database available for ferromagnetic T_{C}, which has been automatically extracted from scientific literature. The implication of this result is that the quality of automatically extracted databases improves markedly with increasing the number of disparate sources used in the extraction and in the quantity of the data extracted. Thus, selecting a wide range of sources is desirable.

As a final evaluation of the usefulness of the database extracted with BERT-PSIE, the RF model trained on the BERT-PSIE dataset was tested to evaluate its efficacy in screening unseen compounds with respect to a certain T_{C} threshold. This test simulates a common goal of applications of ML for materials science, namely that of the high-throughput screening of compounds to target specific properties. For this particular screening test, it is useful to bear in mind that typical magnets employed as part of some ambient technology (e.g. data storage, electrical motors) require a T_{C} of the order of 600 K. Therefore the ability to filter out materials likely to exceed this and other thresholds is of significant technological relevance. The RF models trained on both the BERT-PSIE and the combined BERT-PSIE/ChemDataExtractor datasets were utilised to predict whether materials have a critical temperature exceeding 300 K, 600 K and 900 K, respectively. The test set for this predictor was constructed from compounds that were present in the manually curated database, but not present in the manually constructed ones. The results of this screening for both models can be seen in Fig. 5.9, with the results for the screening done with the BERT-PSIE database in the top panel and the results for the combined dataset in the bottom panel. The shaded blue area of Fig. 5.9 represents the distribution of values of T_{C} that are predicted to be above the dashed line, depicting the screening temperatures of 300 K, 600 K and 900 K, respectively. The recall of this step does remain quite low, indicating that the model will frequently miss compounds that exceed the screening temperature. However, this is not a catastrophic issue as what is more important in screening is that the returned compounds are likely to match the screening criteria with a high precision. Thus, the high precision of the screening biasing the initial distribution set towards distributions only containing compounds exhibiting higher and higher T_{C} values, clearly demonstrates the utility of manually constructed databases in screening for T_{C} with very high fidelity.

5.4.2 Electronic band gap

In order to further validate the BERT-PSIE workflow, a second database was created, this time focusing on the aggregation of compounds and their associated electronic band gaps. In this section, an evaluation of the dataset is once again performed, comparing a separate, manually curated dataset, a dataset obtained with the ChemDataExtractor rule-based model constructed for band-gap extraction and the database obtained using BERT-PSIE. In this instance, the manually curated dataset is taken from the work of Zhuo *et al.* [152], which in turn was constructed using a range of sources [153–156]. The rule-based model arose from the work of Dong *et al.* [157]. The database associated with this work, however, contained data that had been obtained with a combination of the parsing of natural language, combined with data scraped from tables, which is beyond the scope of BERT-PSIE and, therefore, does not constitute a fair comparison. In order to address this, the ChemDataExtractor model from Ref. [157] was run on the same corpus analysed by BERT-PSIE, a step that was not accessible for the case of Curie temperature due to the relative size of the extraction corpora. In this analysis, BERT-PSIE and the rule-based model work on an identical set of publications.

The relative distributions of band gaps can be appreciated in the top panel of Fig. 5.10. In this case, there is a striking similarity between the distributions of the band-gap values that were extracted with automated methods and with the manually curated dataset. The one exception to this strong agreement is the over-representation of values at approximately 2.5 eV in the manually curated dataset when compared with both of the automatically extracted datasets. This discrepancy can potentially be explained by focusing on the bottom panel of Fig. 5.10. There are a series of significant peaks that are present within the manually curated database that are not present in the automatically extracted datasets, namely for sulfur, selenium and a series of metals (Cd, In, Sb, Te, Hg and Pb). These peaks may bias the distribution slightly towards cases that lead to the peak in band-gap values around 2.5 eV. This variance is, again, likely attributable to the difference in data sources used for the manually curated database, with much of the values of the database in Ref. [152] having been obtained from historical literature. This difference in distribution is once again evidence of the bias of curated databases towards the selected sources from which the constituent data is derived.

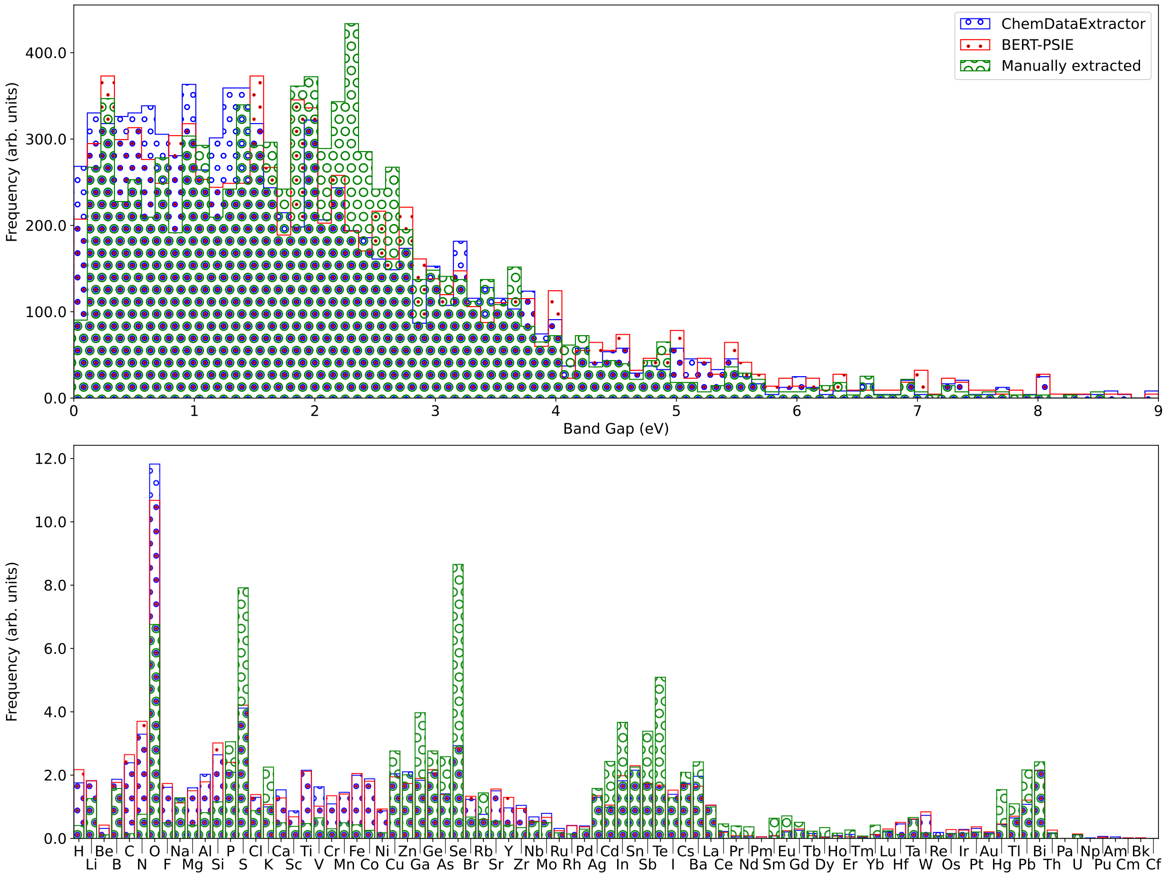

Figure 5.10: Comparison between the content of the different band-gap databases: *(red box)* BERT-PSIE, *(blue box)* ChemDataExtractor and *(green box)* the manually extracted database of Ref. [152]. Top panel: Normalized distribution of the band gaps extracted. Bottom panel: Relative elemental abundance across the compounds present in a database. Note that the automatically extracted datasets and the manually curated one are based on different literature libraries.

For both the database of extracted Curie temperatures (section 5.4.1) and the band gap, there is some level of noise and disagreement present when comparing these with the manually extracted cases. Fig. 5.11, the distribution of band gaps extracted for each of the 5 most common compounds in the BERT-PSIE database (ZnO, TiO_2, C, MoS_2 and Si), may present some clue as to why this disagreement may take place. While there is a wide range of values that are presented for all five compounds, they are not uniformly distributed, implying that this cannot be attributed to random noise, but rather to more systematic features of the underlying data structure. In fact, the distribution of the band-gap values for each of these compounds presents a series of defined peaks, with multiple values reappearing with a high frequency. This data structure can be attributed to different means of obtaining the band gap of a material

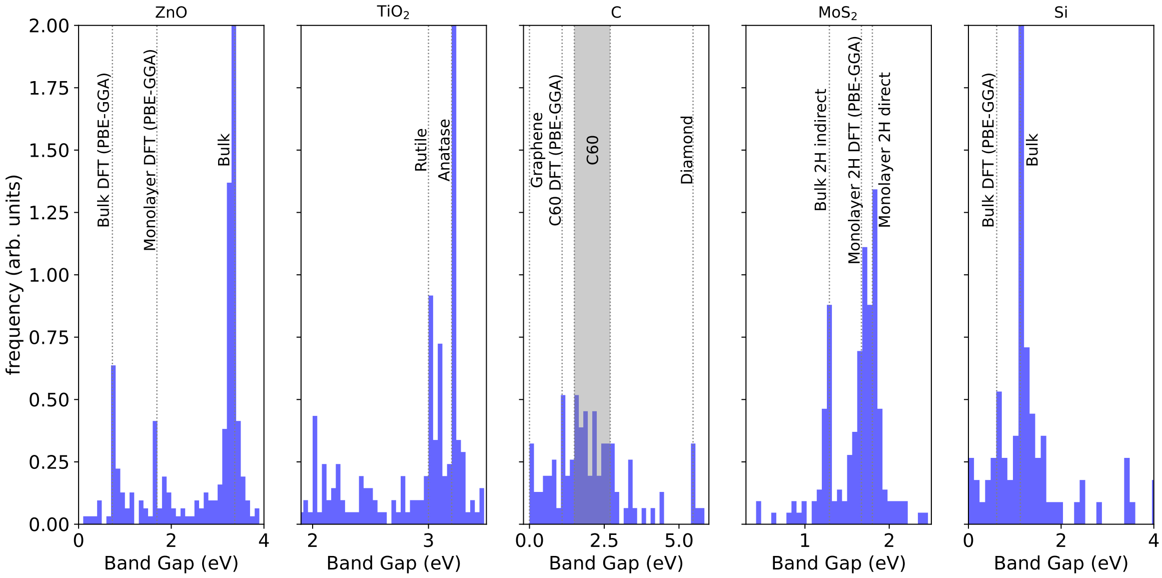

Figure 5.11: The distribution of band-gap values for the five most common chemical formulas found in the BERT-PSIE band-gap database. The histograms report the relative abundance, while dashed lines indicate gap energies corresponding to specific experimental measurements or theoretical calculations.

(experimental optical, experimental transport, theory etc.), and to different polytypes, structures or dopant-varied compounds.

To elaborate more specifically, first consider the left-most panel of Fig. 5.11, which displays the data distribution for ZnO. Within this distribution, there are a total of three clearly defined peaks, which can easily be associated with the experimental bulk band gap (3.37 eV [158]), the value obtained using density functional theory (DFT) for the bulk system (0.73 eV [158] using GGA-PBE) and the DFT-calculated case for monolayer ZnO (1.69 eV [159] again using GGA-PBE). Similarly, the discrepancy between the experimental values and those calculated using DFT can be seen in the right-most panel of Fig. 5.11, which depicts the distribution of values for Si. The most dominant peak is attributable to the indirect experimental gap for bulk silicon (1.1 eV [160]) and the smaller peak is associated with the calculated value of the band gap for bulk silicon using DFT (0.61 eV [37] once again using GGA-PBE). These results are contrasted with the case presented for the second panel of Fig. 5.11, which shows the distribution of band-gap values for TiO$_2$. These two peaks are in fact attributable to the two most abundant polymorphs of TiO$_2$, namely anatase (3.2 eV) and rutile (3.0 eV) [161].

Finally, there is slightly more complexity in the distributions for the remaining two cases, namely, MoS_2 (fourth panel of Fig. 5.11) and C (fifth panel of Fig. 5.11). In the case of MoS_2, there are three dominant peaks in the data distribution. Two of these peaks are related to the experimental band gaps of the material. Specifically, one peak is the experimental direct band gap of the monolayer, 2H polytype of MoS_2 which is 1.8 eV [162] and the second experimental peak is that of the experimental, indirect, bulk band gap at 1.29 eV [163]. The third peak is slightly harder to resolve from the direct monolayer gap, however, there is also a peak at 1.67 eV, corresponding to the DFT-calculated band gap of the monolayer system, for the GGA-PBE functional [164].

The distribution for carbon requires slightly more effort to understand than the other cases. Carbon offers a huge number of polymorphs and therefore it is not trivial to parse the significance of the range of peaks seen in this distribution. There is a clear peak at 0 eV for semimetal graphene [41] and a peak at 5.47 eV, which correlates to the band gap of bulk diamond [165]. In contrast to the other distributions, there is a range of uniformly distributed values. This range is characterised by band-gap values associated with carbon buckminsterfullerenes, C60. These values stretch over the range of 1.5-2.7 eV and there is also a clear peak at the calculated DFT value of 1.09 eV (GGA-PBE) [166].

Table 5.4: Performance comparison between the different datasets against the manually curated one from Ref. [152]. The left-hand side of the table refers to the query test, while the right-hand side refers to the RF band-gap predictor. Together with the databases constructed using BERT-PSIE and ChemDataExtractor, we also consider different BERT-assembled datasets obtained by using different relation-classification strategies (see details in the text). The query benchmark is done over the 231 compounds that are shared by all the datasets, while the RF obtained is done over 2046 compounds that are not present in any of the automatically collated datasets. Values for the best-performing datasets are in bold.

	Entries	Query			RF predictions		
		R^2	MAE (eV)	RMSE (eV)	R^2	MAE (eV)	RMSE (eV)
ChemDataExtractor	2185	0.54	0.78	1.34	0.59	**0.62**	0.87
This work							
Single mentions	1,246	0.65	0.67	1.17	0.61	**0.62**	0.85
Order of appearance	1819	**0.67**	**0.64**	**1.13**	**0.62**	0.63	**0.84**
All combinations	2581	0.63	0.71	1.21	0.60	0.63	0.86
BERT-PSIE	2090	0.64	0.67	1.19	0.61	**0.62**	0.85

The results for the comparison between datasets obtained with different strategies for the execution of downstream tasks are presented in Table 5.4. When both BERT-PSIE and the ChemDataExtractor model are deployed on an equivalent dataset, the situation changes somewhat. For both the pre-defined query test and for the predictor RF model trained on each dataset, the BERT-PSIE-extracted dataset outperforms the ChemDataExtractor dataset by nearly every metric, while extracting a very similar number of compound/property pairs, at 2,021 relationships for the full BERT-PSIE pipeline against 2,185 for ChemDataExtractor.

Interestingly, for sentences containing multiple mentions of band gaps and/or compounds, the best strategy for the association of such pairs seems to be by associating them in order of their occurrence in their respective sentences. This contrasts with the Curie temperature case seen in Table 5.3, for which such a rule implementation degrades the performance. These results indicate an intrinsic difference in the way in which these two quantities are reported in natural language. It appears as if the reporting of the band gap in literature is far more procedural than is the case for the Curie temperature. Thus, the use of a more sophisticated method of establishing the correct associations between compounds and properties introduces a source of noise which would not be present in a simpler system that still captures the relatively more simple associations present in band-gap reporting than Curie temperatures. However, this result is clearly property dependent, and while we can establish that the relationship classification step for the band gap may exhibit a marginally negative impact on our workflow performance, it remains useful to use the relationship resolution for general cases as this relationship cannot possibly be known prior to the extraction and the difference in performance is extremely marginal.

The results for the query test and the RF predictor test are visible in Fig. 5.12. For the query test in the top panel, the results are very similar to those found for the Curie temperature, although we found a more diffuse distribution of band-gap values. This diffuse distribution can be associated with the more ambiguous nature of band-gap definitions (e.g. carbon in Fig. 5.11) when contrasted with the Curie temperature, which tends to be reported for a smaller range of materials. The RF model has a slightly inferior R^2 value to the one constructed for T_{C} but benchmarks similarly to models constructed on manually curated data. Indeed, the MAE is 0.62 eV, against the value reported on MatBench of 0.33 eV [167]. Interestingly, once again, the MAE

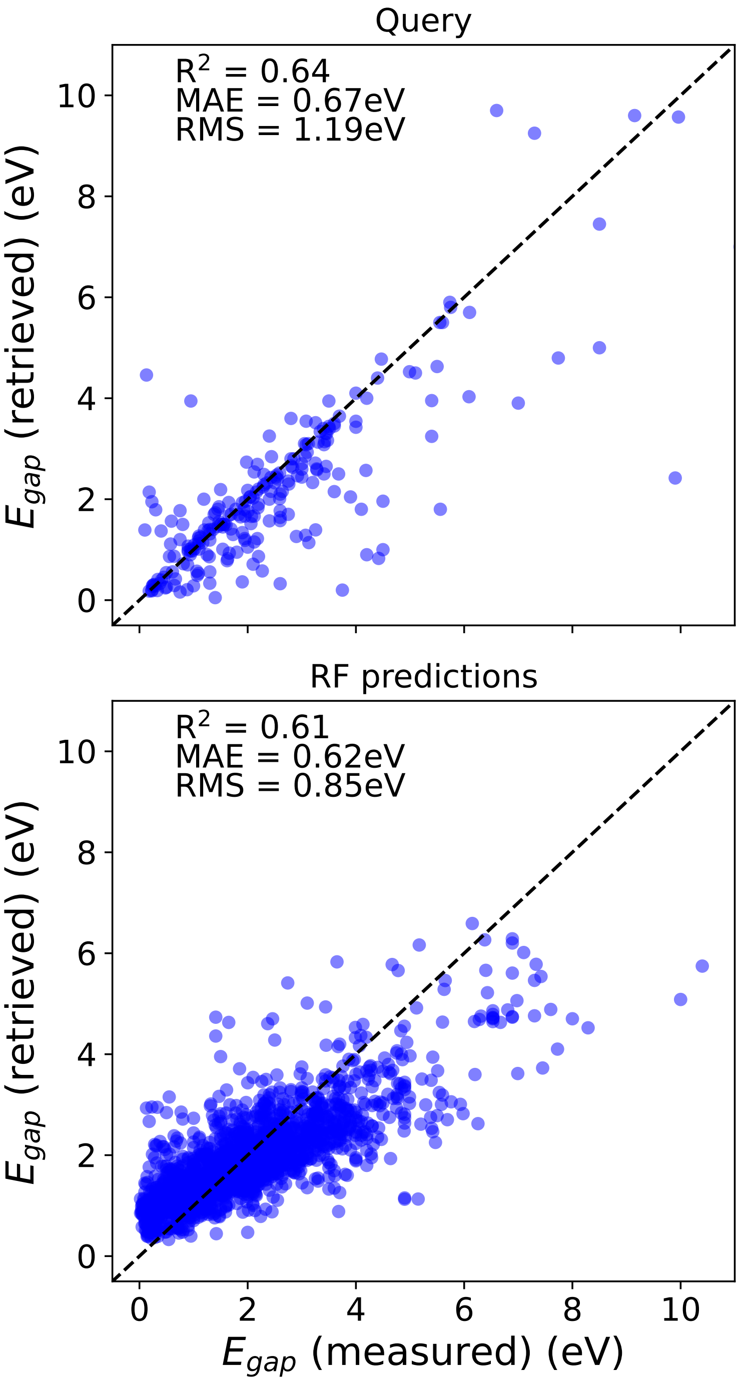

Figure 5.12: Comparison between the band gaps queried in the dataset automatically generated by BERT-PSIE and the values contained in the manually curated dataset (top panel). Parity plot for the best RF compositional model constructed on the BERT-PSIE dataset. The test set consists of the 2046 compounds that are not present in the dataset but for which we have a band gap extracted from the corpus (bottom panel).

of the RF model trained on automatically extracted data is, once again, double the value of a model trained exclusively on manually curated data.

5.5 GPT-Based Systems

It is clear, given the results from this automated extraction using transformer networks, that language models have the potential to completely revolutionize the domain of materials science and beyond. The rapid developments of large-scale autoregressive models further open the door to a wide array of refinements and expansions of existing techniques that have the potential to vastly improve the performance of the models and workflows described thus far in the chapter.

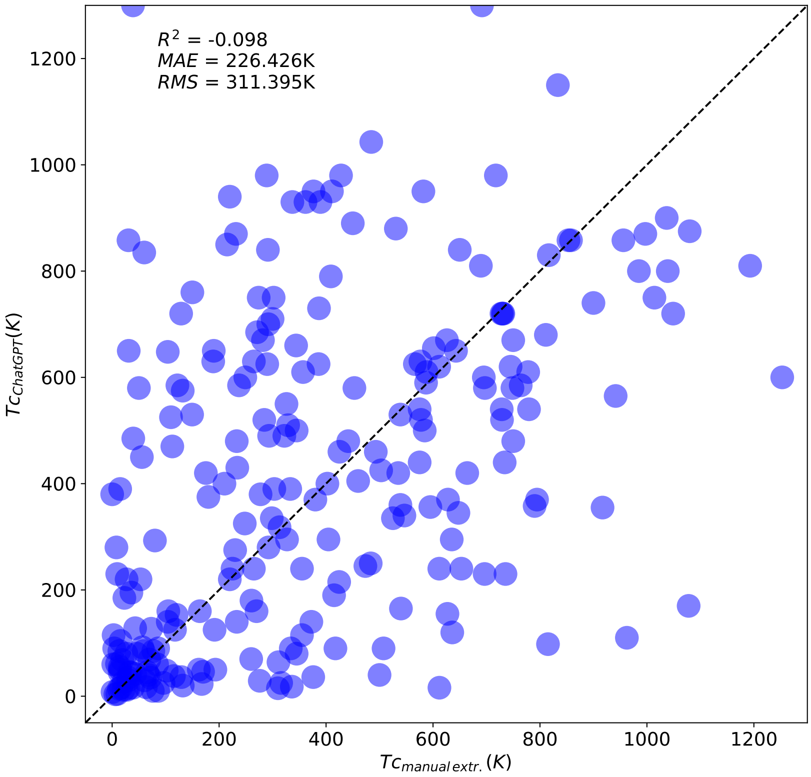

Figure 5.13: Comparison between the T_{C} queried directly from ChatGPT and the values contained in the manually curated dataset. The comparison is performed over the 262 shared compounds between all three datasets, evaluated in section 5.4.1.

With that said, LLMs cannot be treated as a silver bullet for every problem, particularly in fields where the precision and trustworthiness of data are vital commodities. Large-scale GPT models have indeed been shown to be reasonably performant time and time again on general tasks. However, when these large-scale GPT systems were conceived, there was no claim that they were more performant on specific tasks than systems that were pre-trained or fine-tuned for those domain-specific tasks [122, 123]. Thus, such systems are most likely sub-optimal compared to fine-tuned or pre-trained systems for specific domains. In addition, their immense scale makes them very difficult to adapt to specific tasks, compared to smaller models, like BERT. To verify this for the domain of materials science, a test was set up for ChatGPT, which employs GPT-3.5 [124] in the 'LLM for everything' setting. In this test, the assumption is that the LLM has learned all the necessary information for a particular topic in its model weights. This is not a claimed use of the model, however, it is a valuable test as to whether or not there is any knowledge about general trends in magnetic properties for materials.

This test was executed by giving the model the prompt: 'I am going to provide you with a list of chemical compounds and you will generate a list containing the Curie temperature associated with each compound in a JSON file. From now on, you will answer by providing just the requested JSON file and no further information.'

The results of this test can be seen in Fig. 5.13. It should be noted at this point that the initial indications were that the model had a reasonably accurate knowledge of the most common ferromagnetic compounds such as iron and cobalt, returning the correct values for each. As is clear from the plot, however, the model contains little to no information about trends in magnetism for materials, with the zero-shot task producing a R^2 value of -0.098, which indicates absolutely no correlation between the actual material property according to the manually curated database and the value produced by ChatGPT. This implies that the model simply hallucinates with Curie temperature values that appear to be randomised.

This is not at all to say, however, that GPT-based LLMs do not exhibit potential for materials science applications or high-precision data extraction, even without any domain-specific pre-training or fine-tuning. They have already been shown to have uses in certain, limited capacities [132]. Indeed these methods certainly exhibit the potential to enhance aspects of the workflow presented in this chapter, facilitating the extraction of compound/property relationships beyond the sentence level and resolving interdependent properties with an unprecedented precision.

In order to integrate an autoregressive LLM into the information extraction pipeline, there needs to be a mechanism to address the propensity of the model to hallucinate, producing incorrect or inaccurate information and, thereby, increasing the density of noise in the resulting database. A potential mechanism designed to address this can

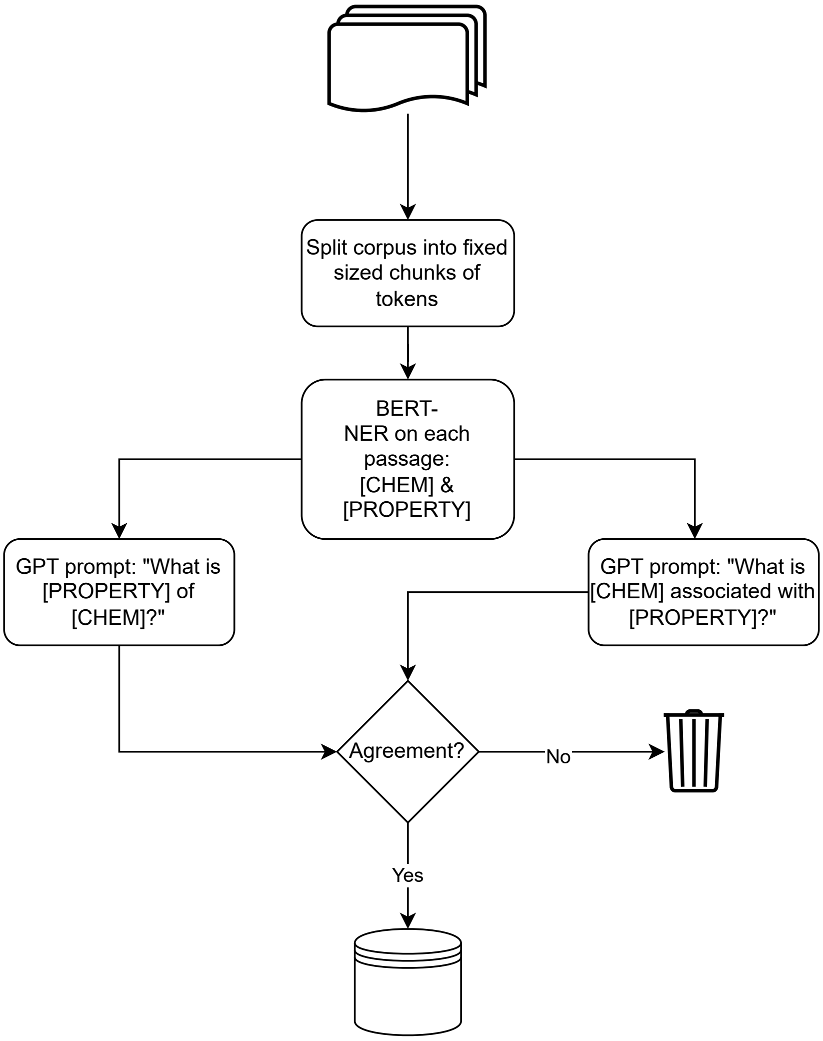
Split corpus into fixed sized chunks of tokens
BERT-NER on each passage: [CHEM] & [PROPERTY]
GPT prompt: "What is [PROPERTY] of [CHEM]?"
GPT prompt: "What is [CHEM] associated with [PROPERTY]?"
Agreement?
No
Yes

Figure 5.14: Flowchart of the proposed workflow for parsing compound/property mentions from a corpus of scientific literature, incorporating the power of autoregressive LLMs. This proposed workflow would involve the training of a single NER system for recognising compounds and properties.

be seen in Fig. 5.14. In this workflow, there is still a reliance on the BERT-NER model described in section 5.3.1, however, in this instance, the model provides its own sanity check to ensure a high probability of the correct information being extracted from the corpus. This process also is not dependent on a sentence-level extraction and can, therefore, resolve more distant interdependencies between compounds and property relationships.

There remain several drawbacks to employing such a workflow, which are mainly attributable to the most powerful of these autoregressive LLMs being proprietary and carrying significant costs associated with API usage. In order to employ such a workflow in-house, there is a need for a variety of specialist hardware prior to workflow execution. Work is ongoing in executing these efforts in order to improve the information extraction workflow.

Further potential of employing LLMs into the existing workflow is in the post-processing of data distributions post-extraction. For example, if we consider the manually annotated data distributions visible in Fig. 5.11. The process of attributing the peak positions could potentially be automated by simply prompting a LLM with the sentences that the data extraction was performed on for the respective peaks and asking that the model finds the similarities between the constituent sentences.

5.6 Summary & Conclusions

This chapter has provided a comprehensive overview of the automated extraction of databases from a vast collection of scientific literature, with a specific focus on materials science. It began by discussing the historical significance of natural language techniques in information extraction within this field. Subsequently, the transformer architecture was introduced, along with an exploration of its main applications in materials research.

The chapter presented the BERT-PSIE workflow, which offers a self-contained series of transformer-based models capable of being trained and deployed without the need for complex grammar and syntactic rules. These models have achieved comparable performance to rule-based techniques in extracting databases for materials properties. Although the resulting databases were found to be of lower quality and predictive power compared to manually extracted databases, they still exhibited impressive accuracy in querying the property values of compounds. Moreover, these databases proved valuable in training predictor models that can precisely screen for high-T_C materials.

Furthermore, the chapter discussed the potential of utilizing massive GPT models to enhance the information extraction paradigm in materials science. Several potential applications of these GPT models were explored, highlighting their ability to extend and improve the automated construction of databases for materials science.

Overall, this chapter has shed light on the advancements in automated database extraction, the transformative potential of transformer architectures, and the possible valuable role of GPT models in the field of materials science. These insights pave the way for further progress and innovation in the automated construction of materials science databases.

Chapter 6

Contextual Representations for Materials

"Tuigeann Tadhg Taidhgín."

Previously in this thesis, the prediction of materials properties and the use of natural language techniques have been treated as entirely separate and unrelated concepts, that constitute two entirely separate pillars of this research work. This chapter is dedicated to bridging the gap between these disparate concepts and endeavours to create embeddings that could potentially enhance machine learning (ML) performance in predicting materials properties. Such a representation could enhance the rapid property prediction step of the inverse-design workflow by creating more robust representations for ML, encoding more information for the model. These embeddings are designed to encode text-based information extracted from literature into contextual representations that exhibit an understanding of a compound's chemical, structural, and property characteristics, all based on the context of their appearance in literature.

The chapter commences with an exploration of techniques for constructing embeddings that effectively capture information and concepts in Euclidean space. These foundational embeddings lay the groundwork for subsequent developments. The latter portion shifts the focus towards previous efforts that have aimed to embed the materials space into context-aware representations using available literature, addressing the limitations and challenges encountered in these endeavours.

The latter part of the chapter is dedicated to reviewing the limited existing proposals to encode the materials space into context-aware embeddings, drawing from the wealth of existing literature. Finally, I will explore the potential of transformer architectures in creating next-generation contextual embeddings. By harnessing the demonstrated superior contextual understanding of transformers, the goal is to unlock new avenues for enhancing the accuracy and effectiveness of models used for predicting materials properties.

6.1 Word Embeddings

A word embedding is a representation of text in a vector space, in which text with similar meanings have similar representations. The term ‘similarity’ in a vector space may correspond to a number of metrics and manipulations that can be performed between two term vectors that give an indication of their relative distance in the vector space. There are many such metrics. However, a simple example that can easily and efficiently be applied to vectors in a given vector space is the Euclidean distance. For vectors $\mathbf{u}$ and $\mathbf{v}$ of dimension n can be written as,

$$(\mathbf{u},\mathbf{v})_{\text{euc}} = \sqrt{\sum_{i=1}^{n} |u_i - v_i|^2}. \tag{6.1}$$

Another simple and, arguably, the most common example of a similarity metric between two vector space embeddings is the cosine similarity,

$$(\mathbf{u},\mathbf{v})_{\cos} = 1 - \frac{\sum_{i=1}^{n} u_i v_i}{||\mathbf{u}||_2 ||\mathbf{v}||_2}, \tag{6.2}$$

where $||\mathbf{u}||_2$ and $||\mathbf{v}||_2$ are the L2-lengths of vectors $\mathbf{u}$ and $\mathbf{v}$, respectively. The L2-length of a vector, $\mathbf{u}$, is given by the formula,

$$||\mathbf{u}||_2 = \sqrt{\sum_{i=1}^{n} u_i^2}. \tag{6.3}$$

The simplest and most intuitive way of constructing a vector-based representation of text that conforms to this requirement is by forming a matrix known as the co-occurrence matrix. The fundamental idea behind this representation is captured by the quote from the eminent linguist, J. R. Firth in 1957, "You shall know a word by the company it keeps." [168]. With this in mind, the co-occurrence matrix counts the co-occurrence of two terms as they appear in a context, as a means of capturing that contextual information. These embeddings are a square, $n \times n$ matrix where n is the number of words in the vocabulary. The vocabulary can be the total number of words in the corpus from which the co-occurrence is derived. Alternatively, pre-processing steps can be applied to filter out a number of pre-defined stop-words that do not contribute any contextual meaning to the co-occurrence and would likely bias the representation, e.g. 'the', 'and', 'a', etc.

Such a process is likely to yield a very high-dimensional representation with a large amount of redundancy, through synonyms, different verb tenses and words sharing semantic similarities, likely to co-occur with similar word distributions. In order to overcome this redundancy, a more condensed version of the same representation can be obtained by funnelling the co-occurrence matrix into a dimensionality reduction algorithm after its construction, such as principal component analysis (PCA) or latent semantic analysis (LSA) [169]. The lighter-weight, resulting representation is more efficient for predicting text associations and relationships, and the reduced redundancy means algorithms can be trained more efficiently, thereby improving the representation and reducing the computational cost of working with these representations.

Co-occurrence matrices and other frequency-based methods serve as valuable initial steps for constructing vector space embeddings. However, they often tend to excessively emphasize the significance of potentially trivial information. Therefore, it is important to explore alternative representations that move beyond mere frequency-based weighting. Although co-occurrence term frequency offers contextual insight and contributes to a good representation, it is crucial to consider methods that reweight the co-occurrence matrix.

When reweighting the co-occurrence matrix, two key factors warrant consideration. Firstly, it is essential to compare the reweighted matrix against the raw count values to assess the impact of the reweighting. Secondly, the resulting distribution of vector objects achieved through the reweighting scheme should reflect the real-world contextual associations among terms in the given vocabulary. By addressing these factors, we can better account for the limitations of frequency-based approaches and enhance the quality of vector space embeddings.

One example of reweighting is normalization, which involves either dividing each vector component by the L2-length of the vector represented in Eq. (6.3), or dividing each vector component by the sum of all components. Another powerful reweighting scheme is known as observed/expected. For a matrix, $\mathbf{X}$, the expected value of element X_{ij} can be represented as,

$$\text{expected}\left(\mathbf{X}, i, j\right) = \frac{\sum_{j'} X_{ij'} \cdot \sum_{i'} X_{i'j}}{\sum_{i'j'} X_{i'j'}}. \tag{6.4}$$

The observed/expected reweighting is then obtained from the co-occurrence matrix by taking the ratio of the observed value X_{ij} with the expected value obtained from Eq. (6.4),

$$\text{oe}\left(\mathbf{X}, i, j\right) = \frac{X_{ij}}{\text{expected}\left(X, i, j\right)}. \tag{6.5}$$

The pointwise mutual information (PMI) [170] scheme takes this idea of observed/expected reweighting of the co-occurrence matrix and places it in log space,

$$\text{PMI}\left(\mathbf{X}, i, j\right) = \ln\left(\frac{X_{ij}}{\text{expected}\left(X, i, j\right)}\right), \tag{6.6}$$

with the imposition that any values of 0 for X_{ij} are set to 0 for the PMI in order to ensure that the quantity is defined. The absolute values of all elements in the resulting matrix are taken in order to avoid a situation in which all elements of the co-occurrence matrix with 0 for the term co-occurrence frequency are the smallest values in the distribution as opposed to being intermingled between positive and negative PMI values. This, improved, co-occurrence frequency-based representation is called the positive PMI or PPMI. The PMI is a powerful means of creating an embedding from a simple co-occurrence matrix as with some simple manipulations, the PMI can be shown to be the log probability of a term co-occurring with another one, when compared to the likelihood of the same terms being independent, thus capturing the semantic similarity between the two terms. For instance, given a word, w_i, and a word, w_j, that have probabilities $P\left(w_i\right)$ and $P\left(w_j\right)$ of occurring, the mutual information

$I\left(w_i, w_j\right)$ is defined as

$$I\left(w_i, w_j\right) = \ln\left(\frac{P\left(w_i, w_j\right)}{P\left(w_i\right)P\left(w_j\right)}\right), \tag{6.7}$$

and $I\left(w_i, w_j\right)$ can be shown to be equivalent to the $\mathrm{PMI}\left(\mathbf{X}, i, j\right)$ in Eq. (6.6) [170]. Other examples of weighting or normalization regimes are term-frequency inverse-document-frequency (TF-IDF) [171] or enforcing a Student t-test distribution over your representations.

6.1.1 Word2Vec

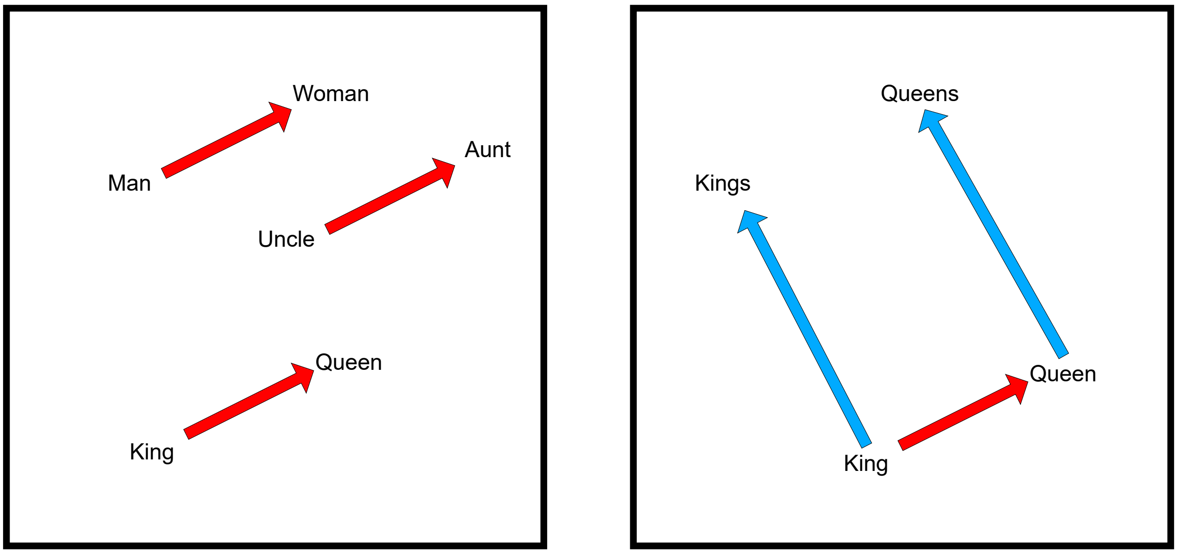

Figure 6.1: *(left panel)* Diagram demonstrating how the same word representation offset vector depicts the relationships between gendered terms. *(right panel)* A similar diagram demonstrating how word representation offset vectors capture the relationship between a term and its pluralised form. *This diagram was adapted from Ref. [172].*

In 2013, it was discovered that applying more sophisticated techniques to the construction of word representations could lead a vector space model to capture, not only word similarities as before but rather, semantic information about the relationship between words, which could then be used to predict or estimate the nature of resulting term-representation vectors [172]. Some examples of this are in the relationship between gender, family relations and monarchs, depicted in the first panel of Fig. 6.1, with the same offset vector describing the difference between gendered words i.e. $x_{\text{king}} - x_{\text{man}} + x_{\text{woman}} \approx x_{\text{queen}}$, where x is a word embedding. Similarly, the same offset vector describing the difference between a word and its plural form (i.e. $x_{king} - x_{\text{kings}} \approx x_{\text{queen}} - x_{\text{queens}} \approx x_{\text{apple}} - x_{\text{apples}}$) is depicted in the second panel of Fig. 6.1.

This work was performed using a recurrent neural network with a single hidden layer to construct the representations, which had been shown to significantly outperform the older, more traditional n-gram models that constructed a probability distribution based on previous words in a training corpus. This idea was further elaborated on with the development of the Word2Vec architecture [173].

The proposed goal of Word2Vec was to scale up the amount of training data and create a system, which could be trained on data sets comprised of billions of words, rather than the datasets of a few hundreds of millions of words that previously were accessible. The dimensionalities of these previous word vectors were also limited to between 50 and 100 terms. In this work, two architectures were proposed in order to construct this new generation of representations. These models were constructed in order to minimize the computational complexity of the training step of the representations, this is achieved by discarding the non-linear hidden layer and instead focusing on the data efficiency of the representation.

Continuous Bag-of-Words (CBOW)

The continuous bag-of-words (CBOW) architecture, depicted in Fig. 6.2, is based on one of the early attempts at constructing a neural network language model (NNLM) [174], where the non-linear activation is simply removed, thus, all words are projected into the same position and their vectors are averaged. Thus, the resulting vector for a given word is constructed by taking the average of the word representation that it appears with. The representation is then optimised based on the criterion of correctly predicting the vector of the output word based on the input representations from a log-linear classifier.

Continuous Skip-Gram

The second architecture, the continuous skip-gram architecture (see Fig. 6.3) is essentially the reverse of the CBOW architecture. The current word is used as an input to a log-linear classifier with a similar projection layer as that of the CBOW method, which

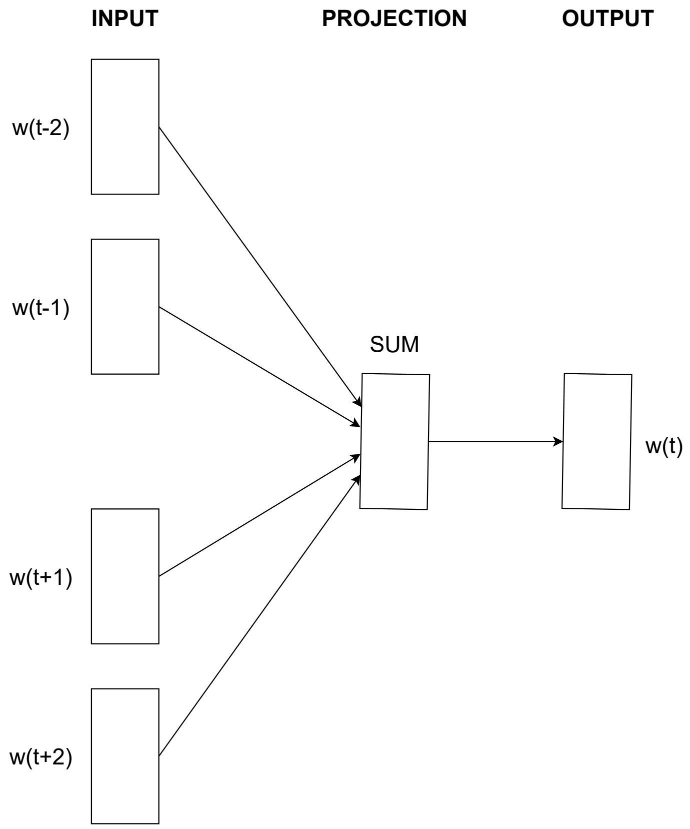

Figure 6.2: A diagram of the CBOW architecture in which the current word is predicted based on the context in which it appears.

is trained to predict words within a certain range before and after the current word. As is evident from the simplicity of the algorithm, both architectures are very computationally lightweight. However, the computational cost of the skip-gram depends on the number of contextual words being predicted. Across a range of tests of the semantic capabilities of the words, the 640-dimensional skip-gram architecture outperformed all prior architectures on every test but one, on which the CBOW architecture slightly outperformed the skip-gram. The Word2Vec architecture was used to construct a total of 1.4 million word-representation vectors, which were trained on more than 100 billion words, representing an early attempt at using large-scale reference databases of word embeddings for general-purpose applications.

6.1.2 GloVe

While Word2Vec exhibits a superior ability to probe the context of words based on the surrounding words with which they appeared, its major drawback is the focus on the local context of the corpus, meaning that it utilises the statistics of the corpus poorly,

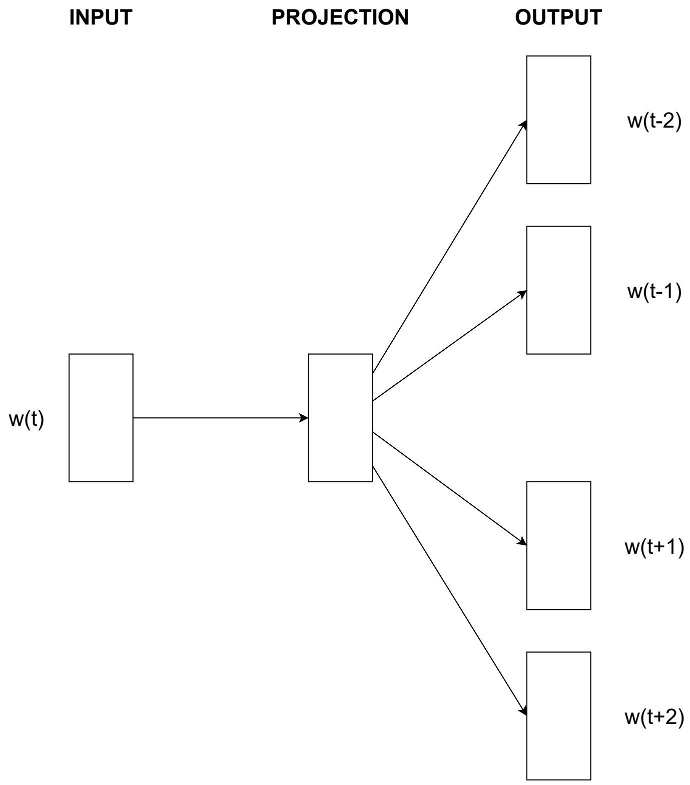

Figure 6.3: A diagram of the skip-gram architecture in which the surrounding context is predicted based on the current word.

with no information whatsoever about the global co-occurrence count. In contrast, methods based on the co-occurrence matrix, such as PPMI, coupled with LSA are shown to leverage statistical information very efficiently. Despite this, however, they generally perform very poorly on word analogy tasks i.e. $x_{king} - x_{man} + x_{woman} \approx x_{queen}$. In order to address these shortcomings, and balance the use of contextual information with global, statistical information, a new algorithm called Global Vectors or GloVe was conceived [175].

GloVe accomplishes this unification of local contextual information with global statistical information by weighting the resulting vectors according to their ability to model the ratio of co-occurrence probabilities between the two words, as a function of the offset between the two vectors. This is taken as a dot product with the word vectors for the context in which those words appear. It does this, in practice, by implementing a matrix factorization technique in which a loss function is minimized in order to find the optimal, lower-dimensional representation of the high-dimensional co-occurrence matrix.

Upon training this new model on a corpus of 42 billion tokens, compared to the

100 billion tokens used for the Word2Vec architecture, GloVe outperformed Word2Vec on every semantic, syntactic and word analogy test, significantly in most cases, using a training corpus less than half the size of that for Word2Vec. This represented a brand-new state-of-the-art for word representations across a wide range of potential applications.

6.1.3 Contexualized Representations

As has previously been described in this thesis (see Section 5.2), transformer architectures, such as BERT (see Section 5.2.3), represent the current state-of-the-art of what is achievable in NLP. This success is mostly reliant on the ability of the representations, resulting from such architectures, to effectively capture the context and latent structure of natural language by leveraging the self-attention mechanism. This representation, however, is entirely contextual and, as a result, the same words or phrases can be represented entirely differently based on the context in which they appear. Thus, in order to construct static embeddings from contextualized embeddings, a number of strategies were proposed. Their goal is to leverage these superior, pre-trained contextualized embeddings, to construct high-quality static embeddings [176], given a word, w, in a context, c. These methods are entirely general and only rely on the assumption that the contextual model maps word sequences to vector sequences.

There are two concepts to consider in order to construct such representations. The first is known as subword pooling, which is the application of a pooling mechanism such as min-, max- or mean-pooling over k subword representations, which are generated for word, w, in context, c, using transformers i.e. $\{\mathbf{w}_c^1, ..., \mathbf{w}_c^k\} \mapsto \mathbf{w}_c$. The nature of the tokenization step when constructing a BERT representation means that often words will be deconstructed into a number of subwords (see Fig. 5.2). Subsequently, the layers of the transformer will compute representations of each subword, $\mathbf{w}_c^1, ..., \mathbf{w}_c^k$. Given these equidimensional vectors, a number of pooling mechanisms can be considered in order to construct our singular word representation, $\mathbf{w}_c$:

$$\mathbf{w}_c = f\left(\mathbf{w}_c^1, ..., \mathbf{w}_c^k\right)$$
$$f \in \{\min, \max, \mathrm{mean}, \mathrm{last}\}$$

where min(·) and max(·) are the element-wise min/max pooling, mean(·) is the arithmetic mean of the vector representations and last(·) simply takes $\mathbf{w}_c^k$ as the word representation. The mean-pooling strategy was shown to perform best in Ref. [176], and therefore this is the main strategy considered.

We must also consider context combination, which is defined as the mapping between representations, $\mathbf{w}_c^1, ..., \mathbf{w}_c^k$, of w in differing contexts, $c_1, ..., c_n$, to a single static embedding, $\mathbf{w}$, which is context-agnostic. The first of these context combination methods is extremely simple in concept. Known as the decontextualized approach, for a word w a single context is taken $c_1 = w$. The pre-trained transformer is fed with a single word and the outputted vector representation is taken as the word representation, $\mathbf{w}$, for w.

The second of these context combination strategies involves the construction of aggregated representations of w as it appears in numerous contexts. In practice, n sentences are sampled across a corpus of documents, each of which contains the word w and the vector for each of these instances of w are computed in different contexts, $\mathbf{w}_{c_1}, ..., \mathbf{w}_{c_n}$. As above, a pooling strategy is applied to the aggregated contextual representations, which is used to construct a single representation, $\mathbf{w}$, for w

$$\mathbf{w} = g\left(\mathbf{w}_{c_1}, ..., \mathbf{w}_{c_n}\right): \quad g \in \{\min, \max, \text{mean}\}, \tag{6.8}$$

where min(·), max(·) and mean(·) have the same meaning as above.

After Word2Vec and GloVe had been the most dominant static word representations for almost 6 years, the new aggregated BERT embeddings constructed in Ref. [176] significantly and consistently outperformed both of the previous state-of-the-art methods when considering the best-performing, aggregated strategy with mean pooling. This, however, happens once a critical threshold of training data over 100,000 contexts has been used to construct the aggregated representation.

6.2 Word Embeddings in Materials Science

The use of static word embeddings and their potential use in materials science remains a critically under-researched field within the domain. There remains a huge amount of unexplored potential in the domain for leveraging the clear power of natural language techniques to enhance our ability to understand relationships between materials and their properties and, furthermore, to use such techniques to potentially improve the predictive power of ML models for materials science.

6.2.1 Mat2Vec

The seminal example of a word-embedding captured from materials science literature is the work of Tshitoyan *et al.* [177], in which a Word2Vec, skip-gram, architecture, called `mat2vec`, is obtained by selectively taking materials-science-relevant abstracts from a corpus of a total of 3.3 million abstracts, obtained from a combination of the Elsevier API, the Springer Nature API and web scraping. This work gave the first indication that chemical intuition could, in fact, be captured by the unsupervised construction of word embedding representations without any explicit encoding of chemical information. `mat2vec` was realised by first taking the cosine similarities of representation vectors of various chemical formulae, a strategy that gave a strong indication as to the similarity of chemical composition and chemical properties of the two compounds. For example, by taking the cosine similarity of the vector representation in the embedding space of $\mathrm{LiCoO_2}$, a well-known lithium-ion cathode compound, the five compositions deemed to have the highest cosine similarity with this compound are $\mathrm{LiMn_2O_4}$, $\mathrm{LiNi_{0.5}Mn_{1.5}O_4}$, $\mathrm{LiNi_{0.8}Co_{0.2}O_2}$, $\mathrm{LiNi_{0.8}Co_{0.15}Al_{0.05}O_2}$ and $\mathrm{LiNiO_2}$, which are all also lithium-ion cathode materials.

Further to this, the potential of static word embeddings to form chemistry-aware analogy resolution was probed for the first time. Constructions such as,

$$x_{\text{ferromagnetic}} - x_{\text{NiFe}} + x_{\text{IrMn}} \approx x_{\text{antiferromagnetic}},$$

were used to explore the capabilities of static-embedding space to understand relationships between compounds and the properties that they exhibit. In order to visualise these embedded relationships more clearly, PCA dimensionality reduction was employed to reduce the 200-dimensional vectors down into two dimensions in order to see the similarity between certain offset vectors. As is evident from Fig. 6.4, there are clear indications that the relationships between these embeddings and their structures can

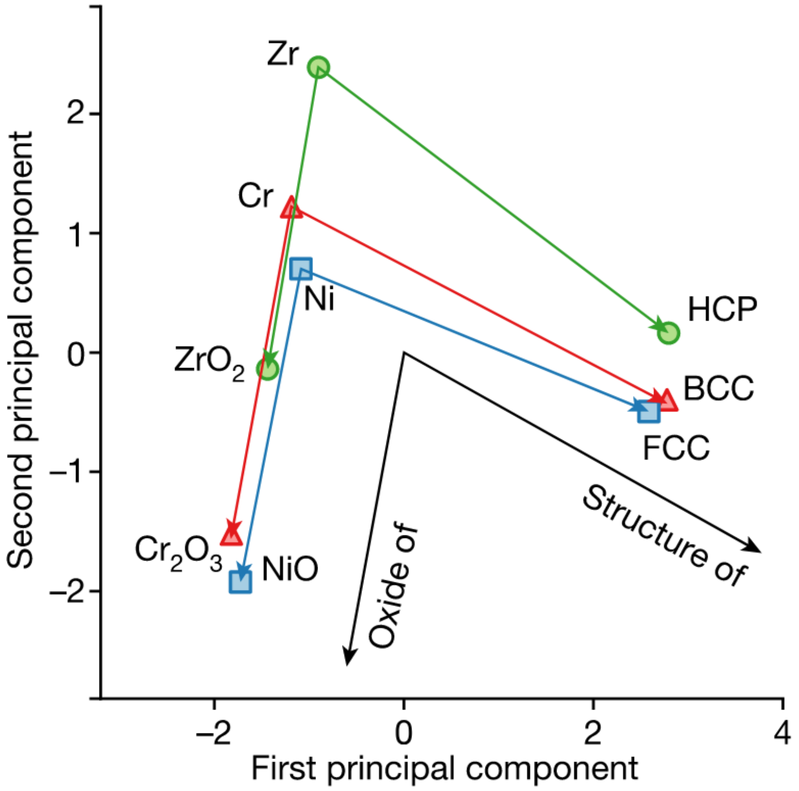

Figure 6.4: Word embeddings for Zr, Cr and Ni, their principal oxides and their standard crystal symmetries, projected onto two dimensions using principal component analysis and represented as points in the embedding space. The offset vectors between these compounds, their oxides and their structures are also visible in the plot. *This plot is adapted from Ref. [177].*

be encoded by the offset vector, which, in turn, can be associated with the relationship ‘structure of’. In the same way, the offset vector between these elements and their oxides are similar enough that it is clear that this offset vector represents the ‘oxide of’ relationship.

An advantage of such an embedding method in the case of materials science was highlighted as being an equivalence in representations of both compound mentions and properties within the embeddings space, meaning that a cosine similarity measure of a compound embedding and the embedding of a property name, could indicate a materials likelihood of being related in some way to the property in question. In the case of Ref. [177], this property was used in order to rank materials embeddings based on their cosine similarity with ‘thermoelectric’. In order to test this hypothesis, compounds that were mentioned in the corpus more than three times, that also appeared in a dataset of thermoelectric power factors [178], were compared with the embedding for the term ‘thermoelectric’, using the cosine similarity.

These compounds were ranked in order of the degree of cosine similarity. It was determined that the top 10 predictions of thermoelectric materials according to this ranking, which do not appear in any literature with thermoelectric keywords at any point, vastly outperform the average material’s thermoelectric performance. For these top 10 compounds based on the embedding similarity rank, the average power factor was 2.4 times larger than the average of known thermoelectric materials and 3.6 times larger than the average taken over all materials present in their database.

The final usage of these embeddings that was demonstrated as being potentially useful was in the ability of the embeddings to predict a material as thermoelectric prior to the time that it was recorded as being thermoelectric in the literature. As such, the algorithm was trained on data that was only available before that point in time. Approximately 40% of the top 50 candidate materials had been determined to be thermoelectric 18 years after the embedding would have predicted it. This test was also performed for terms such as ‘photovoltaics’, ‘topological insulators’ and ‘ferroelectric’ with similar results for each.

6.2.2 Word Embeddings for Materials Property Predictions

There are few pieces of existing literature that explicitly attempt to predict materials properties by employing word embeddings in the prediction pipeline. The original `mat2vec` paper [177], did some limited predictions of materials properties. The work employed a shallow neural network, with a single hidden layer, in order to predict the formation energies of elpasolites, with a composition of ABC_2D_6. The input features of this network were the concatenated embeddings of the elements corresponding to A, B, C and D of the ABC_2D_6 configuration. The `mat2vec` encodings outperformed an input to the network of one-hot encodings. However, embeddings trained according to the GloVe method outperformed all other methods on the prediction. This represents the first indication, in the case of the shallow neural network employed by the work, that the use of literature-based embeddings can potentially enhance the ability of a model to predict physical properties, beyond the use of simple compositional features.

Beyond the original work, the `mat2vec` embeddings were also employed in the architecture of CrabNet [179], which stands for the compositionally restricted attention-based network for materials property prediction. This work used the self-attention mechanism of transformers to train ML models to predict a range of materials properties while relying on a fractional composition-based scaling of the `mat2vec` embeddings for the input features as an efficient means to encode chemical information, beyond simple one-hot encoding methods. It was determined that the transformer self-attention representation employed within CrabNet allowed for insights into the degree to which different elements contributed to the model's predictions of various properties. This highlights the potential value of transformer-based representations in improving the interpretability of ML models for materials science. Additionally, the incorporation of `mat2vec` embeddings facilitated the encoding of some latent materials information within the input representation, thereby improving compound representations.

6.3 Contextual Embeddings in Materials Science

Historically, materials-embedded representations depended on one-hot encoding or incorporating compositional element fraction details [11, 150]. Such representations are limited in the amount of information they are capable of representing. Although there have been initial signs that unsupervised embeddings trained on materials science literature could offer value by encapsulating latent information in compound representations there is still a notable absence of dedicated research attention in this field. This gap persists despite the promising advantages of having a streamlined vector representation capable of encoding a broad spectrum of materials information for applications related to predicting materials properties. Further to this, since the state-of-the-art `mat2vec` embeddings emerged in 2019, a colossal amount of progress has been occurring in the space of general natural language techniques. This section aims to discuss the potential of transformer-based representations in capturing latent materials information, given that the pre-trained contextualized representations of Bommasani *et al.* [176] significantly outperformed both the previous state-of-the-art models for static word representations for general applications.

This will involve studying the performance of pre-trained, contextualized, static embeddings, generated from a variety of algorithms, spanning from BERT architectures, to GPT-3, in predicting the properties of materials. This will also involve a particular focus on the impact of domain-specific pre-training, used for the construction of the static embedding on the overall quality of the prediction, as a means of benchmarking the amount of useful latent information within the model parameters, which could be useful for materials property applications.

Table 6.1: Sizes of the different databases for the evaluation of the static embeddings reduced from contextual representations and their relative split in train, validation and test set.

Database	**All**	**Train**	**Validation**	**Test**
Formation Enthalpy of Elpasolites	10000	8000	500	1500
Ferromagnetic Curie Temperatures	3638	2318	410	910
Ab-Initio Calculations of Bulk Modulus	5578	3555	628	1395

The assessment of a set of embeddings' performance on materials property predictions over another was evaluated by testing the embeddings against a number of different regression tasks. Different shallow neural networks were trained within a supervised training framework to predict the formation energy of elpasolites, the Curie temperature of ferromagnetic compounds and the bulk modulus. The database used for the formation energy of elpasolites was the same one that was used to benchmark the property prediction capabilities of the `mat2vec` embeddings, arising from the work of Faber *et al.* [180]. The database of ferromagnetic Curie temperatures is the same as the one utilised in section 5.4.1, combining the databases of Nelson *et al.* [11] and Byland *et al.* [150]. Finally, the database used for the training of the model predicting the bulk modulus came from the `AFLOW` repository [37]. The respective training, validation and test set sizes of each of the three databases are summarised in Table 6.1.

For each set of embeddings considered, a different shallow neural network architecture was trained to predict the target quantity for each of the tasks considered. The performance of the validation set was employed to implement an early-stopping protocol. Models were trained with varying numbers of nodes and 10 different ones were trained for each embedding system, starting from different random initialised weights in order to account for some of the variability associated with the optimization procedure of the networks. For each of these trained architectures, three metrics were used to evaluate the predictive power, the MAE, the RMSE and the R^2 value. All neural networks were implemented using PyTorch and set to run on GPUs, using instances of Google Colab to run the calculations.

Different strategies were considered to obtain an embedding vector $\mathbf{v}_{el}$ for each element of the periodic table. Given a set of embeddings for each element of the periodic table, a set of input features is generated associated with a given compound $\mathbf{v}_{comp}$ by computing the sum of the elemental embedding, weighted with respect to the atomic fraction w_{el} of that element in the compound,

$$\mathbf{v}_{comp} = \sum_{el \,\in\, comp} w_{el}\mathbf{v}_{el}. \tag{6.9}$$

For example, the feature vector for water ($\mathbf{v}_{\mathrm{H_2O}}$) would be computed as:

$$\mathbf{v}_{\mathrm{H_2O}} = 0.6\bar{6}\mathbf{v}_{\mathrm{H}} + 0.3\bar{3}\mathbf{v}_{\mathrm{O}},$$

where $\mathbf{v}_{\mathrm{H}}$ and $\mathbf{v}_{\mathrm{O}}$ are the embeddings obtained from a language model for hydrogen and oxygen, respectively.

6.3.1 Impact of Domain-Specific Pre-Training

As described in Section 6.1.3, there are two strategies for the construction of static embeddings from contextual representations. The first involved taking a decontextualized output from a contextual model, by feeding single words into the transformer without context, and the other involved pooling the outputs for tokens of interest as they appeared in context, in sentences. In order to assess the relative impact of domain-specific pre-training on the construction of static embeddings, various contextual models were considered. In order of perceived relevance to the domain of materials science, these pre-trained systems were BERT-base [120], PubMedBERT [181], SciBERT [128] and MatSciBERT [129]. The elemental embeddings were obtained for each of these cases in the decontextualized strategy by feeding the elemental symbol through the BERT model in question and extracting a hidden layer embedding. In the original work, outlining the means of construction of such embeddings [176], it was demonstrated that the embeddings extracted from the first quarter of the transformer's layers performed significantly better on tasks than the later layers. A similar trend was observed for this work and the embeddings were extracted from the third layer for all models as a result, based on the BERT architecture considered.

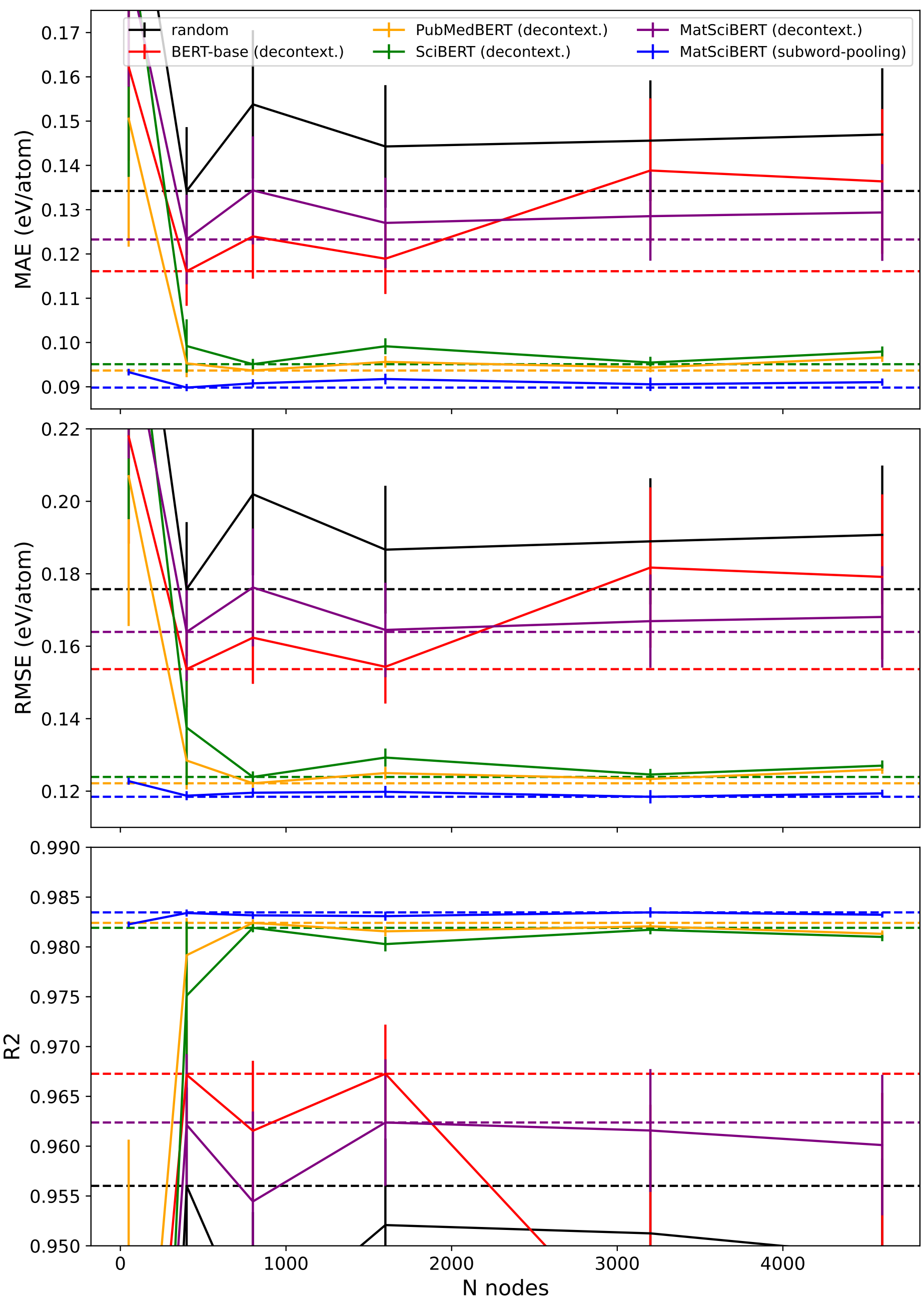

Figure 6.5: Evaluation metrics for the prediction of the formation energies of elpasolite compounds from various decontextualized BERT models, BERT with random initialized weights *(black)*, BERT-base *(red)*, PubMedBERT *(yellow)*, SciBERT *(green)*, MatSciBERT *(purple)* and context-pooled MatSciBERT *(blue)*. The top panel displays the mean average error (MAE), the middle panel displays the root mean squared error (RMSE) and the bottom panel displays the R^2 value of the task against the number of nodes in the neural network. The dashed line represents the best-performing case. The error bars show the standard deviation of the evaluation metrics over multiple restarts with different initial weights.

The best-performing of these decontextualized embeddings, MatSciBERT, was chosen as the best option with which to construct the aggregated, mean-pooled embeddings, which is the second strategy for constructing these static embeddings. A total of 2.2 million cross-domain abstracts were obtained from the arXiv e-print repository [144]. These abstracts were then processed according to the same processing as was implemented when the `mat2vec` embeddings were created, in order to extract sentences from the corpus that contain a chemical entity. From this, embeddings were constructed by mean-pooling the token embeddings from the third hidden layer of MatSciBERT as the chemical compound appeared in each sentence. It should be noted at this point that the efficacy of this method of construction of the embeddings has not reached the critical number of training points for the embedding that would allow it to outperform the Word2Vec method. This is about 100,000 training data points per representation according to Ref. [176]. For example, in the arXiv dataset, the most common element, hydrogen, appears in the database of 2.2 million abstracts only $\sim 39,000$ times in contrast with the 100,000 quantity needed. The rarest of heavy-earth elements only appear in the corpus a handful of times. The performance of this embedding strategy is highly likely to improve further with access to a greater magnitude of literature. Indeed, in the original paper, contextual representations that were aggregated from 10,000 contexts, which is closer to the magnitude of contexts available from the dataset, were outperformed by Word2Vec embeddings on all tasks. In order to obtain a baseline with which to compare the performance of BERT models, a BERT system was initialised with randomised initial weights. This is referred to as 'random' throughout this section.

The evaluation metrics of the neural networks trained for each of the extracted embeddings for the different BERT models are reported in Fig. 6.5 for the case of the prediction of the formation enthalpy of elpasolite compounds. Since the best-performing case according to every metric is the context-pooled MatSciBERT representation, it is clear that the context-pooling of the representation enhances the embedding's ability to capture information pertinent to the neural network's prediction of the formation enthalpy of the compounds. This is made particularly stark when one considers that the decontextualized MatSciBERT embedding is, in fact, the worst-performing embedding aside from the random model, which essentially only learns the fractional composition of the elements in the compound, given the lack of relevant information contained within the model weights. Furthermore, this implies that in pre-training the weights such that it has a better contextual understanding of materials science, it negatively impacts the capability of the model to produce good decontextualized representations. In other words, the MatSciBERT pre-training has made the decontextualized representation worse than the out-of-the-box BERT-base transformer. This is particularly surprising given that other models, pre-trained in other scientific domains, actually outperform the materials-science-specific one and BERT-base, with SciBERT and PubMedBERT performing almost as well as the context-pooled MatSciBERT embeddings.

The improvement in the performance of the context-pooled, aggregated approach, in particular when considering the negative impact of the domain-specific pre-training for the decontextualized representations is likely down to the compensation for this decontextualization that occurs when the contextual representations are pooled. The model's weakness in representing a decontextualized element could well be related to the prevalence of individual elements mainly appearing as dopants in material specifically related to materials science as opposed to in more general texts where they are more likely to appear in general descriptions of the properties of the element in question when they are referenced. This bias may well appear in the model's weights after fine-tuning leading to an increase in noise when attempting to extract meaning from what essentially amounts to querying the model's weights as it attempts to ascribe contextual meaning to a lone token.

This hypothesis can be further developed when observing the results for the prediction of the T_C of the compounds, which can be seen in Fig. 6.6. Interestingly, in this case, according to two out of three metrics, the decontextualized embeddings outperform the aggregated embeddings, with both sets of embeddings representing the best-performing sets. There is a good likelihood that this stark change in the performance of the embedding is once again related to the context in which these tokens would have appeared in the pre-training corpus of MatSciBERT, with the dopant being far more correlated to the Curie temperature and magnetic properties of a system than it would the formation enthalpy of an elpasolite compound. In fact, with the decontextualized embeddings improving over the context-pooled embeddings (according to RMSE and R^2 coefficient), it implies that the decontextualized embedding has a stronger correlation with the embeddings's understanding of magnetism and produces fewer outliers than the context-pooled example. However, the context-pooled

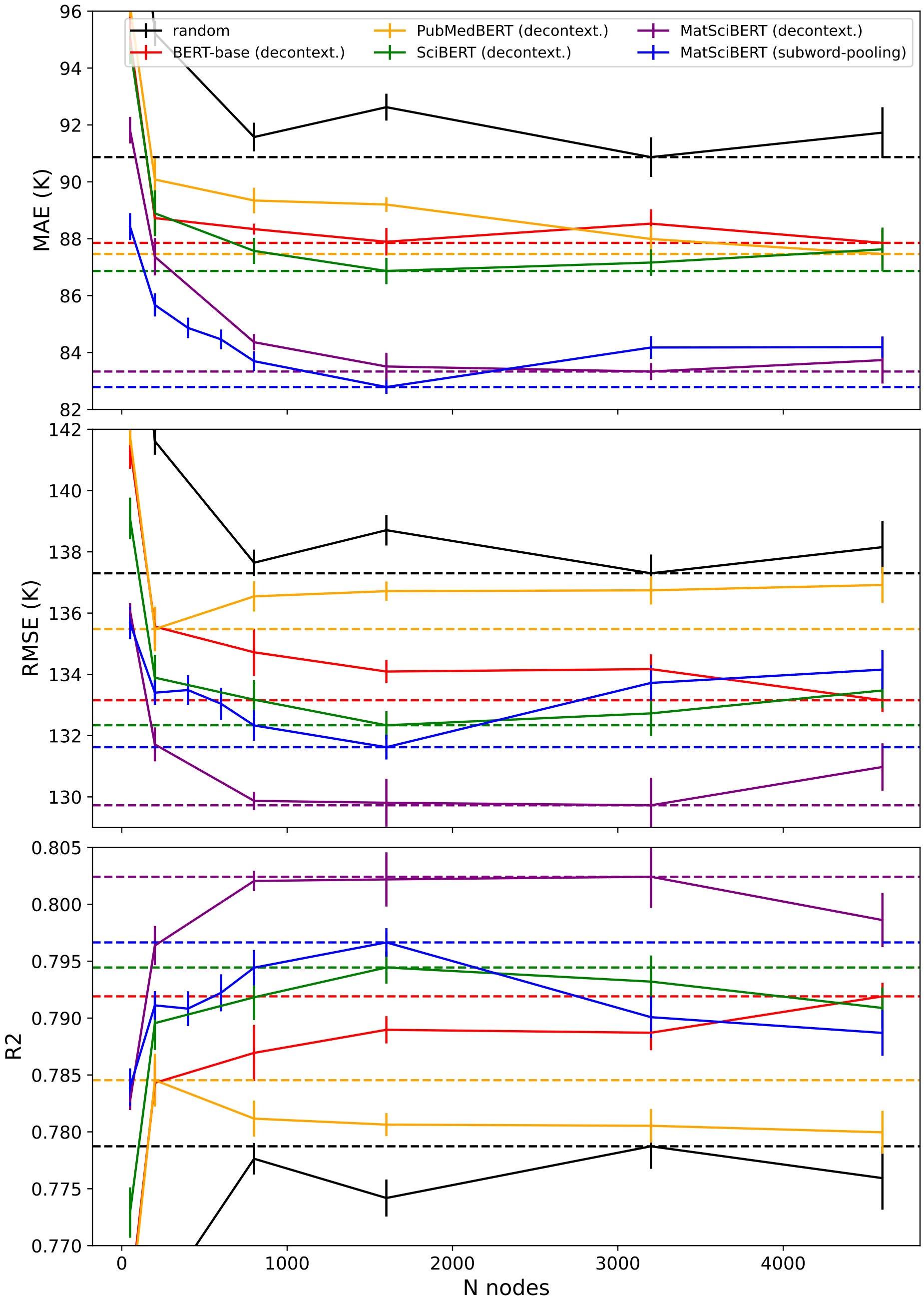

Figure 6.6: Evaluation metrics for the prediction of the Curie temperature of ferromagnetic compounds from various decontextualized BERT transformed representations, BERT with randomly initialized weights *(black)*, BERT-base *(red)*, PubMedBERT *(yellow)*, SciBERT *(green)*, MatSciBERT *(purple)* and context-pooled MatSciBERT *(blue)*. The top panel displays the mean average error (MAE), the middle panel displays the root mean squared error (RMSE) and the bottom panel displays the R^2 value of the task against the number of nodes in the neural network. The dashed line represents the best-performing case for each. The error bars show the standard deviation of the evaluation metrics over multiple restarts with different initial weights.

embedding still manages to outperform the decontextualized representation according to the model's MAE. Generally speaking, the domain-specific pre-training does enhance the ability of the system to predict the Curie temperature of the compounds in the database, with both of the embeddings obtained from the domain-specific models outperforming all other embeddings obtained from models focused on different domains. In this case, pre-training in the biomedical domain detrimentally impacted the transformer's ability to obtain embeddings containing latent information about elemental relevance to magnetism, with the embeddings obtained from PubMedBERT performing better than random but worse than the general BERT-base. The general scientific BERT, SciBERT also outperformed BERT-base, indicating that a more general system, that is still pre-trained in a relevant domain can perform better than a completely general one at predicting the Curie temperature.

In order to account for the different behaviour of the embeddings when being used to describe the different property predictions, the nature of the pre-training data is clearly the primary consideration. In the former case, the transformers that were pre-trained on more chemistry-focused texts, SciBERT and PubMedBERT were only outperformed by a model that had used aggregated representations from literature from the arXiv, which is likely to contain ample information to aid the formation enthalpy. This makes sense given the importance of formation enthalpy to synthetic chemistry. In contrast, the magnetic properties of materials were more precisely predicted by including representations that were likely trained on more data pertinent to magnetism, the transformer with the least relevance to magnetism, PubMedBERT, predictably performed the worst of the domain-specific cases.

The evaluation metrics for the prediction of the bulk modulus of compounds using the various, contextual embeddings can be seen in Fig. 6.7, which presents some very stark results in comparison with the previous examples. Once again, the embeddings resulting from systems, which have been pre-trained on literature relevant to materials science, outperforms all others. The aggregated strategy, once again, appears to be the optimal strategy for the creation of such embeddings for smaller neural network sizes. This is not the case when scaling up to larger network sizes, however, with the decontextualized strategy performing better with larger networks. Notably, for the case of the bulk modulus, every model that was not trained on literature in the domain of materials science produced embeddings that detrimentally impacted the model's ability to predict the bulk modulus accurately, with every such model performing worse

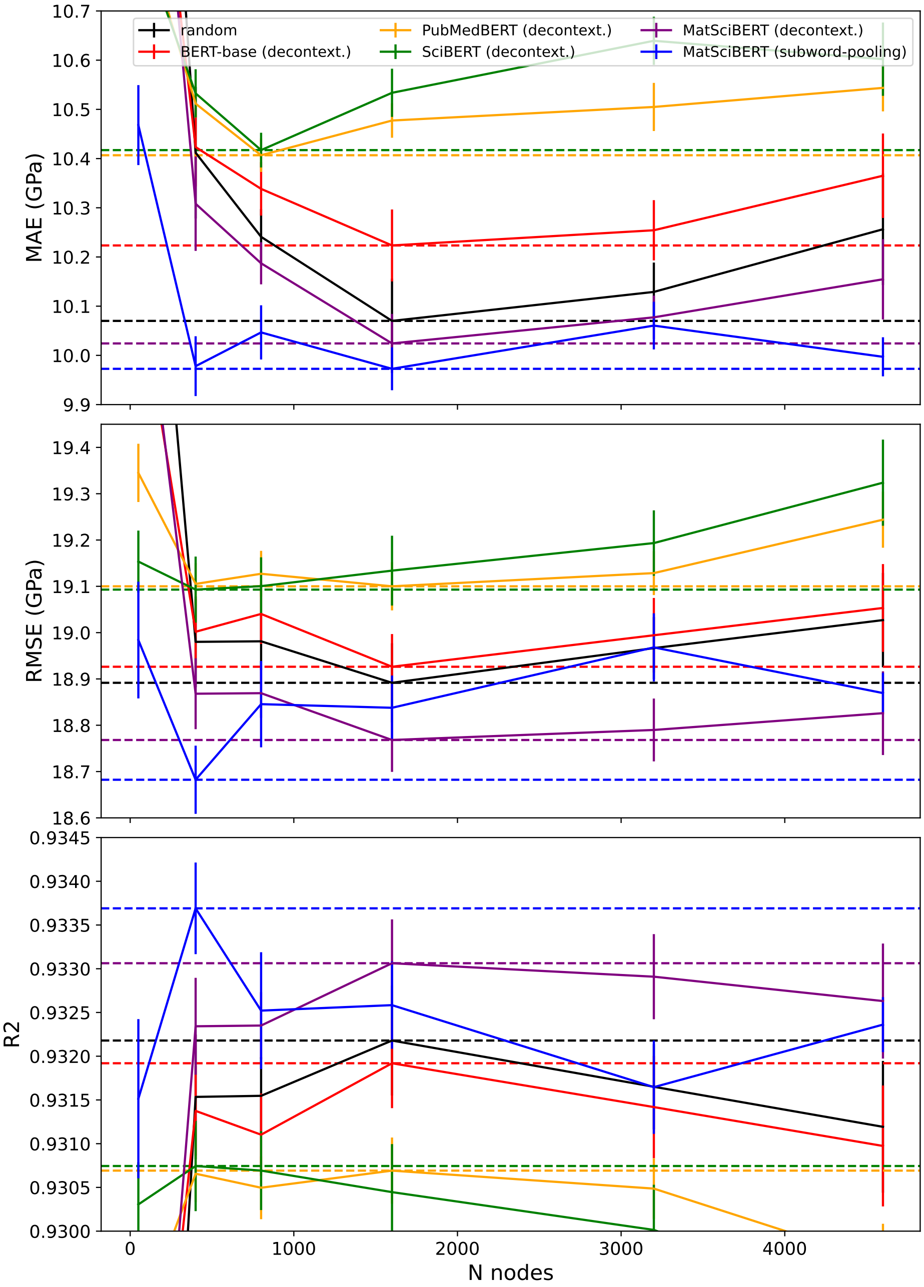

Figure 6.7: Evaluation metrics for the prediction of the bulk modulus of compounds from various decontextualized BERT representations, BERT with random initialized weights *(black)*, BERT-base *(red)*, PubMedBERT *(yellow)*, SciBERT *(green)*, MatSciBERT *(purple)* and context-pooled MatSciBERT *(blue)*. The top panel displays the mean average error (MAE), the middle panel displays the root mean squared error (RMSE) and the bottom panel displays the R^2 value of the task against the number of nodes in the neural network. The dashed line represents the best-performing case for each. The error bars show the standard deviation of the evaluation metrics over multiple restarts with different initial weights.

than even the randomly initialised weights. This suggests that the use of such embeddings actively misleads the predictive model when attempting to predict structural information about a given compound.

The impact of choosing a transformer representation based on the relevancy of the pre-training corpus on constructing is evident based on the trends observed from this analysis. The choice of optimal model for obtaining embeddings is very property-dependent and can be seen to be dependent on the relevancy of different properties to the domain on which the model was trained. The MatSciBERT aggregated embeddings performed consistently well, even with the lack of sentences available in the embedding construction corpus, performing the best out of all BERT cases according to nearly every metric.

6.3.2 Comparison of Contextual Embeddings

Once the value of domain-specific pre-training of the language models used for the construction of the embeddings had been established, the best-performing embeddings were compared with a variety of other models. First, decontextualized embeddings were obtained for the LLM architecture, GPT-3 [123], using the openAI API from the only model capable of producing embeddings, *'text-embeddings-ada-002'*. This test was deemed interesting to compare the capacity of a generative AI with a massive number of parameters to capture elemental information, despite the lack of specific training such a system has in materials science, physics or chemistry. The contextualized embeddings were also compared with the `mat2vec` embeddings, which had previously been shown to have good performance as chemical compound descriptors.

Fig. 6.8 shows all the comparisons of evaluation metrics for the best performing MatSciBERT-based embeddings, those constructed using GPT-3 and those resulting from `mat2vec`. It is evident that the massive LLM, GPT-3 contains sufficient information within the model weights to outperform the domain-specific BERT examples in constructing contextual embeddings effectively. This captures sufficient information on the elemental contributions from the model weights for it to yield a meaningful improvement over the MatSciBERT aggregated and decontextualized models. The only exception to this trend can be observed in the bottom two panels of Fig. 6.8 (b), which displays the RMSE and R^2 coefficient for the Curie temperature prediction. In this case, the decontextualized BERT was able to perform surprisingly well according to these metrics.

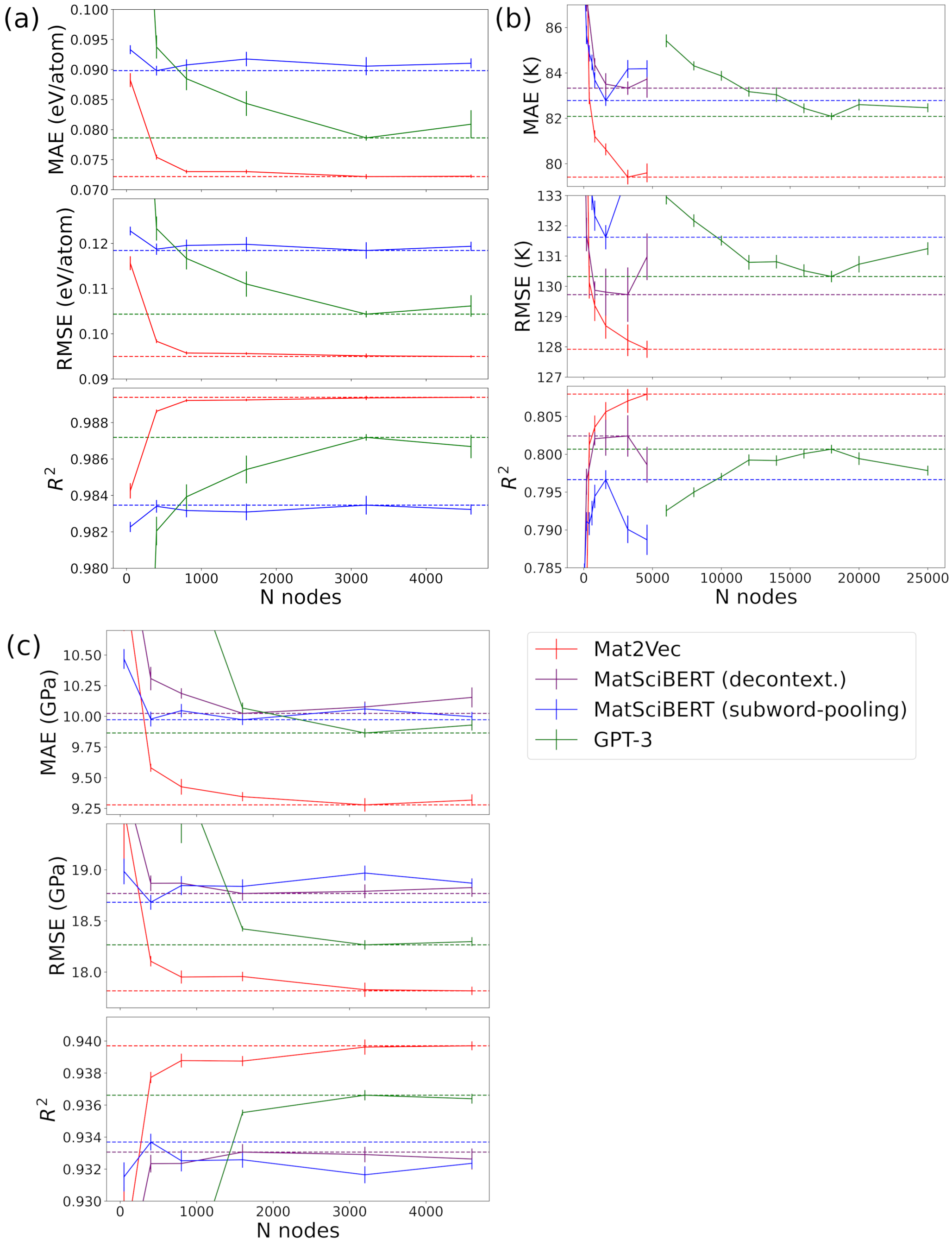

Figure 6.8: Evaluation metrics for the prediction of (a) the formation energies of elpasolite compounds, (b) the Curie temperature of ferromagnetic compounds, and (c) the bulk modulus of compounds with a shallow neural network using descriptors designed from the embeddings of `mat2vec` as input features *(red)*, the context-pooled MatSciBERT representation *(blue)* and GPT-3 *(green)*. The top panel displays the mean average error (MAE), the middle panel displays the root mean squared error (RMSE) and the bottom panel displays the R^2 value of the task against the number of nodes in the neural network. The dashed line represents the best-performing case for each. The error bars show the standard deviation of the evaluation metrics over multiple restarts with different initial weights.

Most importantly, the `mat2vec`-based embeddings perform best according to all metrics across the board and continue as of now to represent the optimal representation for capturing chemical information, constructed from literature. It is important to note that this status may evolve when a new corpus of training data becomes available, enabling the construction of an aggregated representation with ample contextual sentences for a critical number of elements. Such a development could pave the way for a next-generation literature-based representation, further enhancing predictive models for materials property analysis.

6.4 Summary & Conclusions

In this chapter, the potential of literature-based representations in distilling a large amount of complex materials information in a lightweight and efficient representation was investigated. Much of the earlier part of the chapter was focused on the fundamentals of representing concepts and entities in a geometric space such that meaning can be ascribed to the relative positions of these entities in the embedding space.

The balance required between preserving local and global statistical information was emphasised and methods through which to achieve such a balance were demonstrated and discussed. The state-of-the-art static embedding representations were brought from first principles, right up to the most recent strategies of pooling representations based on contextual representations in order to construct static embeddings.

The limited number of previous applications of this subject area of materials science was described and the methods used to construct static embeddings capturing latent information about materials science was discussed. An assessment of the capabilities of domain-specific pre-training to improve performance for the construction of static embeddings from contextual representations in materials science was performed for the first time, proving the value of such pre-training for materials science applications. Further to this, an analysis of the relative value of the use of latent information from the pre-trained weights of LLMs for forming similar contextual embeddings was performed. This led to an improvement in model performance over BERT-based methods.

Although existing `mat2vec` embeddings currently outperform static embeddings based on contextual representations in materials property prediction, this could be due to the lack of sufficient contextual data to construct the most powerful aggregated embeddings. Future research could reveal the potential of such strategies as more data becomes available, opening doors to powerful embeddings yet to be explored in the limited data regime. Such embeddings could be the key to incorporating difficult-to-represent physical information about compounds or chemical behaviours into lightweight representations that are useful for rapid property prediction. Such a tool would further enhance our ability to execute the first screening step of an inverse-design workflow.

Chapter 7

Conclusions and Future Work

"Is fearr súil romhat ná dhá shúil i do dhiaidh."

This thesis has explored and developed a wide range of techniques that have direct applications to each stage of the inverse materials design workflow. Of the three main stages of a computational screening for materials targeting specific properties, the techniques outlined in this work can both enhance and act as the foundations for the first two of these three stages. These stages are the rapid property screening step and the screening for materials stability. This is a desirable outcome of this work considering that the final stage in the computational workflow, the *ab initio* analysis is a field of research, which has entered into maturity and has already demonstrated its use for accurate property calculations from first-principles.

Chapter 2 gave an overview of the key concepts that have allowed density functional theory to be a fundamental component in materials understanding for decades. Further to this, these methods were elaborated upon exclusively for the case of phonon and vibrational properties of materials in Chapter 3. To that end, a theoretical background for phonons was outlined and different methods of obtaining the phononic properties of materials were presented. Certain cases of the methodologies outlined could be used in conjunction with other methods for obtaining the energy and forces of materials, namely that of the finite-displacement method, provided sufficient accuracy. Density functional perturbation theory (DFPT) was also outlined, which could not be used with other total-energy or force calculations, however, this theory was employed extensively for the case of monolayer 2H-NbS$_2$, which was demonstrated to exhibit a dynamical instability as a result of long-range effects when the phonon dispersion was calculated in the harmonic approximation regime. DFPT was also used to calculate benchmarks for a variety of other layered 2D materials, which would be used for comparisons in Chapter 4.

Chapter 2 also gave an outline and a description of the foundations of the field of machine learning (ML), a central theme of this thesis. ML was used extensively throughout this work as the key for bypassing computationally intensive steps that would otherwise be necessary for the creation of a robust inverse-design workflow. The concept of supervised learning was expounded in detail along with key evaluation metrics that are used to verify the efficacy of any supervised ML process. A brief description of unsupervised ML was presented, followed by a discussion of neural networks, which were valuable to understand to clarify the underlying concept of similar networks, such as transformers, another key concept in the later chapters of this thesis.

Once the common theoretical background for the chapters had been established in Chapter 2, and benchmarks had been calculated using the work in Chapter 3, Chapter 4 was the first that concerned the development of tools to aid in the first two stages of the inverse-design workflow. A massive playground of materials for the construction of composite heterostructures, which exhibit different properties to their constituent materials was identified in the family of 2D materials. A library of SNAP potentials was created for 61 non-magnetic 2D materials, encompassing graphene and hBN, along with 59 hexagonal monolayers with a XY_2 composition. These potentials were extended to account for interlayer van der Waal's interactions using a parameterised Lennard-Jones potential. The efficacy of using the resulting potentials for rapidly generating phonon dispersions was demonstrated for both the monolayer and multilayer case. This result constituted the first example of a stage of the inverse-design workflow being enhanced through the use of ML, as this offers a means of rapidly assessing the dynamical stability of a system, with a massive reduction in the computational intensity relative to *ab initio* methods.

The secondary value of these potentials was also demonstrated through the calculation of the thermal conductivity of monolayer and multilayer systems, along with the interfacial thermal conductance of a massive multilayer system, a calculation which would have been infeasible with *ab initio* methods. For the case of thermal conductivity, the ML potential was used to calculate both the second and third-order force-constant matrices, which were subsequently used to determine the direct solution of the linearized phonon Boltzmann equation. The calculated values for the thermal conductivity of the monolayer and heterostructure systems agreed well with both experiment and DFT, with the associated drastic increase in efficiency from the use of the ML potentials as a surrogate for DFT-based methods. Similarly, the interfacial thermal conductance agrees well with that which was obtained in experiment.

This result constitutes a scratching of the surface of the resulting potential of these methods for the rapid prediction of the thermal properties of heterostructure systems. There is an increase in efficiency with the use of these methods. This improvement, however, is not conducive to executing a massive search of the property space for an ideal candidate for a given thermal application. The value of this method is not in its ability to rapidly predict the thermal conductivity or conductance, rather, it is in the ability of these potentials to generate a large-scale database of properties such that the property space for the thermal properties of the composite systems can be reasonably sampled.

Thus, the next direction in which I would take this particular aspect of the project is in the construction of such a database through the systematic construction of heterostructure compounds with the SNAP potentials and the parameterised Lennard-Jones interactions, thus sampling this property space. This space would be rich in the potential to identify candidates for a targeted quantity. In order to construct an ML model to search this space, there are three main considerations, which will have to be taken into account. The first of these is the number of layers of the composite material. The second consideration is that of the order in which the monolayers appear in the heterostructure. The final consideration is that of the composition of each of the layers for the predicted compound. In this instance, the structure is not a consideration as the structures of these particular monolayers are all hexagonal. Therefore, there is no reason to include the structure in the model features. In the longer term, the number of monolayer potentials in the library could be expanded to include monolayers with different structures or, indeed, to include a consideration of magnetic effects, potentially necessitating a change to the feature space described. This work would represent the optimal means of executing the first stage of the inverse-design workflow outlined and the work to that end in this thesis constitutes a valuable first step on that road.

The secondary branch of the research that I performed over the course of this Ph. D is that which involves an application of natural language processing (NLP) to the domain of materials science. Chapter 5 represents an attempt at entirely bypassing the sorts of calculations of properties that I outlined in Chapter 4 to construct a property space for a given class of materials and, instead, isolating the databases directly from scientific literature. This was achieved by leveraging the power of pre-trained transformer networks, a next-generation tool for language modelling, for the purposes of the isolation of arbitrary combinations of compound-property relations, with an attempt at minimizing the amount of human intervention that was necessary in the conception of such models. Two databases were constructed with this transformer-based pipeline, one containing compound-Curie temperature relations, and the other composed of compound-electronic band gap relations. The quality of these databases was compared to those obtained with the previous state-of-the-art in automated database construction, ChemDataExtractor, a grammar rules-based method. The databases were shown to be of almost exactly equivalent quality and the automated transformer databases were used in order to construct a predictor for rapid Curie temperature prediction, which was shown to be able to screen for high-Curie temperature compounds with a remarkably high precision. This fact indicates the clear value of literature-extracted databases for the execution of the rapid property screening stage in the inverse-design workflow to narrow down compositions likely to exhibit the targeted property.

In the future, I would like to expand on the progress that this work represents by including newer-generation, pre-trained transformers, such as a massive pre-trained GPT-based model. The inclusion of such models could ensure the leveraging of their immense power of natural language understanding to improve the efficacy of database extraction. A different workflow for the construction of such a pipeline must be conceived and, therefore, significant work would need to be done to overcome limitations, which are inherent in generative models. Addressing these limitations would be the first step on the road to a more robust workflow. Further along the line, once these limitations have been dealt with, I would like to construct expansive, unified databases of materials properties that have resolved interdependencies between various material phases and their respective properties. Such a database would be an immensely powerful tool for the domain of materials science and would represent a massive step forward in the ability to construct simple, but efficient models for the rapid screening of materials properties for desired applications. Thus, the ability to execute the first stage of the inverse-materials design workflow would, in essence, be unlocked for arbitrary materials properties. Much work, however, is left to do to achieve this aim. Not least of which is gathering more high-quality sources in order to generate these large-scale databases. The open-access movement is, therefore, a valuable ally in this plight.

The final chapter of the results achieved as part of this research is an attempt at unifying these seemingly disparate methodologies, such that the construction of a ML representation for physical compounds, directly using literature can be achieved. This is discussed in detail in Chapter 6. These representations, in essence, embed latent property information into the representation, such that similar compounds, within the literature representation space, are assigned similar vector representations. Such representations could potentially have the ability to, once again, improve the quality of predictions that are made in the rapid property screening step of the inverse-design workflow. These new representations, once again, leverage the superior contextual representative ability of transformer networks. Two strategies were used for the construction of these contextual representations. The first involved extracting the hidden layer representation for elements directly as a result of their input into the model without context. The second was pooling the contextual representations such that the resulting representation aggregated contextual understandings of the term from the sentences it appeared in. Domain-specific pre-training of the transformer networks was shown to improve the representative ability of the decontextualized representations and the pooled representations were generally shown to improve performance again. A critical lack of sufficient data with which to train the pooled representations meant that the ability to generate optimal context-pooled representations was limited. They, therefore, are still not able to outperform similar models for constructing representations from literature. They do, however, show a large degree of promise.

Thus, the future of this particular research area is clear. More data must be gathered with which to pool the context of the elemental embeddings. The contextual representations, beyond the domain of materials science, were shown to improve in their representative ability after several hundred thousand contextual examples, which, therefore, must be the aim to create superior literature-based representations from transformers. Beyond this, using the same representation construction strategy using larger-scale GPT-based models could also prove to be powerful once such models become more generally available. Such a direction could be a promising one to take in the future.

In conclusion, a suite of new techniques and new applications of old techniques have been conceived, all of which have shown to be of value to an inverse-materials design pipeline. These methodologies could prove to be the foundations of future innovation in technological domains, which are of immense value to the development of a sustainable and resilient future for the collective good. It is clear that machine learning and its future directions will prove to play a pivotal role in accelerating the discovery of new materials.

Bibliography

[1] Rossignol, H., Minotakis, M., Cobelli, M. & Sanvito, S. Machine-learning-assisted construction of ternary convex hull diagrams. *arXiv preprint arXiv:2308.15907* (2023).

[2] Minotakis, M., Rossignol, H., Cobelli, M. & Sanvito, S. Machine-learning surrogate model for accelerating the search of stable ternary alloys. *Physical Review Materials* **7**, 093802 (2023).

[3] Mitchell, T. M. *Machine Learning* (McGraw-Hill Science/Engineering/Math, 1997).

[4] Nassif, A. B., Shahin, I., Attili, I., Azzeh, M. & Shaalan, K. Speech recognition using deep neural networks: A systematic review. *IEEE Access* **7**, 19143–19165 (2019).

[5] Erickson, B. J., Korfiatis, P., Akkus, Z. & Kline, T. L. Machine learning for medical imaging. *Radiographics* **37**, 505–515 (2017).

[6] Bojarski, M. *et al.* End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316* (2016).

[7] Khan, A. A., Laghari, A. A. & Awan, S. A. Machine learning in computer vision: a review. *EAI Endorsed Transactions on Scalable Information Systems* **8**, e4–e4 (2021).

[8] Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2 (Springer, 2009).

[9] Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).

[10] Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).

[11] Nelson, J. & Sanvito, S. Predicting the curie temperature of ferromagnets using machine learning. *Physical Review Materials* **3**, 104405 (2019).

[12] Fukushima, K. Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics* **20**, 121–136 (1975).

[13] Maas, A. L., Hannun, A. Y., Ng, A. Y. *et al.* Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th International Conference on Machine Learning*, vol. 30, 3 (2013).

[14] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15**, 1929–1958 (2014).

[15] LeCun, Y., Bengio, Y. *et al.* Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks* **3361**, 1995 (1995).

[16] Li, X. & Wu, X. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4520–4524 (IEEE, 2015).

[17] Lyu, C., Chen, B., Ren, Y. & Ji, D. Long short-term memory rnn for biomedical named entity recognition. *BMC Bioinformatics* **18**, 1–11 (2017).

[18] Martin, R. M. *Electronic Structure: Basic Theory and Practical Methods* (Cambridge University Press, 2020).

[19] Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Physical Review* **136**, B864 (1964).

[20] Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Physical Review* **140**, A1133 (1965).

[21] Perdew, J. P. & Schmidt, K. Jacob's ladder of density functional approximations for the exchange-correlation energy. In *AIP Conference Proceedings*, vol. 577, 1–20 (American Institute of Physics, 2001).

[22] Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters* **77**, 3865 (1996).

[23] Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened coulomb potential. *The Journal of Chemical Physics* **118**, 8207–8215 (2003).

[24] Krukau, A. V., Vydrov, O. A., Izmaylov, A. F. & Scuseria, G. E. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *The Journal of Chemical Physics* **125**, 224106 (2006).

[25] Hartree, D. R. The wave mechanics of an atom with a non-Coulomb central field. Parts I, II, III. In *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, 89–110,111,426 (Cambridge University Press, 1928).

[26] Fock, V. Näherungsmethode zur lösung des quantenmechanischen mehrkörperproblems. *Zeitschrift für Physik* **61**, 126–148 (1930).

[27] Tao, J., Perdew, J. P., Staroverov, V. N. & Scuseria, G. E. Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids. *Physical Review Letters* **91**, 146401 (2003).

[28] Bohm, D. & Pines, D. A collective description of electron interactions: III. Coulomb interactions in a degenerate electron gas. *Physical Review* **92**, 609 (1953).

[29] Blöchl, P. E. Projector augmented-wave method. *Physical Review B* **50**, 17953 (1994).

[30] Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Physical Review B* **47**, 558 (1993).

[31] Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B* **54**, 11169 (1996).

[32] Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science* **6**, 15–50 (1996).

[33] Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B* **59**, 1758 (1999).

[34] Ziman, J. M. *Electrons and Phonons: The Theory of Transport Phenomena in Solids* (Oxford University Press, 2001).

[35] Togo, A. First-principles phonon calculations with phonopy and phono3py. *Journal of the Physical Society of Japan* **92**, 012001 (2023).

[36] Dove, M. T. *Introduction to Lattice Dynamics.* No. 4 in Cambridge Topics in Mineral Physics and Chemistry (Cambridge University Press, 1993).

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

[38] Taylor, R. H. *et al.* A RESTful API for exchanging materials data in the AFLOWLIB.org consortium. *Computational Materials Science* **93**, 178–192 (2014).

[39] Baroni, S., De Gironcoli, S., Dal Corso, A. & Giannozzi, P. Phonons and related crystal properties from density-functional perturbation theory. *Reviews of Modern Physics* **73**, 515 (2001).

[40] Feynman, R. P. Forces in molecules. *Physical Review* **56**, 340 (1939).

[41] Novoselov, K. S. *et al.* Electric field effect in atomically thin carbon films. *Science* **306**, 666–669 (2004).

[42] Geim, A. K. & Novoselov, K. S. The rise of graphene. *Nature Materials* **6**, 183–191 (2007).

[43] Chen, S. *et al.* Raman measurements of thermal transport in suspended monolayer graphene of variable sizes in vacuum and gaseous environments. *ACS Nano* **5**, 321–328 (2011).

[44] Wilson, J. A., Di Salvo, F. & Mahajan, S. Charge-density waves and superlattices in the metallic layered transition metal dichalcogenides. *Advances in Physics* **24**, 117–201 (1975).

[45] Calandra, M. Phonon-assisted magnetic mott-insulating state in the charge density wave phase of single-layer 1T-$NbSe_2$. *Physical Review Letters* **121**, 026401 (2018).

[46] Heil, C. *et al.* Origin of superconductivity and latent charge density wave in NbS_2. *Physical Review Letters* **119**, 087003 (2017).

[47] Zhao, S. *et al.* Two-dimensional metallic NbS_2: growth, optical identification and transport properties. *2D Materials* **3**, 025027 (2016).

[48] Guillamón, I. *et al.* Superconducting density of states and vortex cores of 2H-NbS_2. *Physical Review Letters* **101**, 166407 (2008).

[49] Fisher, W. G. & Sienko, M. Stoichiometry, structure, and physical properties of niobium disulfide. *Inorganic Chemistry* **19**, 39–43 (1980).

[50] Izawa, K. *et al.* A new approach for the synthesis of layered niobium sulfide and restacking route of NbS_2 nanosheet. *Journal of Solid State Chemistry* **181**, 319–324 (2008).

[51] van Loon, E. G., Rösner, M., Schönhoff, G., Katsnelson, M. I. & Wehling, T. O. Competing coulomb and electron–phonon interactions in NbS_2. *npj Quantum Materials* **3**, 32 (2018).

[52] Naito, M. & Tanaka, S. Electrical transport properties in 2H-NbS_2, -$NbSe_2$, -TaS_2 and -$TaSe_2$. *Journal of the Physical Society of Japan* **51**, 219–227 (1982).

[53] Leroux, M. *et al.* Anharmonic suppression of charge density waves in 2H-NbS_2. *Physical Review B* **86**, 155125 (2012).

[54] Güller, F., Vildosola, V. L. & Llois, A. M. Spin density wave instabilities in the NbS_2 monolayer. *Physical Review B* **93**, 094434 (2016).

[55] Bianco, R., Errea, I., Monacelli, L., Calandra, M. & Mauri, F. Quantum enhancement of charge density wave in NbS_2 in the two-dimensional limit. *Nano Letters* **19**, 3098–3103 (2019).

[56] Mounet, N. *et al.* Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nature Nanotechnology* **13**, 246–252 (2018).

[57] Cai, Q. *et al.* High thermal conductivity of high-quality monolayer boron nitride and its thermal expansion. *Science Advances* **5**, eaav0129 (2019).

[58] Buckley, D. *et al.* Anomalous low thermal conductivity of atomically thin InSe probed by scanning thermal microscopy. *Advanced Functional Materials* **31**, 2008967 (2021).

[59] Rongione, N. A. *et al.* High-performance solution-processable flexible SnSe nanosheet films for lower grade waste heat recovery. *Advanced Electronic Materials* **5**, 1800774 (2019).

[60] Li, D. *et al.* Recent progress of two-dimensional thermoelectric materials. *Nano-Micro Letters* **12**, 1–40 (2020).

[61] Deringer, V. L., Caro, M. A. & Csányi, G. Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials* **31**, 1902765 (2019).

[62] Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters* **98**, 146401 (2007).

[63] Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters* **104**, 136403 (2010).

[64] Varshalovich, D. A., Moskalev, A. N. & Khersonskii, V. K. *Quantum Theory of Angular Momentum* (World Scientific, 1988).

[65] Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Physical Review B* **87**, 184115 (2013).

[66] Thompson, A. P., Swiler, L. P., Trott, C. R., Foiles, S. M. & Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *Journal of Computational Physics* **285**, 316–330 (2015).

[67] Talirz, L. *et al.* Materials Cloud, a platform for open computational science. *Scientific Data* **7**, 299 (2020).

[68] Thompson, A. P. *et al.* LAMMPS-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications* **271**, 108171 (2022).

[69] Domina, M., Cobelli, M. & Sanvito, S. Spectral neighbor representation for vector fields: Machine learning potentials including spin. *Physical Review B* **105**, 214439 (2022).

[70] Lennard-Jones, J. E. Cohesion. In *Proceedings of the Physical Society*, vol. 43, 461 (IOP Publishing, 1931).

[71] Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry* **27**, 1787–1799 (2006).

[72] Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* **132**, 154104 (2010).

[73] Tkatchenko, A. & Scheffler, M. Accurate molecular van der waals interactions from ground-state electron density and free-atom reference data. *Physical Review Letters* **102**, 073005 (2009).

[74] Blum, V. *et al.* Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications* **180**, 2175–2196 (2009).

[75] Fang, Z. *et al.* Interlayer binding energy of hexagonal MoS_2 as determined by an in situ peeling-to-fracture method. *The Journal of Physical Chemistry C* **124**, 23419–23425 (2020).

[76] Green, M. S. Markoff random processes and the statistical mechanics of time-dependent phenomena. *The Journal of Chemical Physics* **20**, 1281–1295 (1952).

[77] Green, M. S. Markoff random processes and the statistical mechanics of time-dependent phenomena. II. Irreversible processes in fluids. *The Journal of Chemical Physics* **22**, 398–413 (1954).

[78] Kubo, R. Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems. *Journal of the Physical Society of Japan* **12**, 570–586 (1957).

[79] Kubo, R., Yokota, M. & Nakajima, S. Statistical-mechanical theory of irreversible processes. II. Response to thermal disturbance. *Journal of the Physical Society of Japan* **12**, 1203–1211 (1957).

[80] Togo, A., Chaput, L. & Tanaka, I. Distributions of phonon lifetimes in brillouin zones. *Physical Review B* **91**, 094306 (2015).

[81] Larsen, A. H. *et al.* The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter* **29**, 273002 (2017).

[82] Chaput, L. Direct solution to the linearized phonon boltzmann equation. *Physical Review Letters* **110**, 265506 (2013).

[83] Gu, X. & Yang, R. Phonon transport in single-layer transition metal dichalcogenides: A first-principles study. *Applied Physics Letters* **105**, 131903 (2014).

[84] Zhang, X. *et al.* Measurement of lateral and interfacial thermal conductivity of single-and bilayer MoS_2 and $MoSe_2$ using refined optothermal raman technique. *ACS Applied Materials & Interfaces* **7**, 25923–25929 (2015).

[85] Peimyoo, N. *et al.* Thermal conductivity determination of suspended mono-and bilayer WS_2 by raman spectroscopy. *Nano Research* **8**, 1210–1221 (2015).

[86] Yu, Y., Minhaj, T., Huang, L., Yu, Y. & Cao, L. In-plane and interfacial thermal conduction of two-dimensional transition-metal dichalcogenides. *Physical Review Applied* **13**, 034059 (2020).

[87] Sang, Y. *et al.* Measurement of thermal conductivity of suspended and supported single-layer WS_2 using micro-photoluminescence spectroscopy. *The Journal of Physical Chemistry C* **126**, 6637–6645 (2022).

[88] Zhang, W., Yang, J.-Y. & Liu, L. Strong interfacial interactions induced a large reduction in lateral thermal conductivity of transition-metal dichalcogenide superlattices. *RSC Advances* **9**, 1387–1393 (2019).

[89] Nika, D., Pokatilov, E., Askerov, A. & Balandin, A. A. Phonon thermal conduction in graphene: Role of umklapp and edge roughness scattering. *Physical Review B* **79**, 155413 (2009).

[90] Ghosh, S. *et al.* Extremely high thermal conductivity of graphene: Prospects for thermal management applications in nanoelectronic circuits. *Applied Physics Letters* **92**, 151911 (2008).

[91] Balandin, A. A. *et al.* Superior thermal conductivity of single-layer graphene. *Nano Letters* **8**, 902–907 (2008).

[92] Xu, X. *et al.* Length-dependent thermal conductivity in suspended single-layer graphene. *Nature Communications* **5**, 3689 (2014).

[93] Kong, B. D., Paul, S., Nardelli, M. B. & Kim, K. W. First-principles analysis of lattice thermal conductivity in monolayer and bilayer graphene. *Physical Review B* **80**, 033406 (2009).

[94] Li, H. *et al.* Thermal conductivity of twisted bilayer graphene. *Nanoscale* **6**, 13402–13408 (2014).

[95] Wang, J., Zhu, L., Chen, J., Li, B. & Thong, J. T. Suppressing thermal conductivity of suspended tri-layer graphene by gold deposition. *Advanced Materials* **25**, 6884–6888 (2013).

[96] Lindsay, L. & Broido, D. Theory of thermal transport in multilayer hexagonal boron nitride and nanotubes. *Physical Review B* **85**, 035436 (2012).

[97] Wang, C. *et al.* Superior thermal conductivity in suspended bilayer hexagonal boron nitride. *Scientific Reports* **6**, 25334 (2016).

[98] Lindsay, L., Broido, D. & Mingo, N. Flexural phonons and thermal transport in multilayer graphene and graphite. *Physical Review B* **83**, 235428 (2011).

[99] Liu, Y. *et al.* Thermal conductance of the 2d MoS_2/h-BN and graphene/h-BN interfaces. *Scientific Reports* **7**, 43886 (2017).

[100] Gilligan, L. P., Cobelli, M., Taufour, V. & Sanvito, S. A rule-free workflow for the automated generation of databases from scientific literature. *arXiv preprint arXiv:2301.11689* (2023).

[101] Hawizy, L., Jessop, D. M., Adams, N. & Murray-Rust, P. ChemicalTagger: A tool for semantic text-mining in chemistry. *Journal of Cheminformatics* **3**, 1–13 (2011).

[102] Lowe, D. M., Corbett, P. T., Murray-Rust, P. & Glen, R. C. Chemical name to structure: OPSIN, an open source solution. *Journal of Chemical Information and Modeling* **51**, 739–753 (2011).

[103] Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L. & Murray-Rust, P. OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics* **3**, 1–12 (2011).

[104] Corbett, P. & Copestake, A. Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinformatics* **9**, 1–10 (2008).

[105] Rocktäschel, T., Weidlich, M. & Leser, U. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics* **28**, 1633–1640 (2012).

[106] Zhao, S. Named entity recognition in biomedical texts using an HMM model. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, 87–90 (2004).

[107] McCallum, A., Freitag, D., Pereira, F. C. *et al.* Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, 591–598 (2000).

[108] Lafferty, J. D., McCallum, A. & Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, 282–289 (2001).

[109] He, Y. & Kayaalp, M. Biological entity recognition with conditional random fields. In *AMIA Annual Symposium Proceedings*, vol. 2008, 293 (American Medical Informatics Association, 2008).

[110] Wu, X. *et al.* ChemBrowser: a flexible framework for mining chemical documents. In *Advances in Computational Biology*, 57–64 (Springer, 2010).

[111] Swain, M. C. & Cole, J. M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling* **56**, 1894–1904 (2016).

[112] Kim, E. *et al.* Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials* **29**, 9436–9444 (2017).

[113] Kim, E. *et al.* Machine-learned and codified synthesis parameters of oxide materials. *Scientific Data* **4**, 1–9 (2017).

[114] Court, C. J. & Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Scientific Data* **5**, 1–12 (2018).

[115] Huang, S. & Cole, J. M. BatteryBERT: A pretrained language model for battery database enhancement. *Journal of Chemical Information and Modeling* **62**, 6365–6377 (2022).

[116] Vaswani, A. *et al.* Attention is all you need. *Advances in Neural Information Processing Systems* **30** (2017).

[117] Graves, A. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850* (2013).

[118] Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[119] Sharma, S., Kiros, R. & Salakhutdinov, R. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119* (2015).

[120] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[121] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. *et al.* Improving language understanding by generative pre-training. Preprint at: `api.semanticscholar.org/CorpusID:49313245` (2018).

[122] Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).

[123] Brown, T. *et al.* Language models are few-shot learners. *Advances in Neural Information Processing Systems* **33**, 1877–1901 (2020).

[124] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[125] Lee, P., Bubeck, S. & Petro, J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine* **388**, 1233–1239 (2023).

[126] Sanderson, K. GPT-4 is here: what scientists think. *Nature* **615**, 773 (2023).

[127] Trewartha, A. *et al.* Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns* **3**, 100488 (2022).

[128] Beltagy, I., Lo, K. & Cohan, A. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).

[129] Gupta, T., Zaki, M., Krishnan, N. A. & Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials* **8**, 102 (2022).

[130] Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[131] Shetty, P. *et al.* A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Computational Materials* **9**, 52 (2023).

[132] Dunn, A. *et al.* Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238* (2022).

[133] Isayev, O. *et al.* Materials cartography: representing and mining materials space using structural and electronic fingerprints. *Chemistry of Materials* **27**, 735–743 (2015).

[134] Lederer, Y., Toher, C., Vecchio, K. S. & Curtarolo, S. The search for high entropy alloys: a high-throughput ab-initio approach. *Acta Materialia* **159**, 364–383 (2018).

[135] Sanvito, S. *et al.* Accelerated discovery of new magnets in the heusler alloy family. *Science Advances* **3**, e1602241 (2017).

[136] Xi, L. *et al.* Discovery of high-performance thermoelectric chalcogenides through reliable high-throughput material screening. *Journal of the American Chemical Society* **140**, 10785–10793 (2018).

[137] Jain, A. *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002 (2013).

[138] Kirklin, S. *et al.* The open quantum materials database (OQMD): assessing the accuracy of dft formation energies. *npj Computational Materials* **1**, 1–15 (2015).

[139] Zagorac, D., Müller, H., Ruehl, S., Zagorac, J. & Rehme, S. Recent developments in the inorganic crystal structure database: theoretical crystal structure data and related features. *Journal of Applied Crystallography* **52**, 918–925 (2019).

[140] Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **72**, 171–179 (2016).

[141] Vaitkus, A., Merkys, A. & Gražulis, S. Validation of the crystallography open database using the crystallographic information framework. *Journal of Applied Crystallography* **54**, 661–672 (2021).

[142] Gallego, S. V. *et al.* MAGNDATA: towards a database of magnetic structures. I. the commensurate case. *Journal of Applied Crystallography* **49**, 1750–1776 (2016).

[143] Bird, S., Klein, E. & Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (O'Reilly Media, Inc., 2009).

[144] arXiv.org submitters. arXiv Dataset. Data Available at: `www.kaggle.com/datasets/Cornell-University/arxiv`. Accessed: 2023-02-24.

[145] Soares, L. B., FitzGerald, N., Ling, J. & Kwiatkowski, T. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158* (2019).

[146] Xu, Y., Yamazaki, M. & Villars, P. Inorganic materials database for exploring the nature of material. *Japanese Journal of Applied Physics* **50**, 11RH02 (2011).

[147] Connolly, T. F. *Bibliography of Magnetic Materials and Tabulation of Magnetic Transition Temperatures* (Springer Science & Business Media, New York, US, 2012).

[148] Buschow, K. & Wohlfarth, E. (eds.) *Handbook of Magnetic Materials, Volumes 4-16 and 18* (Elsevier, Amsterdam, Netherlands, 1988-2009).

[149] Coey, J. *Magnetism and Magnetic Materials* (Cambridge University Press, Cambridge, 2010).

[150] Byland, J. K. *et al.* Statistics on magnetic properties of Co compounds: A database-driven method for discovering Co-based ferromagnets. *Physical Review Materials* **6**, 063803 (2022).

[151] Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2**, 1–7 (2016).

[152] Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *The Journal of Physical Chemistry Letters* **9**, 1668–1673 (2018).

[153] Kiselyova, N. N., Dudarev, V. A. & Korzhuyev, M. A. Database on the bandgap of inorganic substances and materials. *Inorganic Materials: Applied Research* **7**, 34–39 (2016).

[154] Strehlow, W. & Cook, E. L. Compilation of energy band gaps in elemental and binary compound semiconductors and insulators. *Journal of Physical and Chemical Reference Data* **2**, 163–200 (1973).

[155] Joshi, N. *Photoconductivity: Art, Science & Technology* (Routledge, 2017).

[156] Madelung, O. *Semiconductors: Data Handbook* (Springer Science & Business Media, 2004).

[157] Dong, Q. & Cole, J. M. Auto-generated database of semiconductor band gaps using chemdataextractor. *Scientific Data* **9**, 193 (2022).

[158] Si, X. *et al.* First-principles investigation on the optoelectronic performance of Mg-doped and Mg–Al Co-doped ZnO. *Materials & Design* **93**, 128–132 (2016).

[159] Chen, L., Wang, A., Xiong, Z., Shi, S. & Gao, Y. Effect of hole doping and strain modulations on electronic structure and magnetic properties in ZnO monolayer. *Applied Surface Science* **467**, 22–29 (2019).

[160] Bludau, W., Onton, A. & Heinke, W. Temperature dependence of the band gap of silicon. *Journal of Applied Physics* **45**, 1846–1848 (1974).

[161] Nosaka, Y. & Nosaka, A. Y. Reconsideration of intrinsic band alignments within anatase and rutile TiO_2. *The Journal of Physical Chemistry Letters* **7**, 431–434 (2016).

[162] Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A. Single-layer MoS_2 transistors. *Nature Nanotechnology* **6**, 147–150 (2011).

[163] Böker, T. *et al.* Band structure of MoS_2, $MoSe_2$, and α-$MoTe_2$: Angle-resolved photoelectron spectroscopy and ab initio calculations. *Physical Review B* **64**, 235305 (2001).

[164] Tang, Q. & Jiang, D.-e. Stabilization and band-gap tuning of the 1T-MoS_2 monolayer by covalent functionalization. *Chemistry of Materials* **27**, 3743–3748 (2015).

[165] Wort, C. J. & Balmer, R. S. Diamond as an electronic material. *Materials Today* **11**, 22–28 (2008).

[166] Jalali-Asadabadi, S. *et al.* Electronic structure of crystalline buckyballs: FCC-C60. *Journal of Electronic Materials* **45**, 339–348 (2016).

[167] Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials* **6**, 138 (2020).

[168] Firth, J. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis* 10–32 (1957).

[169] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* **41**, 391–407 (1990).

[170] Church, K. & Hanks, P. Word association norms, mutual information, and lexicography. *Computational Linguistics* **16**, 22–29 (1990).

[171] Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* **28**, 11–21 (1972).

[172] Mikolov, T., Yih, W.-t. & Zweig, G. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751 (2013).

[173] Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[174] Bengio, Y., Ducharme, R. & Vincent, P. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, vol. 13 (2000).

[175] Pennington, J., Socher, R. & Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543 (2014).

[176] Bommasani, R., Davis, K. & Cardie, C. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4758–4781 (2020).

[177] Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).

[178] Ricci, F. *et al.* An ab initio electronic transport database for inorganic materials. *Scientific Data* **4**, 1–13 (2017).

[179] Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J. & Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *Npj Computational Materials* **7**, 77 (2021).

[180] Faber, F. A., Lindmaa, A., Von Lilienfeld, O. A. & Armiento, R. Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals. *Physical Review Letters* **117**, 135502 (2016).

[181] Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare* **3**, 1–23 (2021).

Appendix A

List of Publications

- L.P.J. Gilligan, M. Cobelli, V. Taufour & S. Sanvito, ‘A rule-free workflow for the automated generation of databases from scientific literature’, *npj Computational Materials* (2023) - Accepted for Publication
- L.P.J. Gilligan, R. Dong & S. Sanvito, ‘Machine-Learned Interatomic Potentials for Rapid Thermal Property Prediction of 2D Materials’ - In Preparation
- M. Cobelli, L.P.J. Gilligan & S. Sanvito, ‘Accelerating Language-Model Training for Materials Science’ - In Preparation